HAL open science

Information Digestion

Gaël Dias

▶ To cite this version:

Gaël Dias. Information Digestion. Machine Learning [cs.LG]. Université d'Orléans, 2010. tel-00669780

HAL Id: tel-00669780

https://theses.hal.science/tel-00669780v1

Submitted on 13 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITY OF ORLÉANS
LIFO
LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLÉANS

H D R T H E S I S

to obtain the title of

Habilitation à Diriger des Recherches

of the University of Orléans
Specialty : Computer Science

Defended by

Gaël Dias

Information Digestion

prepared at the Center of Human Language Technology and Bioinformatics, University of Beira Interior, Portugal

defended on Month Day, 2010

Jury :

President :	Adeline Nazarenko	-	Université Paris-Nord - LIPN (France)
Reviewers :	Brigitte Grau	-	ENSIIE-LIMSI (France)
	Manuel Vilares Ferro	-	Universidade de Vigo (Spain)
	Marie-Francine Moens	-	Katholieke Universiteit Leuven (Belgium)
Examinators :	Isabelle Tellier	-	Université d'Orléans - LIFO (France)
	Mohand Boughanem	-	Université Paul Sabatier - IRIT (France)
	Pierre Zweigenbaum	-	CNRS-LIMSI (France)
	Robert Mahl	-	ENSMP-CRI (France)

Acknowledgments

I would to thank David Everett Rumelhart and James Lloyd McClelland from Stanford University for their work on learning past tense of English verbs with neural networks [Rumelhart 1986]. This paper inspired a young man who was still searching for an exciting area within artificial intelligence. Since then, natural language processing has always been part of my everyday life. The rest is just history, inspiration, hard work, fantastic supervisors, grateful collaborators, magnificent encounters, inspiring reviewers, comprehensive friends but more than anything, family support to cope with ups and downs, successes and failures, doubts and certainties. Thanks to everyone to let this young man still look at new trends in natural language processing with great admiration, but also to motivate him to improve everyday and hopefully reach higher issues. Finally, what really matters is that at the end my family can be proud of me ■

Information Digestion by Extracting Implicit Knowledge about the Language and Explicit Information from Real-World Heterogeneous Texts

Abstract: The World Wide Web (WWW) is a huge information network within which searching for relevant quality contents remains an open question. The ambiguity of natural language is traditionally one of the main reasons, which prevents search engines from retrieving information according to users' needs. However, the globalized access to the WWW via Weblogs or social networks has highlighted new problems. Web documents tend to be subjective, they mainly refer to actual events to the detriment of past events and their ever growing number contributes to the well-known problem of information overload. In this thesis, we present our contributions to digest information in real-world heterogeneous text environments (i.e. the Web) thus leveraging users' efforts to encounter relevant quality information. However, most of the works related to Information Digestion deal with the English language fostered by freely available linguistic tools and resources, and as such, cannot be directly replicated for other languages. To overcome this drawback, two directions may be followed: on the one hand, building resources and tools for a given language, or on the other hand, proposing language-independent approaches. Within the context of this report, we will focus on presenting language-independent unsupervised methodologies to (1) extract implicit knowledge about the language and (2) understand the explicit information conveyed by real-world texts, thus allowing to reach Multilingual Information Digestion.

Keywords: Unsupervised language-independent approaches, Information Digestion, Real-world text environments, Word semantic relations, Explicit and implicit knowledge extraction, Sentiment analysis.

Contents

1 Introduction **1**
1.1 Understanding Language . . . 1
1.2 Understanding Information . . . 4
1.3 Research Environment . . . 6
1.4 Thesis Organization . . . 6

2 Word Similarities **9**
2.1 Symmetric Word Similarities . . . 10
2.1.1 Pattern-based Measures . . . 11
2.1.2 Association Measures . . . 12
2.1.3 Attributional Word Similarities . . . 17
2.1.4 Knowledge-based Word Similarities . . . 23
2.1.5 Relational Word Similarities . . . 24
2.2 Asymmetric Word Similarities . . . 25
2.2.1 Asymmetric Association Measures . . . 26
2.2.2 Asymmetric Attributional Word Similarities . . . 27
2.3 Future Work . . . 29

3 Multiword Units Extraction **31**
3.1 Related Work . . . 34
3.2 Statistical Multiword Units Extraction . . . 38
3.3 Efficient Multiword Units Extraction . . . 42
3.4 Learning Multiword Units Extraction . . . 47
3.5 Hybrid Multiword Units Extraction . . . 52
3.6 Future Work . . . 55

4 Ephemeral Clustering **59**
4.1 Normalized Full Text Hierarchical Overlapping Clustering . . . 66
4.2 Normalized Snippet Hierarchical Overlapping Clustering . . . 73
4.3 Snippet Informative Hierarchical Clustering . . . 76
4.4 Future Work . . . 86

5 Document Summarization and Sentence Reduction **91**
5.1 Topic Segmentation . . . 96
5.1.1 Related Work . . . 97
5.1.2 Informative Topic Segmentation . . . 98
5.2 Construction of Lexical Chains . . . 102
5.2.1 Related Work . . . 103
5.2.2 Construction of a Lexical-semantic Knowledge Base . . . 104
5.2.3 Lexical Chainer Algorithm . . . 105

5.3 Sentence Reduction . . . 110
5.3.1 Related Work . . . 111
5.3.2 Paraphrase Extraction . . . 114
5.3.3 Paraphrase Alignment . . . 120
5.3.4 Reduction Rules Learning . . . 124
5.4 Future Work . . . 129

6 Construction of Lexical-Semantic Resources 135
6.1 Prototype-based Ontologies . . . 139
6.1.1 Discovering Highly Related Words by Similarity . . . 140
6.1.2 Discovering Highly Related Words by Interchangeability . . . 142
6.1.3 Discovering Word Generality . . . 150
6.2 Terminological Ontologies . . . 159
6.3 Future Work . . . 168

7 Subjectivity in Language 173
7.1 Related Work . . . 176
7.2 Automatic Construction of Labeled Data Sets . . . 179
7.3 Single-view Clustering . . . 182
7.4 Multi-view Clustering . . . 188
7.5 Future Work . . . 191

8 Conclusions 193
8.1 Conclusions . . . 193
8.2 Future Projects . . . 197
8.2.1 Temporal Information Retrieval . . . 198
8.2.2 Personalized Information Retrieval . . . 199

A Research Environment 203
A.1 Projects . . . 203
A.2 Academic Staff . . . 207
A.3 Collaborators and Partners . . . 209

Bibliography 211

Chapter 1

Introduction

Contents

1.1 Understanding Language 1
1.2 Understanding Information 4
1.3 Research Environment 6
1.4 Thesis Organization 6

The World Wide Web (WWW) is a huge information network within which searching for relevant quality contents remains an open question. The ambiguity of natural language is one of the main reasons, which prevents search engines from retrieving information according to users' needs. However, the globalized access to the WWW via Weblogs or social networks has highlighted new problems. Web documents tend to be subjective, they mainly refer to actual events to the detriment of past events and their ever growing number contributes to the well-known problem of information overload. In this thesis, we present our contributions to digest information in real-world heterogeneous text environments (i.e. the Web) thus leveraging users' efforts to encounter relevant quality information. However, most of the works related to Information Digestion deal with the English language fostered by freely available linguistic tools and resources, and as such, cannot be replicated directly for other languages. To overcome this drawback, two directions may be followed: on the one hand, building resources and tools for a given language, or on the other hand, proposing language-independent approaches. Within the context of this thesis, we will focus on presenting language-independent unsupervised methodologies to (1) extract implicit knowledge about the language and (2) understand the explicit information conveyed by real-world texts, thus allowing to reach Multilingual Information Digestion.

1.1 Understanding Language

I would like to begin this introduction in a narrative form so that the reader can better understand our research motivations and approaches. This thesis is the outcome of many encounters, readings, suppositions, beliefs, intuitions about natural language processing and information understanding. My first important inspiration was the work by [Rumelhart 1986], who proposed a neural network approach to learn the past tense of English verbs. This work inspired me in the

way that in my early ages I experienced the famous U-shaped learning curve while absorbing the Portuguese language with my grand-parents. Indeed, I never learned Portuguese (my second language). I just assimilated it without any effort, even not knowing it, just by listening to people talks. How could this be possible? I remember to think. While there are people who cannot count nor read, everyone can speak, even many languages at the same time by just being immersed within a speech environment. But, at the same time, language is hard for machines to understand. Indeed, while we are capable of mathematically modeling hard problems form Nature, language still continues to be a great mystery, especially when trying to cross the gap from syntax to semantics. Even if human beings as rude as they can be can speak and communicate, why would language be so hard to model?

Part of the answer comes from a second encounter. This was the speech given by Ido Dagan at the Trans-European Language Resources Infrastructure European seminar in 1999 about the automatic acquisition of multilingual resources. There, he showed his "despair" when arguing how hard it was to outrank simple methodologies based on mere frequency. Indeed complex mathematically well-founded models seemed to fail where simple frequency succeeded. At that time, I was experiencing the exact same issues when trying to model the extraction of multiword units within the scope of my PhD thesis. But is this fact so difficult to understand after all? While mathematical models aim at simulating perfect reproducible experiences, language is far from being perfect and static. As a consequence, language is hard to model by nature. For instance, one can understand the ambiguity of natural language as a consequence of its incompleteness. Indeed, in a perfect world, there would be a unique expression for each concept, as fined-grained as it might be. And ambiguity would automatically disappear. So, much of our work will be based on intuition and heuristics, which we will try to model with mathematically well-founded models, although this issue may not be possible.

Finally, the third important contribution matching my believes about natural language processing came from a study about animal reproduction and the importance of viruses within this process. Because viruses are acellular, they must use the machinery and metabolism of a host cell to reproduce. For this reason, viruses can integrate their genome into the host genome. Surprisingly, some researchers recently found that viruses were playing an important role within the reproduction process. But the most interesting issue of this research is the fact that scientists used to drop the genetic material of viruses from their analysis as they thought they were just impurities within the host genome. As a consequence, they could not understand the role viruses were playing in the reproduction process. But, by taking this small repetitive genetic information lead to new insights within understanding the sight of life. When I was watching this television programme, I immediately remembered the discussions we had with Professor Sylvie Billot and José Gabriel Pereira Lopes about the importance of stop-words. Hence,

I always questioned the issue of "empty" words removal. So, why should we remove words from texts? Just because we are not capable of handling them? This issue was not clear to me. But, to the light of this research within the animal reproduction process, I could not help to make the parallelism with the remotion of stop-words. I would call later this issue as the *corpus integrity principle*.

All these issues, beliefs and intuitions about language are present in this thesis. Consequently, we try to propose unsupervised language-independent methodologies based on complete raw texts to find (1) implicit knowledge about language and (2) explicit information conveyed by texts such as Web page clustering and document summarization. Unsupervised because supervised implies the existence of some "teacher" who tells you what to do. However, we are able to learn languages without teachers. Language-independent because language is not specific to any idiom. Indeed, children learn different languages at the same time without any predisposition for that. Based on raw texts, because this is the only material we have access to in order to learn languages together with speech. And complete texts, because all lexical items count and dropping some textual material may lead to unjustified theories.

Of course, this discourse is not new and has been widely studied by linguists and psycho-linguists. It can even sound purposely corrosive, polemical and challenging. Of course, we clearly assess that linguistic tools and resources may lead to improved results within the scope of natural language understanding. However, we believe that to some extent the research community has lately focused more on results rather than presenting new radical issues. In particular, this is mainly due to the growing number of evaluation conferences such as TREC[1] (Text Retrieval Conference), DUC[2] (Document Understanding Conference), MUC[3] (Message Understanding Conference), CLEF[4] (Cross-Language Evaluation Forum) or RTE[5] (Recognizing Textual Entailment), to name but a few. Although, they are important as they propose standard evaluation data sets, they may narrow the imagination of researchers as the principle objective of the research teams for the next year contest is to improve results based on the same framework of the previous year. As a consequence, these conferences may have a perverse impact by limiting the apparition of new ideas. As a whole, new "risky" challenges are likely to obtain worst results than well-established methodologies. But, on the opposite, they may open new interesting research directions.

Based on these reflections, we clearly believe that a lot remains to be discovered at the basic raw text level before introducing extra linguistic knowledge. Once again, it is clear that enriching texts with (shallow or deep) linguistic

[1] http://trec.nist.gov/ [26th *September*, 2010].
[2] http://duc.nist.gov/ [26th *September*, 2010].
[3] http://en.wikipedia.org/wiki/Message_Understanding_Conference [26th *September*, 2010].
[4] http://www.clef-campaign.org/ [26th *September*, 2010].
[5] http://www.nist.gov/tac/2010/RTE/index.html [26th *September*, 2010].

information must be done to full understanding of natural language. But this need to be done when nothing more seems to apply to improve results on raw texts, as we will demonstrate in this thesis. Further, we will show that for some tasks, we used part-of-speech tagging or shallow-parsing to substantially improve results. However, when introducing extra knowledge within raw texts, we automatically start to think about some ways to overcome this input and push forward our imagination to find new ideas to avoid this extra step. So, understanding natural language based on complete raw texts is obviously a challenging task as most of approaches proposed in the last decades tend to introduce more and more linguistic knowledge from deep parsing to high-level semantic processing. Duly, dealing with raw texts allows language-independency and as a consequence multilingual processing as well as coping with real-word heterogeneous texts (e.g. Web pages). As a summary, we are more interesting in proposing new ideas based on hard constraints rather than following what we call natural language engineering instead of natural language understanding. May be to the detriment of expressive results.

1.2 Understanding Information

Any researcher is always confronted sooner or later with the same question from his friends or relatives. After all what is your research about and what is it for? Once, my father asked me this fatal question. I tried to explain him what I was doing within the scope of extraction of multiword units and why this research was important to reach full understanding of language. But, I saw scepticism in his eyes. I clearly understood that research on its own can be interesting and fulfil your spirit individually. But at some point, it must serve the community. If both language-oriented and application-oriented researches can be mixed to produce innovative solutions, which can change the everyday life of people, somehow you will deeply feel your contribution to the society.

Fortunately, after many years fighting with my father, he finally installed internet at home. Interestingly, I had never understood the difficulties of people searching for information on the Web. But when my parents started to use search engines, they just felt lost with so many documents being retrieved, which were not related to their needs. Moreover, they were confronted to Web pages written in English although their queries were in French. Their first reaction was to give up internet because they did not have search interfaces adapted to their needs nor the retrieval results were satisfactory. In fact, when you know the system in the box, searching for information is natural. However, for most people, searching for information is a difficult process. The recent improvements of the Goggle^{TM} interface are the best proof of this issue, by introducing query suggestion, the magic wheel (a concept between flat and hierarchical clustering), time lines, translation services and so on and so forth. At that time, I clearly understood that much had to be done to ease the search process and that we could take advantage of our

initial works about the acquisition of implicit knowledge about the language to better understand the information conveyed by texts. This would lead later to the construction of our meta-search engine VIPACCESS[6], which aims at automatically digesting information for the user so that the search process is facilitated.

[McCallum 2005] proposed the concept of information distilling, which consists in structuring information from unstructured texts within the context of information extraction. Similarly to [McCallum 2005], we propose to structure information retrieved by Web search engines within the context of information retrieval (IR). We call this new concept, Information Digestion. Information Digestion can be defined as the process of text understanding and its subsequent reduction. As such, Information Digestion includes summarization, selection and organization of relevant information. A typical example has to do with visually impaired people (VIP) in the context of Web search engines. VIP face an overwhelming task when reading texts. Unlike fully capacitated people, blind people cannot read texts by just scanning them transversally. As a consequence, they have to come through all the sentences of any text to understand all the information contained in it. Yet, search results are usually organized as long lists of potentially relevant documents. So, presenting Web page results as information clusters would reduce the user's effort to correctly encounter the intended information by lowering the information to be read.

A similar situation is evidenced by search engines for mobile devices, which must condense text information to fit in small screens so that Web browsing is minimized as much as possible. A direct illation of this situation is that users (eventually VIP) are unlikely to use classical search engines interfaces to search for information on devices such as smartphones or PDA. Instead, they may shift to specifically designed interfaces as illustrated in Figure 1.1 for the case of ubiquitous information retrieval.

Based on these assumptions, Information Digestion within the context of Web search can be seen as all means of reducing potentially readable textual information without loosing the core message. As such, clustering Web search results based on concepts and/or temporal aspects, introducing subjectivity classifiers to remove unreliable Web search results, selecting best results according to user profiles (i.e. personalized IR) or to user contexts (i.e. context-aware IR), or simply summarizing texts such as Web pages or Web snippets can be seen as part of Information Digestion. In summary, we focus in this thesis on unsupervised language-independent methodologies to extract explicit knowledge about the world from real-world texts fostered by the automatic acquisition of implicit knowledge about the language to reach Multilingual Information Digestion.

[6] http://hultig.di.ubi.pt/vipaccess/ [26th *September*, 2010].

Figure 1.1: Ephemeral clustering as Information Digestion [3rd*April*, 2009].

1.3 Research Environment

I initiated my research work at the Centro de Inteligência Artificial (CENTRIA)[7] of the New University of Lisbon (Portugal) with Professor José Gabriel Pereira Lopes and the Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)[8] of the University of Orléans (France) with Professors Sylvie Billot and Jean-Claude Bassano. There, I obtained in 2002 my PhD degree dealing with the extraction of multiword units [Dias 2002]. Then, I moved to the University of Beira Interior (Portugal), where I am now assistant professor. In 2003, I created the Center of Human Language Technology and Bioinformatics (HULTIG)[9] where most of my research has undergone since 2003. However, all the research work presented in this thesis is not the result of a solitary effort by rather the one of a real research team. In particular, bachelor, master and PhD students have been working under my supervision in different research projects since 2003. Moreover, efforts have been made to initiate collaborations with national and international research institutes as well as industrial partners. All this information is listed in appendix A.

1.4 Thesis Organization

This thesis is not organized in two parts i.e. one dealing with the extraction of implicit knowledge about the language and one with the understanding of explicit information conveyed by real-world texts. Instead, we preferred to introduce the

[7] http://centria.di.fct.unl.pt [23rd*September*, 2010].
[8] http://www.univ-orleans.fr/lifo/ [23rd*September*, 2010].
[9] http://hultig.di.ubi.pt [23rd*September*, 2010].

topics of our research as we were confronted to them to the exception of Chapter 2, which is transversal to our research.

Chapter 2 - Word Similarities: Developing language-independent methodologies, which extract implicit and explicit knowledge from natural languages implies working on raw texts as the basic textual information, whether at word, sentence, paragraph or text level. As a consequence, two different types of knowledge can be acquired depending on the basic textual unit under study. On the one hand, analyzing word similarities evidences intrinsic knowledge about the language (i.e. information about the language which is not explicitly encoded in texts). Traditional examples are collocations and word semantic relations such as hypernymy/hyponymy, meronymy/holonymy, synonymy or antonymy, which must be mined from texts. On the other hand, explicit knowledge about the language (i.e. information about the message conveyed by the texts) can be extracted from the evaluation of sentence, passage and text similarities. So, in this chapter, we will define different similarity measures, which contribute to the state-of-the-art and propose new perspectives towards the definition of informative similarity measures.

Chapter 3 - Multiword Units Extraction: Multiword unit extraction is a specific case, where word similarity measures can be applied. Indeed, understanding the implicit information present in texts about the language itself is one of the most important contributions to natural language learning. Within this context, multiword unit identification is a crucial issue towards understanding the essence of the message conveyed by any given text. For that purpose, we will present different language-independent unsupervised strategies following different paradigms to extract MWU: (1) statistical association measures, (2) hybrid association measures and (3) reinforcement learning.

Chapter 4 - Ephemeral Clustering: Ephemeral clustering can be seen as the first reduction step of information reduction within Web search engines. In particular, it requires efficient and accurate algorithms, since clustering is performed on-line and users who are not domain experts are less tolerant to errors. According to [Zamir 1998], multiword units are critical to the success of monothetic and polythetic clustering. This clearly means that MWU best embody the message conveyed by texts. Within this scope, we will propose to apply text normalization (i.e. identifying MWU within texts) to better represent Web snippets and introduce a new paradigm by performing ephemeral clustering through query disambiguation.

Chapter 5 - Document Summarization and Sentence Reduction: Automatic text summarization (ATS) is certainly the straightforward way to reach Information Digestion. Indeed, after selecting the documents, which best fit the users needs, these ones may want to quickly access to their message. As such, ATS takes an important place within our work. In particular, we will propose simple method-

ologies, which can be run "on the fly". However, the obtained results will be far from being satisfactory and more complex methodologies will be proposed to be run off-line. In particular, we will introduce a new algorithm to perform topic segmentation, which will evidence improved results when texts are first normalized. Then, we will develop a new algorithm for the construction of lexical chains based on automatically acquired taxonomies. However, due to the simplicity of the approach, we will first focus our research on the construction of high accuracy lexical-semantic resources (see Chapter 6) and let apart the overall process of ATS. Finally, we will tackle sentence reduction, which can be seen as the ultimate ultra-summarization abstractive approach as opposed to the extractive paradigm.

Chapter 6 - Construction of Lexical-Semantic Resources: If one wants to deeply understand the messages conveyed by texts, extracting implicit knowledge about the language seems compulsory. As a consequence, we will start to look at word similarities to learn tight semantic relationships. In fact, we will propose a new methodology to extract highly semantically related words based on attributional similarity measures and paraphrase alignments. As such, combined with a new method to assess the level of generality of words, we will propose different strategies to build prototype-based ontologies. Moreover, a recent study will be detailed, which allows to automatically build terminological ontologies. Once again, we will show how some characteristics can be unsupervisedly modeled as well as linguistic resources cab automatically be acquired from common sense judgements.

Chapter 7 - Subjectivity in Language: Finally, Information Digestion can be seen as the process of selecting high quality information, or at least warning the user about the objectiveness or the subjectiveness of the retrieved documents. This is a critical issue due to the democratization of the Web through Weblogs and social networks. However, this chapter is certainly the one, which less complies with our ideas about the study of language. Indeed, in the first part of our work, we will use highly specialized linguistic resources both within single-view and multi-view learning frameworks. Although improved results are obtained when compared to the state-of-the-art, we will provide some clues to reach domain- and language independency at the end of the chapter. Especially, we will show how the level of generality of texts can be modeled using the methodologies proposed in Chapter 6.

In the conclusion, we will summarize to what extent we are able to propose unsupervised language-independent methodologies for Information Digestion based on complete raw texts. Of course, we will see that enriching raw texts with extra shallow linguistic information can lead to improved results. Nevertheless, we will evidence that these improvements are only introduced when no advances can be reached by just looking at raw texts. Finally, we will provide some directions for future work within the Information Digestion paradigm, such as temporal clustering and personalized IR ■

Chapter 2

Word Similarities

Contents

2.1 Symmetric Word Similarities **10**
2.1.1 Pattern-based Measures 11
2.1.2 Association Measures 12
2.1.3 Attributional Word Similarities 17
2.1.4 Knowledge-based Word Similarities 23
2.1.5 Relational Word Similarities 24
2.2 Asymmetric Word Similarities **25**
2.2.1 Asymmetric Association Measures 26
2.2.2 Asymmetric Attributional Word Similarities 27
2.3 Future Work **29**

Developing language-independent methodologies, which extract implicit and explicit knowledge from natural languages implies working on raw texts as the basic textual information[1]. The atomic textual information unit is obviously the unicode character. Indeed, character-based languages such as Chinese or Japanese lack word delimiters. Similarly, Thai and Lao languages only use delimiters for phrases and sentences but not for words. For non character-based languages[2], text segmentation is leveraged as word and sentence boundaries are clearly marked. Within the context of our research, the basic textual unit is the word i.e. a sequence of characters delimited by white spaces and some specific delimiters such as punctuation. Consequently, sentences are defined as sets of words, passages as sets of sentences, texts as sets of passages and corpora as sets of texts.

Two different types of knowledge can be acquired depending on the basic textual unit under study. On the one hand, analyzing word similarities evidences intrinsic knowledge about the language (i.e. information about the language which is not explicitly encoded in texts). Traditional examples are collocations and word semantic relations such as hypernymy/hyponymy, meronymy/holonymy, synonymy

[1] In the remainder of this thesis, we will see that some linguistic resources and tools will be used but within limited scopes.

[2] If there exist. This remains an interesting issue. Indeed, we have been carried out different studies at the character level for European languages [Dias 2000d] [Dias 2000c] [Ribeiro 2001] leading to interesting issues worthy to be pursued and further analyzed.

or antonymy, which must be mined from texts. On the other hand, explicit knowledge about the language (i.e. information about the message conveyed by the texts) can be extracted from the evaluation of sentence, passage and text similarities[3]. There are obviously some exceptions. In particular, analyzing sentence similarities in the context of topic segmentation is likely to identify intrinsic knowledge about discourse structure as we show in [Dias 2007a]. As most of our contributions deal with word similarity measures, which in some cases can be extended to sentence similarity measures, we specifically address the issue of word similarity within this chapter and leave for the up-coming chapters our ideas about text similarities (in particular, in Chapter 4 and Chapter 5).

Within the context of word similarity, we have been developing different similarity measures for several years, which contribute to the state-of-the-art in the field and propose new perspectives towards the definition of informative similarity measures. Five main approaches have been proposed so far in the literature: pattern-based similarities, association measures, attributional similarities, knowledge-based similarities and relational similarities. But our research has mainly focused on association measures and informative attributional similarities. In particular, we proposed the Mutual Expectation [Dias 1999a], an association measure, which evaluates the degree of relatedness between all the words present in a positional n-gram and demonstrated improved results for the task of multiword unit extraction compared to other existing measures [Dias 2000e]. We also defined a new word similarity measure called the InfoSimba (IS) [Dias 2005a] based on word context vectors following Harris' distributional analysis paradigm [Harris 1968], which introduces knowledge acquired from word co-occurrences within the definition of attributional similarity measures. The InfoSimba can be seen as a second order similarity measure which can be extended to a N order similarity measure with its recursive version, the RIS proposed in [Cleuziou 2008]. Recently, we have been working on an asymmetric version of the InfoSimba (AIS), and its recursive version (RAIS) to improve ontology construction and textual entailment[4]. We will present these measures in the following sections by focusing first on symmetric word similarity measures and second on asymmetric ones.

2.1 Symmetric Word Similarities

Different approaches have been proposed to evaluate the degree of relatedness between words: pattern-based measures, association measures, attributional similarities, knowledge-based similarities and relational similarities. These different conceptualizations lead mandatorily to different extracted intrinsic word semantic relations. Association measures are more suited for the extraction of collocations in window-based environments but have also been used to track loose semantic

[3]From now on, we will refer to sentences, passages and texts simply as texts.
[4]These works are still under development and have not been published so far.

relations between words in larger contexts such as passages or texts as we show in [Dias 2006b] and [Dias 2007a]. Attributional similarities follow Harris' distributional analysis [Harris 1968] and extract tight semantic relations between words such as hypernymy/hyponymy, meronymy/holonymy or synonymy, although precision remains a great issue. Similarly, pattern-based methodologies are tuned to find co-occurring words within a given pattern[5] and assign a unique semantic relation between the words involved in the relation, thus evidencing high precision but low recall. Knowledge-based word similarities are mainly used in knowledge-rich applications to improve performance such as in text clustering [Xia 2006] [Song 2008], word sense disambiguation [Pantel 2002] [Sinha 2007] or question answering [Lin 2001]. Indeed, the resources used to compute similarities are usually lexical-semantic structures (e.g. Roget's thesaurus [Roget 1852] or WordNet [Miller 1990]) which intrinsically embody the implicit notion of distance between words. Finally, relational similarities measures evaluate the correspondence between relations, in contrast with attributional similarities, which measure similarity between attributes. So, when two pairs of words have a high degree of relational similarity, we can say that their relations are analogous.

2.1.1 Pattern-based Measures

Patterns can be helpful to learn knowledge from texts that can possibly be expressed by constructions known in advance and surely embody the easiest way to induce this knowledge. Most of the works in this area have been dealing with the identification of the hypernymy/hyponymy relation although some other word semantic relations such as synonymy and meronymy/holonymy have been tackled. In order to extract hypernymy/hyponymy relations, [Hearst 1992] first identifies a set of lexical-syntactic patterns that are easily recognizable (i.e. occur frequently and across text genre boundaries). These can be called seed patterns. Based on these seeds, she proposes a bootstrapping algorithm to semi-automatically acquire new more specific patterns such as *such NP as (NP,)* {or | and} NP*. Similarly, [Caraballo 1999] uses predefined patterns such as *X is a (kind of) Y* or *X, Y, and other Zs*, following the discussion in [Riloff 1997] that nouns in conjunctions or appositive relations tend to be semantically related. This information is then integrated in a clustering process where the internal nodes are given labels with respect to the votes for the various possible hypernyms of the words at leaf levels, as caught by the patterns.

A more challenging task is to automatically learn the relevant patterns. Most of the approaches are summarized in [Stevenson 2006]. The most well-known work in this area is certainly the one proposed by [Snow 2005] who use machine learning techniques to automatically replace hand-built knowledge. By using dependency path features extracted from parse trees, they introduce a general-purpose formalization and generalization of these patterns. Given a training set of texts containing known hypernym pairs, their algorithm automatically extracts useful dependency paths

[5] Usually, manually defined.

and applies them to new corpora to identify novel pairs. [Sang 2007] uses a similar approach to derive extraction patterns for hypernymy/hyponymy relations by combining Web search engine counts from pairs of words encountered in WordNet with a bayesian logistic regression. Unlike the other approaches, [Bollegala 2007] tackle the extraction of patterns for the synonymy relation. They find lexical relationships between synonym pairs based on Web snippets counts and apply wildcards to generalize the acquired knowledge. Then, they apply a support vector machine (SVM) classifier to determine whether a new pair shows a relation of synonymy or not, based on a feature vector of lexical relationships. Finally, [Ohshima 2009] present one of the most interesting works in this area. Their motivation is that the close semantic relations are symmetric and the constructions that involve words in such relations are symmetric as well. Such constructions can be appositive relations, conjunctions or enumerations. In order to discover related terms, they instantiate and send to a search engine a number of patterns filled only with one possible candidate. The patterns are then sought through the snippets and the corresponding counterpart is collected, thus constructing two sets of left and right contexts. Those contexts that appear in both sets are taken to be the desired terms. In particular, this method can discover asymmetric relations if asymmetric patterns are available.

Despite the variety of approaches, two common characteristics are transversal to the methodology: (1) the necessity of manual effort as to compose the patterns and (2) the language-dependency of the method. Other drawbacks can be identified. In particular, lexical-syntactic patterns tend to be quite ambiguous as to which relations they indicate and this worsens when ambiguous words are involved. Also, mainly subsets of possible instances of semantic relations are likely to appear, thus imposing the existence of a great number of seed patterns.

2.1.2 Association Measures

Ferdinand de Saussure argues that all concepts are completely *negatively defined* that is, defined solely in terms of other concepts. He maintains that *language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others* and that *concepts are purely differential and defined not in terms of their positive content but negatively by their relations with other terms in the system* [Saussure 1959]. Thus, the assumption that the semantic similarity between two terms can be deduced only by observing their association patterns seems weak. For that reason, different approaches have been emerging, in particular the definition of association measures. Association measures are mathematical models that interpret word co-occurrence frequencies in a given text span (e.g. word windows, sentences, passages or texts). For any pair of words, an association score is computed on a continuous scale, which indicates the amount of (statistical) association between the two words. The association measures can be classified into three different approaches: statistical hypothesis tests, heuristic

combinations of observed joint and marginal frequencies and measures adopted from information theory. In general, the association scores computed by different measures cannot be compared directly, which motivated the work by [Pecina 2006] who present an exhaustive overview of association measures in the context of collocation extraction.

Although [Demonet 1975] and [Labbé 1988] are certainly the first works to study word co-occurrences in texts, they do not propose any association measure but rather an overall methodology. On the contrary, [Church 1990] propose the well-known Pointwise Mutual Information (PMI) adopted from information theory [Fano 1961] and defined in Equation 2.1 for two words x and y where $P(.)$ is the marginal probability function and $P(.,.)$ is the joint probability function.

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x).P(y)}. \tag{2.1}$$

They base their analysis on results of psycholinguistics. Within this context, [Meyer 1975] published an experiment in which the response time of an individual was measured when faced with two specific tasks: (1) classifying successions of letters into words and non words and (2) pronouncing a sequence of characters. This experiment showed that in both cases the answer to a word (e.g. *butter*) was always faster when preceded by an associated word (e.g. *bread*) than a non associated word (e.g. *nurse*). Following this intuition, [Church 1990] show that loose semantic relations can be extracted when counting word co-occurrence frequencies in fixed-size word windows, also known as n-grams. In particular, they evidence that the window size parameter allows to look at different scales such that smaller window sizes are likely to identify collocations and other relations that hold over short ranges, and larger window sizes highlight semantic concepts and other relationships that hold over larger scales. But, in no way, co-occurrence measures can identify the exact word semantic relations standing between two words. In fact, association measures are mainly used in the context of multiword unit extraction and word selection criteria.

Also, within the perspective of information theory, [Cilibrasi 2007] proposes the normalized Web distance (NWD) based on the Kolmogorov complexity [Li 2008]. They justify their approach by the vastness of the Internet and the assumption that the mass of information is so diverse that the frequencies of Web pages returned by a good set of search engine queries can average the semantic information in such a way that one can distill a valid semantic distance between the query subjects. One way to think about the Kolmogorov complexity $K(x)$, where x is any string, is to view it as the length, in bits, of the ultimate[6] compressed version from which x can be recovered by a general decompression program. By extrapolating this idea to search engines, they argue that any search engine such

[6] The lowest bound value.

as Google^{TM} can be seen as a compressor and as a consequence they define the NWD as in Equation 2.2 where $f(x, y)$ corresponds to the number of documents returned by the search engine, which contain both words x and y. Similarly, $f(z)$ corresponds to the number of hits for the query z and N can be approximated by the number of pages indexed by the search engine, which in the case of Google^{TM} is near to 10^{10}.

$$NWD(x, y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - min\{\log f(x), \log f(y)\}}. \tag{2.2}$$

Many association measures are heuristic combinations of observed joint and marginal frequencies, which prove successful results for different tasks. Within the context of construction of bilingual lexicons, [Smadja 1996] proposes the Dice coefficient introduced by [Dice 1945] (see Definition 2.3).

$$Dice(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)}. \tag{2.3}$$

Given a sentence aligned parallel bilingual corpus, they aim at translating collocations (or individual words) in the source language into collocations (or individual words) in the target language. For that purpose, they use a bootstrapping statistical methodology, which incrementally constructs the collocation translations, adding one word at a time. They identify individual words in the target language that are highly correlated through the Dice coefficient with the source collocation, thus producing a set of words in the target language, which are then combined in a systematic, iterative manner to produce a translation of the source language collocation.

Within the context of multiword units extraction, [Silva 1999] and [Dias 1999a] propose two different heuristics which respectively outperform different association measures[7] in the context of contiguous and non-contiguous n-grams. [Silva 1999] propose the Symmetric Conditional Probability (SCP) defined in Equation 2.4 for the contiguous case.

$$SCP(x, y) = \frac{P(x, y)^2}{P(x) \times P(y)}. \tag{2.4}$$

In parallel, in [Dias 1999a] we propose the Mutual Expectation (ME) based on the concepts of support and confidence from association rules [Agrawal 1996] for the non-contiguous case. The ME is defined in Equation 2.5.

$$ME(x, y) = \frac{2 \times P(x, y)^2}{P(x) + P(y)}. \tag{2.5}$$

[7] Both measures were compared to well-known association measures such as the PMI [Church 1990], the Dice [Dice 1945], the log-likelihood ratio [Dunning 1993] and the Φ^2 [Gale 1991]. More details will be given in Chapter 3.

Although heuristic-based association measures evidence successful results in different areas, they lack well-founded mathematical backgrounds[8], thus avoiding correct understanding of their behavior. For that purpose, different works propose association measures based on statistical evidence supported by statistical hypothesis tests, which rely on contingency tables as shown in Table 2.1 where $f(.,.)$ corresponds to observed joint frequencies, $f(.)$ to marginal frequencies and N to the number of words in the corpus. In particular, the decisions are made using null-hypothesis testing. In fact, one use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

words	y	$\bar{y}$	Totals
x	$f(x,y)$	$f(x,\bar{y})$	$f(x)$
$\bar{x}$	$f(\bar{x},y)$	$f(\bar{x},\bar{y})$	$f(\bar{x})$
Totals	$f(y)$	$f(\bar{y})$	N

Table 2.1: Contingency Table

In the field of collocation extraction, [Gale 1991] proposes to test the null hypothesis H_0 stated in Proposition 2.6 as the independence test such that two words x, y are semantically related if it is possible to reject the null hypothesis H_0 i.e. the joint probability is statistically different than the product of the marginal probabilities.

$$H_0 : P(x,y) = P(x).P(y) \tag{2.6}$$

For that purpose, [Gale 1991] proposes to use the Φ^2 test defined in Equation 2.7 based on Pearson's χ^2 statistical test.

$$\Phi^2(x,y) = \frac{(N \times f(w,y) - f(x) \times f(y))^2}{f(x) \times f(y) \times (N - f(x)) \times (N - f(y))}. \tag{2.7}$$

Although many statistical tests exist as listed in [Pecina 2006], [Dunning 1993] claims that all presuppose the assumption of normal distribution, which limits the ability to analyze rare events. Unfortunately, rare events do make up a large fraction of real text. For that reason, he proposes a method based on log-likelihood ratio tests, which yield good results with relatively small samples[9]. In particular, [Dunning 1993] uses the test of maximum likelihood that can successfully analyze contingency tables, which counts are not necessarily high. Thus, he proposes to test the null hypothesis H_0 (i.e. the independence assumption), which states that the probability of occurrence of a word in a given context is independent of the co-occurrence of any other word in its neighborhood. Within this context, it is necessary to determine the alternative hypothesis to H_0 denoted H_1. These two hypotheses are formulated in Proposition 2.8.

[8] An exception is certainly the ME, which shows interesting mathematical properties such as recursivity as shown in [Dias 2002].

[9] It is also well-known that the PMI tends to give good results for rare events.

$$H_0 : P(x|y) = P(x|\bar{y}) = P(x) = \theta \text{ and } H_1 : P(x|y) = \theta_1 \neq P(x|\bar{y}) = \theta_2 \quad (2.8)$$

If we consider a series of Bernoulli experiments, which observe the occurrence or non occurrence of a given word in a n-gram, the validity of the test hypothesis H_0 compared to the alternative hypothesis H_1 is measured with the value $-2log\lambda$ which has a χ^2 Pearson's distribution. Thus, the higher the value $-2log\lambda$, the higher the hypothesis of independence H_0 does not stand and it is likely that two words x and y are in a semantic relation. [Dunning 1993] defines the value $-2log\lambda$ for two words x,y as in Equation 2.9 where $\log(\Theta; n_i; s_i) = \log \Theta^{s_i}(1-\Theta)^{n_i - s_i}$ and $s_1 = f(x,y)$, $s_2 = f(y) - f(x,y)$, $n_1 = f(x)$, $n_2 = N - f(x)$, $\theta_1 = s_1/n_1$, $\theta_2 = s_2/n_2$, $\theta = f(y)/N$.

$$-2\log\lambda = 2 \times (\log(\theta_1; n_1; s_1) + \log(\theta_2; n_2; s_2) - \log(\theta; n_1; s_1) - \log(\theta; n_2; s_2)). \quad (2.9)$$

As we mentioned before, [Pecina 2006] propose a long list of association measures as well as symmetric and asymmetric attributional word similarities. But, one common characteristic remains between all these similarity measures. They are only defined for two words. Indeed, defining association measures between more than two words has not received much attention. Most proposals tend to use similarity measures between two words following the bootstrapping paradigm to acquire n-ary word semantic relations. However, some studies must be referred in this field such as [Salem 1987] [Chartron 1988] [Bécue 1993] [Frantzi 1996] [Schneider 2000] who propose heuristic-based association measures, mainly in the field of multiword unit extraction. But, the most interesting works within this scope are the ones proposed by [Silva 1999] and [Dias 2000e] who introduce general normalization schemes both for contiguous and non-contiguous n-grams. For the contiguous case, we present the new versions of the SCP and the ME respectively in Equation 2.10 and Equation 2.11 as defined in [Silva 1999] where $w_1, ..., w_n$ is a set of contiguous words.

$$SCP(w_1, ..., w_n) = \frac{P(w_1...w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} P(w_1...w_i).P(w_{i+1}...w_n)}. \quad (2.10)$$

$$ME(w_1, ..., w_n) = \frac{2 \times P(w_1...w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} P(w_1...w_i) + P(w_{i+1}...w_n)}. \quad (2.11)$$

Although dealing with contiguous sequences of words is important for multiword units extraction, highly related non-contiguous sequences of words are interesting as they usually catch long span semantic relations as well they strengthen the extraction process of highly related sequences of words by looking at larger contexts than simply the contiguous contexts of words. Within this scope, we propose in [Dias 2000e] a normalization scheme for positional n-grams (i.e. non-contiguous sequences of words) and evaluate it against all association measures proposed so far in this report. The results show that the ME steadily outperforms all competitive

measures based on the GenLocalMaxs algorithm [Dias 1999a] as the evaluation environment for the extraction of multiword units. This issue will be discussed in Chapter 3. So, we define the ME for the non-contiguous case in Equation 2.12 where $\widehat{S} = p_{11}w_1 \ldots p_{1n}w_n$ is a positional n-gram where p_{1n} corresponds to the relative position of word w_n in relation to word w_1, also referred to as the pivot word.

$$ME(\widehat{S}) = \frac{n \times P(\widehat{S})^2}{\sum_{i_1=1}^{2} \cdots \sum_{i_{(n-1)}=i_{(n-2)}+1}^{n} P(p_{i_1 i_1} w_{i_1} p_{i_1 i_2} w_{i_2} \ldots p_{i_1 i_{(n-1)}} w_{i_{(n-1)}})}. \quad (2.12)$$

Although association measures propose a language-independent framework, they are not adapted to encounter tight semantic relations between words, except for the case of collocations. In fact, they may encounter a large spectrum of semantic relations but can not label them. As a consequence, another approach has been proposed in the field based on the attributional paradigm following the distributional hypothesis proposed by [Harris 1968].

2.1.3 Attributional Word Similarities

The distributional hypothesis is introduced by [Harris 1968] and states that words, which occur in the same contexts tend to have similar meanings. But, the underlying idea that *a word is characterized by the company it keeps* was popularized by [Firth 1957]. In fact, the distributional hypothesis is usually referred to as the attributional word similarity paradigm in the context of computational semantics as it can be assimilated to the vector space model, which is an algebraic model for representing words[10] as context feature vectors. In recent years, the distributional hypothesis has provided the basis for the theory of similarity-based generalization in language learning i.e. the idea that children can figure out how to use words they have rarely encountered before by generalizing about their use from distributions of similar words [Yarlett 2008]. The distributional hypothesis suggests that the more semantically similar two words are, the more distributionally similar they will be in turn. As well as for the question of how children are able to learn language so rapidly given relatively impoverished input, computational modeling also tends to be very sensitive to the data-sparsity problem. Moreover, different studies also highlight the polysemy problem such as [Hindle 1990] and [Freitag 2005]. As a consequence, most of the works propose models with some degree of linguistic analysis to reduce data sparseness and polysemy, with some relevant exceptions [Lund 1995] [Landauer 1997] [Sahlgren 2001] [Terra 2003] [Freitag 2005] [Dias 2006b]. As most approaches differ in (1) the context representation (e.g. window-based, document-based and relation-based), (2) the weighting scheme representing the vector features and (3) the underlying mathematical model, we will first present the works, which use shallow to deep linguistic processing and then introduce

[10] And of course, texts.

language-independent methodologies which suit our initial objectives.

[Hindle 1990] proposes a method to determine the similarity between nouns on the basis of a metric derived from the distribution of ⟨subject, verb, object⟩ triples in a large text corpus. Each noun has a set of verbs that it occurs with (either as subject or object), and for each such relationship, its PMI is computed such as in Equation 2.13 where n is the noun, v is the co-occurring verb and r is the given relation (e.g. subject or object). All pairs of nouns are then compared to each other based on a linear interpolation of subject and object similarities based on a generic heuristic defined for two nouns in any given syntactical relation. The results demonstrate the plausibility of the distributional hypothesis as quasi-semantic classification of nouns is achieved.

$$PMI(\langle n|r\rangle, \langle v|r\rangle) = \log_2 \frac{P(n, v|r)}{P(n|r)P(v|r)}. \tag{2.13}$$

[Grefenstette 1993] implements a similar (but more complete) approach than [Hindle 1990] in the attempt to build a draft of a thesaurus of domain specific nouns. First, more syntactical relations are taken into account to capture relevant context words such as adjectives, appositive nouns and prepositional clauses. Each feature is then evaluated in terms of importance with its associated noun based on an entropy-based heuristic defined in [Grefenstette 1992]. Word-pair similarities are then calculated by the Tanimoto coefficient[11] defined in Equation 2.15, an extension of the cosine similarity measure (see Equation 2.14), between syntactical contexts, extracted after the corpus is morphologically analyzed, part-of-speech tagged and finally parsed. In particular, we suppose that $X_i = (X_{i_1}, X_{i_2}, ..., X_{i_p})$ is a row vector of observations on p variables associated with a label i. The similarity measures between two word vectors X_i and X_j are defined as a generic function $f(X_i, X_j)$ where f is some function of the observed values.

$$cos(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} \times X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2} \times \sqrt{\sum_{k=1}^{p} X_{jk}^2}}. \tag{2.14}$$

$$T(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} \times X_{jk}}{\sum_{k=1}^{p} X_{ik}^2 + \sum_{k=1}^{p} X_{jk}^2 - \sum_{k=1}^{p} X_{ik} \times X_{jk}}. \tag{2.15}$$

The produced entries consist of a number of contextually similar words together with salient verb contexts and few frequent common expressions. In order to reduce noise, [Grefenstette 1993] relies on selection of heuristics that are believed to be optimal, although they do not provide formal evaluation of the resulting resource.

[Weeds 2004] noted that due to the lack of a tight definition for the concept of attributional similarity, many different measures had been studied for variety of applications. In order to better understand the behavior of different similarity

[11] It can be seen as a Jaccard coefficient.

measures, they propose an exhaustive analysis of the statistical and linguistic properties of a set of 2000 distributionally similar words returned by ten different similarity measures. In their work, the co-occurrence types are always grammatical dependency relations so that similarity between nouns is derived from their co-occurrences with verbs in the direct object position and similarity between verbs is derived from their subjects and objects. Each attribute (i.e. verb or noun) is associated to its corpus frequency of co-occurrence, which may be used to form probability estimates. From the set of all tested similarity measures, we particularly emphasis Lin's similarity measure[12] [Lin 1998a] defined in Equation 2.16 for two nouns n_1, n_2 with verb v attributes for any relation r where $A = \{\langle v, r\rangle | \exists(n_1, v, r) \wedge \langle v, r\rangle | \exists(n_2, v, r)\}$, $B = \{\langle v, r\rangle | \exists(n_1, v, r)\}$ and $C = \{\langle v, r\rangle | \exists(n_2, v, r)\}$ and can be seen as the ratio between the amount of information needed to state the commonality of the two words and the total information available about them.

$$Lin(n_1, n_2) = \frac{2 \times \sum_{\langle v,r\rangle \in A} \log_2 P(v|r)}{\sum_{\langle v,r\rangle \in B} \log_2 P(v|r) + \sum_{\langle v,r\rangle \in C} \log_2 P(v|r)}. \tag{2.16}$$

The results from their exhaustive evaluation show that there is a large amount of variation in the neighbors selected by different measures and therefore the choice of a given measure in a given application is likely to be important.

Although many studies deal with relation-based word context vectors to evaluate similarity between words following the distributional hypothesis, some other approaches tackle the problem based on raw texts and as such adopt a window-based approach or a document-based approach. Within this context, [Lund 1995] proposes a vector representation of words based on the vector space model. Each word is associated to a vector of its 200 most frequent words[13] co-occurring in its immediate context of ten words over an excerpt of 160 million words of the Usenet newsgroups. A first visual experiment based on multidimensional scaling (MDS) shows that similar words tend to be projected nearby in the two-dimensional space. A second experiment is proposed, which calculates the Euclidean distance between frequency vectors as well as they study the direction of co-occurrences and confirm previous findings that contextual similarity estimates for semantically related words, (e.g. *table* and *bed*) are higher than for associatively related words, (e.g. *coffee* and *cup*) or unrelated ones.

[Landauer 1997] report results with a high-dimensional linear associative model, the Latent Semantic Analysis (LSA), which allows generalization improvements through compact representations of feature vectors. The input to LSA is a matrix of unique words by many individual texts, where a cell contains the number of times a particular word appears in a particular text. After an initial transformation

[12] We will use it in Chapter 6.

[13] Optimal trade-off between completeness and computational cost.

of the cell entries[14], the matrix is analyzed by a statistical technique called Singular Value Decomposition (SVD) closely akin to factor analysis, which allows words and texts to be re-represented as points or vectors in a high dimensional abstract space. The final output is a representation from which one can calculate word context similarities. From the perspective of word similarity, [Landauer 1997] use the cosine measure defined in Equation 2.14. In particular, they test their methodology against 80 test cases[15] from the TOEFL[16] noun synonym portion and report results of 64% correct answers, which is comparable to human guessing from a large sample of applicants to United States colleges. It is important to notice that these results show that the synonyms tend to co-occur in the same documents more often than by chance. This issue will be discussed further in Chapter 6. Similarly, [Sahlgren 2001] proposes a methodology, which uses random indexing of words in narrow context windows to calculate semantic context vectors for each word in text data. In particular, he builds a word-text matrix and reduces it using a technique called random indexing, which outperforms LSA in terms of compression of the resulting matrix. The results are then compared on the same 80 TOEFL test cases and reach 72%.

[Terra 2003] and [Freitag 2005] propose exhaustive evaluations based on different similarity measures and weighting schemes. In particular, [Freitag 2005] propose to test the similarity measure defined in [Ehlert 2003] based on the conditional probability and presented in Equation 2.17. It is important to notice that we keep the same notation as for relation-based similarity measures. Indeed, in this case, the relation between two words, noted r, stands to the fact that both words co-occur in the same document or the same word context window.

$$P(x|y,r) = \frac{P(x,y|r)}{P(y|r)}. \tag{2.17}$$

So, the Ehlert similarity measure (also known as the confusion probability) is defined between two words x and y in Equation 2.18 where $A = \{\langle z,r\rangle | \exists(x,z,r) \wedge \langle z,r\rangle | \exists(y,z,r)\}$.

$$Ehlert(x\|y) = \sum_{\langle z,r\rangle \in A} \frac{P(x,z|r)P(y,z|r)P(z|r)}{P(y)}. \tag{2.18}$$

It is important to notice that the Ehlert similarity measure is asymmetric but it is used in a symmetric way for the sake of the resolution of TOEFL test cases. Indeed, in this case, the target word is always taken as the second argument of the measure and the similarities are evaluated such as $Ehlert$(decoy word‖target word). In their experiments, [Freitag 2005] show that the Ehlert similarity measure, in a

[14]The frequency is changed in a similar way as [Grefenstette 1992] so that the value of each cell is transformed into $\log(1 + \text{cell frequency})/(\text{entropy of the word over all contexts})$.

[15]Each test consists of a problem word and four alternative words (decoys) from which the test taker is asked to choose that with the most similar meaning to the stem.

[16]Test Of English as a Foreign Language.

window-based context, outperforms all other combinations of weighting schemes and similarity measures up-to-then for the 80 TOEFL test cases reaching 82%. However, it drops to 67.6% over a total of 23570 test questions created by the authors (WBST). For the purpose of their evaluation, they also introduce a well-known weighting scheme in information retrieval, which reaches comparable results to the Ehlert measure over the WBST. In fact, it is common to evaluate the cosine measure between two vectors, where attributes are not raw frequency counts, but counts weighted using some version of the inverse document frequency (IDF). Within this context, they propose to weight each word based on a modified version of the IDF as defined in Equation 2.19 where $\langle y, r\rangle$ is a given attribute and N the set of all the words in the corpus.

$$IDF(\langle y, r\rangle) = \log_2 \frac{card(N)}{card(\{x_i \in N | \exists(x_i, y, r)\})}. \tag{2.19}$$

Although many attributional word similarity measures have been proposed so far, all of them are based on the assumption that two words are semantically related if they share common contexts. In practice, this has been interpreted as the fact that two words should contain as many relevant common words as possible, to be declared semantically related. However, due to data-sparseness, it is usually difficult to find overlapping contexts even in huge corpora [Terra 2003]. Moreover, natural languages are particularly ambiguous which may imply that a given word context embodies many word meanings inside. To leverage these issues, we introduced the InfoSimba similarity measure (IS) in [Dias 2006b] within the context of unsupervised learning of lexical-semantic resources. The main idea, which is somehow shared by the generalized vector space model [Wong 1985], is based on a loose definition of context similarity i.e. two words are semantically related if they contain as many relevant related words as possible between contexts. For example, *student*, *pupil* and *educatee* are considered as synonyms in WordNet but their context vectors are unlikely to share many common words as shown in Table 2.2. This list was obtained by our meta-search engine VIPACCESS[17], gathering from the returned Web snippets[18] five of the most relevant context words for each of the three target words considered as query terms.

Target Word	Context Words
student	college, information, loans, education, university
pupil	dictionary, association, teacher, transportation, supervision
educatee	education, student, institution, school, college

Table 2.2: Five relevant context words.

Although the three words almost do not share any common words, their context words are highly correlated. The IS is based on this idea of measuring the corre-

[17] http://193.136.67.141:82/vipaccessnew/vipaccess.aspx [14th $July$, 2010].
[18] More details will be given about VIPACCESS in Chapter 4.

lations between all the pairs of words existing between two word context vectors instead of just relying on their exact match as with the cosine similarity measure. For instance, comparing *student* and *pupil* based on the cosine similarity would lead to a maximum unrelatedness as their contexts do not share any words. On the contrary, the InfoSimba would introduce the level of relatedness between *college* and the five other context words of *pupil* (i.e. *dictionary*, *association*, *teacher*, *transportation*, *supervision*) and then between *information* and all other context words of *pupil* and so on and so forth, thus guaranteeing to some extent a semantic similarity. It is defined in Equation 2.20 where $S(.,.)$ is any symmetric similarity measure and each W_{ij} corresponds to the attribute word at the jth position in the vector X_i.

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.S(W_{ik}, W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.S(W_{ik}, W_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.S(W_{jk}, W_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.S(W_{ik}, W_{jl})\end{array}\right)}. \tag{2.20}$$

As computation of the IS may be hard due to orders of complexity, some experiments have been made in [Cleuziou 2008] with the simplified version of the IS, which we define as the simplified IS, $ISs(.,.)$, in Equation 2.21.

$$ISs(X_i, X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.S(W_{ik}, W_{jl}). \tag{2.21}$$

As defined in Equation 2.20, the IS can be seen as a second order similarity measure. However, a direct application of the definition of the IS can turn it into a N order similarity measure as we propose in [Cleuziou 2008]. Indeed, the IS can be defined recursively as in 2.22, which we call the recursive InfoSimba (RIS), where the initialization is based on the initial version of the IS i.e. $RIS_0(X_i, X_j) = IS(X_i, X_j)$. We also define its simplified version $RISs_N(.,.)$ in 2.23 with the following initialization $RISs_0(X_i, X_j) = ISs(X_i, X_j)$.

$$RIS_N(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RIS_{N-1}(W_{ik}, W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.RIS_{N-1}(W_{ik}, W_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jl}.RIS_{N-1}(W_{jk}, W_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RIS_{N-1}(W_{ik}, W_{jl})\end{array}\right)}. \tag{2.22}$$

$$RISs_N(X_i, X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.RISs_{N-1}(W_{ik}, W_{jl}). \tag{2.23}$$

Small experiments are developed in [Cleuziou 2008] to compare the IS and the RIS based on the classification of similar words by deciles. For that purpose, 38 keywords are gathered from three different conferences (LREC 2002, WWW 2002 and JSAI 1997). Each keyword is then associated to a context vector of its 10 most relevant related words calculated with the Web version of the PMI (i.e. frequency counts are defined as document hits) over a given vocabulary specific of the English language.

A comparison is then made between the SCP and the PMI association measures, the IS with the SCP and PMI as similarity measures with no weighting schemes and a document-based context[19], and finally the RIS with the same initial set-up. The results in Figure 2.1 show that the number of correct decisions increases as well as it shifts to the higher deciles from the simplest measure i.e. the single PMI to the more sophisticated measure i.e. the RIS[20].

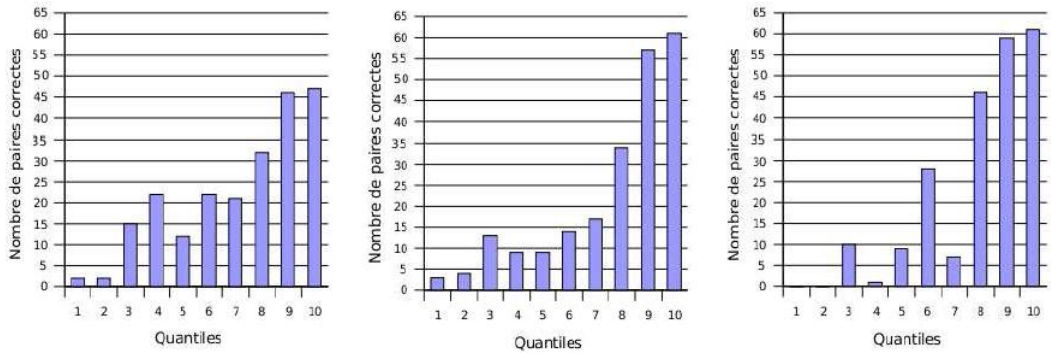

Figure 2.1: Classification of correct answers by decile with (1) the PMI, (2) the IS with PMI and (3) the RIS with the PMI after one iteration.

Although the results are interesting, they are based on a small data set and further investigations will be performed in the near future. Moreover, one remark needs to be made relatively to the RIS, which outperforms all other measures both in the case of the PMI and the SCP. This result is only obtained after the first iteration i.e. RIS_1 and then decreases steadily to reach a limit value. This issue has not yet been studied but is clearly an interesting behavior worthy to analyze.

2.1.4 Knowledge-based Word Similarities

As one of the main purposes of our research is to build lexical-semantic knowledge bases from corpora, our work has not been following the knowledge-based approach. Indeed, we deeply believe that different corpus-based methods can lead to successful results as corpus evidence may fit the semantic component neatly and directly, which will never be possible with general-purpose resources. However, it is interesting to understand the underlying ideas of knowledge-based word similarity measure as some of them will be used in our evaluation schemes, especially in Chapter 6.

Since a taxonomy is often represented as a hierarchical structure, which can

[19]Both the SCP and the PMI are evaluated on Web hits instead of traditional corpus frequencies.

[20]We only show the results with the PMI as similar results were obtained with the SCP with a slight advantage for the PMI.

be seen as a special case of a network structure, evaluating semantic similarity between nodes in the network can make use of the structural information embedded in the network. There are several ways to determine the conceptual similarity of two words in a hierarchical semantic network. Topographically, they can be categorized as node-based and edge-based approaches, which respectively correspond to the information content approach and the conceptual distance approach.

Within a multidimensional space in which a node represents a unique concept with a certain amount of information, and an edge represents a direct association between two concepts, the similarity between two concepts within the information content approach is the extent to which they share information in common. Within this context, many works have been proposed such as [Resnik 1992] [Resnik 1995] [Richardson 1995] [Jiang 1997] [Lin 1998b] [Banerjee 2002]. The most famous piece of work is certainly the one presented by [Lin 1998b] who proposes an information-theoretic definition of similarity defined in Equation 2.24 where x_1 and x_2 are two terms belonging to two concepts C_1 and C_2 respectively and C_0 is the most specific concept that subsumes both C_1 and C_2 in the structure, and $P(C_i)$ is the probability that a randomly selected term belongs to the concept C_i. This measure has already been presented in Equation 2.16 but in a different environment.

$$Lin(x_1, x_2) = \frac{2 \times logP(C_0)}{logP(C_1) + logP(C_2)}. \tag{2.24}$$

The conceptual distance approach is a more natural and direct way of evaluating semantic similarity in a taxonomy. It estimates the distance (e.g. edge length) between nodes which correspond to the concepts/classes being compared. Given the multidimensional concept space, the conceptual distance can conveniently be measured by the geometric distance between the nodes representing the concepts. Obviously, the shorter the path from one node to the other, the more similar they are. Within this context, many researches have also been proposed such as [Rada 1989] [Wu 1994] [Hirst 1998b] and [Leacock 1998]. In particular, we will mention the work done by [Hirst 1998b] in Chapter 5 for the construction of lexical chains.

2.1.5 Relational Word Similarities

Although this paradigm is out of the scope of our work, we give an overview of the method to propose a complete summary of word similarity approaches. [Turney 2006] proposes to study the relational similarity paradigm, which evaluates the correspondence between word relations, which have been proposed by [Medin 1990]. When two pairs of words have a high degree of relational similarity, it can be said that they are analogous. For example, verbal analogies are often written in the form A:B::C:D, meaning that A is to B as C is to D. As such, *traffic*:*street*::*water*:*riverbed* are analogous with respect to the verbs *flow* and *carry* as *traffic flows over a street*, *water flows over a riverbed* and *a street carries traffic*

and *a riverbed carries water*. Within this context, [Turney 2006] defines the latent relational analysis (LRA), which extends the vector space model approach of [Turney 2005] in three ways: (1) the connecting patterns are derived automatically from the corpus, instead of using a fixed set of patterns, (2) singular value decomposition is used to smooth the frequency data and (3) given a word pair such as *traffic*:*street*, LRA considers transformations of the word pair by replacing one of the words by synonyms, such as *traffic*:*road* or *traffic*:*highway*. Basically, the LRA evaluates the similarity between two pairs of words $\langle A : B \rangle$ and $\langle C : D \rangle$ based on the cosine value between their context vectors, where each attribute is the number of times[21] both words co-occur within a given pattern[22]. So, both words are considered analogous if their pattern similarity is high. As a consequence, LRA can be seen as a mix between the pattern-based approach and the distributional hypothesis, as the feature vectors are based on word frequency within a given pattern. As a consequence, it suffers from both drawbacks i.e. ambiguity of patterns and word polysemy. Moreover, LRA relies on a broad-coverage thesaurus of synonyms, which limits its application to any other language and somehow biases the results as they depend on general-purposes external resources (e.g. WordNet [Miller 1990]), which may not neatly and directly embody the semantic component. Lately, an extension of the LRA has been proposed in [Turney 2008], but does not solve any of the drawbacks of the previous model.

2.2 Asymmetric Word Similarities

Most of the metrics, which evaluate the degree of similarity between words are symmetric [Pecina 2006] [Tan 2004], except perhaps pattern-based similarities[23]. But, new trends have recently emerged with the study of asymmetric measures [Michelbacher 2007]. The idea of an asymmetric measure is inspired by the fact that within the human mind, the association between two words or concepts is not always symmetric. For example, as stated in [Michelbacher 2007], "*there is a tendency for a strong forward association from a specific term like adenocarcinoma to the more general term cancer, whereas the association from cancer to adenocarcinoma is weak*". According to [Michelbacher 2007], this idea bears some resemblance to the prototype theory [Rosch 1973], where objects are regarded as members of different categories. Some members of the same category are more central than others making them more prototypical of the category they belong to. For instance, *cancer* would be more central than *adenocarcinoma*. However, we deeply believe that the main background for the direction of association lies in the notion of specific and general terms. Indeed, it is clear that there exists a tendency for a strong forward association from a specific term to the more general term but the backwards association is weaker. Within this scope, seldom new researches have been emerging

[21] Actually, the logarithm of this frequency is used.

[22] Which can be automatically extracted as shown in [Bollegala 2007].

[23] We take the party to classify pattern-based measures as symmetric as they can embody the synonymy relation, but they could also be classified as asymmetric measures.

over the past few years, which propose the use of asymmetric similarity measures, which we believe can lead to great improvements in the acquisition of word semantic relations as shown in [Dias 2008] or [Cleuziou 2010].

2.2.1 Asymmetric Association Measures

Pattern-based measures can embody asymmetry as they were initially defined to discover the hypernymy/hyponymy relation. But, [Ohshima 2009] is certainly the approach that makes the most of asymmetric patterns. Indeed, instantiating and sending to a search engine a number of patterns filled only with one possible candidate may guarantee the extraction of hypernymy/hyponymy or meronymy/holonymy relations if asymmetric patterns exist. However, we know that pattern-based measures are sensitive to word polysemy and pattern ambiguity. Moreover, they are language-dependent techniques which are difficult to replicate for different languages.

In order to keep language-independency and to some extent propose unsupervised methodologies, different works propose to use asymmetric association measures listed in [Pecina 2006] and [Tan 2004] in the domain of taxonomy construction [Sanderson 1999] [Cleuziou 2010][24], cognitive psycholinguistics [Michelbacher 2007] and word order discovery [Dias 2008].

In the domain of taxonomy construction, [Sanderson 1999] is certainly one of the first studies to propose the use of the conditional probability as defined in Equation 2.17, where r stands for the document-based paradigm, to build taxonomies from raw texts. They assume that a term t_2 subsumes a term t_1 if the documents in which t_1 occurs are a subset of the documents in which t_2 occurs constrained by $P(t_2|t_1) \geq 0.8$ and $P(t_1|t_2) < 1$. By gathering all subsumption relations, they build the semantic structure of any domain, which corresponds to a directed acyclic graph (DAG). In [Sanderson 2000], the subsumption relation is relieved to the following expression $P(t_2|t_1) \geq P(t_1|t_2)$ and $P(t_2|t_1) > t$ where t is a given threshold and all term pairs found to have a subsumption relationship are passed through a transitivity module, which removes extraneous subsumption relationships in the way that transitivity is preferred over direct pathways, thus leading to a non-triangular DAG.

[Michelbacher 2007] propose two different measures to model the notion of asymmetric association. Their intent is to determine to what extent these two measures of directed association can be used as a model for directed psychological association in the human mind. These two measures are the plain conditional probability and the ranking measure $R(.\|.)$ based on the Pearson's χ^2 test. In particular, $R(t_2\|t_1)$ returns the rank of t_2 in the association list of t_1 given by the order obtained with the Pearson's χ^2 test for all the words co-occurring with

[24] We will develop this our work in Chapter 6.

t_1. So, when comparing $R(t_2\|t_1)$ and $R(t_1\|t_2)$, the smaller rank indicates the strongest association. The results were evaluated against the results of a large number of free association tasks carried out with human subjects and they found that the new measures were able to distinguish between highly symmetric and highly asymmetric pairs to some extent, but the overall accuracy in predicting the degree of asymmetry was low.

In the specific domain of word order discovery, [Dias 2008] propose a methodology based on directed graphs and the TextRank algorithm [Mihalcea 2004] to automatically induce a general-specific word order for a given vocabulary based on Web corpora frequency counts. For that purpose, we use seven different asymmetric association measures to build an asymmetric word-word matrix for a given vocabulary. A directed graph is obtained by keeping the edge, which corresponds to the maximum value of the asymmetric association measure between two words. Then, the TextRank is applied and produces an ordered list of nouns, on a continuous scale, from the most general to the most specific. Experiments were conducted based on the WordNet noun hierarchy and assessed 65.69% of correct word ordering. As this work will be detailed in Chapter 6, we present the seven asymmetric association measures used in this work: the Braun-blanket (Equation 2.25), the J-measure (Equation 2.26), the Laplace (Equation 2.27), the conviction (Equation 2.28), the certainty factor (Equation 2.29), the added value (Equation 2.30) and the conditional probability (Equation 2.17).

$$BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\bar{x},y)}. \tag{2.25}$$

$$JM(x\|y) = P(x,y) \times \log \frac{P(x|y)}{P(x)} + P(\bar{x},y) \times \log \frac{P(\bar{x}|y)}{P(\bar{x})}. \tag{2.26}$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2}. \tag{2.27}$$

$$CO(x\|y) = \frac{P(x) \times P(\bar{y})}{P(x,\bar{y})}. \tag{2.28}$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)}. \tag{2.29}$$

$$AV(x\|y) = P(x|y) - P(x). \tag{2.30}$$

2.2.2 Asymmetric Attributional Word Similarities

In the context of asymmetric attributional word similarities, researches such as [Lund 1995] and [Freitag 2005] study the directions of co-occurrences but do not propose any solution nor methodology to take into account this phenomenon. They just accept the fact that asymmetries exist between word attractions. It is strange

enough to notice that no further studies have been carried out following these assumptions, at least to our knowledge. Moreover, [Freitag 2005] show that improved results are reached with the asymmetric Ehlert similarity measure defined in Equation 2.18. Other asymmetric distributional similarity measures also exist such as the Kullback Leibler divergence [Kullback 1951] defined in Equation 2.31 where $A = \{\langle z, r\rangle | \exists(x, z, r) \wedge \langle z, r\rangle | \exists(y, z, r)\}$, which has been regularly set apart from the Jensen Shannon divergence [Menéndez 1997], its symmetric counterpart. We can also point at the cross entropy described in [Pecina 2006].

$$KL(x\|y) = \sum_{\langle z,r\rangle \in A} \log P(z|x) \times \frac{\log P(z|x)}{\log P(z|y)}. \tag{2.31}$$

Although there are many asymmetric similarity measures, they evidence problems that may reduce their impact. On the one hand, asymmetric association measures can only evaluate the generality/specificity relation between words that are known to be in a semantic relation such as in [Sanderson 1999] and [Dias 2008]. Indeed, they generally capture the direction of association between two words based on document contexts and only take into account a loose semantic proximity between words. For example, it is highly probable to find that *Apple* is more general than *iPad*, which can not be assimilated to an hypernymy/hyponymy or meronymy/holonymy relation. On the other hand, asymmetric attributional word similarities only take into account common contexts to assess the degree of asymmetric relatedness between two words. To leverage these issues, we introduce the asymmetric InfoSimba similarity measure (AIS), which underlying idea is to say that one word x is semantically related to word y and x is more general than y, if x and y share as many relevant related words as possible between contexts and each context word of x is likely to be more general than most of the context words of y. The AIS is defined in Equation 2.32, where $AS(.\|.)$ is any asymmetric similarity measure, exactly in the same way as for the IS in Equation 2.20 where $S(.,.)$ stands for any asymmetric similarity measure. We also define its simplified version $AISs(.\|.)$ in 2.33.

$$AIS(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}\|W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.AS(W_{ik}\|W_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.AS(W_{jk}\|W_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}\|W_{jl})\end{array}\right)}. \tag{2.32}$$

$$AISs(X_i\|X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}\|W_{jl}). \tag{2.33}$$

Similarly to the case of the RIS, we turned the AIS into a N order similarity measure by proposing its recursive definition as in Equation 2.34, which we call the recursive asymmetric InfoSimba similarity measure (RAIS), where the initialization is based on the initial version of the AIS i.e. $RAIS_0(X_i\|X_j) = AIS(X_i\|X_j)$. We also define its simplified version $RAISs_N(.\|.)$ in 2.33 with the following initialization $RAISs_0(X_i\|X_j) = AISs(X_i\|X_j)$.

$$RAIS_N(X_i \| X_j) = \frac{\sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{jl}.RAIS_{N-1}(W_{ik} \| W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{il}.RAIS_{N-1}(W_{ik} \| W_{il}) + \\ \sum_{k=1}^{p} \sum_{l=1}^{p} X_{jk}.X_{jl}.RAIS_{N-1}(W_{jk} \| W_{jl}) - \\ \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{jl}.RAIS_{N-1}(W_{ik} \| W_{jl}) \end{array} \right)}. \quad (2.34)$$

$$RAISs_N(X_i \| X_j) = \frac{1}{p^2} \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{jl}.RAISs_{N-1}(W_{ik} \| W_{jl}). \quad (2.35)$$

The work on word similarity has been one of the main focuses of our research during the last ten years as we believe that extracting intrinsic knowledge about the language can largely improve the understanding of the implicit information conveyed in texts. Besides, all informative attributional similarity measures presented in this chapter can successfully be extended to deal with text similarity as we will see in Chapters 4, 5 and 6 with the works developed in [Dias 2006b] [Dias 2007a] as well as other on-going studies. However, future works still need to be carried out to assess the success of the proposed word similarity measures. Moreover, new research directions have been emerging, which must clearly be followed.

2.3 Future Work

The work presented in [Cleuziou 2008] is a first experiment, which assesses the potential of the informative paradigm within attributional similarity measures. Nevertheless, an exhaustive evaluation must be carried out taking into account all alternative informative similarity measures such as the $IS(.,.)$ or the $RIS_N(.,.)$ and their respective simplified versions the $ISs(.,.)$ and the $RISs_N(.,.)$ with different symmetric similarity measures $S(.,.)$ and not only association measures. Only a large scale evaluation as proposed in [Terra 2003] [Freitag 2005] [Heylen 2008] will legitimate the success of the informative paradigm. In [Dias 2010], we propose a new automatic methodology to extract TOEFL-like test cases within a language-independent environment. As such, we are able to compute TOEFL-like test cases for any language by identifying likely interchangeable words from aligned paraphrases[25]. This framework will serve as the basis of our exhaustive evaluation, which will allow to test different scenarios such as different word similarity measures within window-based, document-based or relation-based context representations.

But, as evidenced in [Heylen 2008] and [Dias 2010], likely interchangeable words may embody different semantic roles and not only synonymy. For instance, hypernymy/hyponymy, meronymy/holonymy or co-sibling relations are likely to be identified. In the specific case of hypernymy/hyponymy or meronymy/holonymy, there exists the implicit notion of generality. Within this scope, we state that

[25] More details are given in Chapter 6.

general words tend to co-occur with other general words while more specific ones tend to co-occur with other specific words. So, if all general words are semantically related to all specific words, we are likely to discover tight semantic relations, which imply some idea of semantic direction i.e. hypernymy/hyponymy or meronymy/holonymy relations. From this assumption, we may also test in our exhaustive evaluation the $AIS(.\|.)$ and the $RAIS_N(.\|.)$ and their respective simplified versions the $AISs(.\|.)$ and the $RAISs_N(.\|.)$ with different asymmetric similarity measures $AS(.\|.)$ to discover asymmetric word relations. We also plan to introduce these measures in the construction of lexical-semantic resources based on the Pretopology paradigm. This issue will be detailed in Chapter 6.

Finally, most works proposed so far analyze word similarities based on word co-occurring feature vectors, following Harris' distributional analysis. One interesting exception is the Knowledge-based approach, which takes into account pre-existing lexical-semantic knowledge bases to evidence word similarities. However, pre-built lexical-semantic knowledge bases do not exist for the vast majority of languages, thus deeply limiting their scope of action. In order to overcome this drawback, we propose a new research direction, which consists in building word categories feature vectors, based on ephemeral clustering (see Chapter 4). The idea is that two words are similar if they share common or related semantic categories (or ephemeral clusters). In particular, this process can be made easier if query disambiguation is possible as evidenced by the $HISGK$-means algorithm proposed in Chapter 4 and defined in algorithm 5.

It is interesting to see that most methodologies only take into account single words and leave apart phrases to account for word similarity. We deeply believe that the identification of multiword units or phrases is likely to improve word similarity as we show in [Grigonyté 2010]. This is clearly an issue that we need to take into account in future works. As such, in the next chapter, we propose different strategies to identify and extract multiword units ■

Chapter 3

Multiword Units Extraction

Contents

3.1 Related Work 34
3.2 Statistical Multiword Units Extraction 38
3.3 Efficient Multiword Units Extraction 42
3.4 Learning Multiword Units Extraction 47
3.5 Hybrid Multiword Units Extraction 52
3.6 Future Work 55

Understanding the implicit information present in texts about the language itself is one of the most important contributions to natural language learning. Within this context, multiword units (MWU) identification is a crucial issue towards understanding the essence of the message conveyed by any given text. In particular, the identification of MWU proved to improve discourse representation [Dias 2006b], topic segmentation [Dias 2007a], information retrieval [Dias 2009] and textual entailment [Grigonyté 2010] as well as other important research areas such as machine translation [Bilal 2005] or parsing [Nivre 2004]. Most of these improvements are due to the better understanding of texts and as a consequence the better evaluation of text and word similarities. Indeed, most approaches to evaluate text similarities represent texts as bags of words i.e. lists of unordered words. For instance, let's take both sentences (1) and (2).

1. *le ballon d'eau chaude est plein (the boiler is full)*
2. *le ballon est plein d'eau chaude (the ball is full of hot water)*

Although, they contain exactly the same words, their meanings are clearly different as *ballon d'eau chaude* is the French compound word for *boiler*. However, the vector space model [Salton 1975] would yield to a maximum similarity value equal to 1, giving them the exact same meaning as their representation would be exactly the same i.e. [*ballon, chaude, d', eau, est, le, plein*]. In fact, the misinterpretation of the message conveyed by a given text can lead to the miscalculation of text similarities.

Within the context of word similarities, studies based on the distributional analysis paradigm proved that the encountered word semantic relations consist of

contextually similar words (i.e. words in tight semantic relations such as synonymy or words in loose semantic relations such as the usually called *related-to* relation, which can associate *doctor* and *hospital* for example), salient verb contexts and few frequent multiword units [Grefenstette 1993] [Dias 2010]. The attributional paradigm has mainly been introduced to evidence the synonymy relation. As a consequence, the extraction of multiword units is a failure of the approach. This situation can easily be understood. Let's take the well-known Jamaican one hundred meter runner *Usain Bolt*. By taking the respective individual word context vectors of *Usain* and *Bolt*, they are likely to be very close to each other, even exactly the same. So, based on the distributional analysis paradigm, they are likely to be declared synonyms. In fact, *Usain Bolt* is a proper name, which is a sub-type of multiword units, which should be identified as a unique lexical unit in texts as it conveys a unique mental message. We will call this process, the normalization of the corpus. All along this chapter, we will use different examples from different languages to assess that this linguistic phenomenon is transversal to most languages. For example, in Portuguese, a similar example is given by *baba de camelo* literally *camel drivel*, which is a traditional dessert, where both individual words *baba* and *camelo* are unlikely to be encountered in different contexts and are likely to be erroneously identified as synonyms. The same can be found in Bulgarian with the celebration of the 1st of March called *baba marta*.

By definition, MWU are words that co-occur together more often than they would by chance in a given domain and usually convey conceptual information [Dias 2002]. For example, *tomber dans les pommes* (*to faint*) is a sequence of words which meaning is non-compositional i.e. it can not be reproduced by the sum of the meanings of its constituents and thus represents a typical MWU. Multiword units include a large range of linguistic phenomena as stated in [Gross 1996], such as compound nouns (e.g. *chantier naval* meaning in French *shipyard*), phrasal verbs (e.g. *entrar em vigor* meaning in Portuguese *to come into force*), adverbial locutions (e.g. *sans cesse* meaning in French *constantly*), compound determinants (e.g. *un tas de* meaning in French *an amount of*), prepositional locutions (e.g. *au lieu de* meaning in French *instead of*), adjectival locutions (e.g. *a longo prazo* meaning in Portuguese *long-term*) and institutionalized phrases (e.g. *con carne*).

For several years, we have been proposing different contributions, which contribute to the state-of-the-art of MWU extraction and offer new perspectives towards the combination of word statistical relatedness and proper linguistic structures of MWU. We first developed the Software for the Extraction of N-ary Textual Associations (SENTA) [Dias 1999a], which is parameter free and language-independent thus allowing the extraction of MWU from any raw text. It is based on the Mutual Expectation measure defined in Equation 2.12 and the GenLocalMaxs selection algorithm, which does not depend on any threshold. SENTA shows many advantages compared to different methodologies presented so far. It is parameter free, thus avoiding threshold tuning. It can extract relevant sequences

of characters, thus allowing its application to character-based languages and the study of word-based languages on the character level as we show in [Dias 2000d] [Dias 2000c] [Ribeiro 2001]. And, interestingly, it obtains successful results for small texts as it extracts MWU with low frequency with great accuracy without using lists of stop-words or stemming.

Most of the proposed methodologies have mainly been used to build large lexical databases from large corpora. As a consequence, as processing time was not a crucial issue, very few efforts have been made to design efficient algorithms. However, due to the fact that SENTA deals with small amounts of texts from which it is capable to discover low frequency MWU with high accuracy (unlike other models), it was important to adapt it to real-world real-time environments. As a consequence, we proposed an efficient algorithm based on suffix-arrays [Gil 2003b], which improved applications in information retrieval systems such as in [Campos 2005] [Dias 2009]. To reach higher improvements on processing time, we also worked on a parallel version of SENTA [Pereira 2004], which takes advantage of its structure by dividing the overall corpus into several sub-corpora. The algorithm is implemented using the single program multiple data programming methodology and shows that optimum efficiency is obtained for 4 processors.

However, some studies showed that syntactical structure plays an important role in the extraction of MWU. But, all of them apply a two step process. Most of them manually define specific syntactical patterns and then apply association measures to select the best candidates or reduce the search space with association measures and then apply well-known syntactical patterns to select MWU candidates. To answer to the statement made in [Habert 1993], who report insufficiencies of this approach, we proposed to evaluate both syntactical and word tightness over part-of-speech tagged corpora by introducing the combined association measure [Dias 2003] and designed an overall architecture for MWU extraction called Hybrid Extraction of Lexical Associations (HELAS). HELAS uses basic shallow linguistic tools, obtains high levels of accuracy for 3-grams and integrates into a single measure both steps that were disassociated so far.

Within the context of machine learning, it is possible to take advantage of many different clues to decide whether a sequence of words is a potential MWU or not. However, all approaches proposed before the last decade rely on a single statistical measure although many are tested and could improve the extraction process. Indeed, as stated in [Pecina 2006], the association scores computed by different measures cannot be compared directly as they provide different results. For example, it is well-known that the PMI tends to favor rare events, while the log-likelihood ratio extracts more frequent events. To overcome these potential drawbacks, we proposed to combine evolutionary computing and attributional similarity measures to extract MWU from unrestricted texts in [Dias 2001b]. For that purpose, a fitness function is defined whose maximization serves as a basis for the identification of pertinent

word n-grams based on different similarity measures. Thus, we integrate in a reinforcement learning environment clues such as the frequency of the word sequence, its degree of cohesion or its marginal frequency (to name but a few attributes) in order to decide whether a given n-gram must be selected as a potential MWU or not.

In the next sections, we will briefly detail the related work and concentrate on our contributions to the state-of-the-art for MWU extraction in a different form from Chapter 2. Indeed, Chapter 2 is an exception of this thesis in the way that it is deeply developed and detailed. This decision is based on the fact that word (and text) similarities are central issues of our work. As a consequence, the following chapters will introduce our contributions to the respective states-of-the-art but in a lighter way.

3.1 Related Work

Syntactical, statistical, hybrid syntactic-statistical, semantic and machine learning methodologies have been proposed to extract MWU. Although, there exists an important number of approaches, the identification of MWU still remains an open problem within an active research field. Historically, both syntactical and statistical approaches have been privileged. Purely linguistic systems follow the first part of the definition of MWU proposed in [Choueka 1983]: a multiword unit *is defined as a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be directly derived from the meaning or connotation of its components.* Based on the specific syntactical structures embodied by MWU, [David 1990] [Bourigault 1993] [Dagan 1994] propose to extract relevant MWU using techniques that analyze specific syntactical structures in texts (see Expression 3.1). Thus, sequences of words, which match a small set of syntactical patterns defined by regular expressions are searched in part-of-speech tagged corpora and considered as MWU. The set of syntactical patterns is usually adapted to a specific domain by experts and most of them only match sequences of nouns, which seems to yield the best hit rates in most environments. More recently, [Debusmann 2004] proposes a more sophisticated framework to model multiword expressions as dependency subgraphs, using the grammar formalism of Extensible Dependency Grammar (XDG). In particular, he extends XDG to lexicalize dependency subgraphs and shows how to compile them into simple lexical entries. However, these methodologies suffer from their monolingual basis as the systems require highly specialized linguistic analyzes to identify clues that isolate possible MWU candidates and as a consequence are hardly transposable to other languages or domains.

Purely statistical methodologies are based on the definition given by [Smadja 1993] who states that MWU *are recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usage.* So,

[Church 1990] [Gale 1991] [Smadja 1993] [Dunning 1993] [Shimohata 1997] extract discriminating MWU from text corpora by means of association measure regularities, respectively the PMI (see Equation 2.1), the Φ^2 (see Equation 2.7), the t-score, the log-likelihood ratio (see Equation 2.9) and the entropy. One very interesting work is also proposed by [Tomokiyo 2003] who use the Pointwise Kullback Leibler divergence 2.31 between multiple language models for scoring both phraseness and informativeness, which can be unified into a single score to rank extracted digrams. However, while [Church 1990] [Gale 1991] [Dunning 1993] [Tomokiyo 2003] only extract digrams (i.e. sequences of two words), [Smadja 1993] and [Shimohata 1997] propose a bootstrapping methodology, which consists in encountering recursively relevant digrams and as a consequence may extract MWU of any size. The bootstrapping methodology shows an interesting solution to extract relevant long MWU, but depends on the first iteration of the methodology. Indeed, if erroneous digrams are retrieved, longer sequences of words are unlikely to embody correct MWU. To solve this problem, different approaches have been proposed such as [Salem 1987] [Chartron 1988] [Kim 1994] [Frantzi 1996] [Maynard 1999] [Schneider 2000] who define n-ary association measures so that relevant n-grams can directly be extracted as MWU candidates. Following this idea, [Dias 1999a] is certainly the most complete work as it leads to improved results compared to most other approaches as shown in [Dias 2000e] based on the GenLocalMaxs selection algorithm. As they use plain text corpora and only require the information appearing in texts, such approaches are highly flexible and extract relevant compound words independently of the domain and the language of the input text. However, these methodologies can only identify textual associations in the context of their usage. As a consequence, many relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures and further linguistic treatment is necessary as explained in [Dias 2001a] and [Dias 2005c].

To solve some of the problems presented by purely syntactical or statistical methodologies, the hybrid syntactic-statistical approach defines co-occurrences of interest in terms of syntactical patterns and statistical regularities. Some studies first apply syntactical patterns and then apply association measures to extract potential candidates, while others prefer to reduce the search space by using association measures and then filtering the MWU candidates, which embody given syntactical patterns. Within the first paradigm, many studies can be listed such as [Enguehard 1993] [Justeson 1993] [Daille 1996] [Feldman 1998] [Bannard 2007] [Fazly 2007]. The basic idea is to apply syntactical patterns, mostly noun patterns, which produce a list of MWU candidates, which are then filtered by association measures. The different studies only vary on the degree of syntactical analysis and the used association measures. A typical syntactical pattern can be found in Expression 3.1 extracted from [Justeson 1993]. Some other approaches adopted the opposite scheme i.e. reducing the search space by applying any statistical technique and then selecting the MWU, which fit specific syntactical patterns as in

[Smadja 1993] [Venkatsubramanyan 2004] [Dias 2006b].

$$((A|N)^{+}|((A|N)^{*}(NP)^{?})(A|N)^{*})N \tag{3.1}$$

However, most approaches deal with the extraction of compound nouns and do not cope with other linguistic phenomena expressed by MWU. For the particular case of verbal locutions, we can mention the studies of [Dias 2001a] [Bannard 2003] [Stevenson 2004] [Bannard 2007]. In particular, [Dias 2001a] are the first to tackle the extraction of verbal locutions, in this case for the Estonian language. In particular, we perform a morphological analysis and the disambiguation of the corpus. Then, the verbs are transformed in their lemma while the other words keep their original form. The following step aims at only selecting verbal n-grams i.e. those n-grams containing verbs and SENTA [Dias 1999a] is finally run over this data set. The extracted phrases are then manually checked. The results showed that 865 verbal locutions were added to the existing Estonian dictionary, which contained 10816 entries i.e. 8% of the overall lexicon. Besides the extraction of compound nouns and verbal locutions, very few other linguistic phenomena have been tackled. We can mention [Utsuro 2007] who proposes an approach to process Japanese compound functional expressions and [Sharoff 2004] who detects expressions starting with a preposition, which cover not only prepositional phrases, but also fixed syntactical constructions such as *in the course of*. One major drawback of such methodologies is that they do not deal with a great proportion of interesting MWUs (e.g. adverbial locutions, compound determinants). Moreover, they lack flexibility as the syntactical patterns have to be revised whenever the targeted language changes. To solve part of these problems, [Dias 2003] proposes a new idea, which combines both statistical and syntactical clues into a single association measure, called the combined association measure. Interestingly, a recent work by [Ramisch 2008] follows this same idea.

Although MWU depict conceptual entities, which usually embody a unique mental concepts, a few studies have been tackling MWU extraction through a semantic angle. [Piao 2003] is certainly one of the first works to attack the identification of MWU based on semantic analysis. They use the USAS semantic tagger developed at the University of Lancaster to semantically tag their corpus and produce a set of rules for the third task of the tagger (the overlapping MWU resolution) to identify MWU, which substantiate single semantic concepts. Normally, semantic MWU take priority over single word tagging, but in some cases a set of templates may produce overlapping candidate tags for the same set of words. So, six heuristics are applied to enable the most likely template to be treated as the preferred one for tag assignment. Following this approach, [Piao 2003] claim that low frequency MWU can be identified unlike most of other methodologies, with the exception of [Dias 2002], and that efficient algorithms are needed to distinguish between free word combinations and relatively fixed, closely affiliated word bundles. [Van de Cruys 2007] propose a loose semantic-based

approach based on the non-compositionality paradigm intrinsic to MWU. Their intuition is that a noun within a MWU cannot easily be replaced by a semantically similar noun. To implement this intuition, they apply the K-means[1] algorithm using attributional similarity measures[2], which result is a set of clusters of semantically related nouns. A number of statistical measures based on selectional preferences (i.e. replacing the initial words by similar ones) are then applied to formalize the intuition of non-compositionality. These approaches are certainly promising. However, the one proposed by [Piao 2003] requires semantic resources and tools, which are not available for a majority of languages and as a consequence limits its scope of action. On the opposite, the methodology proposed by [Van de Cruys 2007] can easily be reproduced for other languages, but they only deal with compound nouns and may be biased by low accuracy of attributional similarity measures. However, this approach is promising for new informative similarity measures such as the IS or the RIS and we will discuss this issue in Chapter 8.

Previous works on automatic identification or extraction of MWU generally make use of certain statistical measures to test the significance of association between words to yield the n-best MWU candidates for human scrutiny, optionally with linguistic preprocessing or filtering. The drawback of these approaches is that they can only rely on one single statistical measure (optionally in association with a given threshold, except for the case of the GenLocalMaxs algorithm). However, as [Dias 2001b] [Yang 2003b] [Pecina 2006] mention, different association measures lead to different and eventually complementary results. Within this context, [Dias 2001b] is certainly one of the first studies to propose a machine learning environment to extract MWU. In particular, we propose an architecture based on a floating point representation genetic algorithm, where each positional n-gram is represented by a vector of different numerical clues about phraseness (e.g. frequency, degree of cohesion, context analysis) and a fitness function, which maximization is likely to provide the "best" representation for MWU. In order to extract relevant MWU, a set of different similarity measures are used to evidence the relatedness between a specific positional n-gram and the elected "best" individual. As a consequence, very close genotypes are listed as pertinent word associations whereas unrelated chromosomes are discarded. Later, [Yang 2003b] proposes a supervised learning environment based on the C4.5 algorithm. Each digram of a lemmatized corpus is associated to a vector of seven statistical measurements: the simple frequency, the t-test score [Smadja 1993], the PMI (see Equation 2.1), the Dice coefficient (see Equation 2.3), the log-likelihood ratio (see Equation 2.9), the χ^2 and the I-score [Pon 2007]. Following the same idea, an exhaustive study is proposed in [Pecina 2006], where several classification methods are described for combining association measures. In particular, they propose a feature selection algorithm, which significantly reduces the number of combined measures with only

[1] With K=1000.

[2] In particular, the cosine similarity measure is applied over the dependency relations feature vectors weighted by the PMI.

a small performance degradation. Many other works could be explored such as [Tsuchiya 2006] [Xu 2006], but the main problem evidenced by all machine learning approaches, as well as most other methodologies, is the definition of thresholds. This issue is particularly well-studied in [Dias 2001c] and [Yang 2003b], who evidence that threshold definition is always a complicated task as it must be tuned for each particular experiment.

3.2 Statistical Multiword Units Extraction

In order to overcome the problems previously highlighted by purely statistical methodologies, we combine the Mutual Expectation (see Equation 2.12) with a new acquisition process called the GenLocalMaxs [Dias 1999a]. On the one hand, the ME evaluates the degree of cohesiveness that links together all the words contained in a positional n-gram. On the other hand, the GenLocalMaxs retrieves the candidate MWU from the set of all the valued n-grams by evidencing local maxima of association measure values. The combination of the ME with the GenLocalMaxs proposes an innovative integrated solution to the problems of enticement techniques and global thresholds defined by experimentation evidenced by most approaches.

Unlike other studies, we propose to segment the text into positional n-grams. Although most works have been dealing with contiguous MWU, MWU may also be non-contiguous rigid sequences of words interrupted by one or more gaps that are filled in by a small number of interchangeable words. For instance, in French, the negation is usually realized by non-contiguous sequences such as *ne ... pas*, *ne ... plus* or *ne ... jamais*[3]. By analyzing long non-contiguous text spans, we also give more credits to the selected MWU by the GenLocalMaxs. Indeed, to be selected, a MWU must show that no other words, in its right and left contexts, can strengthen its intrinsic degree of association. So, each positional n-gram $\widehat{S} = p_{11}w_1 \ldots p_{1n}w_n$, where p_{1n} corresponds to the relative position of word w_n in relation to word w_1, also referred to as the pivot word, is evaluated in terms of cohesiveness by $ME(\widehat{S})$ as shown in Equation 2.12.

Electing MWU among the sample space of all the valued positional n-grams may be defined as detecting combinations of features that are common to all the instances of the concept of MWU. In the case of purely statistical methods, frequencies and association measure values are the only features available to the system. Consequently, most of the approaches base their selection process on the definition of global frequency thresholds and/or on the evaluation of global association measure thresholds. This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether a positional n-gram is a pertinent word association or not. However, these thresholds are prone to error as they depend on experimentation. Furthermore, they highlight evident

[3]More examples can be found in [Dias 2002] for Portuguese and French.

constraints of flexibility, as they need to be re-tuned when the type, the size, the domain and the language of the document change. To overcome these drawbacks, we propose a more flexible and fine-tuned approach for the selection process as it concentrates on the identification of local maxima of association measure values, the GenLocalMaxs algorithm. Specifically, the GenLocalMaxs elects MWU from the set of all the valued positional n-grams based on two assumptions. First, the association measures show that the more cohesive a group of words is, the higher its score will be. Second, MWU are localized associated groups of words. So, we may deduce that a positional n-gram is a MWU if its association measure value is higher or equal than the association measure values of all its sub-groups of $(n-1)$ words and if it is strictly higher than the association measure values of all its super-groups of $(n+1)$ words. Let's $assoc(.)$ be any association measure, W a positional n-gram, Ω_{n-1} the set of all the positional (n-1)-grams contained in W, Ω_{n+1} the set of all the positional (n+1)-grams containing W and $size(.)$ a function that returns the number of words of a positional n-gram. The GenLocalMaxs is defined in algorithm 1.

Algorithm 1 The GenLocalMaxs algorithm.

$\forall W_{n-1} \in \Omega_{n-1}, \forall W_{n+1} \in \Omega_{n+1}$
if $size(W) = 2 \wedge assoc(W) > assoc(W_{n+1})$ **then**
  return MWU
else
  if $size(W) \neq 2 \wedge assoc(W) \geq assoc(W_{n-1}) \wedge assoc(W) > assoc(W_{n+1})$ **then**
    return MWU
  else
    return NO-MWU
  end if
end if

Among others, the GenLocalMaxs shows two interesting properties. On the one hand, it allows to test different association measures on the same basis without having to define specific thresholds, which could yield to biased results. So, we performed many experiments with different association measures over several languages. In particular, we tested the following normalized mathematical models in [Dias 2000e] adn [Dias 2002] i.e. the PMI (see Equation 2.1), the Dice coefficient (see Equation 2.3), the Φ^2 (see Equation 2.7), the log-likelihood ratio (see Equation 2.9), the SCP (see Equation 2.10) over French, Portuguese, English and Italian [Dias 1999b]. Experiments were also made for Slovenian [Dias 1999c] and Estonian [Dias 2001a] just with the Mutual Expectation. In all cases the ME associated to the GenLocalMaxs lead to better results than all other models.

On the other hand, the GenLocalMaxs allows to extract MWU obtained by composition. Indeed, as the algorithm retrieves pertinent units by analyzing their immediate contexts, it may identify MWU that are composed by one or

more other MWU. This situation is illustrated in Figure 3.1. In this example, the GenLocalMaxs elects the MWU *operating system windows vista* built from the composition of both MWU *operating system* and *windows vista*. Roughly exemplifying, one can expect that there are many operating systems (e.g. Linux, MacOS, Windows). Therefore, the association measure value of *operating system windows* is likely to be lower than the one of *operating system*. But, if the first word occurring after *operating system* is *windows*, the expectation to appear *vista* is very high and the association measure value of *operating system windows vista* is then likely to be higher than the association measure values of all its sub-groups and super-groups, as no word can be expected to strengthen the overall unit.

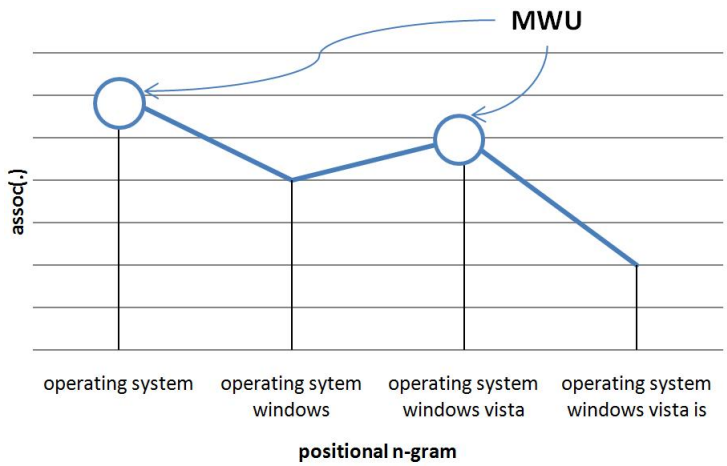

Figure 3.1: Extracting MWU by Composition.

Although the GenLocalMaxs proved to lead to improved results compared to the definition of thresholds [Dias 2002], it shows one major drawback. Indeed, based on its definition, the GenLocalMaxs cannot extract MWU with successive numbers of words. For example, the extraction of both MWU *soft contact lenses* and *contact lenses* or *saturated carbon dioxide* and *carbon dioxide* is impossible with the GenLocalMaxs. So, we proposed to study the recursive application of the GenLocalMaxs following the bootstrapping strategy in [Lourenço 2009] over a portion of 1.2 millions words of the Reuters RCV1 corpus[4] for the case of contiguous n-grams. We present the results in Table 3.1, where N is the size up-to-which the MWU are re-introduced in the corpus (e.g. if N=5 then all extracted MWU with size equals to 2, 3, 4 and 5 are re-integrated in the corpus and are looked for in the next iteration), F is the size of the right and left context windows, *Acc.* is the overall accuracy and *Acc. Extra* is the accuracy of the extracted MWU after the first iteration. The results show interesting issues as a great number of extra MWU can be extracted although with less accuracy than the original algorithm, except for the case of $N = 3$. For instance, the following MWU were extracted: *Antoinette*

[4] http://trec.nist.gov/data/reuters/reuters.html [14th $July$, 2010].

Kennedy and *Judge Antoinette Kennedy*, *hero Colin Powell* and *War hero Colin Powell*, *retail industry association* and *HDE retail industry association*.

The exhaustive study of SENTA proposed in [Dias 2002] shows that it is possible to extract MWU from raw texts without defining any threshold or using lists of stop-words or stemming by combining the ME association measure with the GenLocalMaxs algorithm. A recent study [Lourenço 2009] demonstrates that the bootstrapping strategy can lead to higher recall than the original SENTA. Moreover, all MWU types are covered and not only compound nouns are extracted. However, the statistical approach is based on the study of text corpora and identifies textual associations in the context of their usage. As a consequence, many terminologically relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures for that purpose, which is supported by low accuracy results (see Table 3.1)[5] as all purely statistical approaches.

N	*F*	# MWU	It.1	It.2	It.3	It.4	# Extra	*Acc.*	*Acc. Extra*
2	3	132	111	19	2	-	21	78.1%	62.5%
2	5	108	97	11	-	-	11	83.7%	75.0%
3	5	259	174	69	14	2	85	71.0%	75.9%
4	5	287	209	67	10	1	78	65.3%	55.6%
2	7	102	90	12	-	-	12	83.6%%	66.7%
3	7	215	158	57	10	2	69	61.6%	71.0%
4	7	265	192	73	8	-	81	63.1%	55.2%
5	7	287	207	80	6	-	86	64.2%	55.6%
6	7	290	211	79	6	2	87	62.2%	50.0%

Table 3.1: GenLocalMaxs by Bootstrapping.

But, even with low precision figures, SENTA can extract a great deal of correct MWU and shows interesting properties that could be of great interest in real-world real-time environments such as information retrieval. In particular, SENTA obtains high accuracy results for digrams (which are most of the MWU for the English language), its precision does not degrade with the decreasing size of the corpus as shown in [Dias 2005c] and its no-threshold basis allows its application to word-based or character-based languages without any necessary tuning. As a consequence, we decided to develop an efficient version of SENTA so that we could use it in our works towards information digestion [Campos 2005] [Dias 2006b] [Dias 2007a] [Dias 2009].

[5] In the remainder of this chapter, we will propose an original alternative to SENTA called HELAS.

3.3 Efficient Multiword Units Extraction

With the explosion of gigabytes of Web data, new efficient models of natural language processing (NLP) have been appearing during the last decade [Dias 2006c]. Within the context of word association discovery, many models have been proposed to evaluate word dependencies. One of the most successful statistical models is certainly the n-gram model [Jelinek 1991]. However, in order to overcome its conceptual rigidity, [Kuhn 1994] defined the polygram model that estimates the probability of a n-gram by interpolating the relative frequencies of all its k-grams (k $\leq$ n). Another way to account for variable length dependencies is the n-multigram model designed by [Deligne 1995]. All these models have in common the fact that they need to compute continuous string frequencies. This task can be colossal when gigabytes of data need to be processed. Indeed, [Yamamoto 2001] show that there exist $N(N+1)/2$ substrings in a N-size corpus. That is the reason why low order n-grams have been commonly used in NLP applications.

In the specific field of MWU extraction, we introduced the positional n-gram model, which evidenced successful results for the extraction of continuous and discontinuous collocations from large corpora. Unlike previous models, positional n-grams are ordered sequences of words that represent continuous or discontinuous substrings of a corpus computed in a $(2.F+1)$-word size context window (F represents the context in terms of words on the right and on the left of any word in the corpus). As a consequence, the number of generated substrings rapidly explodes and reaches astronomic figures. We show in Equation 3.2 that Δ positional n-grams can be computed for a N-size corpus in a $(2.F+1)$-size context window.

$$\Delta = (N - 2.F) \times \left(1 + F + \sum_{k=3}^{2.F+1} \sum_{i=1}^{F} \sum_{j=1}^{F} C_{j-1}^{i-1} C_{j}^{k-i-1} \right). \tag{3.2}$$

For instance, 4.299.742 positional n-grams (n=1..7) would be generated from a 100.000-word size corpus in a seven-word size context window. In comparison, only 700.000 n-grams would be computed for the classical n-gram model. It is clear that huge efforts need to be made to process positional n-gram statistics in reasonable time and space. In [Gil 2003b] and [Gil 2003a], we propose an implementation that computes positional n-grams statistics in $\mathcal{O}(h(F)NlogN)$ time complexity and $\mathcal{O}(N)$ space complexity, where $h(.)$ returns the number of masks[6] for a given F. The architecture is based on the definition of masks that allow to virtually represent any positional n-gram in the corpus. Thus, we adapt the virtual corpus approach introduced by [Kit 1998] and build a suffix-array-like data structure, which can be seen as an adaptation, for the case of positional n-grams, of the structures proposed by [Kit 1998] and [Yamamoto 2001] for the continuous case.

One way to represent positional n-grams is to use a set of masks, which

[6] Masks are defined later in this paragraph.

identify all valid types of word sequences i.e. positional n-grams in a given right and left context of size F. Thus, each mask is nothing more than a sequence of 1 and 0 (where 1 stands for a word and 0 for a gap) and any positional n-gram can be identified by a position in the corpus and a given mask, which corresponds to the virtual approach as shown in figure 3.2. It is clear that the number of masks increases exponentially with the size of the word context window as more potential correct sequences of words exist in larger contexts. This number is computed by the $h(.)$ function. Once the virtual approach has been defined, it is necessary to count positional n-gram frequencies and propose a data structure, which allows to quickly search for specific positional n-grams to both compute the ME and apply the GenLocalMaxs, as sub- and super-positional n-grams frequencies are needed.

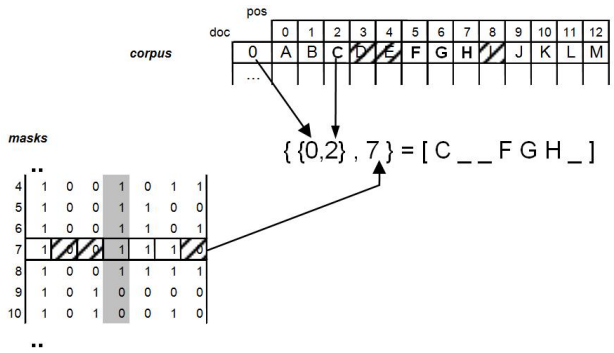

Figure 3.2: The virtual approach.

Counting positional n-grams can be computed following the virtual corpus approach. A suffix-array structure is sorted using lexicographic ordering for each mask following the idea proposed in [Kit 1998] and [Yamamoto 2001]. After sorting, the count of a positional n-gram in the corpus is simply the count of adjacent indices that stand for the same sequence, i.e. for the same mask as illustrated in figure 3.3.

So, the efficiency of the counting mainly resides in the use of an adapted sort algorithm. [Kit 1998] propose to use a bucket-radixsort although they acknowledge that the classical quicksort [Hoare 1962] performs faster for large vocabulary corpora. Around the same perspective, [Yamamoto 2001] use the Manber and Myers's algorithm proposed in [Manber 1990], an elegant radixsort-based algorithm that takes at most $\mathcal{O}(NlogN)$ time and shows improved results when long repeated substrings are common in the corpus. For the specific case of positional n-grams, we chose to implement the multikey quicksort algorithm [Bentley 1997], which can be seen as a mixture of the ternary-split quicksort [Bentley 1993] and the MSD radixsort [Anderson 1998]. Different reasons lead us to use the multikey quicksort algorithm. First, it performs independently from the vocabulary size. Second, it shows $\mathcal{O}(NlogN)$ time complexity in our specific case. Third, [Anderson 1998] show that it performs better than the MSD radixsort and proves comparable results

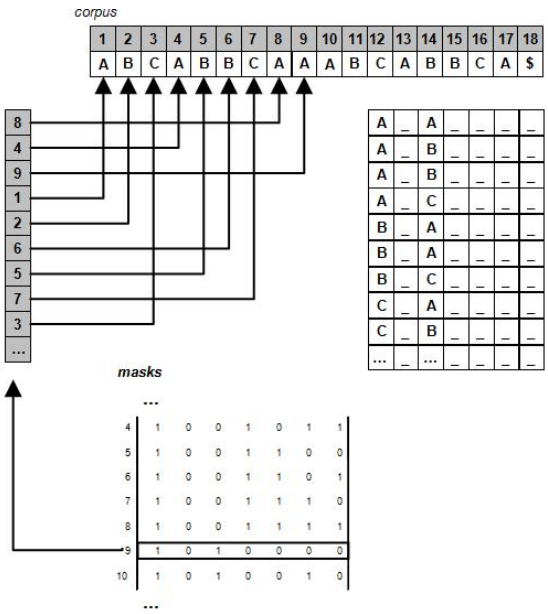

Figure 3.3: The suffix-array data structure.

to the newly introduced forward radixsort.

The MWU acquisition process is based on the ME and the GenLocalMaxs algorithm, for which sub-positional n-grams (sub-groups of (n-1) words) and super-positional n-grams (super-groups of (n+1) words) of any positional n-gram need to be accessed. As a consequence, it is necessary to build a data structure that allows efficient search over the space of all positional n-grams. For that purpose, we proposed to build classical data structures, which are well explained in [Gil 2003b] and [Gil 2003a] and allow quick access to the search space.

In order to evaluate the improvements of this new architecture in terms of processing time, we conducted a number of experiments of our C++ implementation over the CETEMPúblico Portuguese corpus[7]. The experiments were realized on a laptop with an Intel pentium 900MHz processor and 390Mb of RAM running the Windows XP operating system. From the original corpus, we randomly defined five differently sized sub-corpora and ran SENTA to assess its theoretical time complexity. The results are presented in Table 3.2.

This implementation of SENTA is freely available under GPL license and can be downloaded from the website of the HULTIG Center. As we will show in the next chapters, this implementation will allow us to use SENTA in real-world real-time environments as proposed in [Campos 2005] and [Dias 2009] as well as

[7] http://www.linguateca.pt/cetempublico/ [14th $July$, 2010].

Corpus	1	2	3	4	5
# words	114.373	342.734	864.790	1.092.723	1.435.930
# n-grams	4.917.781	14.737.304	37.185.712	46.986.831	61.744.732
Time (h:mm:ss)	0:00:24	0:02:44	0:07:52	0:10:14	0:18:23

Table 3.2: SENTA processing time.

in many other applications such as in [Dias 2006b] [Dias 2007a] [Lourenço 2009] [Rodrigues 2009]. But, improvements can still be obtained using distributed and parallel computing. This is our proposal in [Pereira 2004].

In [Pereira 2004], we propose a parallel multikey quicksort algorithm that allows faster computation of positional n-grams frequencies taking into account the processing power of different central units spread over a network. In particular, we propose a parallel algorithm based on parallel sorting by regular sampling (PSRS) as described in [Shi 1992] that apply their method to randomly generated 32 bit integers and use the classical quicksort as the sequential sorting algorithm. Indeed, for a variety of shared and distributed memory architectures, their results displayed better than half linear speedups.

Our PSRS algorithm sorts positional n-grams using the multikey quicksort as the sorting algorithm and can be divided into three distinct phases: (1) a parallel multikey quicksort phase, (2) a reorganization by global pivots phase and (3) a merge sort phase. Before presenting the overall architecture, it is important to understand the multikey quicksort algorithm. The multikey quicksort works as follows: the array of words is first partitioned into three parts based on the first symbol of each word. In order to process the split, a pivot element is chosen[8] just as in the classical quicksort giving rise to one part with elements smaller than the pivot, one part with elements equal to the pivot and one part with elements larger than the pivot. Then, the smaller and the larger parts are recursively processed exactly in the same manner as the whole array and the equal part is also sorted recursively but with partitioning starting from the second symbol of each word. Finally, the process goes on recursively i.e. each time an equal part is being processed, the considered position in each word is moved forward by one symbol.

So, the first phase of our architecture consists in partitioning the corpus into p contiguous lists, one per node (i.e. a central unit), and uses the multikey quicksort algorithm to sort each local contiguous list. The second phase consists in (1) determining the $(p-1)$ local pivots on each node, (2) determining the global pivots from the $p*(p-1)$ local pivots and (3) reorganizing the local lists in terms of the global pivots. The idea is to join and sort all local sorted contiguous lists

[8] We use the median of three modification method to improve the average performance of the algorithm while making the worst case unlikely to occur in practice as suggested by [Lan 1992].

in a parallel way with good load balancing. For that purpose, we use the regular sampling approach suggested by [Shi 1992]. Finally, the third phase consists in creating on each node one final locally sorted list using a merge sort list. For that purpose, each node splits its sorted list into p sorted sub-lists based on the values of the global pivots. The overall process is defined in Algorithm 2. The algorithm has been implemented using the single program multiple data (SPMD) programming methodology using the ANSI C programming together with the freely available MPICH implementation of the Message Passing Interface (MPI) library. A network of up to ten identical workstations was used. Each workstation consisted of a single Pentium IV 2.4GHz processor with 512Mb of RAM running the Windows XP operating system and connected via a 100Mb Ethernet network.

Algorithm 2 The parallel multikey quicksort algorithm.

The original data file (size N) is copied to all p processor nodes of the cluster.
Each of the p nodes reads the data file to its local memory and builds the suffix-array.
Each of the p nodes determines a contiguous list of size $w = N/p$ from the original data.
Each node creates the valid masks.
for Each valid mask **do**
  Each node sorts the contiguous list using the multikey quicksort algorithm.
  Each node determines $(p - 1)$ local pivots from its sorted list.
  The node 0 gathers all local pivots.
  The node 0 calculates the global pivots from the list of all local pivots.
  Node 0 broadcasts the global pivots to all nodes.
  Each node splits its sorted contiguous list into p sorted sub-lists based on the values of the global pivots.
  Each node keeps one sorted sub-list and passes the others to the appropriate nodes.
  Each node merges the p sorted sub-lists into one local sorted list using the merge sort algorithm.
  Each node communicates the position of first instance of each n-gram plus its frequency to node zero.
end for
The node 0 constructs the matrix of all n-grams frequency.

We conducted a series of experiments for two sub-corpora of the CETEMPúblico Portuguese corpus using a seven-word size context window. The first corpus consisted of 1.000.000 words corresponding to 42.999.742 positional n-grams (Corpus 1) and the second one gathered 3.000.000 words corresponding to 128.999.742 positional n-grams (Corpus 2). In Table 3.3, we illustrate the results for both corpora. The performance of the parallel algorithm is similar for both corpora, slightly better in the larger case as would be expected due to communications of data. Initially with a small number of processors reasonable speedups and efficiency are obtained

but this parallel performance deteriorates with the augmenting number of processors due to high levels of communications. However the overall time taken for the sort is still a monotonically decreasing function when using up to ten processors. So, the combination between speedup and efficiency shows that the best architecture is found for four processors as shown in [Pereira 2004]. This implementation is also freely available under the GPL license and can be downloaded from the website of the HULTIG Center.

Units	Time (h:mm:ss) Corpus 1	Time (h:mm:ss) Corpus 2	Speedup Corpus 2	Efficiency Corpus 2
1	0:03:10	0:11:42	-	-
2	0:02:06	0:07:38	1.53	0.77
3	0:01:47	0:05:47	2.02	0.67
4	0:01:27	0:04:47	2.44	0.61
5	0:01:19	0:04:13	2.78	0.56
6	0:01:10	0:03:54	3.00	0.50
7	0:01:09	0:03:41	3.19	0.46
8	0:01:01	0:03:30	3.34	0.42
9	0:01:00	0:03:20	3.50	0.39
10	0:00:59	0:03:16	3.58	0.36

Table 3.3: Parallel SENTA processing time.

Developing efficient architectures for SENTA was an extremely important work as it allows to introduce implicit knowledge in texts, which we will see in the next chapters tends to improve word and text similarities. Indeed, the so-called text normalization is an important step towards information understanding and as a consequence its digestion. However, facing low accuracy, especially for 3-grams, we experimented a completely different paradigm by applying reinforcement learning to extract MWU and take advantage of different clues, which identify MWU.

3.4 Learning Multiword Units Extraction

Purely statistical models for MWU extraction only rely on one single statistical measure. However, as [Dias 2001b] [Yang 2003b] [Pecina 2006] mention, different clues (e.g. different association measures, right and left contexts) can lead to different and eventually complementary results. As a consequence, in [Dias 2001b] we decided to apply a machine learning approach to MWU extraction. Unlike [Yang 2003b] and [Pecina 2006], who employ supervised learning, we propose reinforcement learning to extract MWU. Indeed, supervised learning implies the labeling of MWU examples. However, there is a lack of a well-established definition of MWU within the field as well as the absence of golden standards,

which complicates the labeling task[9][10]. Moreover, labeling examples can be a tedious task as many examples of different types must be provided, e.g. compound nouns, verbal locutions, compound determinants or adverbial locutions. Although semi-supervised learning could leverage this task, the definition of different classifiers for MWU extraction is not an easy task, but may be worthy investigating in the near future. Based on these arguments, we proposed to apply a floating point representation genetic algorithm coupled with a set of similarity measures to extract MWU from raw texts, without the demand of labeling.

The basic idea of the algorithm is simple. First, the corpus is segmented into a set of positional n-grams from which significant individuals will have to be identified. For that purpose, each positional n-gram is associated to a set of attribute values, which represent a specific chromosome of the overall population. Once the population is defined, the maximization of the fitness function via a genetic algorithm provides the "best" genotype that hopefully is a global maximum. Finally, in order to extract relevant MWU from the original population, a set of different similarity measures evidence the relatedness between a specific positional n-gram in the population and the elected "best" individual. As a consequence, very close genotypes are listed as pertinent word associations whereas unrelated chromosomes are discarded.

A genetic algorithm (GA) is a stochastic algorithm whose search method models two basic natural phenomena: genetic inheritance and Darwinian strife for survival. Within this context, a GA performs a multi-directional search over a sample space by maintaining a population of potential solutions and also encourages information formation and exchange between individuals. As a consequence, the population in consideration undergoes a simulated evolution so that, at each generation, the relatively "good" solutions reproduce and the relatively "bad" solutions die. The binary representation traditionally used in genetic algorithms revealed some drawbacks when applied to multidimensional high precision numerical problems [Michalewicz 1996]. As a consequence, experiments have been realized for parameter optimization problems with real-coded genes together with specific genetic operators designed for them. On one hand, the conducted experiments indicate that the floating point representation is faster, more consistent from run to run and provides better precision than the binary representation for large domains [Goldberg 1989]. On the other hand, as intuitively closer to the problem space, the floating point representation allows a one-gene-one-variable correspondence thus easing the codification process. Consequently, each chromosome can easily be represented by a vector of real numbers, each one corresponding to a specific variable of the problem. In the context of MWU extraction, we define seven variables that have been proposed in different studies as good heuristics for the

[9] To our point of view, the most serious work for this task is the one developed in [Gross 1996] for the French language. However, its transfer to other languages may not be straightforward.

[10] As well as the evaluation of any MWU extractor.

identification of highly cohesive sequences of words. For a given positional n-gram, gene x_0 is its ME, gene x_1 is its frequency[11], gene x_2 is the number of longer positional n-grams in which it appears[12], gene x_3 is its marginal frequency i.e. the average frequency of all its words taken individually[13], gene x_4 and gene x_5 are respectively the highest ME value of all its sub-groups and the highest ME value of all its super-groups[14], and gene x_7 evidences the frequency of its most frequent super-group[15].

To distinguish between different chromosomes, we use an objective (evaluation) function which plays the role of the environment. This function is called the fitness function. Within the context of our research, we need to select pertinent individuals in terms of word associations among the set of attribute-valued positional n-grams. From the previous assumptions, a simple fitness function is directly suggested in Equation 3.3 for a given positional n-gram $\widehat{S}$.

$$g(\widehat{S}) = x_0 + x_1 + x_2 - x_3. \tag{3.3}$$

However, as stated in [Cooper 1970], *a little observation and reflection will reveal that all optimization problems of the real word are, in fact, constrained problems.* As a consequence, this assumption leads to the introduction of four constraints to penalize infeasible solutions as defined in Proposition 3.4.

$$x_0 \geq x_4 \wedge x_0 > x_5 \wedge x_6 < x_1 \wedge x_3 \geq x_1. \tag{3.4}$$

The definition of constraints implies the introduction of penalty functions whose goal is to penalize infeasible solutions. Indeed, if new individuals do not guarantee the constraints, they must be penalized in terms of fitness so that their probability to reproduce is lowered. As a consequence, the fitness function $g(.)$ is transformed in a more generic one called $eval(.)$ defined in Equation 3.5 where $penal_k(\widehat{S})$ is the penalty function attributed to the chromosome $\widehat{S}$ when the constraint k is not verified.

$$eval(\widehat{S}) = \begin{cases} g(\widehat{S}), \text{ all constraints are respected.} \\ g(\widehat{S}) - \sum_{k=1}^{4} penal_k(\widehat{S}), \text{ some constraints are not respected.} \end{cases} \tag{3.5}$$

In order to penalize the fitness function, we use the same methodology for each constraint. The penalty function is based on the distance of the current value of the variable compared to its limit value. For instance, for the constraint on gene x_4, the penalty function would be defined as in Equation 3.6.

[11] This heuristic is proposed by [Justeson 1993].
[12] This heuristic is proposed by [Frantzi 1996].
[13] This heuristic is proposed by [Dias 2000b].
[14] This heuristic corresponds to the GenLocalMaxs.
[15] This heuristic is proposed by [Frantzi 1996].

$$penal_1(\widehat{S}) = \begin{cases} \frac{x_4 - x_0}{x_0}, \text{ if } \ x_4 > x_0. \\ 0, \text{otherwise.} \end{cases} \tag{3.6}$$

Once the fitness function and its constraints are defined, the goal of the genetic algorithm is to find its global maximum. For that purpose, different genetic operators must be defined as well as a sampling mechanism. Within this context, different operators were implemented as to provide a complete evaluation platform. As a consequence, three crossover operators (i.e. single point, two points and uniform) and two different mutation strategies (uniform and non-uniform) were considered. As sampling mechanism, the elitist model was selected [De Jong 1975] together with a stochastic selection process, the Russian roulette. The overall procedure is illustrated in algorithm 3.

Algorithm 3 The genetic algorithm.

```
Build the initial population from the corpus.
Select a random sample of the initial population.
while Still generations to run do
    Evaluate population (fitness).
    Stochastic selection of individuals.
    Apply Crossover.
    Apply Mutation.
    Apply Elitism.
end while
Return the best genotype.
```

The application of the genetic algorithm over the initial population is likely to provide the "best" genotype that is supposed to evidence the best "typical" MWU. However, work still need to be done in order to identify pertinent word associations. For that purpose, we use a similarity measure, which goal is to evaluate the relatedness between each positional n-gram built from the initial population and the "typical" selected MWU. Different similarity measures were implemented. Suppose that $X_i = (X_{i_1}, X_{i_2}, ..., X_{i_p})$ is a row vector of observations on p variables associated with a label i. The distance between two units X_i and X_j is defined as $f(X_i, X_j)$ where f is some function of the observed values. The following functions were proposed: the Euclidean distance in Equation 3.7, the Divergence distance in Equation 3.8, the Bray/Curtis distance in Equation 3.9 and the Soergel distance in Equation 3.10.

$$E(X_i, X_j) = \frac{1}{p} \sum_{k=1}^{p} (X_{i_k} - X_{j_k})^2. \tag{3.7}$$

$$D(X_i, X_j) = \frac{1}{p} \sum_{k=1}^{p} \frac{(X_{i_k} - X_{j_k})^2}{(X_{i_k} + X_{j_k})^2}. \tag{3.8}$$

$$BC(X_i, X_j) = \frac{\sum_{k=1}^{p} |X_{i_k} - X_{j_k}|}{\sum_{k=1}^{p} (X_{i_k} + X_{j_k})}. \tag{3.9}$$

$$S(X_i, X_j) = \frac{\sum_{k=1}^{p} |X_{i_k} - X_{j_k}|}{\sum_{k=1}^{p} max(X_{i_k}, X_{j_k})}. \tag{3.10}$$

Different evaluations were made in [Dias 2001b] [Dias 2001c] [Dias 2004b] based on different languages, different similarity measures, different parameters for the GA and the necessary thresholds. The common result to all experiments confirms that the Bray/Curtis provides better results both in terms of precision and recall than all other similarity measures mostly due to the fact that the definition of a suitable threshold is easier in this case. However, accuracy is still low even for the best thresholds. Indeed, in [Dias 2001c], we clearly show the difficulty to propose appropriate thresholds as minimum changes can imply drops of more than 30% accuracy. The most interesting evaluation is certainly the one proposed in [Dias 2001b], where a multilingual evaluation (English, French, Portuguese) is performed over a corpus of debates of the European Union parliament with 200.000 words. The results are presented in Table 3.4 only for the contiguous positional n-grams as the extracted non-contiguous sequences were no sense sequences in almost all cases.

Rank	English	French	Portuguese
100	81%	62%	79%
200	80%	63%	72%
300	79%	68%	70%
400	73%	65%	69%

Table 3.4: Accuracy results by language and rank.

It is clear that the results strongly depend on the language into consideration as they can vary from 81% to 62% between English and French respectively i.e. a 19% difference. Two major causes can be pointed at. First, this architecture tends to favor the extraction of positional 2-grams. This situation clearly benefits the results for English. Indeed, as French and Portuguese make great use of prepositions to construct complex terms, MWU are usually long sequences of words between three and six words. The other important reason has to do with the fact that French and Portuguese are languages where flexion plays an important role. As a consequence, the same linguistic unit can be evidenced in different graphical forms thus complicating the observation of regularities.

Studying the automatic extraction of MWU based raw texts shows its limits as word associations can only be discovered within the context of their usage. In fact, many wrong associations are retrieved because in corpora, one has also to deal with the writing style, the genre of the text, which may induce wrong regularities as in law texts and the coverage of a given topic in the overall corpus.

To overcome these drawbacks, existing methodologies have been proposing to use lists of stop-words, define high frequency or association measure thresholds or lemmatize the corpora to improve accuracy. However, these are simple techniques that may produce higher accuracy results but with no well-founded theories from which one cannot expect more than little improvements but surely not a solution to extract MWU from corpora of any domain, language or genre. So, for more than a decade, most of the research in MWU extraction has been applying *ad hoc* heuristics in specific domains of expertise with well-behaved corpora (i.e. not real-world texts) so that MWU extractors succeed to extract interesting MWU, but constantly failing to deal with the reality of language, i.e. its diversity and dynamics. In the next section, we propose a new model based on the observation of MWU in terms of syntactical constructions, which we believe can approximate the limits of the problem of the automatic extraction of MWU from corpora.

3.5 Hybrid Multiword Units Extraction

In order to extract MWU, hybrid strategies define co-occurrences of interest in terms of syntactical patterns and statistical regularities. However, by reducing the search space to specific word sequences, which correspond to *a priori* defined syntactical patterns (e.g. Noun+Adj, Noun+Prep+Noun), such systems do not deal with a great proportion of MWU. Moreover, they lack flexibility as the syntactical patterns have to be revised whenever the targeted language changes. Finally, as stated in [Habert 1993], *existing hybrid systems do not sufficiently tackle the problem of the interdependency between the filtering stage [the definition of syntactical patterns] and the acquisition process [the scoring and the election of relevant sequences of words] as they propose that these two steps should be independent.*

In order to overcome these difficulties, we propose in [Dias 2003] a model, which combines word statistics with endogenously acquired linguistic information. We base our study on two assumptions. On one hand, a great deal of studies in lexicography and terminology assess that most of the MWU evidence well-known morphosyntactical structures [Justeson 1993] [Gross 1996]. On the other hand, MWU are recurrent combinations of words. Indeed, according to Habert and Jacquemin in [Habert 1993], the MWU may represent a fifth of the overall surface of a text. As a consequence, it is reasonable to think that the syntactical patterns embodied by the MWU may be endogenously identified by using statistical scores over corpora of part-of-speech tags exactly in the same way word dependencies are identified in corpora of words. So, the global degree of cohesiveness of any sequence of words may be evaluated by a combination of its degree of cohesiveness of words and the degree of cohesiveness of its associated part-of-speech tag sequence. For that purpose, a new association measure called the combined association measure (CAM) is introduced in Equation 3.11 where t_i corresponds to the part-of-speech tag of word w_i of the positional n-gram $\widehat{S} = p_{11}w_1t_1 \ldots p_{1n}w_nt_n$.

$$CAM(\widehat{S}) = ME(p_{11}w_1 \dots p_{1n}w_n)^{\alpha} \times ME(p_{11}t_1 \dots p_{1n}t_n)^{1-\alpha}. \quad (3.11)$$

The parameter α allows to tune the model towards total focus on part-of-speech tags (i.e. the relevance of a word sequence is based only on the relevance of its part-of-speech sequence with $\alpha = 0$) to total focus on words (i.e. the relevance of a word sequence is defined only by its word dependencies with $\alpha = 1$).

Based on this new association measure, we propose the following architecture to extract MWU. First, a part-of-speech tagged corpus is divided into two sub-corpora: one containing words and one containing part-of-speech tags. Each sub-corpus is then segmented into a set of positional n-grams. Third, the ME independently evaluates the degree of cohesiveness of each positional n-gram i.e. any positional n-gram of words and any positional n-gram of part-of-speech tags. The CAM is then used to evaluate the global degree of cohesiveness of any sequence of words associated with its respective part-of-speech tag sequence. Finally, the GenLocalMaxs retrieves all the MWU candidates by evidencing local maxima of association measure values thus avoiding the definition of global thresholds. The overall architecture called Hybrid Extraction of Lexical Associations (HELAS) is illustrated in Figure 3.4.

A first approach to combine part-of-speech tags and word regularities had been proposed in [Dias 2000a] but with less formalism. However, the reader can find interesting experiments in this study. The work proposed in [Dias 2003] is more complete and well-founded, so that we will reproduce its main results in this thesis. In order to test HELAS, we conducted a number of experiments with eleven different values of α (i.e. $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$) over a portion of the part-of-speech tagged Brown Corpus containing 249.578 words. As many authors assess [Smadja 1993] [Justeson 1993], deciding whether a sequence of words is a MWU or not is a tricky problem. For that purpose, different definitions of MWU have been proposed. One of the most successful attempts can be attributed to Gross [Gross 1996], who classifies MWU into six groups and provides techniques to determine their class. Although his study was mainly developed for the French language, we tried to transfer his rules for the English language in [Dias 2003] and the Slovene language in [Dias 2005c]. As a consequence, a MWU is considered to be any compound noun, compound determinant, verbal locution, adverbial locution, adjectival locution or prepositional locution. Based on this proposal, the best results were obtained for $\alpha = 0.5$ and reached 62%. However, a detailed analysis shows that accuracy up to 85% can be obtained for positional 3-grams but HELAS fails somehow to extract longer and smaller sequences as shown in Table 3.5, although they represent a small proportion of the extracted MWU.

Although these results were interesting, further experiments were necessary. Indeed, the Brown Corpus is a balanced corpus and as such embodies the most

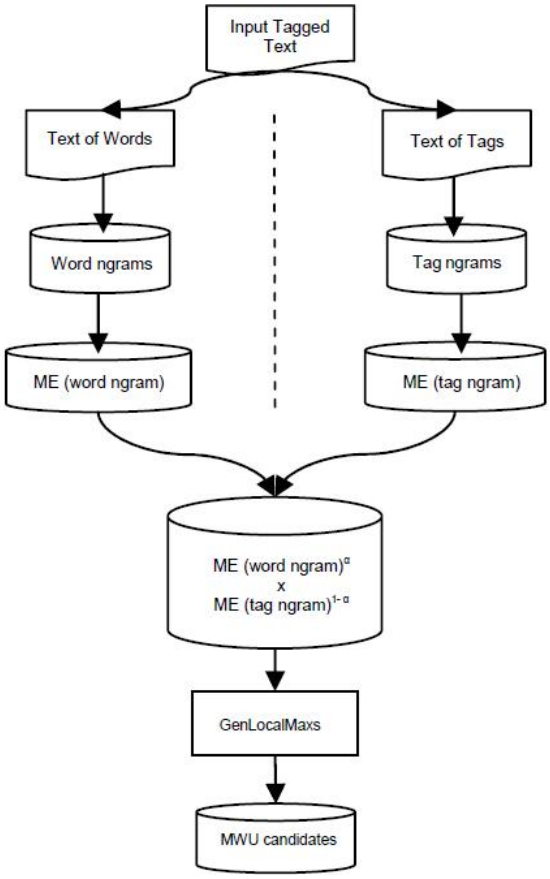

Figure 3.4: The HELAS architecture.

difficult challenge for any MWU extractor based on word or part-of-speech tag regularities. As a consequence, we proposed in [Dias 2005c] a multilingual (English and Slovene) evaluation over three domain specific corpora. In particular, we chose three sub-corpora of the multi-domain bilingual Slovene-English IJS-ELAN corpus [Erjavec 2002]: the annex II of the Europe agreement about EU legislation and politics of 25.000 words, the Slovenian economic mirror about economics of 239.000 words and the Linux installation and getting started about computing of 173.000 words. The results showed interesting characteristics of the architecture.

First, while the overall precision regardless of n-gram type and text type shows that the best result for English is obtained for $\alpha = 0.6$, the precision for Slovene gradually rises as α increases, with the highest value for $\alpha = 0.9$. The part-of-speech sequence apparently plays a less important role with a highly inflectional language like Slovene, due to the reduced observed regularities.

Second, the best value of α for positional 2-grams is regularly higher than

n-grams	Accuracy
2-grams	60%
3-grams	85%
4-grams	41%
5-grams	34%
6-grams	38%

Table 3.5: Accuracy results by n-gram for $\alpha = 0.5$.

for the other positional n-grams for both languages, meaning that positional 2-grams seem to show higher degrees of cohesiveness between words than between part-of-speech tags. These results had already been evidenced in [Dias 2000a] and are confirmed in this evaluation.

Third, unlikely observed in [Dias 2003], the results for positional 4-grams, 5-grams and 6-grams are steady and high independently of the given α. To some extent, this means that positional 4-grams, 5-grams and 6-grams are some kind of institutionalized phrases and domain specific collocations, which show both word and part-of-speech tags associative regularities. This is clearly due to the fact that the corpora stress a given domain.

Finally, by comparing overall precisions by positional n-gram and text types it becomes clear that the size of the corpus plays a substantial role in the accuracy of the MWU extractor. In fact, the bigger the corpus is, the lower the precision is, especially for 2-grams. These results are very interesting as it has always been mentioned as an evidence that huge corpora would automatically lead to better results for statistical methodologies. It seems that this assumption does not stand for our model. So, what could be seen as a problem of scalability is in fact a providential result for many real-world NLP applications, which can now integrate MWU recognition "plug-ins" capable to process texts in real-time and provide applications with a text normalization module, thus achieving enhanced performances as shown in [Dias 2006b] [Dias 2007a] [Dias 2009].

3.6 Future Work

Although a lot of studies have been carried out in the field of MWU extraction, many works still need to be produced to deeply understand the phenomenon of MWU. In particular, the introduction of the $CAM(\widehat{S})$ rises different important issues, which may be tackled. First, fine-grained or coarse-grained part-of-speech tagging may lead to different results. Indeed, the more specific the part-of-speech tagging will be, the more difficult it will be to extract relevant sequences of part-of-speech tags and the extraction process will probably suffer from this situation. Second, we propose a geometric combination to evaluate the strength of any positional n-gram

$\widehat{S}$. But, we could propose a linear interpolation such as in Equation 3.12 to leverage the impact of any low ME value.

$$CAM_{lp}(\widehat{S}) = \alpha \times ME(p_{11}w_1 \ldots p_{1n}w_n) + (1-\alpha) \times ME(p_{11}t_1 \ldots p_{1n}t_n). \quad (3.12)$$

It is also clear that different values of α evidence different accuracies depending on the size of the positional n-grams. For instance, positional 2-grams obtain higher accuracies for high values of α (i.e. giving almost no strength to sequences of part-of-speech tags), while positional 3-grams need equally weighted word and part-of-speech tag sequences to evidence high precision. Learning these values of α would be interesting, so that we could really rely on a black-box software for the extraction of MWU, which would be intrinsically optimized. However, as many authors assess [Smadja 1993] [Justeson 1993], deciding whether a sequence of words is a MWU or not is a tricky problem, which hardens the development of golden standards. Although there have been efforts in this sense with the works of [Yang 2003b] and [Pecina 2006], who manually tag MWU to apply supervised machine learning algorithms, they usually tackle compound nouns or compound names but not the full spectrum of MWU as described in [Gross 1996]. Developing a golden standard based on the work made by [Gross 1996] would certainly be a great contribution to the field of MWU, but would be a time consuming task, which may not reach its goals at the end. So, a way to leverage the task of manual annotation of MWU for learning purposes would be to apply semi-supervised learning. By building different views based on different clues to extract MWU, semi-supervised learning could incrementally produce new labeled examples, which would serve as a basis for further learning purposes as well as it would propose a new model to extract MWU based on a few initial seeds. We clearly intend to follow both research directions as we are aware of the potential of HELAS and also believe that MWU embody many different characteristics that may be taken into account within a multi-view learning environment.

From a linguistic point of view, MWU should be non-compositional, non-substitutable and non-modifiable. While the non-substitutable and non-modifiable characteristics are embodied by most MWU, compositionality is usually a loose continuum between complete semantic opacity and perfect transparency. As a consequence, many researches proposed to evaluate the degree of compositionality of MWU, while just a few tackled non-substitutability and non-modifiability, although these two research directions are clearly promising. Within the scope of non-substitutability, [Van de Cruys 2007] propose an interesting work, which we intend to extend by (1) applying a new methodology to extract synonyms [Dias 2010] and (2) introducing informative similarity measures to improve accuracy over non-informative measures[16]. The basic idea is to automatically find potential synonyms of constituents within any n-gram as we propose in [Dias 2010] and to

[16]Indeed, we will show in Chapter 6 that non-informative similarity measures show low precision.

compare its strength to the sequence, where the initial word has been replaced by its synonym using the ME for example.

Within the scope of non-modifiability, we aim at relaxing the position constraint of the positional n-grams. Indeed, positional n-grams are built based on the idea that MWU are recurrent sequences of words, which cannot be modified in terms of position. However, verbal locutions are a clear exception of this rule. As a consequence, we propose to compare any n-gram with and without position constraints to evaluate to which extent this text segment (contiguous or non-contiguous) is modifiable or not.

Finally, knowing whether a MWU is compositional or non-compositional is a great issue for the construction of lexical-semantic knowledge databases as we will see in Chapter 6. For that purpose, many works have been proposed but largely rely on linguistic clues, thus limiting their scope of action to specific languages. To overcome this drawback, we propose to evaluate compositionality versus non-compositionality based on informative level of generality of positional n-grams by applying the $AIS(.\|.)$ (see Equation 2.32) or the $RAIS_N(.\|.)$ (see Equation 2.34) asymmetric informative similarity measures. For example, let's take the MWU *saturated carbon dioxide* and drop its first constituent to give rise to the following sequence *carbon dioxide*. If *carbon dioxide* and *saturated carbon dioxide* are semantically related and *carbon dioxide* is more general than *saturated carbon dioxide*, we may conclude that *saturated carbon dioxide* is non-compositional and that *carbon dioxide* is an hypernym of *saturated carbon dioxide* with some degree of certainty. If we continue this process, we compare *carbon dioxide* to *carbon* and *dioxide*. As it is likely that *carbon* evidences loose semantic similarity with *carbon dioxide* compared to *dioxide*, and that *dioxide* is more general than *carbon dioxide*, we may also conclude that *carbon dioxide* is non-compositional and that *dioxide* may be its direct hypernym.

In Chapter 4, we will present how new similarity measures introduced in Chapter 2 and efficient MWU extractors can improve information retrieval applications towards our main goal of information digestion ■

Chapter 4

Ephemeral Clustering

Contents

4.1 Normalized Full Text Hierarchical Overlapping Clustering 66
4.2 Normalized Snippet Hierarchical Overlapping Clustering . 73
4.3 Snippet Informative Hierarchical Clustering 76
4.4 Future Work 86

With so much information available on the Web, in particular with the explosion of weblogs and social networks, looking for relevant information on the internet is more and more difficult. Indeed, traditional Web search engines still return lists of ranked documents represented by their titles and corresponding Web snippets[1], from which users have to go through extensively to find the document(s) that most satisfy their needs. To avoid what has turned to be a tedious task, some search engines such as Yippy[2], Carrot[3], iBoogie[4], SnakeT[5] or VIPACCESS propose to help users in their process of seeking for information, by digesting Web page results through the dynamic generation of taxonomic structures. Known as post-retrieval document browsing or ephemeral clustering [Maarek 2000], this process constructs flat or hierarchical taxonomies from sets of Web page results, which evidence a short life span and are usually used for interactive browsing purposes.

Ephemeral clustering, for interactive use, introduces several new challenges. It requires an efficient algorithm, since clustering is performed on-line. It also requires high precision, because users who are not domain experts are less tolerant to errors and as a consequence, the fully automatically generated taxonomies cannot show any imprecision, as opposed to off-line clustering for which the (mostly hierarchical) structures are often manually modified [Croft 1980] [Willett 1988]. Finally, interactive clustering requires a presentation layer that enables users to effectively browse the flat or hierarchical structure, including visualization techniques and automatic labeling of clusters.

[1] Short meaningful descriptions of Web pages.
[2] http://www.yippy.com [14th $July$, 2010].
[3] http://carrot2.org [14th $July$, 2010].
[4] http://www.iboogie.com [14th $July$, 2010].
[5] http://snaket.di.unipi.it [14th $July$, 2010].

Although it can be strange to address Web page results clustering as an information digestion paradigm, it becomes clear when illustrating two different user situations. First, although a lot has been done for visually impaired people (VIP) to access information with braille screens, braille keyboards, braille handled devices and speech-to-speech interfaces, very little has been done to reduce the amount of information they have to deal with. Indeed, blind people face an overwhelming task when reading texts. Unlike fully capacitated people, blind people cannot read texts by just scanning them quickly i.e. they cannot read texts transversally. As a consequence, they have to come through all the sentences of a text to understand its level of importance. In the specific context of information retrieval, Web search must be particularly tackled. Indeed, Web search engines usually provide long lists of unorganized search results. But, this way of presenting information is a clear obstacle for VIP to quickly access information as long lists of results take a long time to scan.

Second, the shift in human computer interaction from desktop computing to mobile interaction highly influences the needs for future user-adaptive systems. Indeed, small size screens of handled devices are a clear limitation to display long lists of relevant documents resulting in time consuming scrolling to encounter the relevant information. A direct illation of these two situations is that users (eventually VIP) are unlikely to use classical search engines to search for information on devices such as smartphones or PDA. But, they may shift to specifically designed interfaces as illustrated in Figure 4.1 for the case of ubiquitous information retrieval, which shows the mobile interface of our meta-search engine VIPACCESS[6] for the query *madonna*.

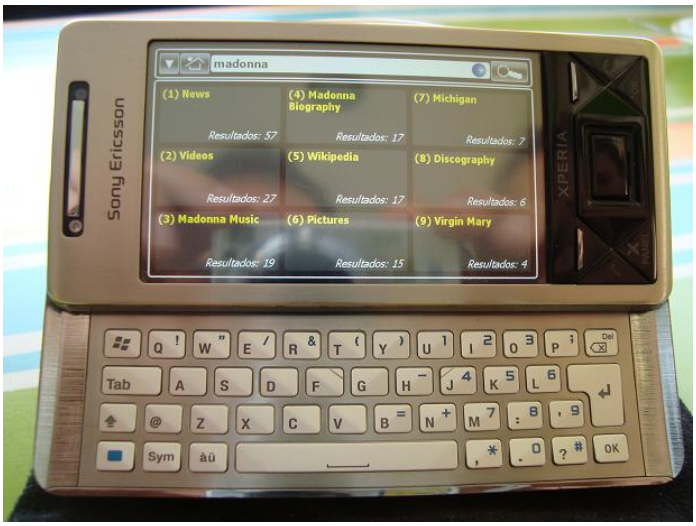

Figure 4.1: Ephemeral clustering as information digestion [21st *August*, 2010].

[6]The mobile implementation of VIPACCESS is freely available at the HULTIG Web page.

The idea of regarding ephemeral clustering as a way of summarizing information was first cited in [Lawrie 2003] but unfortunately received little attention. In particular, they stated that hierarchical structures provide two kinds of information. The first one is a method of navigating to particular sub-clusters, which may contain information of interest to a user. The second one is a summarization of all Web search results as the labels in the hierarchy should describe the documents that are found in the corresponding clusters. We deeply support these ideas and believe that the most interesting applications for ephemeral clustering are certainly the development of specific interfaces for mobile devices and for visually impaired people, which are capable to digest information and afford easier access to it.

Ephemeral clustering has been studied for more than a decade and many studies have been proposed as summarized in Table 4.1[7]. As stated in [Kummamuru 2004], there exist two main different approaches: monothetic clustering (also known as Label-centered clustering) and polythetic clustering (also known as Document-centered clustering). Monothetic algorithms are those in which a document is assigned to a cluster based on a single feature, whereas polythetic algorithms assign documents to the clusters based on multiple features. On the one hand, monothetic clustering has mainly been studied as it is well suited for generating hierarchies for summarization and browsing search results because each cluster is described by a single feature or concept and all the documents present in a cluster contain this feature. Hence, the user can easily understand clusters generated by monothetic clustering algorithms. On the other hand, polythetic algorithms usually need an extra step to process cluster labels, which is a difficult task as mentioned in [Maarek 2000] as clusters are defined by extension (i.e. by enumeration of their members), rather than by intention (i.e. by membership rules). Therefore, there is no direct way to name the identified clusters and specific algorithms must be developed for this purpose. However, it is important to avoid simplistic conclusions. Indeed, monothetic algorithms have mainly been used as the created labels can easily be understood by the users, unlike in most works proposed so far based on Document-centered clustering. For instance, [Hearst 1996] and [Jiang 2002] do not even propose cluster labeling. However, this result does not mean that the contents of the clusters are satisfactory. On the contrary, most Label-centered algorithms rely on *ad hoc* heuristics, which are likely to give meaningful labels but do not guarantee the intrinsic quality of clusters. With that respect, we deeply believe that applying the polythetic strategy is the way to produce high quality results founded on well-known clustering algorithms. Moreover, there is no clear evidence that the Label-centered approach outperforms the Document-centered strategy as the only comparative work proposed so far is the one in [Jiang 2002], who conclude, based on a qualitative and to some extent subjective analysis, that *the n-gram based approach seems to perform better than the vector space based approach.*

[7] For a complete state-of-the-art, the reader can find a complete survey in [Carpineto 2009].

Although the main difference between all approaches is the clustering strategy, many other characteristics can classify ephemeral clustering as shown in Table 4.1. These are all exhaustively defined in [Hearst 1996] and [Zamir 1998], who clearly settle the foundations of post-retrieval document browsing. According to [Zamir 1998], *both cluster overlap and multi-word phrases are critical* to the success of their suffix tree clustering algorithm (STC), which means that MWU best embody the message conveyed by texts as well as documents may belong to different clusters as they may focus on different topics.

Work	Taxonomy	Algorithm	Overlap
[Hearst 1996]	Flat and Hierarchical	Document-centered	No
[Zamir 1998]	Flat	Label-centered	Yes
[Maarek 2000]	Hierarchical	Document-centered	No
[Zhang 2001]	Hierarchical	Document-centered	No
[Jiang 2002]	Flat	Doc./Label-centered	Yes
[Fung 2003]	Hierarchical	Label-centered	No
[Lawrie 2003]	Hierarchical	Label-centered	No
[Zeng 2004]	Flat	Label-centered	Yes
[Osinski 2004]	Flat	Label-centered	Yes
[Carpineto 2004]	Lattice	Label-centered	Yes
[Kummamuru 2004]	Hierarchical	Label-centered	Yes
[Campos 2005]	Hierarchical	Document-centered	Yes
[Ferragina 2008]	Hierarchical	Label-centered	Yes
[Dias 2009]	Hierarchical	Document-centered	Yes
[Machado 2009b]	Flat	Label-centered	Yes
Work	**Text**	**MWU**	**Labels**
[Hearst 1996]	Document	No	No
[Zamir 1998]	Snippet	Yes	Yes
[Maarek 2000]	Document	No	Yes
[Zhang 2001]	Snippet	Yes	Yes
[Jiang 2002]	Snippet	No	No/Yes
[Fung 2003]	Document	No	Yes
[Lawrie 2003]	Document	Yes	Yes
[Zeng 2004]	Snippet	Yes	Yes
[Osinski 2004]	Snippet	Yes	Yes
[Carpineto 2004]	Snippet	No	Yes
[Kummamuru 2004]	Snippet	Yes	Yes
[Campos 2005]	Document	Yes	Yes
[Ferragina 2008]	Snippet and KB	Yes	Yes
[Dias 2009]	Snippet	Yes	Yes
[Machado 2009b]	Snippet	Yes	Yes

Table 4.1: Classification of ephemeral clustering algorithms.

Moreover, they show that applying clustering algorithms based on the overall documents instead of their corresponding Web snippets leads to improved results both in the case of monothetic and polythetic strategies[8], except for the well-known K-means algorithm[9] as illustrated in Figure 4.2 extracted from their work in [Zamir 1998]. However, the decrease in quality of the clusters is apparent but relatively small. As a consequence, they argue that Web snippets are likely to provide the correct clustering of the documents as they embody the excerpts of the documents mostly related to the query terms.

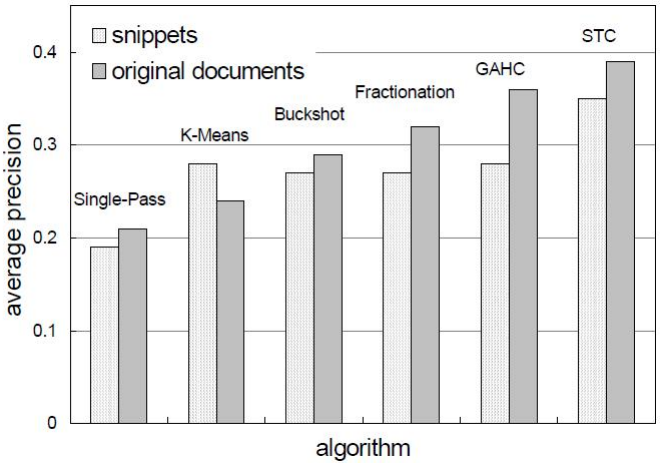

Figure 4.2: Results of the STC algorithm (extracted from [Zamir 1998]).

Another important issue is stated in [Hearst 1996] and confirmed later in [Maarek 2000]. Indeed, [Hearst 1996] compare both flat and hierarchical clustering based on the classical vector space model [Salton 1975] with a partitioning algorithm called fractionation and show that *the best cluster is actually the result of two clustering steps (the best of the best)*. As a consequence, based on the statements of [Hearst 1996] and [Zamir 1998], ephemeral clustering should hopefully tackle normalized[10] full-text hierarchical overlapping clustering to produce relevant results. Within this scope, we are the first to propose such a study in [Campos 2005] and [Campos 2008].

In [Campos 2005], we propose a full-text parameter free methodology, which (1) integrates the better understanding of documents via the introduction of identified MWU through the SENTA software [Dias 1999a], (2) implements the Pole-Based Overlapping Clustering algorithm (PoBOC) proposed in [Cleuziou 2003] and (3) applies Web content mining techniques to represent Web pages in a similar way as [Zeng 2004]. As a consequence, we propose a language-independent

[8] In this case, they use the classical vector space model suggested by [Salton 1975].
[9] A result they cannot explain.
[10] With the identification of MWU.

parameter free architecture, which implements normalized full-text hierarchical overlapping clustering. It is important to point at that more than being the first work to propose the full spectrum of ephemeral clustering, we also tackle language-independency and parameter independence, which none of the works listed in Table 4.1 proposes. Indeed, all works depend on some thresholds and apply at least stemming and the removal of stop-words, which we avoid in our overall architecture.

However, developing a real-world application on a full-text basis is a time consuming task as the analysis of Web page results must be done "on the fly" of execution. This issue had already been mentioned in [Zamir 1998] as processing time is an important factor for the success of any information retrieval application. However, while their incremental STC algorithm is well-suited to leverage network latency, our solution needs a distributed architecture to produce results in due time. Indeed, with a usual amount of 30 to 40 documents per query it takes for a single CPU more than 40 to 50 minutes to produce the results. This situation is clearly unbearable. As a consequence, we proposed a peer-to-peer distributed version of our architecture in [Campos 2008]. As a result, using a small network of only 3 computers, we were able to compute tasks about 10 times faster than when based on one single server. But results still remained in the order of minutes, which clearly limited the evaluation of our model. Although this architecture proved to lead to interesting results that could not be extended to a large exhaustive evaluation due to the lack of processing power, a large number of clusters were created in all the experiments we made. These results somehow negated the compactness paradigm expressed in [Kummamuru 2004]. Indeed, since the main purpose of taxonomy generation is to summarize and provide a better browsing experience, the taxonomy should be as compact as possible. If a taxonomy becomes too wide or too deep, then the basic purpose of taxonomy generation may not be served[11].

To be able to easily evaluate ephemeral clustering both in terms of user satisfaction and theoretical issues (namely the clustering phase), we proposed a snippet-based approach in [Dias 2009]. In particular, we compared the PoBOC algorithm to the well-known single-link, complete-link, group-average hierarchical agglomerative clustering algorithms (HAC) [Jain 1999] and the QT algorithm [Heyer 1999] to understand if the over-generation of clusters could be avoided[12]. Experiments were also made with visually impaired professor Henrique Amorim[13] as we can see in Figure 4.3, who showed great enthusiasm about this new way to search for information on a mobile device.

Within the snippet-based approach, two other works have been proposed

[11] We will address further this issue in Section 4.3.

[12] Only the PoBOC algorithm deals with overlapping, but for the sake of the evaluation of cluster generation, non-overlapping algorithms could be applied as freely available implementations can easily be found.

[13] With an adapted speech-to-speech module.

[Kummamuru 2004] [Ferragina 2008] tackling the overall ideas of [Hearst 1996] and [Zamir 1998]. However, both use lists of stop-words and stemming, and [Ferragina 2008] uses specific knowledge resources such as the DMOZ categories[14] to rank sentences based on category distributions and increase Web snippets sizes. In order to overcome language-dependency, we proposed a polythetic methodology based on (1) SENTA to normalize Web snippets and (2) a surface-based text similarity measure called the Sumo-metric, which we proposed in [Cordeiro 2007a] and defined in Equation 5.9 (see Chapter 5). The language-independency paradigm follows the *corpus integrity principle* enounced in [Dias 2000a], which states that the text must be treated as a unique piece of data, which must not be transformed in any way, at least in a non *ad hoc* manner. This paradigm is also shared by [Osinski 2004] who states that *while stemming and stop-words removal are very common operations in information retrieval, interestingly, their influence on results is not always positive yielding no improvement to the overall quality of ephemeral clustering.*

Figure 4.3: Experiments of ephemeral clustering on mobile devices with VIP.

However, although this affirmation may be true for Label-centered algorithms, it must be applied with care for Document-centered algorithms as Web snippets are usually malformed sentences, which processing must be adapted to ephemeral clustering to guarantee language-independency. For that purpose, we computed Web snippet similarities based on a surface-based text similarity measure called the Sumo-metric (as the classical vector space model failed to produce high precision results) and showed that part-of-speech tagging does not lead to improved results compared to surface-based text similarity measures. Moreover, performing different clustering strategies did not lead to the reduction of clusters. Indeed, all clustering algorithms evidenced a great number of generated clusters leading to loose taxonomies, thus contradicting the compactness paradigm. In fact, this

[14] http://www.dmoz.org/ [14th $July$, 2010].

situation is shared by most algorithms, both monothetic or polythetic, as illustrated in Figures 4.4 (iBoogie meta-search engine), 4.5 (Yippy meta-search engine), 4.6 (Carrot2 meta-search engine implementing [Osinski 2004] algorithm) and 4.7 (Carrot2 meta-search engine implementing [Zamir 1998] STC algorithm) for the ambiguous query *ballon* meaning both *ball* and *balloon* in French. A clear conclusion of this research, is that both polythetic and monothetic strategies over-generate clusters because they are unable to capture the semantics of the Web snippets.

Although it has been proved that hierarchical clustering including phrase detection leads to improved browsing results, users still feel reluctant to use such search engines. We deeply believe that one of the main reasons is the fact that too many clusters are presented to the users which prefer to scan long lists of Web pages rather than going through long lists of (possibly misdescriptive) labels of clusters. For instance, to avoid this problem [Hearst 1996] [Zamir 1998] [Kummamuru 2004] respectively show the top 5, 10, 5 clusters. Indeed as stated in [Kummamuru 2004] and [Zeng 2004], since the main purpose of the taxonomy generation is to summarize and provide a better browsing experience, the taxonomy should be as compact as possible. Based on this idea, we started to work on a Document-centered solution, which (1) integrates the semantic dimension by using the InfoSimba similarity measure (see Chapter 2) between Web snippets and (2) proposes a new selection process based on a recursive global K-means algorithm to produce a hierarchy of concepts. We call this algorithm the hierarchical InfoSimba global K-means ($HISGK$-means). In this way, we aim at reaching query-based disambiguation as a means of reducing the number of potential clusters as illustrated in Figure 4.8 from our meta-search engine VIPACCESS. This is an on-going work and the introduction of an overlapping version of the $HISGK$-means and the identification of MWU within Web snippets are already being studied in order to deal with the full spectrum of ephemeral clustering.

4.1 Normalized Full Text Hierarchical Overlapping Clustering

In [Campos 2005], we propose a meta-search engine called WISE, which builds soft hierarchical clusters "on the fly", without pre-defined categories or pre-existing knowledge bases, by applying an overlapping clustering algorithm called PoBOC introduced by [Cleuziou 2003]. To represent Web page contents, we use Web content mining techniques, introduced in the context of the Webspy software[15] [Dias 2004a], and statistical methodologies for phrase detection, with the use of the SENTA software. Unlike most existing works, WISE fully analyzes Web page contents, is parameter free, threshold independent and does not use lists of stop-words or stemming. As a consequence, it can be applied to any language tackling the reality of the Web. The overall architecture is defined in algorithm 4.

[15] Webspy is freely available at the HULTIG Web page.

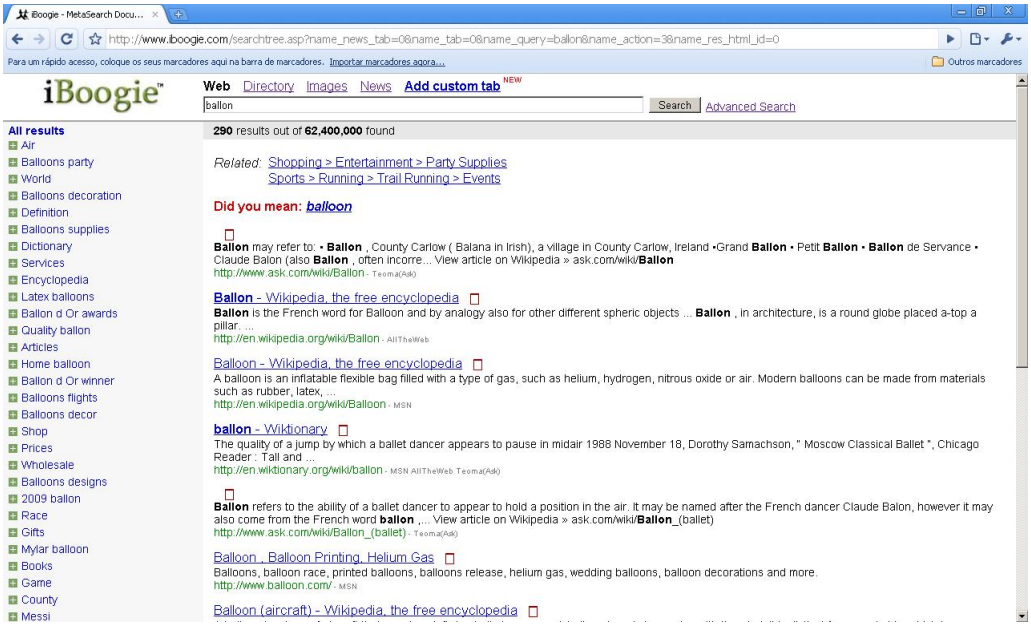

Figure 4.4: The iBoogie algorithm for the query *ballon* [$14^{th} July, 2010$].

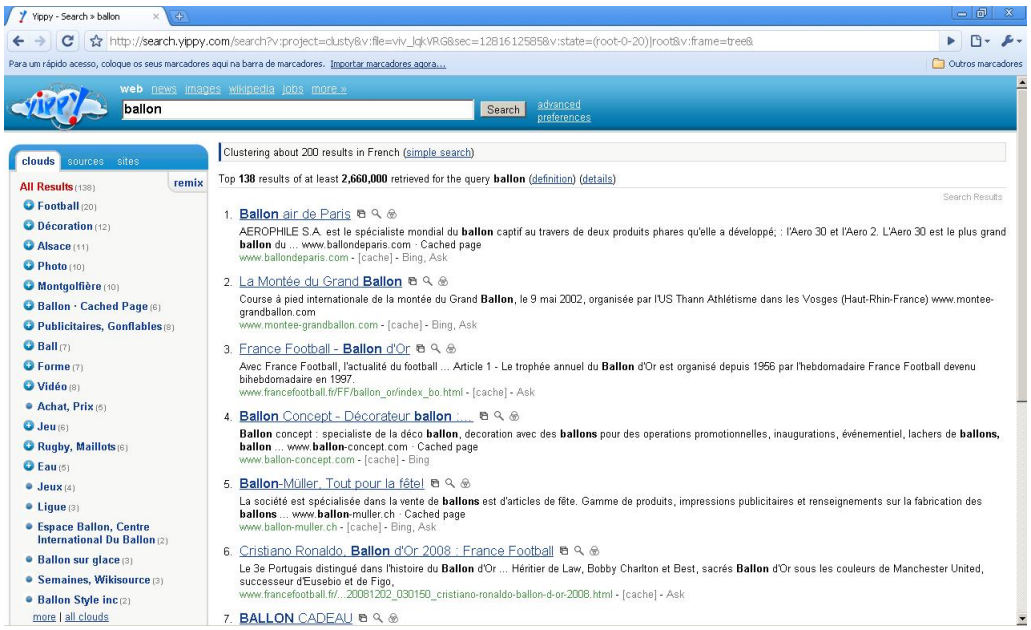

Figure 4.5: The Yippy algorithm for the query *ballon* [$14^{th} July, 2010$].

Existing methodologies usually treat retrieved Web page results as if they have equal relevance with respect to the query disregarding the fact that the estimated relevance of a document decreases as more results are gathered [Jiang 2002],

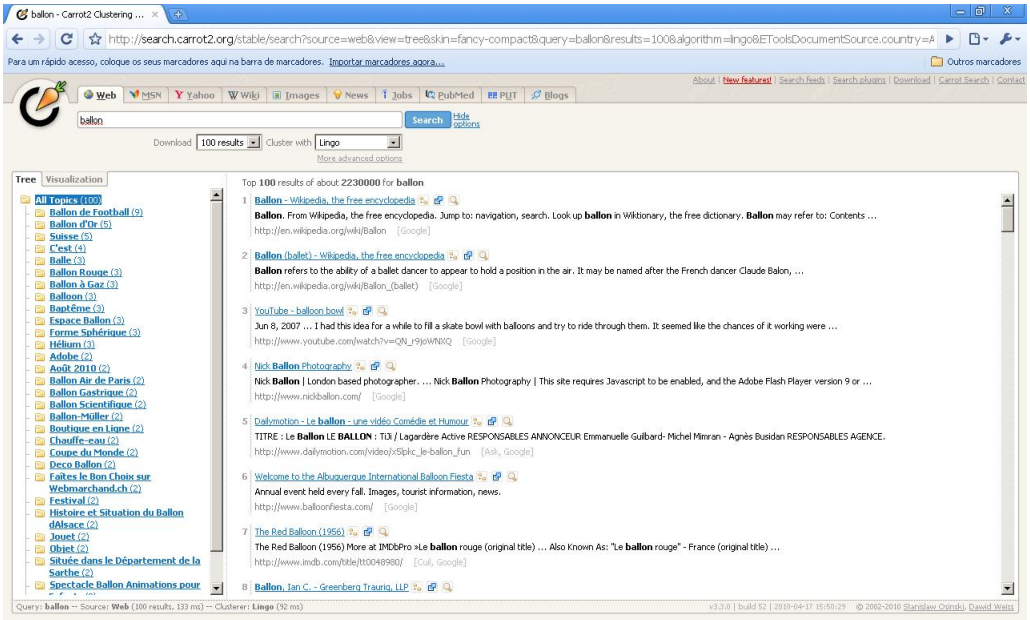

Figure 4.6: The Carrot2 algorithm for the query *ballon* [$14^{th} July, 2010$].

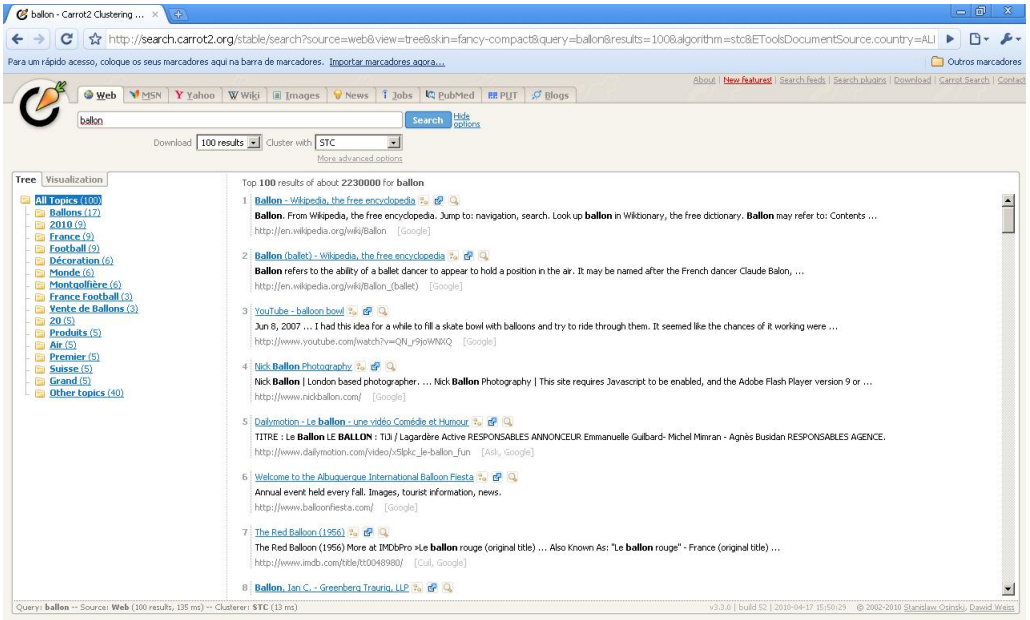

Figure 4.7: The STC algorithm for the query *ballon* [$14^{th} July, 2010$].

probably decreasing the quality of the final clusters. For that purpose, we propose a new method for the extraction of relevant Web pages, which ignores some of the retrieved documents and adds some others. In particular, we only keep Web page

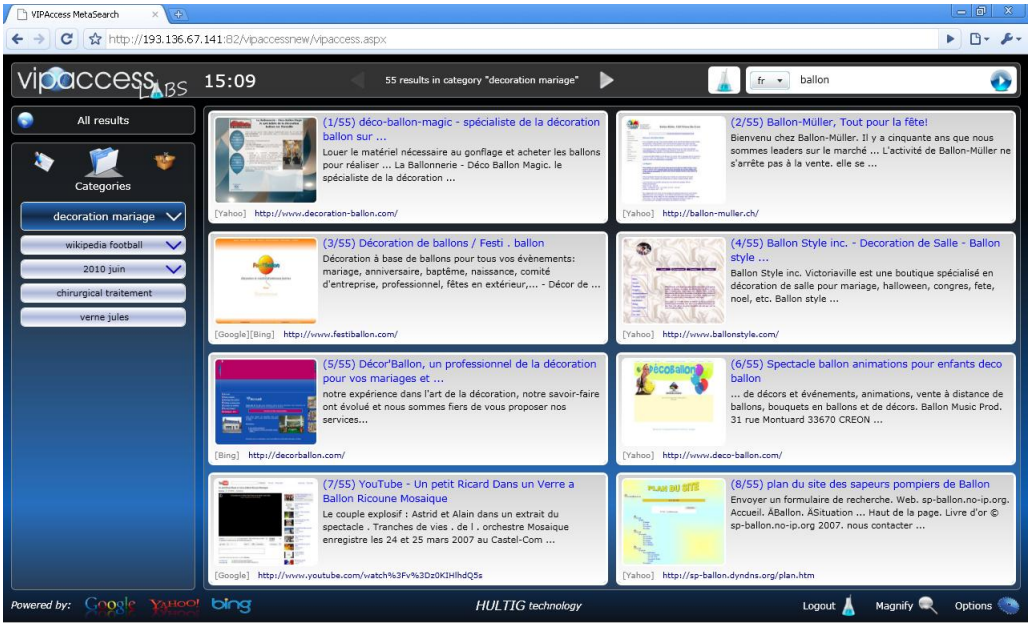

Figure 4.8: The HISGK-means algorithm for the query *ballon* [$14^{th} July, 2010$].

Algorithm 4 The WISE architecture.

Retrieve Web page results from Google Web Services.
Select relevant Web pages.
Identify MWU within all selected Web pages with SENTA.
Build flat clusters using Webspy.
Cluster flat clusters with PoBOC.
Label clusters.

results, which (1) are domains (e.g. http://www.vodafone.com) and (2) any relative url, which number of occurrences of its Web domain is greater than the average of all domains returned by the search engine. In addition, we extend the number of relevant Web pages returned by the requested search engine by considering a set of Web pages, which were not caught initially by the system, but related to the query. For that purpose, for any relevant retrieved domain, we catch its N best Web pages re-running the search engine over the domain alone with the same query.

Next, we identify all contiguous[16] MWU from all the Web page results selected from the previous step. The key idea is to better understand the message conveyed by any Web page and as a consequence benefiting the clustering process by

[16]Dealing with non-contiguous sequences of words is a complicated problem as many nested combinations may exist, implying the description of a decision module. This issue is important as [Ferragina 2008] obtains improved results with gapped sentences, but will be treated in future work.

integrating implicit knowledge about language into Web pages as well as identifying potential meaningful cluster labels. For that purpose, we used the SENTA software [Dias 1999a] implemented efficiently with the virtual corpus approach in [Gil 2003b] and [Gil 2003a].

As [Carpineto 2004] refer to, search engines suffer from lack of a concise representation of the retrieved documents. [Lawrie 2003] consider that the main challenge in creating hierarchical clusters is to select terms that will accurately describe the documents. Within this context, most approaches have been trying to extract meaningful words or repeated sequences of words within documents or Web snippets. However, it is important to keep in mind that not all the words or n-grams present in the texts are related to the query. Indeed, in the specific case of search engines, the representation of the content of any given url must be evaluated in correlation with the query. This is the case for Web snippets, which evidence the words or sentences of the Web page relevant with respect to the query. However, when dealing with full texts, the relevant text segments must be identified based on the query. For that purpose, we propose to use Web content mining techniques to extract (possibly hidden) knowledge correlated to the query as reported in [Dias 2004a] within the context of the Webspy software.

The Webspy software implements a C5.0 decision tree [Quinlan 1996] based on 8 characteristics (i.e. inverse document frequency, normalized text frequency, normalized density, normalized first position, size in characters, capital letters, SCP (see Equation 2.4) and normalized distance to the query), which identifies all the words and phrases correlated to the query terms in a given document. Many authors showed, such as [Turney 2000], that decision trees apply particularly well to textual data. Moreover, as [Zeng 2004] refer to, although a supervised learning method requires additional training data, it makes the performance of search results clustering significantly improved. In particular, the C5.0 decision tree was built over 5 pruned decision trees previously trained using a 5-fold cross validation and applying over-sampling to avoid data sparseness. The model was trained over a set of newspapers article gathering three distinct domains (sports, politics and society) in order to assure its generality, domain and topic independence, obtaining 82.49% of average precision over the five test data to classify positive and negative examples, and 60.72% average precision over the same five test data to classify only positive examples. As a consequence, Webspy retrieves the set of the terms within any Web page, which more correlate to the query with a probability of relevance and as a consequence represent as best as possible the semantic content of each text constrained by the query terms.

The representation of any Web page is then a vector containing its most relevant scored phrases or words, which can be called key concepts. From this vector representation, we could directly apply any polythetic clustering algorithms. However, by inverting the vector representation, each element of the key concept

vector can be the label of a flat overlapping cluster with a list of related urls as shown in Table 4.2 for the query *benfica*, a football team and a neighborhood of Lisbon, where *T1* is a flat with one room, *Aluguer* means *for rent* in Portuguese and all other key concepts are named entities related to the football team. In fact, in this first step of the algorithm, we propose a Label-centered strategy to create flat clusters, as the leaves of the hierarchical taxonomy must be of high precision and usually contain the same meaningful words or phrases. As a consequence, we transform a classical document clustering problem into a problem of flat cluster clustering in the next step of our architecture.

Flat cluster label	Web page results
Vilarinho	http://www.abola.pt http://www.slbenfica.pt
Eusébio	http://www.abola.pt
T1	http://www.era.pt
Aluguer	http://www.era.pt
Nuno Gomes	http://www.slbenfica.pt
Luís Filipe Vieira	http://www.slbenfica.pt

Table 4.2: Flat clustering.

In order to process two-levels hierarchical clustering, the Webspy software is applied to each label of a flat cluster together with its set of Web pages. As a consequence, each flat label is represented by a set of key concepts with a given relevance probability, where each key concept is not only related to the initial query terms but also to the label of the flat cluster, thus allowing a fine-grained representation of flat clusters. So, a square flat cluster label $\times$ flat cluster label similarity matrix is built based on the cosine measure (see Equation 2.14) over attribute vectors representing the relevance probability of each key concept given by the decision tree. This similarity matrix is the input of the PoBOC algorithm [Cleuziou 2003].

By applying the PoBOC algorithm to the flat cluster label matrix, we propose a disambiguation methodology, as key concepts with different meanings are likely to be gathered in different soft clusters. As mentioned earlier, unlike other clustering algorithms and in particular the clustering by committee (CBC) [Pantel 2002], PoBOC can be used "on the fly" as it does not depend on any input parameter. Moreover it has shown encouraging results in [Cicurel 2006] compared to other clustering algorithms when applied to textual data, building a hierarchy of concepts. In particular, the final hierarchy contains only two levels, which tends to lead to the best ephemeral clustering architecture according to [Hearst 1996] and [Maarek 2000].

Finally, each cluster must be labeled. Unfortunately, labeling has not received its due attention as many different issues appeared during the construction

of the overall architecture, mainly the processing time, which limited our evaluation. So, each cluster is labeled using a simple heuristics that chooses the key concept that occurs more often within the flat label vectors of the cluster, taking into account the sum of scores in case of ties[17]. An example of the overall architecture is presented in Figure 4.9 for the query *benfica* from a set of 100 documents retrieved from the Google^{TM} Web service, which resulted in a set of 23 clusters.

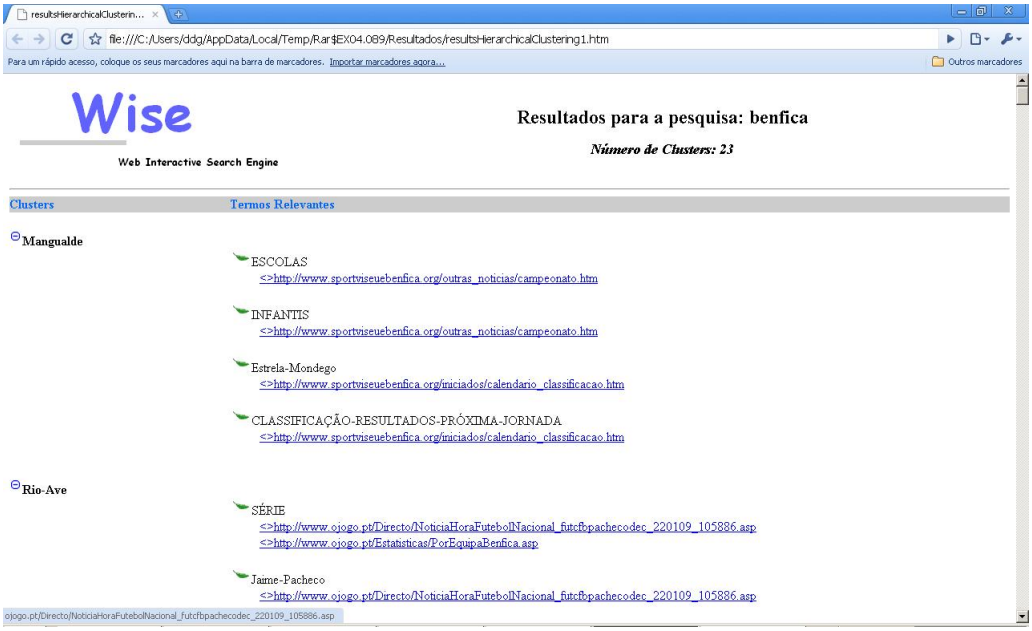

Figure 4.9: The results of WISE for the query *benfica* [$15^{th} May, 2005$].

Although interesting results were obtained, the processing time was unendurable, taking more than 40 minutes to answer any user query on a single processor unit, mainly due to repetitive access to full texts and network latency. To be real-world adaptable, we proposed in [Campos 2008] a peer-to-peer architecture to handle queries in affordable time. As a result, by using a small network of only 3 computers, WISE was able to compute the hierarchical taxonomy 10 times faster than when based on one single server, thus reducing processing time to minutes. However, this was still unbearable within the context of information retrieval. But, if more peers were installed, the processing time would automatically decrease and reach acceptable response speed. Moreover, Webspy could easily be parallelized as all features are independent from each other. As a consequence, it would be easy to compute them in parallel and gather the final results on one main peer in seconds. Regrettably, building such an infrastructure is costly and requires large efforts in terms of technical manpower, which are usually far from the research objectives of Universities. As a consequence, we decided to orient our research towards Web snippets instead of dealing with full texts and to focus on the problem of the

[17] We will show more interesting works in the following two sections about cluster labeling.

over-generation of clusters, which relaxes the compactness paradigm proposed in [Kummamuru 2004].

4.2 Normalized Snippet Hierarchical Overlapping Clustering

Although it was to be expected that document-based and snippet-based approaches would be similar in terms of results with small advantages to using the whole document contents as evidenced in [Zamir 1998], the experiments showed that dealing with Web snippets is a much more difficult task than one could expect in the context of the polythetic strategy. Indeed, Web snippets are not well-formed sentences as most do not contain verbs and are usually just lists of related words as shown in Figure 4.10.

Figure 4.10: Web snippet for the query *ballon* [$14^{th} July, 2010$].

This structure clearly affects the clustering phase and the labeling stage, within the polythetic paradigm. Indeed, as stop-words are not frequent in Web snippets, they can not be eradicated by the usual idf score [Spärck Jones 1972]. But, more important, they are erroneously given high relevance scores due to their high discriminant values within Web snippets, which tend to bias clustering algorithms, as well as cluster labeling. A direct consequence of the structure of Web snippets over the polythetic algorithms, is that the classical vector space model proposed by [Salton 1975], based on the tf.idf and the cosine similarity measure to evaluate the similarity between Web snippets, and evaluated in [Hearst 1996] and [Jiang 2002], is likely to lead to erroneous results. This is certainly the main reason why new polythetic algorithms have not been proposed since [Zhang 2001] to the exception of our works. Indeed, all studies proposed since then are monothetic ones, which first focus on the extraction of potential label candidates and then agglomerate documents, which contain the given labels.

To avoid the over-evaluation of empty words within the vector space model, we propose to evaluate the similarity between Web snippets with surface-based text similarity measures. Such measures evaluate the similarity between texts based on the counts of their overlapping word or character n-grams. Within this context, we can mention the well-known edit distance [Levenshtein 1966], the word n-gram overlap [Barzilay 2003a], the exclusive longest common prefix word n-gram

overlap [Cordeiro 2007a], the BLEU measure [Papineni 2002] and the ROUGE measure [Lin 2004]. However, recently, we proposed in [Cordeiro 2007a] a new surface-based similarity measure, called the Sumo-metric (see Equation 5.9), which outperforms all other measures in the context of paraphrase identification. So, after normalizing all returned Web snippets with the SENTA software, we build a Web snippet × Web snippet matrix, where each cell corresponds to the surface similarity between two specific Web snippets based on the Sumo-metric. As a consequence, no stemming or lists of stop-words are used to process the similarity between Web snippets. Based on this similarity matrix, we then applied the PoBOC algorithm as well as the single-link HAC, the complete-link HAC, the group-average HAC and the QT algorithms[18] in order to evaluate if different algorithms would lead to more compact taxonomies. Finally, the labeling phase was studied.

For that purpose, we implement the differential cluster labeling proposed in [Manning 2008] and present a new methodology called the centroid cluster labeling. On the one hand, the differential cluster labeling is based on the same idea as the tf.idf but taking into account the distribution of words among clusters instead of documents. As a consequence, a word is likely to be a cluster label if it is frequent in the cluster and loosely spread over all other clusters. On the other hand, the centroid cluster labeling is based on the idea that the document nearest to the centroid of the cluster is more likely to contain the cluster label. So, each Web snippet is represented by the vector of its words or phrases weighted by their cluster frequencies and the centroid is just the mean vector representing the center of gravity of the cluster. The Euclidean distance (see Equation 3.7) is then applied between all the Web snippets and the centroid to determine the Web snippet, which best represents the cluster. The word or phrase with the highest cluster frequency is then selected for cluster label.

From this study, different conclusions could be drawn. On the one hand, the different clustering algorithms did not manage to reduce the number of generated clusters to a reasonable size and the PoBOC algorithm reached best results within this scope, thus confirming a recent comparative study by [Cicurel 2006], who show that both the CBC and the PoBOC algorithms lead to relevant results for text data clustering. On the second hand, the surface-based text similarity approach outperformed the classical vector space model due to the structure of Web snippets, which over-evaluates the strength of empty words. On the third hand, the centroid cluster labeling logically provided better results than the differential strategy once again due to the structure of Web snippets. However, it is clear that Web snippet segmentation is one of the most important task for snippet-based ephemeral clustering as mentioned in [Lawrie 2003]. Indeed, due to the specific structure of Web snippets, SENTA was not able to extract high quality phrases as well as extracting topical words was a difficult task without the use of lists of stop words.

[18] Keeping in mind that only two hierarchical levels were kept.

As a consequence, following [Lawrie 2003] assumption, we decided to run an experiment involving part-of-speech tagging of Web snippets as recent works such as [Kummamuru 2004] and [Ferragina 2008] had been proposing to part-of-speech tag Web snippets to improve the precision of text segmentation. For that purpose, after identifying the language of each Web snippet using the methodology introduced in [Beesley 1988], the Treetagger from [Schmid 1994] was applied to all French, Portuguese and English Web snippets. SENTA was then run and all text segments corresponding to given patterns (such as in Proposition 4.1 for the English language) were kept as the attributes of the vector representation of Web snippets.

$$[\text{Noun} \mid \text{Verb} \mid \text{Adj} \mid \text{Noun Noun+} \mid \text{Adj Noun+} \mid \text{Noun Prep Noun}] \tag{4.1}$$

In particular, word attributes were weighted based on a linear interpolation of their relative frequency and relative position in the Web snippets, while phrase attributes were weighted based on a linear interpolation of their relative frequency, the relative position of their first word in the Web snippets and the average frequency of all their word constituents. Finally, the cosine similarity measure was computed over these vectors to build a Web snippet $\times$ Web snippet similarity matrix.

Interestingly, the clustering results were only slightly improved by the introduction of linguistic information, leading to almost similar results as the one obtained with the surface-based strategy but still over-generating clusters. However, the labeling phase clearly improved for both methodologies, although the centroid cluster labeling reached more satisfactory results. To some extent, these results confirmed the claims in [Lawrie 2003] as the identification of relevant text segments effectively leads to improved results. But, applying part-of-speech tagging is not the ultimate solution to real-world ephemeral clustering as it turns any application into a language-dependent solution, thus limiting its scope of action. Moreover, part-of-speech tagging is difficult over Web snippets as their structure does not allow high precision results and may induce incorrect decisions in the extraction of potential labels as can be evidenced in Figure 4.11, where *includes animal* is clearly the result of incorrect part-of-speech tagging of the word *includes* into an adjective or a noun. Moreover, well-known syntactical patterns to identify MWU proposed by [Justeson 1993] or [Daille 1996] fail when dealing with Web snippets. Indeed, *apple movie*, *movie theater* or *consciousness stanford* are clearly not MWU, but recurrent text segments within Web snippets.

At this stage of our research, we clearly understood that work had to be done on Web snippet segmentation. Indeed, on the one hand, using lists of stop-words and stemming algorithms are nothing more than *ad hoc* engineering tricks, which easily eliminate a given problem and go against the *corpus integrity principle* enounced in [Dias 2000a]. On the other hand, part-of-speech tagging can improve ephemeral

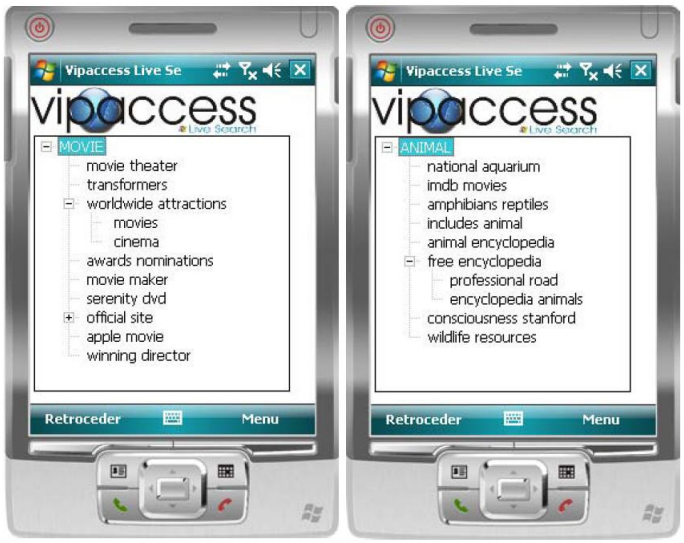

Figure 4.11: VIPACCESS mobile [$26^{th} January, 2007$].

clustering but limits its scope to a small number of languages for which it is possible to build linguistic tools. Moreover, specific part-of-speech taggers would have to be developed to deal with the specificity of Web snippets, as their structure is still a clear obstacle to reach high precision tagging. The second problem evidenced by both polythetic and monothetic strategies is the over-generation of clusters as existing similarity measures are unable to capture the semantics of the Web snippets and thus leading to a great number of small clusters instead of a small number of great clusters. These two issues are addressed in the following section.

4.3 Snippet Informative Hierarchical Clustering

As mentioned in [Kummamuru 2004], the main purpose of ephemeral clustering is to summarize and provide a better browsing experience. As a consequence, the generated taxonomy should be as compact as possible. Within this context, we would like to go even further. Indeed, if the generated taxonomy is compact but does not embody all the possible meanings of the query terms, all efforts in building adapted user interfaces will be useless. In fact, the main intent of post-retrieval document browsing is to produce hierarchical taxonomies as dense as possible embodying all the possible meanings of the query terms. Although this definition has never been proposed so far, many researchers have shown interests in dealing with polysemous queries. [Zeng 2004] mention that going through long lists of Web search results and examining the titles and Web snippets sequentially to identify required results, is a time consuming task for any user when multiple sub-topics of a given query are mixed together. In particular, they exemplify this situation with the query *jaguar* for which the user should go to the 10th, 11th, 32nd and

71st results to get search results related to *big cats*. [Kummamuru 2004] also point at the importance of polysemous queries by carrying out a user study with both ambiguous and non-ambiguous queries and show that their algorithm DisCover outperforms both CAARD [Kummamuru 2001] and DSP [Lawrie 2003], especially in the case of ambiguous queries, where the difference is higher. Within this scope, [Kummamuru 2004] state that *the utility of a good hierarchy is more evident when the queries are ambiguous*. However, so far, all methodologies have failed to reach query disambiguation as too many clusters are generated and different sub-topics are mixed together with different meanings as illustrated in Figures 4.4, 4.5, 4.6 and 4.7 for the query *ballon*, which can embody different meanings as the ones listed in Table 4.3 based on WordNet and much other ones not listed in WordNet as the *balloon tamponade*.

Word	Sense	Gloss
ball	1	round object that is hit or thrown or kicked in games
ball	3	an object with a spherical shape
balloon	1	mall thin inflatable rubber bag with narrow neck
balloon	2	large tough nonrigid bag filled with gas or heated air

Table 4.3: Different meanings for *ballon* in English based on WordNet.

But to reach high precision results and query-based disambiguation, work on Web snippet segmentation must be done to be able to clearly understand the contents of Web snippets. In [Machado 2009b], we propose to extract relevant meaningful words from Web snippets, while eliminating empty words automatically based on the study of word context distributions. As a consequence, both monothetic and polythetic algorithms will be likely to show improved results, as Web snippets will be represented by highly relevant meaningful words. In order to identify potential relevant text segments, most methodologies have been proposed in the context of Label-centered algorithms. Most of them are based on the extraction of frequent sets of words that appear together in more than a minimum fraction of the whole document set. For that purpose, different approaches have been proposed. [Zamir 1998] implement a suffix tree structure, [Zhang 2001] and [Osinski 2004] propose a suffix-array methodology, [Fung 2003] use association rules to extract itemsets, [Zeng 2004] learn a linear regression and [Ferragina 2008] propose to extract common gapped sentences from linguistically enriched Web snippets. As one may want to search over the entire Web in any language, it is important that the clustering algorithm only depends on language-independent features. Within this scope, the identification of relevant text segments is mainly based on frequency of occurrence as the unique clue for extraction. However, this methodology suffers from the poor quality of Web snippets, which mainly contain ill-formed sentences with many repetitions. As a consequence, meaningless text segments are likely to be extracted and must be post-processed to reach satisfactory results. Moreover, all these methodologies eliminate empty words, thus facilitating the extraction process.

To overcome these drawbacks and keep to the *corpus integrity principle*, we propose to weight strings based on three different word analyzes and consequently extract meaningful words without relying on lists of stop-words.

First, we evaluate the internal value of a text segment. If a string appears alone in a text span separated on both sides by any given delimiter, this string is likely to be meaningful. This characteristic $W_1(.)$ is defined in Equation 4.2 where w is any string, $A(w)$ is the number of occurrences where w appears alone in a chunk and $F(w)$ is the total number of occurrences of w.

$$W_1(w) = \frac{A(w)}{\log F(w)}. \tag{4.2}$$

Then, we assess the external value of a string following the idea proposed in [Frantzi 1996] based on the study of word context distributions. This idea is stated as follows. The bigger the number of strings that co-occur with any string w both on its left and right contexts, the less meaningful this string is likely to be. This characteristic $W_2(.)$ is defined in Equation 4.3 where $WIL(w)$ (resp. $WIR(w)$) is the number of strings which appear on the immediate left (resp. right) context of the string w.

$$W_2(w) = \frac{WIL(w) + WIR(w)}{2 \times F(w)}. \tag{4.3}$$

However, when a word w is not at the beginning or at the end of a delimiter, $W_2(.)$ leads to incorrect evaluation. So, we propose this new characteristic. The bigger the number of different strings that co-occur with any string w both on its left and right contexts compared to the number of co-occurring strings on both contexts, the less meaningful this string is likely to be. This characteristic is defined in Equation 4.4 where $WDL(w)$ (resp. $WDR(w)$) is the number of different strings which appear on the immediate left (resp. right) context of the string w and $FH(w)$ is equal to $max[F(w)]$, $\forall w$.

$$W_3(w) = \begin{array}{l} \left(\frac{WDL(w)}{WIL(w)} + \frac{WDL(w)}{FH(w)}\right) \times \frac{WIL(w)}{F(w)} + \\ \left(\frac{WDR(w)}{WIR(w)} + \frac{WDR(w)}{FH(w)}\right) \times \frac{WIR(w)}{F(w)}. \end{array} \tag{4.4}$$

Based on these three characteristics, we propose to weight all strings from the Web snippets as in Equation 4.5 such that a small $W(w)$ value corresponds to a meaningful string w.

$$W(w) = \begin{cases} W_2(w) \times W_3(w), W_1(w) < 0.5 \\ \frac{W_2(w) \times W_3(w)}{1+W_1(w)}, W_1(w) \geq 0.5. \end{cases} \tag{4.5}$$

In order to illustrate the efficiency of the $W(.)$ score, we present in Table 4.4 the 30 most relevant words for the query term *programming* searched over Google^{TM}, Yahoo!^{TM} and Bing^{TM} accessed via their respective Web services by our VIPACCESS meta-search engine.

Words (1-6)	Words (7-12)	Words (13-18)	Words (19-24)	Words (25-30)
articles	java	wiki	home	internet
wikibooks	php	security	advanced	tips
computers	training	database	documentation	science
compilers	forums	cgi	news	object-oriented
subject	tutorials	category	net	site
perl	c	knuth	unix	downloads

Table 4.4: The first 30 words for the query *programming* [$3^{rd} March, 2009$].

Based on Equation 4.5, it is possible to isolate single words from Web snippets with relatively high accuracy, thus allowing to represent Web snippets contents with high precision. This work on Web snippet segmentation answers to some extent to [Lawrie 2003] assumption. However, the identification of MWU is a crucial issue for the success of ephemeral clustering. As a consequence, not only meaningful words should be extracted but also phrases, which may convey the message of Web snippets with great accuracy. Within this scope, as we showed in the previous section that SENTA fails to extract meaningful n-grams with high accuracy due to Web snippets odd text structures, we started to work on the extraction of frequent itemsets containing base words (i.e. single meaningful words) based on a suffix-array data structure, which seems to open interesting directions as promising results are illustrated in Figures 4.12 for English, 4.13 for French and 4.14 for Bulgarian.

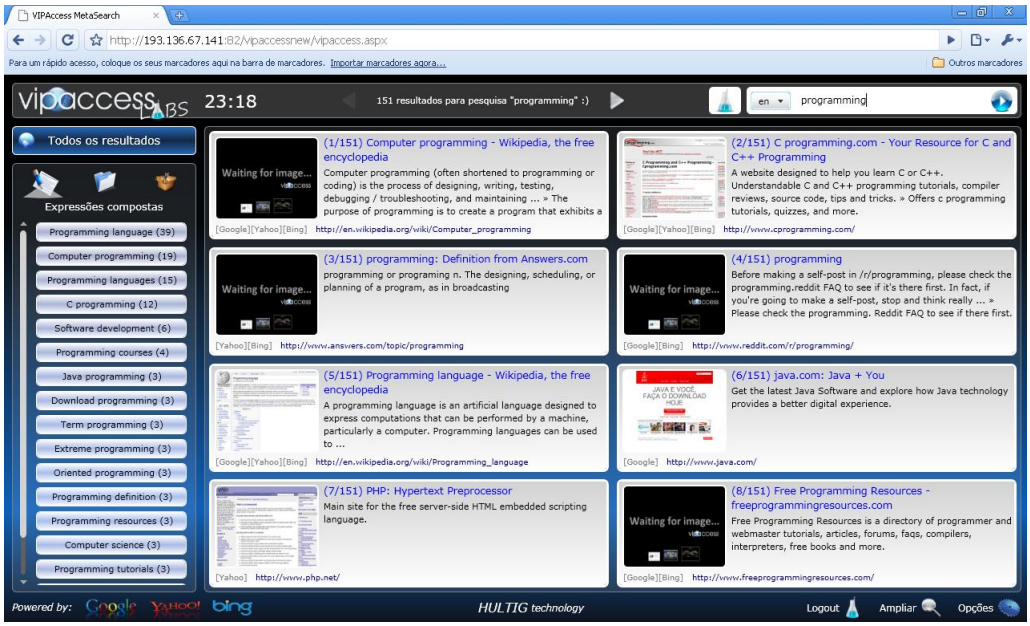

Figure 4.12: VIPACCESS for the query *programming* [$16^{th} August, 2010$].

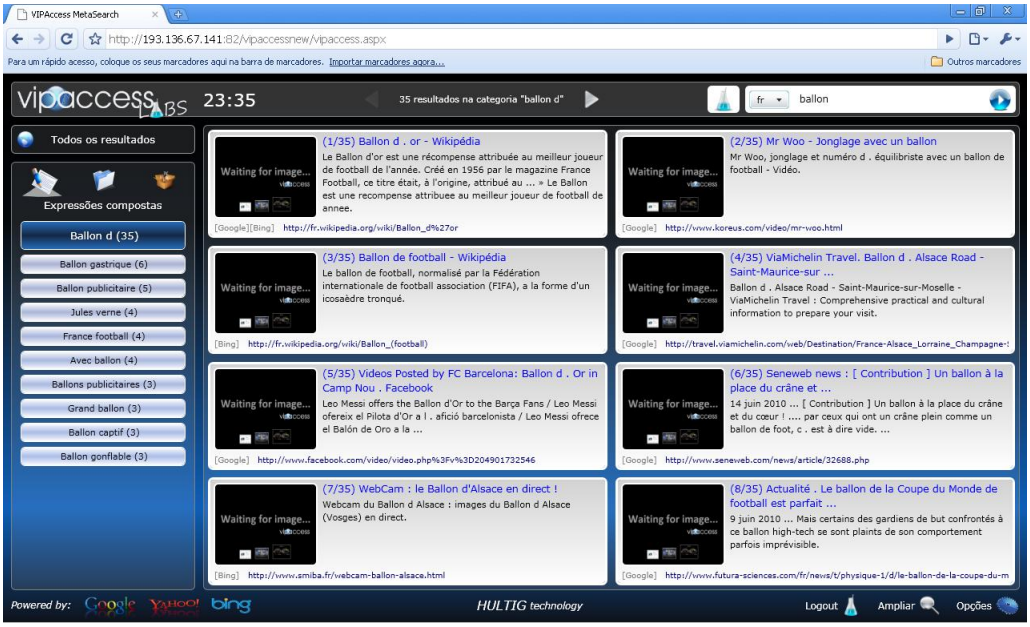

Figure 4.13: VIPACCESS for the query *ballon* [16th *August*, 2010].

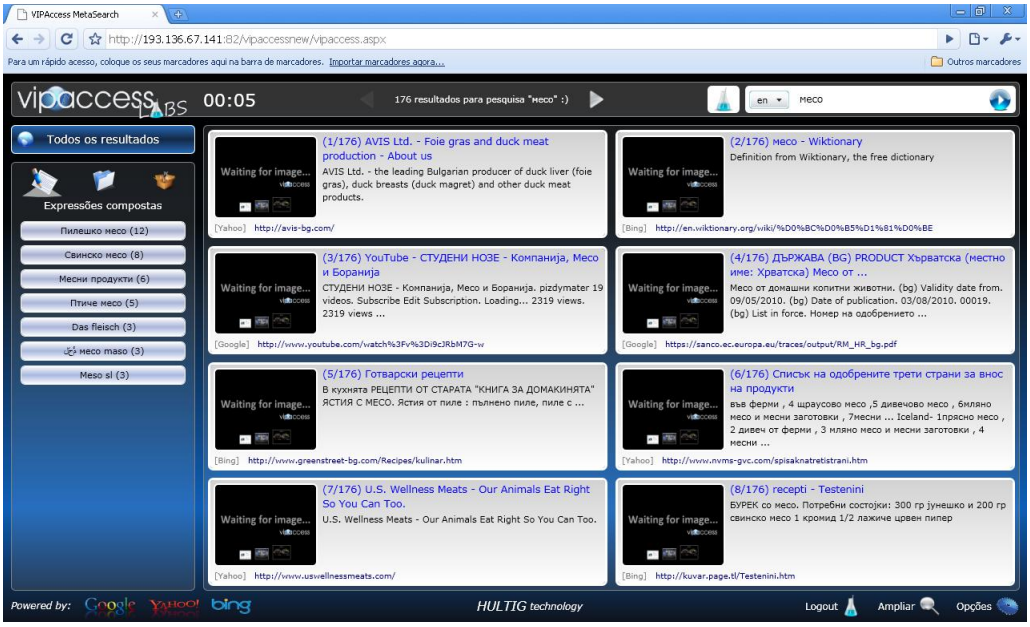

Figure 4.14: VIPACCESS for the query MECO (meaning *meat*) [16th *August*, 2010].

Another constatation from our work in the field of ephemeral clustering is that all methodologies proposed so far have failed to reach query disambiguation as too many clusters are generated and different sub-topics are mixed together with

different meanings. To overcome this situation, we started to work on a new architecture with Guillaume Cleuziou from the University of Orléans (France) and David Machado from the University of Beira Interior (Portugal), which allows to build compact hierarchical taxonomies, which best embody the different possible meanings of any query terms. The basic idea is simple. While existing methodologies, both polythetic or monothetic, evaluate the similarity between Web snippets based on the exact match of constituents, we propose that two Web snippets are highly related if both share highly related (eventually different) constituents. So, similarity is not any more based on exact matching of constituents. This idea has already been mentioned in Chapter 2 in the context of the InfoSimba similarity measure, which we propose to extend to text similarity.

Although the PoBOC algorithm is well-suited to text data, as it does not depend on any threshold nor parameter, it lacks from theoretical background. As a consequence, PoBOC does not not ensure that an optimum clustering is found. To overcome this drawback, we propose a new hierarchical divisive hard clustering algorithm called the hierarchical InfoSimba-based global K-means ($HISGK$-means) for which we provide a well-founded mathematical background, which guarantees optimal clustering. In particular, the hierarchical process is a top-down approach which splits recursively a set of Web snippets based on a variant of the global K-means algorithm (GK-means) [Likasa 2003] combined with the InfoSimba informative similarity measure (see Equation 2.20), which we call the InfoSimba-based global K-means ($ISGK$-means).

So, given a set of Web snippets, each one being described by a word context vector (i.e. a set of words selected based on the lowest $W(.)$ scores), the main goal of our approach is to hierarchically organize the Web snippets into a compact taxonomy, guaranteeing that at each level of the hierarchy, we automatically find the most suitable number of clusters and extract for each cluster a small set of representative words (i.e. the cluster labels). The algorithm runs as in algorithm 5. Starting with all Web snippets, we first split the data set into non-overlapping clusters using the $ISGK$-means, which is then recursively applied to all the clusters obtained at previous steps. It is interesting to notice that both online and batch procedures of the $HISGK$-means are possible. The online process consists in splitting only the clusters when the user formulates his demand on a given cluster. Oppositely, the batch process consists in first building all the hierarchy before exploring it according to the user's request.

In order to understand all the procedure, we first need to describe the well-known K-means algorithm and its adaptation to the InfoSimba similarity measure. The K-means method is a well known geometric clustering algorithm based on work by [Lloyd 1982]. Given a set of n data points, the algorithm uses a local search approach to partition the points into K clusters. A set of K initial cluster centers is chosen. Each point is then assigned to the center closest to it, and the centers are

Algorithm 5 The $HISGK$-means algorithm

Input: A set of Web snippets S and a stopping criterion C
Output: A hierarchy
Initialize the root h_0 of the hierarchy to S
Initialize the level of the hierarchy to 1 i.e. $l = 1$
Initialize the number of representative words for the centroid to 2 i.e. $p = 2$
Apply $ISGK$-means at level h_0
Retrieve K_0 clusters $h_{1,1}, \ldots, h_{1,K_0}$
Link all clusters $h_{1,k}$ to their parent h_0
Label all clusters $h_{1,k}$ and h_0 based on their p-sized centroids
$l = l + 1$
$p = p + 1$
for Each cluster $h_{l-1,k}$ and C is true **do**
  Apply $ISGK$-means at level h_{l-1}
  Retrieve K_l clusters $h_{l,1}, \ldots, h_{l,K_l}$
  Link all clusters $h_{l,k}$ to their parent h_{l-1}
  Label all clusters $h_{l,k}$ and h_{l-1} based on their p-sized centroids
  $l = l + 1$
  $p = p + 1$
end for

recomputed as centers of mass of their assigned points. This is repeated until the process stabilizes. It can be shown that no partition occurs twice during the course of the algorithm, and so the algorithm is guaranteed to terminate. So, the K-means procedure consists in the following algorithm 6.

Algorithm 6 The K-means algorithm

Input: Number of K, Data set X, List of Centroids L_{in}
Output: K partitions, List of Centroids L_{out}
Initialize K cluster centers in the data set X, randomly and/or using L_{in}
while convergence is not obtained **do**
  Assign each data $x_i \in X$ to its nearest cluster
  Update each cluster center by computing its centroid
end while

In order to assure convergence, an objective function Q must be defined which can be proved to decrease at each step of the algorithm. The K-means algorithm uses the objective function in Equation 4.6 to be minimized with $E(.,.)$ the Euclidean distance (see Equation 3.7), π_k the cluster labeled k, $x_i \in \pi_k$ an object in the cluster and m_{π_k} the centroid of the cluster π_k.

$$Q = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} E(x_i, m_{\pi_k})^2. \tag{4.6}$$

In the particular context of Web snippets clustering, the K-means algorithm needs to be adapted in order to use the InfoSimba similarity measure. Indeed, a Web snippet is not defined as a numerical vector but as a set of p words (i.e. a word context vector of size p) over which a proximity coefficient is defined, in this case, the simplified InfoSimba $ISs(.,.)$ defined in 2.21. In particular, all the words contained in the word context vector are given a score of 1 and first experiments were made based on the SCP (see Equation 2.4) and the PMI (see Equation 2.1) association measures as the similarity coefficient of the $ISs(.,.)$. As a consequence, we define the objective function Q_{IS} to maximize during the clustering process in Equation 4.7.

$$Q_{IS} = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} IS_s(x_i, m_{\pi_k}). \tag{4.7}$$

Notice that a cluster centroid m_{π_k} is now defined by a p-context vector of words $(w_1^{\pi_k}, \ldots, w_p^{\pi_k})$. As a consequence, we must define a way to update cluster centroids in such a way that Q_{IS} increases at each step of the clustering process. The choice of the best p words representing each cluster is a way of assuring convergence. For that purpose, we define the procedure UPDATE(π_k) which consists in selecting p words from the global vocabulary V in such a way that Q_{IS} is improved. The global vocabulary is defined by the set of all the words which appear in the context vectors of at least one Web snippet[19]. So, for each word $w \in V$ and any proximity coefficient PC (in this case, the SCP or the PMI), we compute its interestingness $\lambda^k(w)$ as regards to cluster π_k as defined in Equation 4.8 and we select only the p words with higher interestingness value to construct the cluster centroid. We can easily show that in such a way, Q_{IS} is maximized.

$$\lambda^k(w) = \frac{1}{p} \sum_{s_i \in \pi_k} \sum_{w_q^i \in s_i} PC(w_q^i, w). \tag{4.8}$$

So, the adaptation of the K-means within the context of Web snippets clustering is straightforwardly defined in algorithm 7 and called the InfoSimba-based K-means (ISK-means).

Algorithm 7 The ISK-means algorithm

Input: Number of K, Data set X, List of Centroids L_{in}
Output: K partitions, List of Centroids L_{out}
Initialize K cluster centers in the data set X, randomly and/or using L_{in}
while convergence is not obtained **do**
  Assign each data $s_i \in X$ to its nearest cluster using $IS_s(.,.)$
  Update each cluster center by computing its centroid using UPDATE(π_k)
end while

[19] Notice that V can be high.

Finally, now that the well-known K-means has been adapted to the case of Web snippet clustering, we introduce the GK-means clustering algorithm [Likasa 2003], which is at the basis of our overall $HISGK$-means algorithm. The GK-means constitutes a deterministic effective global clustering algorithm for the minimization of the clustering error that employs the K-means algorithm as a local search procedure. The algorithm proceeds in an incremental way. As such, to solve a clustering problem with M clusters, all intermediate problems with $1, 2, ..., M-1$ clusters are sequentially solved. The basic idea underlying the proposed method is that an optimal solution for a clustering problem with M clusters can be obtained using a series of local searches using the classical K-means algorithm. At each local search the $M-1$ cluster centers are always initially placed at their optimal positions corresponding to the clustering problem with $M-1$ clusters. The remaining M^{th} cluster center is initially placed at several positions within the data space. Since for $M = 1$ the optimal solution is known, it is possible to iteratively apply the above procedure to 2nd optimal solutions for all K-clustering problems $K = 1, ..., M$. In addition to effectiveness, the method is deterministic and does not depend on any initial conditions or empirically adjustable parameters. Moreover, its adaptation to the specific case of Web snippet clustering is direct as shown in algorithm 8. We call this algorithm the InfoSimba-based global K-means ($ISGK$-means), which is the core of our new hierarchical divisive hard clustering algorithm $HISGK$-means.

Algorithm 8 The $ISGK$-means algorithm

Input: Number of K, Data set X
Output: K partitions, List of Centroids L_{out}
Run ISK-means(1, X, [])
$Lcentroids_1 \leftarrow$ centroid of ISK-means(1, X, [])
for Each $k = 2$ to $k = K$ **do**
  Run ISK-means(k, X, $Lcentroids_{k-1}$)
  $Lcentroids_k \leftarrow$ centroids of ISK-means(k, X, $Lcentroids_{k-1}$)
end for

Once the clustering process has been handled, selecting the best number of clusters still remains to be decided. In most real life clustering situations, selecting the number of clusters in the final solution is a hard and still opened problem. Usually the user requires to define *a priori* the desired number of clusters. As a consequence, numerous procedures to determine the "best" number of clusters dividing a data set have been proposed [Milligan 1985]. However, none of the listed procedures where effective or adaptable to our specific problem. As a consequence, we decided to propose a new methodology based on the definition of a rational function which models the quality criterion Q_{IS} in the context of the $ISGK$-means algorithm. To better understand the behavior of the Q_{IS} at each step of the $ISGK$-means algorithm, we present in Figure 4.15 its values for $K = 10$, the query *scorpions* and 100 retrieved Web snippets from Google^{TM}, Yahoo!^{TM} and Bing^{TM} search engines. Indeed, Q_{IS} can be modeled as a rational function given in Equation 4.9 which

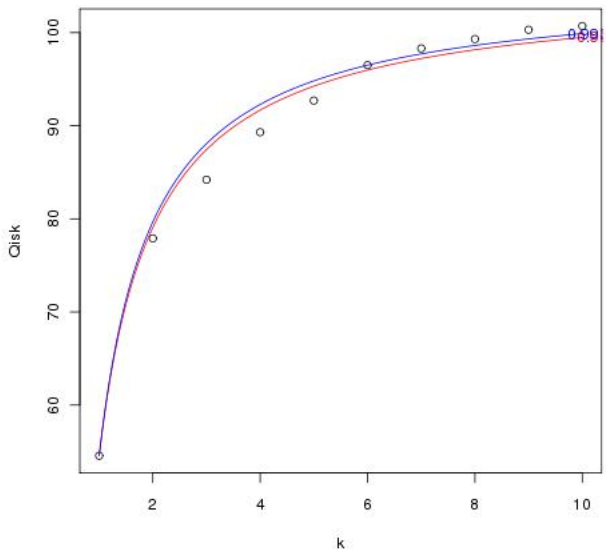

Figure 4.15: Q_{IS} and its model.

converges to a limit that we will call α when K increases and starts from Q^1_{IS} (i.e. Q_{IS} at $K = 1$). The basic idea of our approach is that the best number of clusters is given by the parameter β of the rational function which maximizes the difference with the average β^{mean}. For that purpose, we need to express α, β and γ parameters independently of unknown variables.

$$\forall K, f(K) = \alpha - \frac{\gamma}{K^{\beta}}. \tag{4.9}$$

As α can theoretically or operationally be defined as well as it can be easily proved that $\gamma = \alpha - Q^1_{IS}$, we need to represent β based on γ or α. This can also be easily calculated and the given result is expressed in Equation 4.10.

$$\beta = \frac{log(\alpha - Q^1_{IS}) - log(\alpha - Q^K_{IS})}{log(K)}. \tag{4.10}$$

Now, we only need to obtain the value for α, which best approximates the limit of the rational function. Within this scope, we computed its maximum theoretical value, its maximum experimental value and its approximated maximum experimental value based on the δ^2-Aitken [Aitken 1926] procedure to accelerate the convergence as explained in [Kuroda 2008]. The best results were obtained with the maximum experimental value. In particular, the maximum experimental value of α can be defined by building the cluster centroid m_{π_k} for each Web snippet individually. Indeed, the maximum value of α is obtained when each snippet is a cluster. For that purpose, we use the procedure UPDATE(π_k), which consists in

selecting p words from the global vocabulary V defined by the set of all the words in all the retrieved snippets. So, for each word $w \in V$, its interestingness $\lambda^k(w)$ is calculated as regards to cluster π_k and only the p words with higher interestingness value are selected to construct the cluster centroid. Then, Q_{IS} is evaluated as in Equation (4.7) and equals to α.

Finally, the selection process is based on the following idea. The best number of clusters is given by parameter β of the rational function which maximizes the difference with the same function which takes as parameter β the average value of all β values, i.e. β^{mean}. The procedure is defined in algorithm 9.

Algorithm 9 The best K selection procedure.

Calculate β^K for each K.
Evaluate the mean of all β^K i.e. β^{mean}.
Select β^K which maximizes $\beta^K - \beta^{mean}$.
Return K as the best number of partitions.

This situation is illustrated in Figure 4.15 where the red line corresponds to the average $\beta^{mean} = 0.96$ and the blue line to the maximum $\beta^{max} = \beta^6 = 0.99$ which selects the best partition $K = 6$[20]. In order to illustrate the soundness of the procedure, we present in Figure 4.16, the different values for β at each K iteration and the differences between consecutive values of β. We clearly see that the biggest inclination of the curve is between clusters 5 and 6, which also corresponds to the highest difference between consecutive values of β.

The $HISGK$-means shows interesting results as illustrated in Figure 4.8 for French, but also for Portuguese in Figure 4.17 and English in Figure 4.18 in the case of ambiguous queries but also for non-ambiguous queries as in Figure 4.19 for which different sub-topics of *New York* are displayed. Indeed, both compactness and query disambiguation seem to be approached in a much better way than existing reference works. The $HISGK$-means also shows interesting properties. First, it is mathematically well-founded so that optimum clustering is guaranteed. Second, the labeling step is embodied in the clustering process, thus avoiding an extra step to label clusters. Third, it is applied on a language-independent architecture. But more important, the $HISGK$-means opens many new research directions.

4.4 Future Work

The $HISGK$-means is certainly the clustering algorithm, which will deserve most attention from us in the next few months and years. Indeed, it is the only one to propose a compact hierarchical taxonomy in a language-independent framework as no linguistic information is introduced, thus keeping to the *corpus integrity*

[20] These values are calculated for $\alpha = 105$.

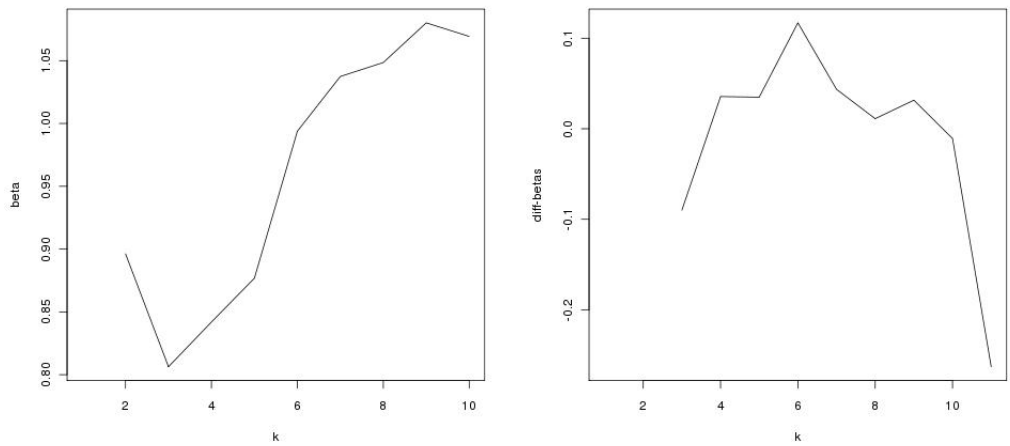

Figure 4.16: Values of β (on the left) and Differences between consecutive values of β (on the right).

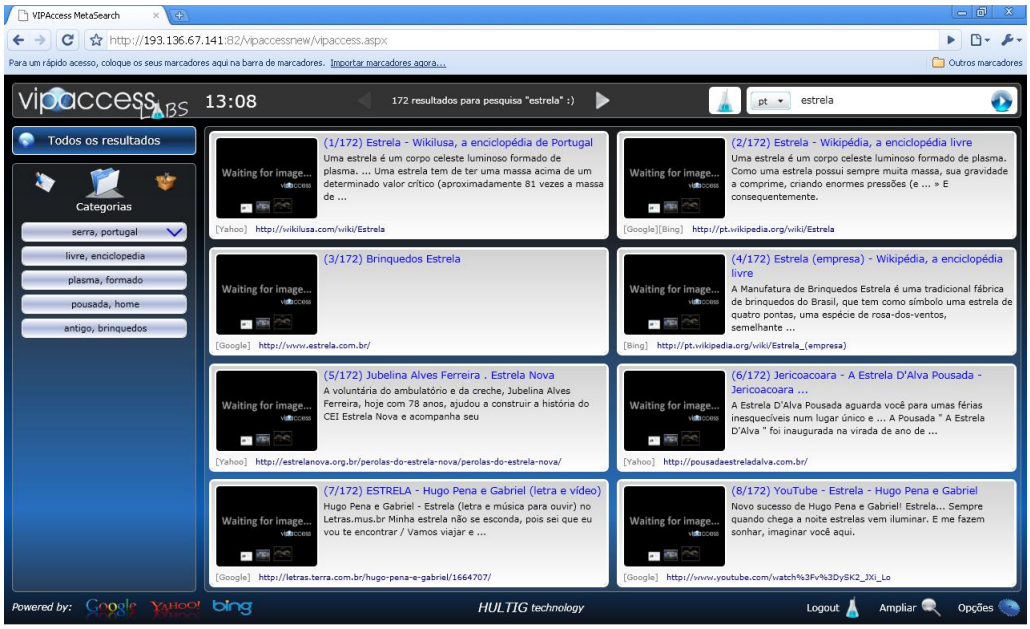

Figure 4.17: $HISGK$-means for the query *estrela* (meaning *star*) [17th $August$, 2010].

principle. However, although stimulating results are obtained, there is still much to improve about the overall methodology in the near future. In particular, there are many alternative informative similarity measures, which can be adapted to text similarity as defined in Chapter 2 and may lead to improved results. Moreover,

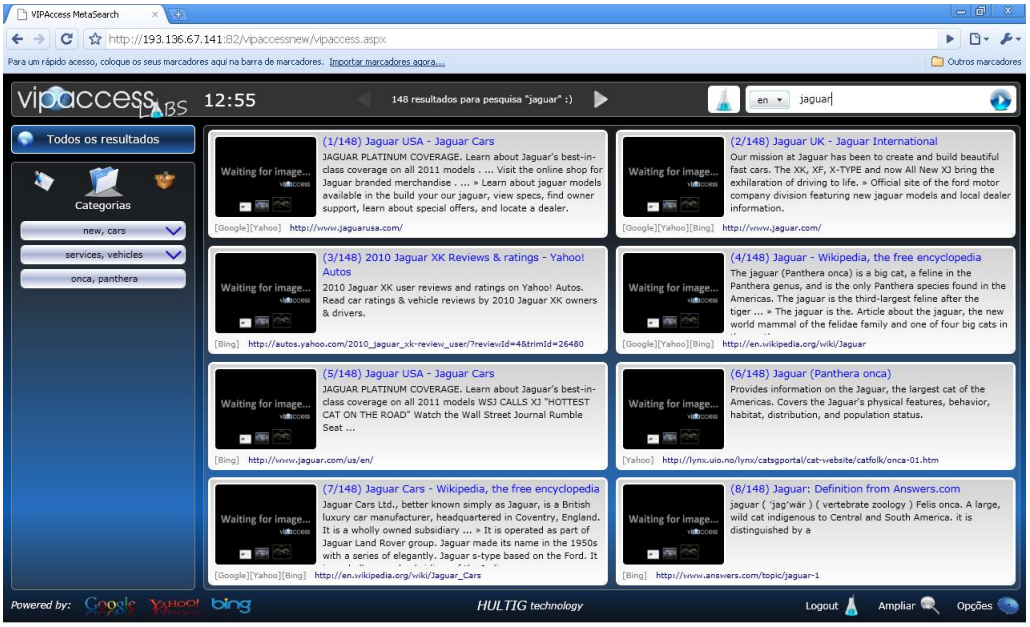

Figure 4.18: $HISGK$-means for the query *jaguar* [17th $August, 2010$].

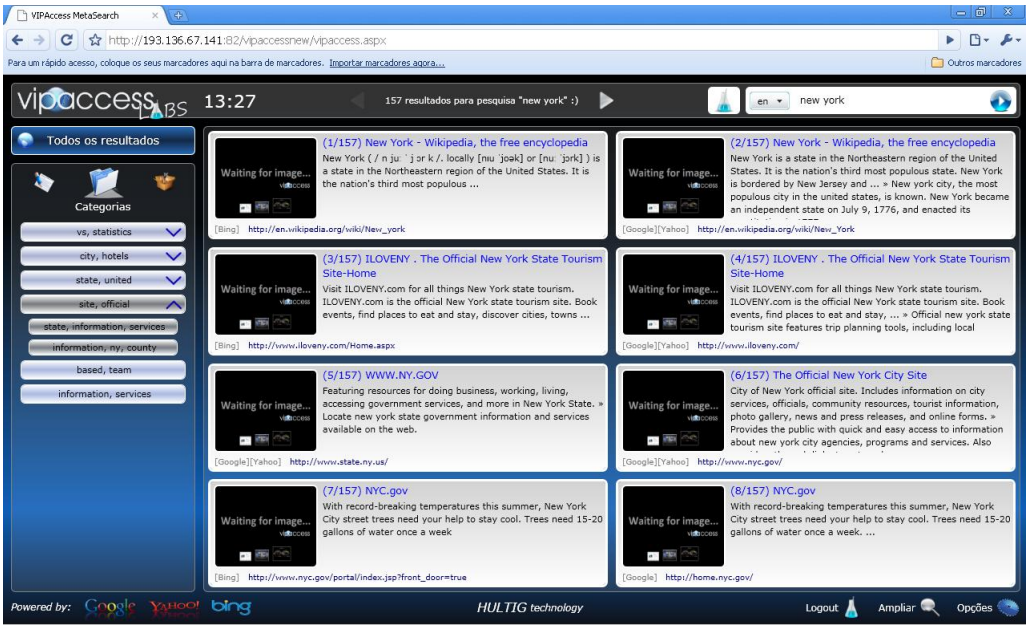

Figure 4.19: $HISGK$-means for the query *new york* [17th $August, 2010$].

the scores of the words can be taken into account as attribute values of the word context vectors and participate in the clustering process as well as the labeling phase. We also noticed that the SCP tends to provide best clusters while the

PMI is well-suited to encounter meaningful labels. Based on this observation, we are thinking about a new way to encounter meaningful cluster labels by building SCP-based centroids for clustering purposes and PMI-based centroids for labeling purposes[21]. MWU extraction is also an important issue for ephemeral clustering [Zamir 1998]. Although interesting results have been obtained from the extraction of frequent sequences of meaningful words, this process is still not as reliable as the SENTA software for well-structured texts. To overcome this difficulty, two different strategies can be followed: (1) relying on a local search engine indexing a large excerpt of the Web, such as the Open Directory Project database[22], and controlling the process of Web snippets generation or (2) deeply analyzing the Web snippets structures and adapt the GenLocalMaxs to extract meaningful frequent itemsets. The first strategy is likely to be followed as it will turn our architecture independent from commercial search engines and will allow to evaluate different algorithms based on the same indexed corpus, without having to deal with the dynamic Web i.e. the set of Web page results differing every minute for the same given query. Besides, if quality Web snippets can be produced, we may be able to extract meaningful non-contiguous sequences of words with SENTA, which may improve results as proposed in [Ferragina 2008]. Finally, an overlapping version of the $HISGK$-means algorithm must also be proposed to allow a given document to belong to different clusters. This issue is already being studied.

Although many results have been reported, all are difficult to analyze as they are not based on golden standards or standard measures of accuracy. In fact, very little has been done for the evaluation of ephemeral clustering. Some works propose automatic evaluations but each time on different data sets thus preventing possible comparisons such as in [Hearst 1996] [Zamir 1998] [Maarek 2000] [Fung 2003] [Lawrie 2003]. Some others base their evaluations on user studies such as in [Jiang 2002] [Zeng 2004] [Osinski 2004] [Kummamuru 2004] [Ferragina 2008], which may be biased as user subjectivity is always to be taken into account. Moreover, they are difficult to run for different algorithms as user studies are usually overwhelming tasks. Finally, some works do not even propose any evaluation such as in [Zhang 2001] [Carpineto 2004] [Campos 2005] [Dias 2009], who simply show some of their results. As a consequence of all these arguments, very few works propose comparative evaluations between different algorithms. [Zamir 1998] [Jiang 2002] [Kummamuru 2004] are the few exceptions. [Zamir 1998] compare their STC algorithm to the one proposed in [Hearst 1996] and four other well-known clustering algorithms as shown in Figure 4.2. [Jiang 2002] propose a visual comparison with the STC algorithm and [Kummamuru 2004] compare their Discover algorithm with their previous CAARD [Kummamuru 2001] and the DSP algorithm proposed in [Lawrie 2003]. It is clear that great efforts must be carried out to assess the quality of all existing algorithms by proposing both automatic evaluations on

[21] This idea is already being tested.
[22] http://www.dmoz.org/ [14th $July$, 2010].

golden standards as well as user studies for specific cases of user's need. Within this context, we have started a large spectrum user study for mobile devices a few months ago[23] and we expect to work on a golden standard data set in the near future.

Ephemeral clustering is an important step towards information digestion in the context of information retrieval. However, traditionally, text summarization is the ultimate way to digest information. In the next chapter, we present different works we have been developing during the last years both on sentence compression and text summarization, including topic segmentation and discourse representation via lexical chains ■

[23] Without results yet.

Chapter 5

Document Summarization and Sentence Reduction

Contents

5.1 Topic Segmentation 96
5.1.1 Related Work 97
5.1.2 Informative Topic Segmentation 98
5.2 Construction of Lexical Chains 102
5.2.1 Related Work 103
5.2.2 Construction of a Lexical-semantic Knowledge Base 104
5.2.3 Lexical Chainer Algorithm 105
5.3 Sentence Reduction 110
5.3.1 Related Work 111
5.3.2 Paraphrase Extraction 114
5.3.3 Paraphrase Alignment 120
5.3.4 Reduction Rules Learning 124
5.4 Future Work 129

Although powerful search engines exist to find relevant documents within millions of Web pages, judging the relevance of the information available on the Web is an overwhelming task. This situation is mainly due to the explosion of the Web associated to its democratization with Weblogs, on-line information services and social networks emerging everyday. As a consequence, the user is surrounded with information, which he must digest within a small amount of time. This situation is known as the information overload problem. In order to leverage this problem, the development of systems to automatically summarize texts has become the focus of considerable interests and investments for the last two decades. For instance, Newsblaster [McKeown 2003] allows the users to be updated about the interesting events happening around the world, without the need to spend time searching for the related news articles. IBM Remail [Rohall 2004] summarizes the threads of e-mail messages based on simple sentence extraction techniques. And, in order to avoid repetitive Web browsing on mobile devices such as PDA or smartphones, we recently proposed in [Dias 2009] an integrated software, which summarizes any Web page (as shown in Figure 5.1) based on simple extractive methodologies [Dias 2006a] [Dias 2007b].

Figure 5.1: Summarization for handled devices [$26^{th} January, 2007$].

According to [Mani 2001], the goal of automatic text summarization (ATS) is *to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.* ATS has traditionally been classified into two main categories: extractive and abstractive [Hahn 2000]. While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences, abstractive summarization puts strong emphasis on the form, aiming at producing a grammatical summary, which usually requires advanced language generation techniques.

The abstractive approach is much closer to the kind of summarization made by humans and is naturally much more difficult to automate, since it implies better understanding of texts. SUMMONS [McKeown 1995] [Radev 1998] is one of the first attempts following this direction by synthesizing a summary from filled template slots. In particular, its architecture consists of two major components: a content planner that selects the information to include in the summary through the combination of the input templates and a linguistic generator that selects the right words to express the information in a grammatical and coherent text. [Mani 1998] also describe an information extraction framework for summarization, a graph-based method to find similarities and dissimilarities in pairs of documents. Albeit no textual summary is generated, the summary content is represented via concepts and relations that are displayed respectively as nodes and edges of a graph. Rather than extracting sentences, they detect salient regions of the graph via a

spreading activation technique. Since then, existing works have been quite limited and can be broadly categorized into two categories: (1) approaches using prior knowledge [McKeown 1999] [Barzilay 1999] [Finley 2002] and (2) approaches using natural language generation systems [Jing 2000b] [Min-Yen 2002] [Carenini 2008]. Recently, [Ganesan 2010] proposed a new issue in abstractive summarization, which may open new research directions in the field as they assume no domain knowledge and use shallow NLP, leveraging mostly the word order in the existing texts and their inherent redundancies to generate informative abstractive summaries.

In parallel, the extractive approach aims at creating summaries by selecting the sentences, which best convey the initial text message[1]. This idea was first explored by [Luhn 1958], [Baxendale 1958] and [Edmundson 1969], who respectively proposed paradigms to extract salient sentences from texts using features like word and phrase frequency, position in the text and key phrases. This approach is usually referred as the Edmundsonian paradigm [Mani 2001] and is defined as the process of scoring each sentence of a source text and keeping the highest ranked ones, with respect to a given compression rate. Although individual clues may lead to interesting results, combining relevant characteristics is likely to benefit the scoring function. For instance, [Radev 2004] evidence that position and length are useful surface features to extract meaningful sentences in the context of multi-document summarization. A logical approach to take into account different features to rank sentences, is to propose machine learning environments. As a consequence, many studies appeared following this direction such as in [Kupiec 1995] [Aone 1999] [Conroy 2001] [Radev 2004] [Steinberger 2005] [Svore 2007] [Wong 2008]. However, most of these methodologies represent texts as bags of words and omit semantic structural information, which is clearly an important aspect of the summarization process as understanding the semantic structure of texts is likely to lead to coherent and cohesive summaries as stated in [Barzilay 1997]. In particular, [Morris 1991] point out that cohesion relates to the fact that the elements of a text *tend to hang together*, while coherence refers to the fact that *there is sense (or intelligibility) in a text.* As a consequence, different approaches to ATS appeared, which can be categorized into two classes: the lexical chain paradigm [Barzilay 1997] [Silber 2000] and the structural strategy [Ono 1994] [Marcu 1997] [Mithun 2010] [Uzêda 2010], who follow the rhetorical structure theory [Mann 1988]. More recently, semi-structure events have been investigated [Filatova 2004] [Liu 2007]. In particular, they balance document representation with words and structures and usually treat summarization as a three-component problem, involving (1) the identification of the textual units into which the input text should be broken and which are later used as the constituent parts of the final summary, (2) the textual features which are associated to the important concepts of the input text and (3) the algorithm for selecting the textual units to be included into the summary.

[1] Although most systems deal with extracting the best sentences, different granularity is possible as phrases or paragraphs can be the object of a summarization system.

It is generally agreed that automating the summarization procedure should be based on text understanding that mimics the cognitive process of humans. However, it may take some time to reach a level where machines can fully understand documents. In the interim, we must take advantage of other properties of texts, such as lexical cohesion analysis, which does not rely on full comprehension of text but provides a good indicator for the discourse structure of a document as well as its content. Within this scope, we started to work on an extractive summarization approach, which aims at identifying the most semantically and structurally important portions of texts to ensure both coherence and cohesiveness. For that purpose, we first proposed a new informative topic segmenter, called ITOS, which introduces new insights in the field of topic segmentation [Dias 2007a] and a new lexical chainer, which does not depend on pre-existing lexical-sematic knowledge bases but on an automatically built hierarchical taxonomy of concepts [Dias 2006b]. It is important to notice that, at this stage of our research, the whole extractive process has not yet been studied. Indeed, the selection of topically relevant sentences and their structuring aiming at guaranteeing cohesive and coherent summaries has not yet been achieved. This issue is mainly due to the fact that we have put all our efforts and focus on the automatic construction of lexical-semantic ontologies for the last several months and years. All this research is described in Chapter 6.

Best results to extract representative sentences from texts have been reached using topic segmentation, which automatically identifies topical portions of texts [Barzilay 1997] [Silber 2000] [Boguraev 2000] [Angheluta 2002] [Farzindar 2004] [Dias 2007a]. However, within the scope of topic segmentation, most of the works reach reasonable results when texts are artificially created with passages sharing no topics at all. For example, most experiments have been tested on the data set proposed in [Choi 2000], which concatenates passages from religion, fiction and humor. Of course, this situation is artificial and these systems drastically drop in accuracy when dealing with real-world coherent documents [Moens 2001] [Angheluta 2002] [Xiang 2003]. Besides, the systems proposed so far in the literature show three main other problems: (1) systems based uniquely on lexical repetition [Hearst 1994] [Reynar 1994] [Choi 2000] show reliability problems as common writing rules prevent from using lexical repetition, (2) systems based on lexical cohesion, using existing linguistic resources that are usually only available for dominating languages like English, French or German, do not apply to less favored languages [Morris 1991] [Kozima 1993] and (3) systems that need previously existing harvesting training data [Beeferman 1997] do not adapt easily to new domains as training data is usually difficult to find or build depending on the domain being tackled. To overcome these different drawbacks, we propose in [Dias 2007a] a new algorithm called ITOS which tackles informative topic segmentation. In particular, we evaluate sentence similarity with the InfoSimba informative similarity measure (see Equation 2.20) after text normalization is processed with SENTA [Dias 1999a] as well as we present a new heuristic to select

relevant topical changes. As we keep to our *corpus integrity principle* enounced in [Dias 2000a], we propose a language-independent methodology which does not use lists of stop-words nor linguistic resources or tools. Only word relevance and text similarities are analyzed, which allow total flexibility, reuse and language adaptation.

Topic segmentation evidences the topical text structure but not the semantically salient portions of texts. For that purpose, lexical chains proved to lead to successful results as evidenced in [Barzilay 1997] [Silber 2000]. Lexical chains represent the lexical cohesion of a text as they identify sets of words that are semantically related (i.e. have a sense flow), generally inferred from existing linguistic resources. As such, using lexical chains in text summarization is efficient, because these relations are easily identifiable within the source text and are likely to evidence salient text segments. By using lexical chains, it is then statistically possible to find the most important concepts by looking at the semantic structure of the document rather than relying on deep semantic analysis. However, their construction is always based on lexical-semantic resources for English such as WordNet [Miller 1990] or Roget's thesaurus [Roget 1852]. Within the context of our research based on language-independence and as a consequence on multilinguality, the use of resources specifically defined for given languages must be avoided. As a consequence, in [Dias 2006b], we proposed to build a lexical-semantic resource based on the combination of the InfoSimba to evaluate the similarity between words/phrases[2] pairs and the PoBOC hierarchical overlapping clustering algorithm proposed in [Cleuziou 2003], which gives rise to a prototype-based ontology (with no words or phrases inside the nodes of the hierarchy). Then, we introduced a new algorithm to build lexical chains based the generated taxonomy and a forced adaptation of the Lin similarity measure (see Equation 2.24) to take into account the absence of labeled nodes. In particular, we relied on a part-of-speech tagger to build a taxonomy of nominal expressions, as lexical chains represent the nominal lexical cohesion of a text. This way, we broke our "total" language independency paradigm. However, it has been shown that powerful part-of-speech taggers can be obtained from small training sets, i.e. in the order of 5000 words [Marques 2001], which lessens the impact on language dependency.

Once relevant sentences have been extracted and summarized coherently and cohesively, automatic text summarization can go even further in terms of information reduction i.e. sentence reduction[3]. Very few studies have been proposed, which tackle sentence reduction. Most proposals are prototypes which can not reasonably be applied in real-world applications. Some need deep linguistic analysis [Jing 2000a] [Knight 2002] [Turner 2005], while others would need terabytes of text to learn rewriting rules with limited coverage and accuracy [Le Nguyen 2004] [Specia 2010]. Within this context, we proposed a methodology based on three

[2]Identified by SENTA.

[3]Sentence reduction can be classified within the abstractive approach.

different steps. First, we use the Sumo-metric (see Equation 5.9) to extract corpora of asymmetric paraphrases from Web news stories. In particular, the Sumo-metric showed improved results compared to existing surface text similarities as illustrated in [Cordeiro 2007a]. It is important to notice that this process is completely language-independent as well as the used thresholds are learned using a bisection strategy over standard corpora, which adds to the universality of the methodology. Second, the obtained asymmetric paraphrases paraphrases are aligned using a mixture of global and local alignment algorithms borrowed from Bioinformatics [Needleman 1970] [Smith 1981]. Within this scope, we proposed a new algorithm, which tests the degree of word movements inside sentences to discover the best local or global alignments as presented in [Cordeiro 2007b]. Once again, no linguistic resources or manually defined heuristics are introduced in the overall process. Third and finally, rewriting rules are learnt based on inductive logic programming (ILP) [Muggleton 1991]. For that purpose, we used the Aleph system [Srinivasan 2000] based on shallow parsed aligned paraphrases. In particular, ILP shows different advantages compared to other machine learning techniques as its formalism easily allows to encode the studied problem and may induce rules with high accuracy without negative learning examples. For these reasons, it exactly fitted our requirements. Different experiments were carried out in [Cordeiro 2009], from taking words as the single learning information[4] to the combination of words, part-of-speech tags and syntactical chunks. The results showed that the introduction of linguistic features outperforms the raw text approach mainly due to data sparseness reduction.

5.1 Topic Segmentation

Improving access to information by dividing lengthy documents into topically coherent sections is a research area commonly called topic segmentation. It can be defined as the task of breaking documents into topically coherent multi-paragraph subparts. In order to provide solutions to access useful information from the ever-growing number of documents on the Web, topic segmentation is a crucial issue as people who search for information are now submerged with unmanageable quantities of texts. For that purpose, topic segmentation has extensively been used in information retrieval and automatic text summarization. In the context of information retrieval, it is clear that some users would prefer to find a document in which the occurrences of the query terms are concentrated into one or two paragraphs rather than loosely spread over the whole document. This particular research domain is usually called passage retrieval and proposes techniques to extract fragments of texts relevant to a query such as studied in [Salton 1993] [Kaszkiel 1997] [Kaszkiel 1999]. In the context of ATS, topic segmentation is usually used as the basic text structure in order to reach coherent sentence extraction as proposed in [Barzilay 1997] [Silber 2000] [Boguraev 2000] [Angheluta 2002] [Farzindar 2004] [Dias 2007a].

[4] To remain language-independent.

5.1.1 Related Work

[Hearst 1994] [Reynar 1994] [Choi 2000] [Sardinha 2002] proposed different architectures based on lexical item repetition: respectively, TextTiling, Dotplotting and the Link Set Median Procedure. However, it has been proved that systems based on lexical repetition are not reliable when applied to non-technical texts without small controlled vocabularies. For instance, articles in newspapers tend to avoid word repetition. In fact, good writing should avoid word repetition. As a consequence, these techniques can only be applied to technical texts where synonyms rarely exist for a given concept so that word repetition is almost compulsory.

In order to avoid these limitations, [Kozima 1993] proposed an architecture based on a semantic network built from the Longman dictionary of contemporary English (LDOCE[5]) from which lexical cohesion can be fine-grained induced. First, [Morris 1991] had proposed a discourse segmentation algorithm based on lexical cohesion relations called lexical chains using Roget's thesaurus. However, such linguistic resources are not available for the vast majority of languages so that both methodologies are drastically limited and as a consequence do not apply to less favored and emerging languages.

As a consequence, [Beeferman 1997] proposed a technique to identify document boundaries using statistical techniques. In particular, they built statistical models within a framework which incorporates a number of clues about the story boundaries such as the appearance of particular words before a boundary and the appearance of cue words in the beginning of the previous sentence of a boundary. Unfortunately, this work is limited by the need of previously existing harvesting training data as it proposes a supervised solution to the problem of topic segmentation. Once more, this lacks in flexibility as new training is necessary when the genre/domain/language change.

It is clear that unsupervised language-independent techniques, which automatically induce some degree of semantics propose a promising solution to solve all the exposed problems. [Phillips 1985] and [Ponte 1997] proposed such techniques. [Phillips 1985] proposes to identify a lexical network based on word collocation frequency statistics and cluster analysis. However, he does not propose a classical topic segmentation technique but rather a topic detection system as he does not output boundaries in the text. [Ponte 1997] propose a topic segmentation technique based on the local content analysis (LCA) [Xu 1996] allowing to substitute each sentence with words and phrases related to it. A pairwise similarity measure is then calculated between all transformed sentences and then introduced into a final score in order to find at each point in the corpus the block that maximizes the score function. The important point to focus on is the use of the LCA, which introduces some degree of semantics to the system without requiring harvesting

[5] http://www.ldoceonline.com [26th *August*, 2010].

linguistic resources and thus reducing the problem of word repetition. In order to introduce endogenously acquired semantic knowledge, [Ferret 2002] also proposed to automatically extract collocations from texts in order to compute semantic similarity measures.

Although our approach tends to stand to the basic ideas of these unsupervised methodologies, our main contributions are twofold. First, we apply the informative similarity measure InfoSimba, which takes into account word co-occurrence endogenously and avoids the extra step in the topic identification process as it is the case in [Ponte 1997]. Second, we clearly pose the problem of word weighting for topic segmentation and show that the usual term frequency or the tf.idf measures as proposed in [Hearst 1994] [Reynar 1994] [Choi 2000] are not the best heuristics to achieve improved results for this specific task. And finally, we confirm the conclusions of [Ferret 2002], as improved results are obtained with the identification of MWU.

5.1.2 Informative Topic Segmentation

Our algorithm ITOS is based on the vector space model, which determines the similarity of neighboring groups of sentences and places subtopic boundaries between dissimilar blocks. In our specific case, each sentence in the corpus is evaluated in terms of similarity with the previous block of k sentences and the next block of k sentences. The simplest form of the vector space model treats a document (in our case, a sentence or a group of sentences) as a vector whose attribute values correspond to the number of occurrences of the words appearing in the document as in [Hearst 1994]. Although [Hearst 1994] showed successful results with this weighting scheme, we strongly believe that the importance of a word in a document does not only depend on its frequency. Indeed, frequency can only be reliable for technical texts where ambiguity is drastically limited and word repetition largely used. But unfortunately, these documents are an exception in the global environment of the internet. According to us, three main factors must be taken into account to define the relevance of a word for the specific task of topic segmentation: (1) its relevance within a text collection based on its $tf.idf(.,.)$ [Salton 1975], (2) its relevance within a document based on its $tf.isf(.,.)$ (the adaptation of the tf.idf measure for sentences) [Dias 2007a] and (3) its density $dens(.,.)$, which measures the concentration of each word in a given document. [Dias 2007a]. A clear example of our motivation is given in Figure 5.2, where it is clear that the focus of the document changes from *moon* to *star*.

While the $tf.idf(.,.)$ is a well-known measure, we clarify the ideas of both the $tf.isf(.,.)$ and the $dens(.,.)$. On the one hand, the basic idea of the $tf.isf(.,.)$ score is to evaluate each word in terms of its distribution over the document. Indeed, it is obvious that words occurring in many sentences within a document may not be useful for topic segmentation purposes as they do not embody a specific concept in a

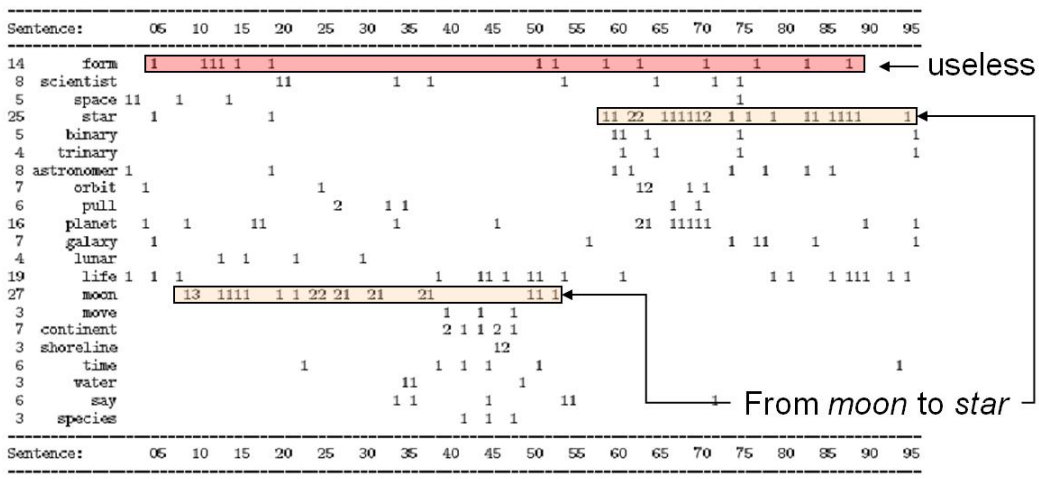

Figure 5.2: Word distributions in texts.

portion of the text but rather a continuum along the document. On the other hand, the idea of the word density measure is to evaluate the dispersion of a word within a document. For instance, if a word w appears in consecutive or near consecutive sentences it will have a strong influence on what is said in this specific region of the text, whereas if it occurs in distant sentences, its importance will be negligible. As a consequence, each word w present in a document d, is weighted by a linear interpolation of these three measures as in Equation 5.1.

$$score(w,d) = \alpha \times \|tf.idf(w,d)\| + \beta \times \|tf.isf(w,d)\| + \gamma \times \|dens(w,d)\|. \quad (5.1)$$

Once all words in the document to segment have been evaluated in terms of relevance and distribution, the next step of ITOS is to determine similarities between a focus sentence and its neighboring groups of k sentences. Within this scope, the most interesting approach is proposed by [Ponte 1997], who present a topic segmentation technique based on the LCA, allowing to substitute each sentence with words and phrases related to it. Our methodology is based on this same idea but differs from it as the word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences thus avoiding an extra-step in the topic boundaries discovery. Another direct contribution is that, unlike [Ponte 1997], we propose a well-founded mathematical model that deals with the word co-occurrence factor. For that purpose, we propose to evaluate the similarity between each sentence of a document and its surrounding k sentences contexts with the InfoSimba, $IS(.,.)$ (see Equation 2.20) associated to the SCP (see Equation 2.4) as its symmetric similarity measure $S(.,.)$. Let's take the focus sentence F_i and a block of surrounding sentences $\breve{F}_j$. For each word in the focus sentence F_i, then for each word in the block of sentences $\breve{F}_j$, we calculate the product of their weights and then multiply it by the degree of cohesiveness existing

between those two words calculated by the SCP. As a result, the more relevant the words will be and the more cohesive they will be, the more they will contribute for the cohesion within the text and will not contribute for a topic shift. As a summary, for each sentence F_i in a given document, we compute the InfoSimba with its k previous sentences noted $IS(F_i, \breve{F}^k_{i-1})$ and its k following sentences noted $IS(F_i, \breve{F}^k_{i+1})$.

Once we know all the similarities between sentences and surrounding sentence contexts, the next step of the algorithm aims at placing subtopic boundaries between dissimilar blocks. For that purpose, we propose a detection methodology based on the standard deviation algorithm proposed by [Hearst 1994]. Different methodologies have been proposed to place subtopic boundaries between dissimilar blocks [Kozima 1993] [Hearst 1994] [Ponte 1997] [Beeferman 1997] [Stokes 2002]. For that purpose, we propose a new methodology based on ideas expressed by different research. Taking as reference the idea of [Ponte 1997] who take into account the preceding and the following contexts of a segment, we calculate the informative similarity of each sentence in the corpus with its surrounding sentence context. The idea is to know whether the focus sentence is more similar to the preceding block of sentences or to the following block of sentences. In order to evaluate this preference in an elegant way, we propose a score for each sentence in the text in the same way [Beeferman 1997] compare short and long-range models. Our preference score (PS) is defined in Equation 5.2.

$$PS(F_i) = \log \frac{IS(F_i, \breve{F}^k_{i-1})}{IS(F_i, \breve{F}^k_{i+1})}. \tag{5.2}$$

So, if $PS(F_i)$ is positive, it means that the focus sentence F_i is more similar to the previous block of k sentences, $\breve{F}^k_{i-1}$. Conversely, if $PS(F_i)$ is negative, it means that the focus sentence F_i is more similar to the following block of k sentences, $\breve{F}^k_{i+1}$. In particular, when $PS(F_i)$ is near 0, it means that the focus sentence F_i is similar to both blocks and so we may be in the continuity of a topic. In order to better understand the variation of the $PS(.)$ score, each time its value goes from positive to negative between two consecutive sentences F_i and F_{i+1}, there exits a topic shift. We will call this phenomenon a downhill. In fact, it means that the previous sentence F_i is more similar to the preceding block of sentences $\breve{F}^k_{i-1}$ and the following sentence F_{i+1} is more similar to the following block of sentences $\breve{F}^k_{i+2}$ thus representing a shift in topic in the text. So, the amplitude of a downhill is evaluated as in Equation 5.3.

$$downhill(F_i, F_{i+1}) = PS(F_i) - PS(F_{i+1}). \tag{5.3}$$

However, not all downhills identify the presence of a new topic in the text. Indeed, only deeper ones must be taken into account. As a consequence, most relevant topic shifts are selected based on a threshold, which is a function of the average $\overline{downhill}(.)$

and the standard deviation σ of the downhills depths as defined in Equation 5.4, where c is a constant to be tuned.

$$downhill(F_i, F_{i+1}) \geq \overline{downhill}(.,.) + c \times \sigma. \tag{5.4}$$

Evaluating a task such as topic segmentation consists in determining if the topic shifts are well identified. This can be quite subjective unless correct boundaries are *a priori* known. To avoid subjectivity, the evaluation task is usually supervised, by using texts for which we are sure about the topic boundaries. This is usually achieved by building an artificial single document (the one to be segmented) from a collection of texts pieces dealing with different issues as in [Choi 2000] [Ferret 2002] [Angheluta 2002]. In particular, [Choi 2000] runs his c99 algorithm over a concatenation of text segments, each one extracted from a random selection of the brown corpus, which consists of 500 texts sampled from 15 different text categories, such as religion, fiction, and humor. According to many authors [Moens 2001] [Angheluta 2002] [Xiang 2003], this test set eases the identification of the boundaries as the terms used differ drastically from domain to domain. Instead, [Hearst 1994] proposes a segmentation algorithm with a different goal: to find subtopic segments i.e. to identify, within a single-topic document, the boundaries of its subparts. A similar experiment is performed in [Xiang 2003].

So, following the work of [Hearst 1994] [Xiang 2003], we built our own benchmark based on real-world texts retrieved from the Web from the soccer domain. We automatically gathered 100 articles of approximatively 100 words and then built 10 test corpora, by choosing randomly 10 articles from our database of 100 articles leading to 10 texts of around 1000 words-long. This choice is not casual. Independently of the topic of any article (e.g. a soccer player being transferred to a different club, a report about a certain game or a championship), it is usual to find many common words in all texts. As a consequence, we aim at dealing with fine-grained topic segmentation rather then coarse-grained. ITOS was evaluated against TextTiling [Hearst 1994] and c99 [Choi 2000], using three different evaluation metrics (the F-Measure, the P_k estimate [Beeferman 1997] and the WindowDiff [Pevzner 2002]) with and without integrating the normalization of the corpus. The final results are well-described in [Dias 2005b] [Dias 2007a] and show that our algorithm obtains improved results both with and without the identification of MWU compared to the state-of-the-art algorithms. They are summarized in Table 5.1 for the following parameters of ITOS: $c = -1.5$, $k = 2$ and the window size for the evaluation of the SCP equals to 10[6].

The first result is that ITOS outperforms all other algorithms with and without the normalization of the corpus. The introduction of MWU clearly benefits the detection process compared to the non normalized version. Moreover, it is interesting to see that for ITOS without MWU, the weights of the words are useless and just

[6] We do not present the results of the WindowDiff as they are correlated to the P_k.

Algorithm	F-meas.	P_k	Parameters
ITOS with MWU	76%	0.17	$\alpha = 1$, $\beta = 0$, $\gamma = 1$
ITOS without MWU	72%	0.21	$\alpha = 0$, $\beta = 0$, $\gamma = 0$
TextTiling with MWU	53%	0.24	default
TextTiling without MWU	47%	0.33	default
c99 with MWU	44%	0.31	default
c99 without MWU	44%	0.31	default
ITOS with Cosine tf.idf without MWU	14%	0.53	no parameters
ITOS with Cosine tf.idf with MWU	9%	0.55	no parameters

Table 5.1: Comparative results for topic segmentation.

the SCP is enough to understand the similarity between sentences. Second, the c99 algorithm is the one that worst performs over our the test corpus. This goes against the evaluation of [Choi 2000], which evidences improved results when compared to the TextTiling algorithm over the c99 corpus. This result clearly shows that the c99 data set can not be taken as a golden standard for topic segmentation evaluation schemes as it has been done in many works. The reason why the TextTiling algorithm performs better than the c99 on our benchmark is the fact that [Hearst 1994] use the appearance of new lexical units as a clue for topic boundary detection whereas [Choi 2000] relies more deeply on lexical repetition which drastically penalizes the topic boundary detection process. Finally, the difference between using the InfoSimba and the cosine similarity measure (see Equation 2.14) is huge, which clearly confirms our idea that similarity should not be computed as exact matches of attributes but as the amount of correlation attributes share.

5.2 Construction of Lexical Chains

Topic segmentation aims at discovering the sequential structures of texts based on topical shifts. As such, they provide extractive summarization with meaningful clues to guarantee the coherence of summaries. But, cohesiveness must also be ensured. Cohesiveness can be evidenced by deep semantic understanding of texts [Heim 1983] [Kamp 1995] [Muskens 1996] or by the identification of loose lexical-semantic relations between words [Morris 1991] [Hirst 1998a] [Stokes 2004] [Terra 2005] [Vechtomova 2006]. This second interpretation of cohesiveness is usually referred to as lexical cohesion [Halliday 1976] and is based on the idea that the complete meaning of a word in a text can only be realized when it is interpreted in combination with the surrounding words, forming lexical cohesive ties with them.

Lexical cohesion is a method for the identification of semantically connected sub-parts of a text based on the discovery of lexical-semantic relations between words along the document. One method of uncovering these relationships between words is called lexical chaining, where lexical chains are defined as sequences of

semantically related words spread over the entire document. Within this scope, different studies have been proposed [Morris 1991] [Barzilay 1997] [Hirst 1998a] [Silber 2000] [Galley 2003], which showed that lexical chains are strong intermediate representations of documents in comparison to the bag-of-words approach.

5.2.1 Related Work

Lexical chaining requires the identification of the semantic relations between words to determine the compatibility of a word with respect to a chain. Since lexical cohesion is realized in texts through the use of related vocabulary, knowledge resources such as thesauri, ontologies or dictionaries have been used as a means of identifying related words.

[Morris 1991] were the first to propose the concept of lexical chains to explore the discourse structure of a text. However, at the time of writing their paper, no machine-readable thesaurus was available. So, they manually generated lexical chains using Roget's Thesaurus [Roget 1852]. A first computational model of lexical chains was introduced by [Hirst 1998a]. Their biggest contribution was the mapping of WordNet relations [Miller 1990] and paths (transitive relationships) to the word relationship types proposed in [Morris 1991]. However, their greedy algorithm was not using a part-of-speech tagger. Instead, the algorithm only selected those words, which contained noun entries in WordNet to compute lexical chains. But, as [Barzilay 1997] point at, the use of a part-of-speech tagger can eliminate wrong inclusions of words such as *read*, which has both noun and verb entries in WordNet.

So, [Barzilay 1997] proposed the first dynamic method to compute lexical chains. They argue that the most appropriate sense of a word can only be chosen after examining all possible lexical chain combinations that can be generated from a text. Because all possible senses of a word are not taken into account, except at the time of insertion, potentially pertinent context information that is likely to appear after the word is lost. However, this method of retaining all possible interpretations until the end of the process, causes the exponential growth of the time and space complexity. As a consequence, [Silber 2000] proposed a linear time version of [Barzilay 1997] lexical chaining algorithm. In particular, their implementation creates a structure, called meta-chains, which implicitly stores all chain interpretations without actually creating them, thus keeping linear both space and time usage of the program.

Finally, [Galley 2003] proposed a chaining method, which disambiguates nouns prior to the processing of lexical chains. Their evaluation shows that their algorithm is more accurate than the one proposed in [Barzilay 1997] and [Silber 2000] ones.

5.2.2 Construction of a Lexical-semantic Knowledge Base

One common point of all these works is that lexical chains are built using WordNet as the standard linguistic resource. Unfortunately, systems based on static linguistic knowledge bases are limited. First, such resources are difficult to find. Second, they are largely obsolete by the time they are available. Third, linguistic resources capture a particular form of lexical knowledge which is often very different from the sort needed to specifically relate words or sentences. In particular, WordNet is missing a lot of explicit links between intuitively related words. [Fellbaum 1998] refers to such obvious omissions in WordNet as the "tennis problem" where nouns such as *nets*, *rackets* and *umpires* are all present, but WordNet provides no links between these related tennis concepts.

In order to solve these problems, we first propose in [Dias 2006b] to automatically construct a lexical-semantic knowledge base from a collection of documents. The basic idea is borrowed from the works developed for the construction of prototype-based ontologies such as in [Hindle 1990] [Pereira 1993] [Caraballo 1999] [Paaß 2004] (see Chapter 6). The hierarchical taxonomy is built based on the PoBOC algorithm provided with an informative similarity matrix through the evaluation of word/phrase pairs similarities with the InfoSimba similarity measure. In particular, this process clusters words with similar meanings and allows words with multiple meanings to belong to different clusters.

Lexical chains are text semantic representations based on nominal expressions. As a consequence, the hierarchical taxonomy should only contain nouns, names and nominal MWU. For that purpose, the reference corpus[7] is first part-of-speech tagged with the TnT tagger [Brants 2000]. Then, SENTA is run over the entire corpus and a linguistic filter is applied to obtain high quality nominal MWU as proposed in [Justeson 1993] and [Daille 1996]. The list of identified nominal words and phrases (i.e. terms) is called the vocabulary, which must be structured into a hierarchical taxonomy. For that purpose, we apply Harris's distributional hypothesis over selected word context vectors (See Chapter 2).

To improve the quality of the evaluation of the similarity between the terms of the vocabulary, we propose to use the InfoSimba similarity measure over term context vectors. The term context vectors are not built over the whole vocabulary to avoid data sparseness. Instead, we associate to each term its best N[8] co-occurrent terms based on the SCP similarity measure calculated from the reference corpus in a context window of 20 words[9] (see Equation 2.4). Each of the N terms t are then weighted as in Equation 5.5 based on the combination of their average $tf.idf(.,.)$

[7] In our experiments, we use a set of texts extracted from the DUC 2004 text collection.
[8] In our experiments, $N = 10$.
[9] Best results were obtained for this size.

and their average density measure $dens(.,.)$[10] over the reference text collection, in order to weight their importance within a given domain.

$$weight(t) = \overline{tf.idf}(t, \forall d) \times \overline{dens}(t, \forall d). \tag{5.5}$$

Finally, to build the term $\times$ term informative similarity matrix as input for the PoBOC algorithm, the similarity between each term is evaluated with the InfoSimba measure, $IS(.,.)$ (see Equation 2.20) associated to the SCP (see Equation 2.4) as its proximity coefficient $S(.,.)$ calculated in the same context window of 20 words.

As mentioned earlier, the PoBOC algorithm shows interesting properties as (1) it requires no parameters i.e. its input is restricted to a single similarity matrix, (2) the number of final clusters is automatically found and (3) it provides overlapping clusters allowing to take into account the different possible meanings of terms. Moreover, a recent comparative study [Cicurel 2006] shows that CBC (Clustering by Committee) [Pantel 2002] and PoBOC both lead to relevant results for the task of word clustering. However, CBC requires parameters hard to tune whereas PoBOC is free of any parametrization. The last argument encouraged us to use the PoBOC algorithm. So, from the term $\times$ term informative similarity matrix, PoBOC first groups terms into overlapping clusters (or soft-clusters) such that the final clusters are likely to correspond to conceptual classes[11], which are then hierarchically structured in order to capture semantic links between them.

5.2.3 Lexical Chainer Algorithm

The second step of the process aims at automatically extracting lexical chains from texts based on our lexical-semantic knowledge base. For that purpose, we propose a new greedy algorithm, which can be seen as an extension of [Hirst 1998a] and [Barzilay 1997] algorithms, but which allows polysemous words to belong to different chains thus breaking the *one sense per discourse* paradigm expressed in [Gale 1992]. In particular, as we want to deal with real-world applications, this characteristic may show interesting properties to summarize Web multi-topic documents. For example, documents like Web news stories are likely to contain different topics as different news stories may appear in a unique Web page. In some way, we follow the idea of [Krovetz 1998], who found significantly more occurrences of multiple-senses per discourse than reported in [Gale 1992] (33% instead of 4%). Our algorithm is presented in algorithm 10.

Based on algorithm 10, we just need to explain how the relatedness criterion is evaluated to perfom the construction of lexical chains. Indeed, in order to assign a term to a given lexical chain, we need to evaluate the degree of relatedness of the given term to the terms in the chain. In fact, this is done by evaluating the

[10] Proposed earlier in section 5.1.2.

[11] In the remainder of this chapter, clusters will now on be assimilated as concepts.

Algorithm 10 The lexical chainer algorithm.

```
Begin with no chain.
for all distinct terms taken in the order of the text do
  for all its concepts do
    Find a chain for which the relatedness criterion is satisfied.
    if no chain is found then
      start a new chain.
    else
      Link the word to the chain.
      Remove unappropriate concepts from the chain
    end if
  end for
end for
```

relatedness between all the clusters (or concepts) present in the lexical chain and all the concepts the term embodies. For that purpose, our algorithm implements Lin's information-theoretic definition of similarity [Lin 1998b] (see Equation 2.24). However, unlike [Barzilay 1997] [Silber 2000] [Galley 2003] who build lexical chains based on WordNet, which foundations rely on inter-connected non-empty synsets, our soft hierarchical taxonomy only contains terms in its leaves[12]. For that purpose, we propose that all upper clusters (i.e. nodes) in the taxonomy gather all distinct terms, which appear in the clusters they subsume as illustrated in Figure 5.3. For example, clusters C_{305} and C_{306} of our hierarchical tree, for the domain of Economy, are represented by the following sets of terms C_{305} ={*life*, *effort*, *stability*, *steps*, *negotiations*} and C_{306} ={*steps*, *restructure*, *corporations*, *abuse*, *interests*, *ministers*}.

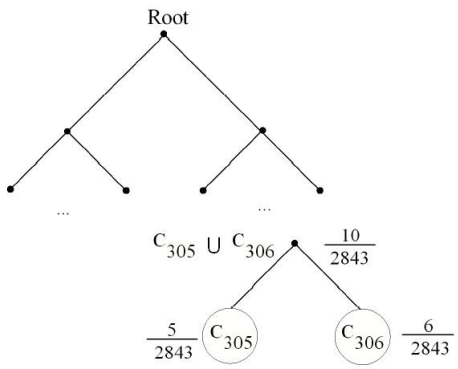

Figure 5.3: Fragment of our taxonomy.

As a consequence, in order to apply Lin's information-theoretic definition of sim-

[12]We will see that most of our work in Chapter 6 deals with populating upper-levels nodes in a taxonomy.

ilarity, we redefine $P(C_i)$ (which is the probability that a randomly selected term belongs to the concept C_i) as in Equation 5.6[13].

$$P(C_i) = \frac{\#\text{ of words in the cluster } C_i}{\#\text{ of distinct words in all clusters } C_j, \forall j = 1..n}. \tag{5.6}$$

We are now able to define the relatedness criterion, which is nothing more than the threshold, which needs to be respected in order to assign a term to a lexical chain or not. Basically, if the semantic similarity between a candidate concept (i.e. a cluster C_k) respects the relatedness criterion, the term associated to cluster C_k is assigned to the lexical chain. The idea for this threshold is that the average value of the concept pairs similarities within a given lexical chain must be superior to the average value of all concept pairs similarities in the taxonomy to be considered a valid term for the lexical chain. This situation is defined in Equation 5.7, where c is a constant to be tuned and V is the number of concepts in the taxonomy. So, if the relatedness criterion expressed as in Equation 5.7 is satisfied, the term t with cluster C_k (i.e. the interpretation of t as C_k) is added to the lexical chain.

$$\frac{\sum_{k=1}^{m} Lin(C_k, C_l)}{m} > c \times \frac{\sum_{i=1}^{V-1} \sum_{j=i+1}^{V} Lin(C_i, C_j)}{\frac{V^2 - V}{2}}. \tag{5.7}$$

In order to better understand algorithm 10, we propose the following example. Let's consider that a lexical chain is created for the first term encountered in the text e.g. *crisis* with its sense (C_{31}). Imagine, the next appearing term is *recession*, which has two senses (C_{29} and C_{34}). Considering a relatedness criterion equal to 0.81 (i.e. the right part of Equation 5.7) and the following similarities, $Lin(C_{31}, C_{29}) = 0.87$ and $Lin(C_{31}, C_{34}) = 0.82$, the choice of the sense for *recession* splits the lexical chain into two different interpretations as shown in Figure 5.4, as both similarities respect the relatedness criterion.

Figure 5.4: Interpretation 1 (left) and Interpretation 2 (right) of the first lexical chain.

The next candidate term *trouble* appearing in the text has also two senses (C_{29} and C_{32}). As all the terms in a lexical chain influence each other in the selection of the respective senses of the new term into consideration, we have the following situation for both interpretations of the first lexical chain illustrated in Figure 5.5.

[13]The value 2843 in Figure 5.3 is the total number of distinct terms in our concept hierarchy.

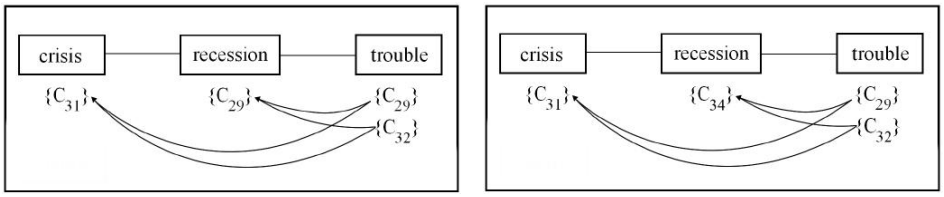

Figure 5.5: Interpretation 1 (left) and Interpretation 2 (right) of the first lexical chain.

Now, three cases can happen. First, the average of all similarities in both interpretations respects the relatedness criterion and we must consider both representations. Second, only one interpretation observes the relatedness criterion and we only consider this representation. Third, none of the interpretations preserves the relatedness criterion and we create a new lexical Chain. Let's continue with the following figures. Interpretation 1 shows the following similarities $Lin(C_{31}, C_{29}) = 0.87$, $Lin(C_{31}, C_{32}) = 0.75$, $Lin(C_{29}, C_{29}) = 1.0$, $Lin(C_{29}, C_{32}) = 0.78$ and interpretation 2 the following ones, $Lin(C_{31}, C_{29}) = 0.87$, $Lin(C_{31}, C_{32}) = 0.75$, $Lin(C_{34}, C_{29}) = 0.54$, and $Lin(C_{34}, C_{32}) = 0.55$. By computing the average similarities for interpretations 1 and 2, we reach the following results: $\overline{Interpretation1} = 0.85 > 0.81$ and $\overline{Interpretation2} = 0.68 > 0.81$. As a consequence, the word *trouble* is inserted in the lexical chain with the appropriate sense (C_{29}) as it maximizes the overall similarity of the chain and the chain members senses are updated. In this example, the interpretation with (C_{32}) is discarded as is the cluster (C_{34}) for *recession*. This processing is illustrated in Figure 5.6.

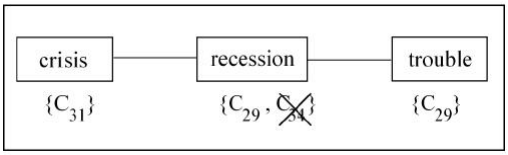

Figure 5.6: Selection of appropriate senses.

Once all chains have been computed, only the high-scoring ones must be picked up as representing the important concepts of the original document as proposed in [Barzilay 1997]. Therefore, one must first identify the strongest chains. For that purpose, we define a chain score following the same idea as [Barzilay 1997], which is defined in Equation 5.8 for our specific case, where $|chain|$ is the number of terms in the chain, $C_i \in t_i$ represents all the concepts C_i associated to the term t_i in the lexical chain and $|C_i \in t_i|$ is the number of these concepts.

$$score(chain) = \frac{\sum_{i=1}^{|chain|-1} \sum_{j=i+1}^{|chain|} Lin(C_i \in t_i, C_j \in t_j)}{\sum_{i=1}^{|chain|-1} \sum_{j=i+1}^{|chain|} |C_i \in t_i| \times |C_j \in t_j|}. \tag{5.8}$$

As we mentioned in the beginning of the chapter, our contribution to the field of document summarization is limited to the ITOS algorithm and a new lexical chainer based on an automatically built lexical-semantic knowledge base. As such, we did not reach the overall process of extractive summarization and will let for future work many issues such as the combination of topic segmentation and lexical chaining to produce coherent and cohesive summaries. Nevertheless, although we managed to produce a relevant and exhaustive evaluation for our topic segmenter, the evaluation of lexical chains is a much more difficult task. Indeed, even if they can be effectively used in many practical applications like automatic text summarization, topic segmentation and others, lexical chains are seldom desirable outputs in a real-world application. Moreover, it is unclear how to assess their quality independently of the underlying application in which they are used [Budanitsky 2006]. For example, in summarization, it is hard to determine whether a good or bad performance comes from the efficiency of the lexical chaining algorithm or from the appropriateness of using lexical chains in that kind of application.

It is also true that some work has been done to evaluate lexical chains intrinsically [Budanitsky 2006] by collecting human lexical chains to compare against automatically built lexical chains. However, this type of evaluation is logistically impossible to perform as we aim at developing a system that does not depend on any language or topic. So, we only present some results generated by our lexical chainer (like in [Barzilay 1997] and [Teich 2004] do). In particular, we processed four prototype-based ontologies from four different domains (Sports, War, Politics and Economy) taken from the DUC 2004[14] text collection and compared our results with different parameters. The results showed the great potential of the methodology as informative chains were acquired. For example, we present the five highest-scoring chains for the best threshold that we experimentally evaluated to be $c = 7$ for the Sport domain in Table 5.2.3. It is clear that the obtained lexical chains show a desirable degree of representativeness of the text in analysis. For instance, the lexical chain #16 clearly exemplifies the tragedy of climbers that were killed in a sudden change of weather in the mountains and who could not be rescued by the authorities. Nevertheless, some other chains show lists of related terms with limited semantic content for document understanding as chains #0 and #9 due to their reduced size. More examples can be found in [Dias 2006b] and [Santos 2006].

[14] http://duc.nist.gov/duc2004/ [23rd *September*, 2010].

Domain=Sport, Document=3, c=7
- #0, 1 cluster and score=1.0: {United States, couple, competition}
- #6, 3 clusters and score=1.0: {boats, Sunday night, sailor, Sword, Orion, veteran, cutter, Winston Churchill, Solo Globe, Challenger, navy, Race, supposition, instructions, responsibility, skipper, east, Melbourne, deck, kilometer, masts, bodies, races, GMT, Admiral's, Cups, Britain, Star, Class, Atlanta, Seattle, arms, fatality, sea, waves, dark, yacht's, Dad, Guy's, son, Mark, beer, talk, life, Richard, Winning, affair, canopy, death}
- #9, 1 cluster and score=1.0: {record, days, hours, minutes, rescue}
- #16, 3 clusters and score=1.0: {Snow, shape, north, easters, thunder, storm, change, knots, west, level, maxi's, search, Authority, seas, helicopter, night vision, equipment, feet, rescues, Campbell, suffering, hypothermia, safety, foot, sailors, colleagues, Hospital, deaths, bodies, fatality}
- #19, 2 clusters and score=1.0: {challenge, crew, Monday, VC, Offshore, Stand, Newcastle, mid morning, Eden, Rescuers, aircraft, unsure, whereabouts, killing, contact}

Table 5.2: The 5 best lexical chains for the Sport domain.

Although encouraging results were obtained, a serious evaluation must still be performed. In particular, our algorithm must be compared on the same basis to state-of-the-art algorithms, which unfortunately are not freely available [Barzilay 1997] [Hirst 1998a] [Galley 2003]. Moreover, it is clear that prototype-based ontologies are structures with loose semantic connections which may not sufficiently fit the necessary requirements to build lexical chains. As a consequence, work must be done to build fine-grained lexical-semantic resources. While the first point has not still been tackled, the second has greatly focused our attention and different strategies have already been proposed as exposed in Chapter 6.

5.3 Sentence Reduction

During the last decade, research in ATS has been continuing its objective to move from the traditional extractive methodology to a more abstractive resembling approach. Within this context, sentence reduction is certainly one of the new approaches, which has received great attention from the research community [Jing 2000b] [Knight 2002] [Le Nguyen 2004] [Vandeghinste 2004] [Daelemans 2004]

[Turner 2005] [Clarke 2006] [Unno 2006] [Specia 2010]. Sentence reduction is also known as sentence compression as well as sentence simplification. Within the scope of our research, we will prefer to use sentence reduction since sentences are simplified through adequate content removal, while sentence compression and sentence simplification have traditionally tackled more complex syntactical transformations than just content remotion. An example of a sentence reduction is given in sentences (1) and (2), respectively the original sentence and one of its possible reductions. It is clear that the reduced sentence preserves the most relevant information of the original one and maintains grammaticality.

1 `In Louisiana, the hurricane landed with wind speeds of about 120 miles per hour and caused severe damage in small coastal centers such as Morgan City, Franklin and New Iberia`

2 `In Louisiana, the hurricane landed and caused severe damage`

Sentence reduction is not a marginal issue. Indeed, from our experiments made from Web news stories, we evidenced that the average sentence length is equal to 20 words, which can motivate imagination for different solutions. As a consequence, different approaches have been experimented essentially based on three main strategies: deep knowledge-driven methodologies [Chandrasekar 1997] [Jing 2000a] [Knight 2002] [Turner 2005] [Unno 2006], poor-knowledge systems [Vandeghinste 2004] [Daelemans 2004] [Clarke 2006] and machine translation (MT) methodologies [Le Nguyen 2004] [Specia 2010]. Despite their individual virtues, all works proposed so far embody important drawbacks and practical limitations that should be addressed. Most of them are based on supervised machine learning techniques, which require large data sets of handcrafted training examples. Within this context, considerable amounts of manual labor and tuning experiments are necessary, which are also subject to incompleteness and imperfection. Some other methodologies rely on deep linguistic knowledge, which are hugely language-dependent as they rely on the availability of rich linguistic resources or tools that are scarce in many non English languages. Only the MT methodologies propose a language-independent framework. However, they need terabytes of parallel texts to learn rewriting rules with some coverage and accuracy. However, these resources do not exist. Our approach follows a different strategy, by using minimum linguistic resources and an unsupervised machine learning strategy. Thus, we intend to minimize human inputs by automatically constructing training examples from automatically extracted corpora of aligned asymmetric paraphrases and applying ILP as the learning paradigm. As such, our framework can be divided into three separate modules: (1) paraphrase extraction, (2) paraphrase alignment and (3) reduction rules learning.

5.3.1 Related Work

One of the earliest works is proposed in [Chandrasekar 1997]. Their idea is to transform a long sentence into a set of few shorter and consequently simpler sentences

from where an automatic system may pick the most relevant one(s). This method is described as a two stage process where the first one provides a structural representation of the sentence, and the second one applies a sequence of rules to identify and extract the components that can be simplified. Special sentence splitting points as phrasal extremes, punctuation, subordinate/coordinate conjunctions, and relative pronouns are identified based on defined rules and the grammatical formalism used. The main particularity of this work is that all rules are manually encoded heuristics.

Later, [Jing 2000a] presented a sentence reduction system, which automatically removes extraneous phrases by combining four linguistic knowledge resources with six major operations previously identified for sentence editing in [Jing 2000b]. In particular, a full parse tree is first built to identify critical undeletable sentence components as well as usually deleted components, such as prepositional and to-infinitive phrases or adjectives. Then, the system computes a score for each phrase, which represents the amount of relatedness to the main topic based on WordNet lexical relations. Finally, a probabilistic model is trained over a set of sentence pairs $\langle F_{original}, F_{reduced} \rangle$ to learn sentence subtree removal likeliness. Unfortunately, besides the use of deep language understanding resources and tools, the results did not satisfy the authors due to the small amount of training data.

Maybe the most well-known work in sentence reduction is described in [Knight 2002] as they are the first to apply supervised machine learning techniques over deeply linguistically enriched training examples. Their work shows two different sentence reduction experiences based on two learning models (i.e. the noisy-channel and the decision trees). While, the noisy-channel model infers a statistical sentence reduction model, the decision trees are directed towards learning a set of syntactic transformation rules. To evaluate both algorithms, they randomly selected 32 sentence pairs from their parallel corpus and used the other 1035 sentence pairs for training. The results show that the noisy-channel tends to best keep grammaticality, while decision trees reach better results of importance. Moreover, they state that *when applied to sentences of a different genre, the performance of the noisy-channel compression algorithm degrades smoothly, while the performance of the decision-based algorithm drops sharply.* As such, it seems that the noisy-channel is likely to provide improved results with some modifications. Both [Turner 2005] and [Unno 2006] will follow this direction. [Turner 2005] proposed an extension of the noisy-channel showing an improved noisy-channel model by incorporating supervised and semi-supervised learning methods and additional linguistic constraints to improve compression. [Unno 2006] also proposed an extension of the noisy-channel by applying a maximum entropy model to introduce machine learning features that are defined not only for context free grammars (CFG) rules but also for other characteristics in a parse tree, such as the depth from the root node or the words contained in a sub-tree. They also introduced a bottom-up method to learn complicated relations of two unmatched trees.

The methods proposed so far rely on deep linguistic analysis, which are not always available for other languages. As a consequence, poor-knowledge resources or tools must be used instead. [Vandeghinste 2004] is one of the first works to propose a sentence compression tool for the Dutch language. The idea is similar to the one proposed in [Jing 2000a] but with much less resources. In particular, they used a parallel corpus, which consists of transcripts of television programs and their corresponding subtitles and a shallow rule-based parser called ShaRPA. No other knowledge resources are used. So, after the parallel corpus is shallow-parsed and chunk-aligned, chunk and clause removal, non-removal and reduction probabilities are estimated. But, as the statistical information allows the generation of ungrammatical sentences, a number of rules were added to respect grammaticality. Finally a word compressor was also proposed to reduce the size of compound nouns based on lexicon look-up [Vandeghinste 2002]. Although reasonable results were obtained, [Vandeghinste 2004] stated that a full syntactic analysis of the input sentence would lead to better results, as ShaRPA usually misinterprets coordinating conjunctions, which leads to chunking errors and consequently to misestimations of the compression probabilities, thus introducing noise in the system.

Following the same idea, [Daelemans 2004] applied a memory-based learner based on a shallow-parsed parallel aligned corpus of TV transcripts and subtitles. The supervised learner is fed with words, lemmas, part-of-speech tags, chunk tags, relation tags and proper name tags. Apart from the focus word, they also include information regarding a context of two words to the left and right. Thus, the learner had 30 features to its disposal, which were passed through a feature selection process with bidirectional hill-climbing. Similarly to [Vandeghinste 2004], the performances were low and, most importantly, the approach frequently made nonsensical errors, like removing sentence subjects or deleting a part of MWU. As a consequence, handcrafted rules were introduced and combined with the learner to reach acceptable results.

[Clarke 2006] proposed to view sentence reduction as an optimization problem, where the goal is to find the optimal reduced sentence version within a set of integer programming constraints. A sentence is characterized as a sequence of n words and the space of all possible reductions is equal to 2^n reduced versions obtained by word elimination. As in any optimization problem, the set of constraints guides the search by narrowing the subset of valid solutions (here reductions), until the optimum is reached. For that purpose, they codified a set of linguistic constraints as linear inequalities, in order to ensure sentence structural and semantic validity for the generated reductions. The authors claimed that comparable results were reached compared to [Knight 2002], following the same evaluation scheme with human annotators. Although no parallel corpus was used, their model is knowledge-driven as sentence reduction is reached through a set of handcrafted rules. As all other works proposed so far, this approach is not easily transportable to other languages, since a redefinition of linguistic knowledge is nec-

essary. Furthermore, it is not able to cover all the variety of linguistic phenomena behind the sentence reduction process. As a consequence, different machine translation approaches appeared, which can be easily reused and are language-independent.

In [Le Nguyen 2004], the sentence reduction task is described as the process of translating a sentence from a verbose (source) language into a succinct (target) language. In particular, it is claimed that it is simpler than traditional machine translation as it is not necessary to capture every senses and nuances of the source sentence. In particular, they first adapted a translation-template learning method from the example-based paradigm. But, after after stressing out the major drawback of the method (i.e. complexity inefficiency for rule combinatorics), they proposed an optimization based on the computation of an hidden markov model over a parallel corpus, in order to efficiently find the best template reduction rule combinations for a given sentence. The authors described similar evaluation procedures as in [Knight 2002] and claimed human comparable performance.

Lately, [Specia 2010] proposed a standard framework for statistical machine translation based on the noisy-channel model from information theory. The goal is to learn a probabilistic dictionary of phrases and their corresponding simplified versions (the translation model) along with a model that indicates how likely a segment is a correct simplified segment in Portuguese (the language model). Three other models are also proposed. The reordering model evaluates how to best distribute phrases in a simplified segment. The word penalty model penalizes translations that differ in length as compared to the source segment. And, the distortion model measures the limit on the amount of allowed reorderings. Each translation candidate is then scored according to a linear interpolation of all these five models. The authors concluded that while translating between variations of the same language should alleviate the need for large parallel corpora, the fact that several types of simplification were considered invalidates this statement. Indeed, both [Le Nguyen 2004] and [Specia 2010] clearly need terabytes of parallel texts to learn rewriting rules with some degree of accuracy. However, these resources do not exist and only small data sets are always used for evaluation. For that purpose, we proposed in [Cordeiro 2007a] a new unsupervised methodology to automatically build corpora of asymmetric paraphrases, which can be viewed as pairs of sentences that express the same meaning or coincide in almost all their semantic constituents yet are usually written in different styles.

5.3.2 Paraphrase Extraction

Paraphrase corpora are golden resources for learning monolingual text-to-text rewritten patterns. However, such corpora are expensive to construct manually and will always be an imperfect and biased representation of the language paraphrase phenomena. Therefore, reliable automatic methodologies able to extract paraphrases from text and subsequently corpus construction are crucial, enabling better

pattern identification. In fact, text-to-text generation is a particularly promising research direction given that there are naturally occurring examples of comparable texts that convey the same information but are written in different styles. Web news stories are an obvious example. Thus, presented with such texts, one can pair sentences that convey the same information, thereby building a training set of rewriting examples i.e. a paraphrase corpus.

Three different approaches have been proposed for paraphrase detection: unsupervised methodologies based on lexical similarity [Barzilay 2003c] [Dolan 2004], supervised methodologies [Barzilay 2003b] [Brockett 2005] and methodologies based on linguistic analysis of comparable corpora [Hatzivassiloglou 1999]. In order to keep to our language-independency unsupervised strategy, we specifically tackled unsupervised methodologies based on lexical similarity. Within this context, [Dolan 2004] created the first corpus of paraphrases by automatically extracting monolingual paraphrases from massive comparable news stories. For that purpose, they used the Edit distance (also known as Levenshtein distance [Levenshtein 1966]) between pairs of sentences. In parallel, [Barzilay 2003c] used the simple word n-gram overlap function in the context of paraphrase lattices learning. However, these unsupervised methodologies show a major drawback by extracting quasi-exact or even exact match pairs of sentences as they rely on classical surface similarity measures. These pairs are clearly useless for the purpose of sentence reduction. Indeed, we are interested in extracting asymmetric paraphrases i.e. pairs of sentences, which convey the same major semantic message although written in a different style or length.

To overcome this problem, we investigated new paraphrase identification functions in [Cordeiro 2007a], which give no credit to paraphrases where both sentences are equal and also reject pairs of sentences which do not share any word. These are called hill-shape functions such as the triangular, the parabolic, the gaussian and the entropic functions. Within this context, we proposed a new surface-based similarity measure called the Sumo-metric, which outperforms, though slightly in some cases, all mathematically well-founded hill-shape functions. The Sumo-metric was inspired by the entropy function of information theory. By viewing asymmetric paraphrasing as a one way entailment, we may see the entailed sentence as a compressed output obtained from an input sentence (the entailer) through a noisy channel model transmission, similarly to what is done in [Knight 2002]. Therefore, one may think about the information gain value of the compressed sentence in relation to the input expanded one. Following this idea, we defined the Sumo-metric, which is a kind of entropic function calculated from the exclusive links connecting two sentences. For a given sentence pair, an exclusive link is a connection between two equal words, from each sentence. When such a link holds then each word is bounded to its counterpart and can not be linked to any other word. This is illustrated in figure 5.7, where, for example, the determinant *the* in the first sentence has only one link to the first determinant *the* in the second sentence and the second determinant *the* in that sentence remains unconnected.

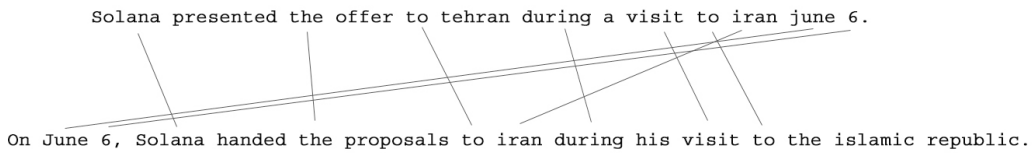

Figure 5.7: Exclusive links between a sentence pair.

So, the Sumo-metric is defined in Equation 5.9 based on the entropy-like function defined in Equation 5.10. In particular, λ is the number of exclusive lexical links connecting two sentences F_j and F_j, l the number of words of the longest sentence and s the number of words of the smallest one. Moreover, in order to adjust the Sumo-metric the best as possible to sentence reduction, we can tune the α, β and k parameters (with $\alpha, \beta \in [0, 1]$ such that $\alpha + \beta = 1$).

$$Sumo(F_i, F_j) = \begin{cases} L(\lambda, l, s) & if\ L(\lambda, l, s) < 1 \\ 0 & if\ \lambda = 0 \\ e^{-k*L(\lambda,l,s)} & otherwise. \end{cases} \tag{5.9}$$

$$L(\lambda, l, s) = -\alpha \log_2(\frac{\lambda}{l}) - \beta \log_2(\frac{\lambda}{s}). \tag{5.10}$$

In order to evaluate the Sumo-metric against eight other surface-based metrics, we used two standard corpora: the Microsoft research paraphrase corpus $\{MSRPC\}$ created by [Dolan 2004] and a paraphrase corpus $\{KMC\}$ supplied by Daniel Marcu based on his research on sentence reduction [Knight 2002]. Although, these two corpora were a good basis for evaluation, they presented some biased characteristics. For that purpose, we created three new paraphrase corpora to propose a well-founded and complete benchmark.

The $\{MSRPC\}$ corpus is the first freely available paraphrase corpus containing a total of 5801 sentences pairs, with 3900 annotated as true paraphrases and the remaining 1901 as negative paraphrase examples. One of its particularity is to embody mostly symmetric paraphrases (i.e. each sentence of the paraphrase completely entails the other one), which are not so well-suited for sentence reduction. Contrarily, the handcrafted $\{KMC\}$ corpus contains 1087 paraphrases, in which one of the sentence is the reduced version of the other one. As such, it represents a golden standard to learn asymmetric paraphrases. Nevertheless, these two corpora had to be conveniently adapted and combined in order to provide relevant test sets for the task of paraphrase identification.

One major limitation of the $\{KMC\}$ corpus is that it only contains positive examples and therefore should not be taken as such to perform any evaluation.

Indeed, it is necessary to add an equal number of negative examples in order to obtain balanced evaluations. Similarly, the $\{MSRPC\}$ corpus is fairly unbalanced. As a consequence, we decided to expand both corpora by adding negative examples i.e. sentence pairs randomly selected from Web news stories. To balance the $\{MSRPC\}$ corpus, we added 1999 negative examples and noted this new corpus as $\{MSRPC \cup X^{-}_{1999}\}$. Similarly, we transformed the $\{KMC\}$ corpus into the $\{KMC \cup X^{-}_{1087}\}$ corpus. Finally, we built the biggest paraphrase corpus so far by gathering the 3900 symmetric positive examples of the $\{MSRPC\}$ corpus and the 1087 positive asymmetric paraphrase pairs from the $\{KMC\}$ corpus, thus summing a total of 4987 positive pairs, 21.8% being asymmetric. To balance the corpus, an equal number of negative pairs were added to give rise to the new $\{MSRPC^{+} \cup KMC \cup X^{-}_{4987}\}$ corpus. It is important to notice that the negative examples of the $\{MSRPC\}$ corpus were not retained as most of them are correct asymmetric paraphrases.

Based on these three balanced corpora, we could evaluate the performance of each one of the nine surface-based metrics (i.e. Edit distance (Edit) [Levenshtein 1966], word n-gram overlap (Overlap) [Barzilay 2003c], exclusive word n-gram overlap (Excl. Overlap) [Cordeiro 2007a], BLEU [Papineni 2002], Sumo-metric[15] [Cordeiro 2007a], trigonometric [Cordeiro 2007a], parabolic [Cordeiro 2007a], entropy [Cordeiro 2007a] and gaussian [Cordeiro 2007a])[16]. Extracting paraphrases is clearly a classification problem. Indeed, assuming that we are evaluating a generic paraphrase function $pf(.,.)$, which calculates the likelihood of any two sentences F_i and F_j being a paraphrase, a threshold methodology would classify the sentence pair $\langle F_i, F_j \rangle$ as a paraphrase if and only if the following expression is satisfied: $pf(F_i, F_j) > threshold$. However, thresholds are parameters that unease the process of a fair evaluation. Indeed the best possible threshold parameter should be first determined for any given function. However, this is not always the case and, very often, wrong experimental evaluations are followed and reported. In order to avoid the aforementioned drawback, we proposed an evaluation framework, which automatically scans the best threshold of the evaluated function value based on the golden section search method [Anita 2002] for a given test corpus. Therefore, any function is tested with its best threshold. Table 5.3 evidences the obtained thresholds as well as their standard deviations within a 10-fold cross validation type test over the $\{MSRPC^{+} \cup KMC \cup X^{-}_{4987}\}$ corpus. By analyzing the low variation ranges, we can agree that the golden section search method efficiently approximates the global optimum threshold for each function.

Based on the approximated thresholds, we performed an exhaustive evaluation for the nine surface-based similarity measures based on the well-known F-measure (see Table 5.4) and the accuracy metric usually used in machine learning (see Table 5.5)

[15] The thresholds were calculated for $\alpha = \beta = 0.5$ and $k = 3$.

[16] For all the hill-shape functions, the input argument is $\frac{\lambda}{l} * \frac{\lambda}{s}$.

Function	$\{MSRPC \cup X^-_{1999}\}$	$\{KMC \cup X^-_{1087}\}$	$\{MSRPC^+ \cup KMC \cup X^-_{4987}\}$
Edit	17.003 ± 0.0000	18.800 ± 1.1353	17.000 ± 0.0000
Overlap	0.1668 ± 0.0000	0.2096 ± 0.0138	0.1298 ± 0.0008
Excl. Overlap	0.5000 ± 0.0000	0.7205 ± 0.0382	0.5000 ± 0.0000
BLEU	0.6887 ± 0.0015	0.0204 ± 0.0051	0.6369 ± 0.0077
Sumo-metric	0.0718 ± 0.0033	0.0049 ± 0.0010	0.0070 ± 0.0001
Trigometric	0.2889 ± 0.0803	0.2685 ± 0.0000	0.3421 ± 0.0000
Parabolic	0.5091 ± 0.0146	0.3242 ± 0.0001	0.3255 ± 0.0048
Entropy	0.6166 ± 0.0065	0.3851 ± 0.0011	0.4455 ± 0.0319
Gaussian	0.5250 ± 0.0095	0.3670 ± 0.0006	0.3986 ± 0.0105

Table 5.3: Thresholds mean and standard deviation.

over three different corpora. In particular, we compared four similarity measures specially defined to identify symmetric paraphrases (i.e. Edit, Overlap, Excl. Overlap and BLEU) with five similarity measures specially defined to identify asymmetric paraphrases (i.e. Sumo-metric, trigonometric, parabolic, entropy and gaussian). It is also interesting to notice that with our three paraphrase corpora, we also tested the functions in a wider range of paraphrase types, from a "pure" symmetric type set $\{MSRPC \cup X^-_{1999}\}$ to an exclusive asymmetric type set $\{KMC \cup X^-_{1087}\}$ as well as a corpus gathering pairs from both types $\{MSRPC^+ \cup KMC \cup X^-_{4987}\}$.

Function	$\{MSRPC \cup X^-_{1999}\}$	$\{KMC \cup X^-_{1087}\}$	$\{MSRPC^+ \cup KMC \cup X^-_{4987}\}$
Edit	74.42%	71.06%	80.97%
Overlap	78.94%	95.34%	94.72%
Excl. Overlap	76.54%	91.06%	86.27%
BLEU	78.72%	69.26%	85.86%
Sumo-metric	**80.93**%	**98.29**%	**98.52**%
Trigonometric	77.12%	61.40%	87.02%
Parabolic	80.18%	97.55%	98.40%
Entropy	80.25%	97.50%	98.35%
Gaussian	80.20%	97.56%	98.37%

Table 5.4: F-measure results.

It is clear that the Sumo-metric outperformed all other proposed metrics, and curiously even for the $\{MSRPC \cup X^-_{1999}\}$ corpus, which mainly contains symmetric paraphrases. A large evaluation of all the metrics can be found in [Cordeiro 2007a], always leading to the same conclusion. As a consequence, we are able to automatically extract paraphrase corpora based on any input corpus in a complete unsupervised way. In particular, we developed a prototype, which daily crawls Web news stories[17] and automatically outputs likely asymmetric paraphrases. However,

[17] From Google^{TM} News Web services.

Function	$\{MSRPC \cup X^{-}_{1999}\}$	$\{KMC \cup X^{-}_{1087}\}$	$\{MSRPC^{+} \cup KMC \cup X^{-}_{4987}\}$
Edit	67.68%	69.26%	79.27%
Overlap	73.85%	95.16%	94.48%
Excl. Overlap	70.51%	90.36%	84.49%
BLEU	74.00%	57.88%	86.18%
Sumo-metric	**78.18**%	**98.29**%	**98.52**%
Trigonometric	71.63%	69.22%	87.99%
Parabolic	75.74%	97.58%	98.39%
Entropy	75.88%	97.24%	98.34%
Gaussian	75.92%	97.56%	98.37%

Table 5.5: Accuracy results.

as one could expect, results drastically drop when dealing with real-world texts. For example, for a corpus of Web news stories compiled on October 2006 containing 166 documents about three main topics $\{WNS\}$, the Sumo-metric reached 61.8% precision[18].

In order to improve paraphrase extraction, [Barzilay 2003c] proposed to apply clustering techniques. On the one hand, they argue that clusters of paraphrases can lead to better learning of text-to-text rewriting rules compared to just pairs of sentences. On the other hand, clustering algorithms may lead to better performance than stand-alone similarity measures as they may take advantage of the different structures of sentences in the cluster to detect a new similar sentence. However, [Barzilay 2003c] did not propose any evaluation about which clustering algorithm should be used. As a consequence, we experimented different clustering algorithms based on a similarity matrix calculated with the Sumo-metric. Contrarily to [Barzilay 2003c], our conclusions were such that clustering is not a worthy effort [Cordeiro 2007b] for sentence reduction. In particular, we tested four clustering algorithms i.e. the single-link and complete-link hierarchical agglomerative clustering algorithms (HAC)[19] [Jain 1999], the quality threshold algorithm (QT) [Heyer 1999] and the expectation maximization algorithm (EM) [Dempster 1977], based on the $\{WNS\}$ corpus. The results were then manually cross-validated and they are exposed in Table 5.6.

Sumo-metric	S-HAC	C-HAC	QT	EM
61.8%	57.7%	56.9%	64.0%	48.9%

Table 5.6: Precision of clustering algorithms.

The main conclusion is that clustering tends to achieve worst results than simple

[18] A manual cross-validated evaluation was performed for this case.

[19] Taken as conceptual clustering i.e. with a quality criterion to find the best number of clusters.

paraphrase extraction to the exception of the QT algorithm compared to the single Sumo-metric. However, the results with the QT algorithm were obtained with a very restrictive value for cluster attribution as it is shown in Table 5.7 with an average of almost two sentences per cluster. In fact, this leads to similar results as the classical threshold extraction method with the Sumo-metric.

Algorithm	Average sentences per cluster
S-HAC	6.23
C-HAC	2.17
QT	2.32
EM	4.16

Table 5.7: Average number of sentences per cluster.

Moreover, Table 5.7 shows that most of the clusters have less than 6 sentences, which leads to question the results presented in [Barzilay 2003c], who only keep clusters with more than 10 sentences, which showed to be of very bad quality in our experiments. To summarize, we think that clustering is not a worthy effort especially for the purpose of sentence reduction as alignment methods for more than two sentences are more keen to error. However, for other purposes such as automatic extraction of semantically related words, clustering may lead to interesting results as shown in [Dias 2010] and discussed in Chapter 6.

5.3.3 Paraphrase Alignment

The next natural step in our research is to investigate paraphrase alignment techniques. By paraphrase alignment, we mean that equal or even similar[20] word pairs, one from each sentence in the paraphrase, are aligned and dissimilar words are lined up with empty gaps. For example, the following paraphrase (i.e. sentences (3a) and (3b)) may globally be aligned by sentences (4a) and (4b).

3 Extracted paraphrase

3a `Police used tear gas to scatter protesters`

3b `Police reacted by firing tear gas to scatter protesters`

4 Aligned paraphrase

4a `Police _____ __ _____ used tear gas to scatter protesters`

4b `Police reacted by firing ____ tear gas to scatter protesters`

As many alignments are possible for a given paraphrase, we investigated methods for global and local alignments. Finally, we proposed a dynamic strategy, which chooses the best global or local alignment algorithm at run time. Sequence

[20] We will specify later in this section what type of similar words we use.

alignment algorithms have been extensively explored in Bioinformatics since the beginning of the human genome project and aim at aligning two sequences of genes to find structural similarities, differences or transformations between them. In NLP, alignment has recently received some attention in fields such as text generation [Barzilay 2003c] [Cordeiro 2007b] [Cordeiro 2009] and the extraction of semantically related words [Dias 2010] [Grigonyté 2010]. Sequence alignment algorithms usually fall into two main categories: global and local. Global algorithms try to align all the symbols of both sequences, while local algorithms aim at finding relevant sub-alignments. The appropriateness of each category depends on sentence structure constraints. Usually, global alignments are more useful when sequences are similar both in structure and size, while local alignments are more appropriate for paraphrases, which embody common sub-sequences restructurings. In fact, we will see that both global and local alignment algorithms are necessary within the context of paraphrase alignment due to different structures of extracted paraphrases. Examples are given in [Cordeiro 2007b].

The Needleman-Wunsch algorithm [Needleman 1970] is the first to propose a solution to rapidly determine sequence homology. It uses dynamic programming to find the best possible global sequence alignment between two sequences, with respect to the scoring system being used, which includes a substitution matrix and a gap-scoring scheme. The substitution matrix defines the cost of aligning two sequence symbols, either equal or different, and the gap-scoring scheme specifies the alignment cost between a symbol and a gap. The Smith-Waterman algorithm [Smith 1981] is similar in many aspects to the Needleman-Wunsch algorithm. The Smith-Waterman algorithm is specially conceived to extract at least one optimal sub-alignment from a given sequence pair. According to [Smith 1981], their algorithm is more adequate than global alignment algorithms for heterogeneous sequences, either evidencing size differences or/and a great proportion of symbol dissimilarity.

While global alignment should be preferred over local alignment, it may not be the case for asymmetric paraphrases, which may evidence size and structure dissimilarities. For instance, while a global alignment is well-suited for the example presented below, it may not be the case for the following paraphrase.

5 `During his brilliant speech, the president remarkably praised IBM`

6 `The president praised IBM, during his speech`

This type of chunk rotations or alternations may occur with some frequency in a paraphrase corpus. In such cases, global alignment algorithms may over-align text segments to word gaps leading to poor-quality alignments. Contrarily, local alignments algorithms would lead to more interesting results. For this example, two local alignments would be identified.

7 First local alignment

7a `the president remarkably praised IBM`

7b `the president _____ praised IBM`

8 Second local alignment

8a `during his brilliant speech`

8b `during his _____ speech`

Based on these studies between global and local alignments, we proposed in [Cordeiro 2007b] a dynamic algorithm, which chooses at run time the best alignment to perform. For that purpose, we used the notion of link-crossing between sequences as illustrated in Figure 5.8, where the 4 crossings are signalized with small squares.

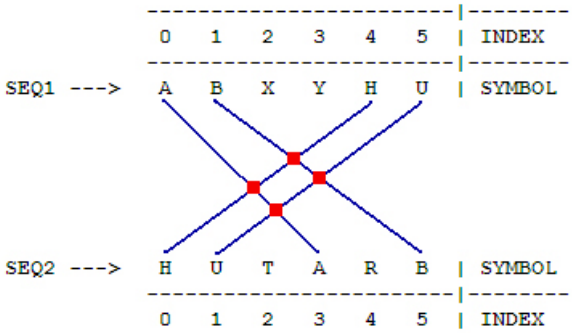

Figure 5.8: Crossings between a sequence pair.

It is easily verifiable that the maximum number of crossings between two sequences with n exclusive links is equal to $\theta = \frac{1}{2} \times n \times (n-1)$. So, we propose that the choice of the global or the local strategy for the alignment should be decided whether the number of crossings x within a paraphrase overtakes a threshold depending on a proportion of θ or not i.e. $c \times \theta$ with $c \in [0, 1]$. Remark that the more c tends to 1 the more unlikely the global strategy will hold. So, after different experiments, we proposed that $c = 0.4$ i.e. if $x > 0.4 \times \theta$, then the Smith-Waterman should be applied (local alignment), otherwise the Needleman-Wunsch algorithm (global alignment) should be used.

Biology-based sequence alignment algorithms are usually guided by a scoring function, defining the gene mutation probability. In fact, a scoring function is a matrix (usually called the mutation matrix), which evaluates the mutation likelihood between all the constituents of the vocabulary. As such, any mutation matrix models a set of concepts according to which there may be a probable evolutionary gene mutation. Hence, different matrices are likely to generate different alignments, depending on the encoded concepts. Within our context, a

straightforward and natural possibility to define a word mutation representation function was to use the well-known Edit distance $edist(.,.)$ as a negative reward for word alignment. Indeed, for a given word pair $\langle w_i, w_j \rangle$, the greater the Edit distance is, the more different w_i is from w_j. As a consequence, the more unlikely w_i would be aligned with w_j. So, to take into account the well-known drawbacks of the Edit distance, we proposed a new function $costAlign(.,.)$ defined in Equation 5.11 for word mutation penalization where ε is a small value[21] and $maxseq(.,.)$ returns the length of the longest common subsequence between two words divided by the word with maximum length.

$$costAlign(w_i, w_j) = -\frac{edist(w_i, w_j)}{maxseq(w_i, w_j) + \varepsilon}. \tag{5.11}$$

So, finally, we are able to present the results of our methodology to align paraphrases. As no golden standards exist, we proposed to carry out a manual evaluation of random samples of aligned paraphrases. In particular, we tested our alignment methodology using two sets of aligned paraphrases obtained from (1) the $\{WNS\}$ corpus and (2) the DUC 2002[22] text collection. First, we extracted a random set of 100 paraphrases from the $\{WNS\}$ corpus. Second, we automatically extracted a set of 220 asymmetric paraphrases from the DUC 2002 corpora of $\langle text, summary \rangle$ pairs. Both data sets were obtained using the Sumo-metric.

For each dataset, two human judges cross-evaluated the quality of each paraphrase alignment according to four quality labels, two positives and two negatives: excellent, good, flaws and bad. An excellent label was assigned to alignments without any error at all. The good label was used for alignments with at most two misalignments without having major implications in the overall quality of the alignment. The first negative label flaws was assigned to those cases with several misalignments, and despite the fact that good matches might still exist, the amount of wrong word alignments had an overall negative effect. Finally, a bad label was obviously assigned to completely erroneous alignments. In order to get a full spectrum evaluation of the methodology, we also asked the human judges to evaluate whether the decision of choosing the global or the local alignment algorithm was adequate or inadequate. The results obtained are shown in Table 5.8.

As it can be seen, the results of all data sets are close to each other. The best results obtained for the DUC 2002 data set are mainly due to the fact that the $\{WNS\}$ data set is noisy as it was automatically gathered from the Web without any specific tokenization process. As a consequence, many errors were due to bad text pre-processing rather than theoretical issues. Finally, the overall performance is satisfactory since we achieved 87.5% of excellent and good alignments.

[21] We took $\varepsilon = 0.01$ in our experiments.

[22] http://duc.nist.gov/duc2002/ [23rd $September, 2010$].

Data set	excellent	good	flaws	bad	adequate
DUC	144	50	18	8	15/18
220	65.5%	22.7%	8.2%	3.6%	83.3%
WNS	54	32	9	5	-
100	54%	32%	9%	5%	-
DUC ∪ WNS	198	82	27	13	15/18
320	61.9%	25.6%	8.4%	4.1%	83.3%

Table 5.8: Alignments precision.

5.3.4 Reduction Rules Learning

Aligned paraphrases evidence clear examples of the kind of knowledge that can be learnt for sentence reduction. For instance, local alignments (7) and (8) show that adverbs and adjectives may be dropped from the original sentence without great syntactical or semantic loss. Therefore, we first investigated specific aligned regions for the formulation of learning instances for the induction of sentence reduction rules. These specific aligned text segments are called bubbles [Cordeiro 2009], which are provided to an ILP framework, the Aleph system [Srinivasan 2000], as learning instances. In particular, ILP [Muggleton 1991] evidences relevant characteristics such as the capacity to generate symbolic and relational knowledge, the possibility to securely avoid negative instances [Muggleton 1996], the ability to mix different types of attributes and to control the theoretical search process. Moreover, while most learning algorithms require a complete definition of the feature set, prior to the learning process, ILP allows to define a set of possible features from which the induction process will search for the optimum solution. As a consequence, the ILP paradigm well-suits our purposes as we need to combine lexical, morpho-syntactic and shallow-syntactic attributes and may control the generalization process by only taking into account these attributes individually or combined.

A bubble is a non-empty segment aligned with an empty segment from an aligned paraphrase, sharing "strong" left and right contexts. In particular, the heterogeneous alignment is called the kernel. Some examples are shown as follows.

9a `the situation here in chicago with the workers`

9b `the situation ____ in chicago with the workers`

10a `america is in the exact same seat as sweigert and`

10b `america is in ___ _____ same seat as sweigert and`

11a `after a while at the regents park gym, the president`

11b `after a while at ___ _______ ____ gym, the president`

To extract a bubble, left and right contexts must be "strong". In fact, the idea of strength aims at discarding meaningless bubbles i.e. bubbles where the kernel is too large compared to the size of the left and right contexts. More precisely, in Equation 5.12, we define the condition to decide whether a bubble should be extracted or discarded. So, if this condition is respected, the bubble is kept, otherwise it is removed. In particular, L and R respectively stand for the left and right contexts, K is the kernel and the $Sz_{(.)}$ function computes the length of a given segment in terms of number of words.

$$Sz_{(L)} - Sz_{(K)} + Sz_{(R)} \geq 0. \tag{5.12}$$

It is important to notice that so far we only considered bubbles where the kernel implies an alignment with a void segment ($K \xrightarrow{transf} \emptyset$). However, more general transformations may be investigated. Indeed, any transformation ($K \xrightarrow{transf} Y$), where $Y \neq \emptyset$, respecting $Sz_{(K)} > Sz_{(Y)}$, may be a relevant compression example. This will remain for future work.

Once "strong" bubbles have been extracted, they must be transformed into learning instances to serve as input to the Aleph system. For that purpose, each bubble is transformed into a first-order-logic predicate, which embodies all the possible features to be considered during the induction process. More precisely, each bubble is represented by a 5-ary term (`bub/5`) where the first argument is a sequential index and the second argument is a 2-ary term (`t/2`) indicating the kernel size transformation. The fourth argument is a 2-ary term (`X ---> Y`) with `X` and `Y` being the lists of tagged words from the kernel transformation. Finally, the third and fifth arguments are respectively the left and right contexts, represented as lists of shallow-parsed words. For example, paraphrase (9a-9b) would give rise to the following predicate after shallow-parsing is processed and context and kernel sizes are limited to 3 words[23].

```
bub(1, t(1,0),
               [situation/nn/np, the/dt/np],
               [here/rb/advp] ---> [],
               [in/in/pp, chicago/nn/np, with/in/pp]
).
```

Aleph is an empirical ILP system and the natural successor of several older ILP systems, such as Progol [Muggleton 1999], FOIL [Quinlan 1990] and Indlog [Camacho 1994]. In particular, the Aleph system can be appropriately parameterized to emulate any of those older systems. Within the context of our research, one major advantage of Aleph is the possibility to learn exclusively from positive

[23] These bubble types represent 83.46% of the extracted "strong" bubbles from a corpus of 30 days Web news stories, the $\{WNS30\}$ corpus. As a consequence, we limited our study to these specific bubbles.

instances, contrarily to what is required by most learning systems. Moreover, there is theoretical research work [Muggleton 1996] demonstrating that the increase in the learning error tend to be negligible with the absence of negative examples, as the number of learning instances increases. This is a relevant issue, for many learning domains, and specially our, where negative examples are not available. In order to understand all the parametrization of the Aleph system, the reader will find useful information in [Cordeiro 2009]. To the specific case of this thesis, the only interesting issue is to understand the kind of rules, which are learnt based on the predicate representation of bubbles. One of these rules is illustrated as follows.

```
rule(A) :- transfdim(A, n(1,0)),
           chunk(A,left,np),
           chunk(A,center:x,advp),
           inx(A,right,1,pos(pp)),
           inx(A,right,2,pos(nn)).
```

In this case, the induced rule states that whenever (1) the kernel is an adverbial phrase (`chunk(A,center:x,advp)`), (2) the left context is a noun phrase chunk (`chunk(A,left,np)`), (3) the first part-of-speech tag in the right context is a preposition (`inx(A,right,1,pos(pp))`) and (4) the second part-of-speech tag in the right context is a noun (`inx(A,right,2,pos(nn))`), then only the word in the center of the kernel can be eliminated (`transfdim(A, n(1,0))`). This situation can be applied to the aforementioned predicate (`bub(1,X,[K]--->[],Y)`) representing the paraphrase (9a) and (9b) and leading to the removal of the adverb *here*.

In any rule-based system, many rules may apply at the same time to the same example. For instance, we evidenced that, on average, three rules might fire for any sentence. As a consequence, a decision module must determine which rewriting rule is the most likely to be applied given a set of constraints. In our case, we decided (1) to privilege grammaticality and (2) to favor rule coverage in case of ties[24]. So, in order to avoid syntactical errors and sentence leaps or discontinuities, we built a part-of-speech language model with the Carnegie Mellon University statistical language modeling toolkit (CMU-Toolkit)[25] over a set of 1Gb Web news stories previously part-of-speech tagged with the Open NLP project[26]. In particular, we computed the 2-gram model and then used this model to ensure that the reduced sentences still maintained a reasonable expectable part-of-speech sequence.

In order to exemplify the procedure, let's assume that there exists some reduction rule ρ such that the bracketed sequence in sentence F (see Proposition 5.13 where w_i stands for a word and t_i for its corresponding part-of-speech tag) can be

[24]The coverage of each rule is given by Aleph.
[25]http://www.speech.cs.cmu.edu/SLM_info.html [23rd*September*, 2010].
[26]http://opennlp.sourceforge.net [23rd*September*, 2010].

eliminated. In this case, ρ may be applied to F only if some model conditions are satisfied.

$$F = w_1/t_1 \dots w_i/t_i \quad [w_{i+1}/t_{i+1} \dots w_k/t_k] \quad w_{k+1}/t_{k+1} \dots w_n/t_n \tag{5.13}$$

In this case, ρ will only be applied if the sequence embodied by both consecutive part-of-speech tags, after the removal of the bracketed sequence (i.e. $[t_i, t_{k+1}]$), is more probable in terms of part-of-speech language model than the average probability of the two border sequences, which are defined as follows: the first sequence is the first word occurring before the removable sequence and the first occurring word in the bracketed sequence (i.e. $[t_i, t_{i+1}]$) and the second sequence is the last word occurring in the removable sequence and the first occurring word appearing just after the removable sequence ($[t_k, t_{k+1}]$). As a summary, any rewriting rule ρ can be applied only if condition in Equation 5.14 is respected[27].

$$P([t_i, t_{k+1}]) > \frac{P([t_i, t_{i+1}]) + P([t_k, t_{k+1}])}{2}. \tag{5.14}$$

Finally, we reach the evaluation of all our pipeline since the extraction of paraphrases to the application of reduction rules. For that purpose, we compiled a corpus of 30 days Web news stories, the $\{WNS30\}$ corpus, which gathers 133.5 Mb of raw text. By applying the previous steps of our framework, we extracted and aligned 596.678 paraphrases, which were then shallow-parsed[28], before 143.761 bubbles were extracted. This situation contrasts with previous evaluations, which were performed over very small data sets. For example, [Knight 2002] used a set of 1.035 sentences to train their system and only 32 sentences to test it. As far as we know, we are the first to propose such a large evaluation. In order to assess as clearly as possible the performance of our methodology on large data sets, we also propose a set of qualitative and quantitative evaluations based on three different measures: utility, n-gram simplification and correctness.

A relevant issue, not very commonly discussed, is the utility of a learned theory. In real life problems, people may be more interested in the volume of data processed than the quality of the results. Maybe, between a system which is 90% precise and processes only 10% of data, and a system with 70% precision, processing 50% of data, the user would prefer the last one. This is the idea of utility. Therefore, we defined the utility as the geometrical mean of the processed elements percentage ($pproc(\mathfrak{m})$) and the previously estimated precision, as shown in equation 5.15, where $\mathfrak{m}$ represents the system or theory being evaluated.

$$utility(\mathfrak{m}) = \sqrt{precision(\mathfrak{m}) \times pproc(\mathfrak{m})}. \tag{5.15}$$

[27] A similar calculation can be made for 3-gram and 4-gram and then combine values in order to make the decision.

[28] We used the tools of the Open NLP project.

The n-gram simplification methodology is an automatic intrinsic test performed to evaluate to what extent grammaticality is respected after sentence reduction. For that purpose, we computed a 4-gram part-of-speech language model exactly in the same way as proposed for the decision module. But, in this case, we normalized the model to take into account different sentence sizes as expressed in Equation 5.16, where $\vec{T_s} = [t_1, t_2, ..., t_m]$ represents the part-of-speech sequence of a given sentence of size m and $P\{t_k \mid t_{k-1}, ..., t_{k-n}\}$ is the conditional probability of the part-of-speech tag t_k given the previous sequence $t_{k-1}, ..., t_{k-n}$ for the general model.

$$P(\vec{T_s}) = \Big(\prod_{k=n}^{m} P\{t_k \mid t_{k-1}, ..., t_{k-n}\} \Big)^{\frac{1}{m-n+1}}. \tag{5.16}$$

The third evaluation is qualitative and has been followed by most of the works in sentence reduction [Knight 2002] [Le Nguyen 2004] [Clarke 2006]. It aims at measuring the correctness of the learnt rules when applied to sentence reduction. For that purpose, a human judge must evaluate the adequacy of each compression rule applied to a given sentence segment. This is usually performed over a scale from 1 to 5, where 1 stands for total inadequacy and 5 for perfect fit.

To perform our evaluation, a sample of 300 sentences were randomly extracted, where at least one reduction rule had been applied. This data set was then subdivided into three subsets of 100 instances: the $\{BD1\}$ subset, which involves the removal of just one word, the $\{BD2\}$ set for which two words are removed and $\{BD3\}$, which implies the reduction of exactly three words. Finally, another 100 sentences random sample was extracted to evaluate our base-line $\{BL\}$, which consists in the direct application of a bubble set to reduce sentences. This means that no learning process is performed. Instead, we store the complete bubble set as if they were rules by themselves, in a similar way as [Le Nguyen 2004] do. In particular, this represents the unsupervised language-independent solution, which keeps to the *corpus integrity principle* i.e. the "optimum" solution that we are seeking all along our different research works.

Table 5.9 compiles the comparative results for precision, utility, n-gram simplification and correctness for each data set. In particular, precision is the normalized correctness and the percentage for n-gram simplification is the proportion of test cases where $P(reduced(\vec{T_s})) \geq P(\vec{T_s})$.

Table 5.9 evidences the improvement achieved by the introduction of shallow-parsed features in comparison to the base line, on each test parameter. On the one hand, this result is due to data sparseness, although a large number of paraphrases were extracted. Indeed, a learning process based on just raw text testifies difficulties in the process of generalization. On the other hand, sentence reduction is an abstractive approach of ATS, which involves a deeper degree of linguistic processing. Nevertheless, the overall process of sentence reduction is totally unsupervised, language-

Parameter	$\{BL\}$	$\{BD1\}$	$\{BD2\}$	$\{BD3\}$
$precision(\mathfrak{m})$	58.60%	71.20%	80.60%	80.20%
$pproc(\mathfrak{m})$	8.65%	32.67%	85.72%	26.86%
$utility(\mathfrak{m})$	22.51%	48.23%	83.12%	46.41%
n-gram simplification	47.39%	89.33%	90.03%	89.23%
correctness	2.93	3.56	4.03	4.01

Table 5.9: Results of the reduction process.

independent for the first two modules and shallow-linguistically motivated. Within these conditions, best results overall are obtained for $\{BD2\}$ with 80.6% precision, 83.12% utility and 90.03% n-gram simplification, which means that we can expect a reduction of two words with high quality for a great number of sentences.

5.4 Future Work

This chapter is certainly the one, which most opens new directions for future research. It is also the one for which more concepts and methodologies were introduced. First, we will refer to the work that has been done on mobile text summarization in [Dias 2006a] [Dias 2007b] [Dias 2009]. Real-time ATS is still far from being possible for complex strategies. Indeed, the construction of lexical chains and the discovery of topical shifts are time-consuming processes, which so far do not allow their introduction in real-time applications. For that purpose, some authors have been dealing with efficient algorithms [Silber 2000]. But, not enough to our point of view, including ourselves. As a consequence, most real-time ATS solutions for mobile devices are still based on poor-knowledge-driven algorithms and rely on the extraction of representative sentences within texts using just a few simple clues as in [Buyukkokten 2001] [Yang 2003a]. This is also our case in [Dias 2006a] [Dias 2007b] [Dias 2009], where we proposed different weighting (word and sentence) schemes and different extractive strategies based on simplistic heuristics such as the ones proposed in [Luhn 1958] and [Edmundson 1969] but introducing deeper understanding features such as the identification of MWU. Of course, the results are not satisfactory as incoherent and cohesionless summaries are mostly retrieved or even erroneous ones, not only due to simplistic algorithms but also because Web pages embody specific page structures based on frames or table structures, which complicate the extraction of relevant text information. Within this scope, an interesting work is proposed in [Cai 2004], who designed a vision-based page segmentation algorithm (VIPS), which automatically extracts text content structures. This work is particularly interesting for mobile text summarization compared to the ones proposed by [Buyukkokten 2001] based on semantic textual units[29] and [Zhang 2003], who learn a C5.0 classifier to

[29] Which we follow in our works.

differentiate narrative paragraphs from non narrative ones based on 34 features. We took the party not to develop our works on real-time mobile text summarization as they do not evidence major innovations in the field although they evidenced the major difficulties to be overtaken. In future work, it is clear that our attention will have to be focused on (1) the Web page pre-processing stage, (2) the design and development of efficient algorithms[30] based on the extraction of meaningful text concepts i.e. MWU and (3) the structure of the returned summaries for better Web browsing. But our immediate work for mobile text summarization is clearly the implementation of our sentence reduction module into our mobile VIPACCESS application to reach "full" information digestion within the context of information retrieval.

Although summarizing Web pages "on the fly" is one of the most important applications for information digestion, there are other ways to digest information within the scope of ATS. In particular, we recently endeavored an interesting work, which objective is to propose a summary of each generated cluster within the scope of ephemeral clustering (see Chapter 4). Indeed, most of the time, the cluster label is not enough expressive to propose a clear understanding of the cluster content. For that purpose, we proposed an innovative solution, which is based on the discovery of the most expressive and general snippet within a cluster based on the notion of textual entailment. Our idea is simple. The Web snippet, which best embodies a given cluster is the one, which entails all other ones without loss of information. So, we started some experiments based on the simplified asymmetric InfoSimba informative similarity measure $AISs(.\|.)$ (see Equation 2.33), which showed promising results. This work is still unpublished as more experiments as well as an exhaustive evaluation must be concluded to confirm all our believes. Moreover, we intend to go further within the scope of this research. Indeed, once a user knows more about a cluster, he may find useful to understand what are the slight differences embodied by each Web page within the cluster. As such, each Web snippet could be highlighted (or ultra-summarized) by its differences and not its commonalities. These issues are very interesting for mobile information retrieval as well as for VIP users, as they may allow fast access to relevant information. Moreover, they can easily be computed in real-time based on our initial ideas.

Although real-time ATS systems show interesting research issues, we are likely to continue our work on more complex summarization algorithms to the detriment of computation efficiency. Indeed, the summarization of a given Web page can be processed off-line within a local search engine, i.e. without major processing time issues. Within the general scope of enhanced ATS, the ITOS algorithm proposed new insights for topic segmentation [Dias 2007a] by introducing a sharper evaluation of inter-sentence similarity. Although exhaustive evaluations have been

[30] Although, we are convinced that complex systems such as the one proposed in [Barzilay 1997] can only be processed off-line.

carried out for different word weighting schemes, work still needs to be done with different alternatives of the InfoSimba, i.e. the recursive InfoSimba $RIS_N(.,.)$ (see Equation 2.22) and different proximity coefficients such as other association measures than the SCP or attributional similarity measures, or even knowledge-based similarity measures[31]. But, maybe a more important challenge resides in the topic boundary detection algorithm. Our solution is based on an adaptation of the works proposed in [Ponte 1997] and [Beeferman 1997]. Although, these heuristics evidence interesting results, they rely too much on tuning parameters. As a consequence, fine-grained or coarse-grained segmentations can be obtained depending on the definition of the parameters. Although it can be an interesting asset of ITOS, we know that parameter tuning is not always easy and does not provide a total independent solution. Indeed, we evidenced interesting downhills within detected segments, which could help for fine-grained summarization. An interesting issue would be to work on a solution, which would dynamically find the best "local" threshold based on a intrinsic quality criterion to be maximized over the all document. This could be achieved by an iterative solution, which would find the best parameter to segment the text after one iteration by maximizing a global quality criterion and then recursively apply the maximization of the same criterion for each segment until convergence is reached. We clearly think that this methodology is likely to provide topic segmentation with a new unsupervised parameter-free mathematically well-founded solution.

The work that has been carried out for the automatic construction of lexical chains has mainly been instructive for the sake of our research on ATS. Indeed, although lexical chains present great assets to statistically discover the semantic structures of documents, our first experiments showed some limitations due to (1) the specific paradigm of lexical chains and (2) the insufficient world knowledge representation. On the one hand, the lexical chain #9 from section 5.2.3, which relates words *days*, *hours* and *minutes* is a clear example of existing lexical chain algorithms. Indeed, although it is clear that these words are semantically related, it is not evident that they express any relevant semantic document structure. In fact, this may be due to the fact that existing chainers do not sufficiently tackle lexical cohesion between long span word relationships. To overcome this drawback, most works [Barzilay 1997] [Silber 2000] propose to extract lexical chains within topical segments. To our point of view, although this is an interesting issue to definitely conform to, the way existing solutions propose to combine topic segmentation and lexical chains is not satisfactory. Indeed, actual lexical chain algorithms propose a poor combination of long chains, which traverse the overall document, with small local chains. It is clear that important work must be carried out within this scope i.e. studying the interaction between lexical chains over text segments. However, this is not the only reason why lexical chains may fail in representing the semantic structure of texts. Indeed, the words *days*, *hours* and *minutes* may appear in the

[31] On the condition, that they are based on automatically built lexical-semantic knowledge bases.

same segment and even in the same sentence. So, lexical chain algorithms may also suffer from short span word relationships. This kind of problem has usually been avoided by defining an *ad hoc* number of words within a chain to be accepted as a representation of the semantic content of a text segment. This solution is certainly not acceptable as there is no linguistic evidence that a lexical chain should contain a certain amount of words to be meaningful. Let's take the following chain, which would artificially contain the words *Byelorussia*, *defeat* and *France*. It is clear that work needs to be done in this area. Some research directions could include a combination of word semantic relatedness within a static knowledge base with corpus evidence based on specific word weighting schemes to evaluate the relatedness criterion. The other important drawback of our solution is certainly the pre-processed prototype-based ontology. Although soft clustering allows a given word to belong to different clusters, the absence of subsumptionable words in the nodes of the hierarchy is a great handicap to capture semantic relatedness with great accuracy. Indeed, most of the obtained lexical chains contain the words of one or two clusters only. This could be solved to some extent by taking overlapping words as upper-level nodes. However, due to simplistic evaluation of word similarity based on noun-noun relationships within the attributional similarity paradigm, these words were not enough to propose a satisfactory solution. For that purpose, we started to work on different issues to build high quality lexical-semantic knowledge based, which are developed in Chapter 6. In particular, we proposed (1) a new methodology to discover word generality order, so that upper-nodes can be populated based on well-founded theories [Dias 2008], (2) a new framework to find semantically related words with high precision, so that semantically tight clusters can be found [Moralyiski 2007] [Dias 2010] and (3) a new unsupervised language-independent methodology to build terminologically-based ontologies [Cleuziou 2010]. Finally, sentence extraction algorithms will have to be proposed based on the combination of automatically identified topical shifts with lexical chains with high document semantic structure representativeness.

Although sentence reduction has widely been studied, many important issues remained opened. On the first hand, the attentive reader may have noticed that we did not performed text normalization for paraphrase extraction and alignment. However, this experiment has recently been studied in [Grigonyté 2010] and showed great improvements[32]. However, the normalization was linguistically motivated within a domain specific corpus. As a consequence, we aim at experimenting SENTA over the $\{WNS30\}$ corpus to acknowledge how much can be reached with language-independent methodologies. The reduction process may also be revisited. Indeed, the decision module only applies the best scoring rule to any sentence. However, this does not mean that this is the maximum reduction. Ideally, a sentence may fire different but compatible rules, thereby producing more compacted sentences. Two solutions are possible. On the one hand, we may

[32]The normalization was associated to a simple adaptation of the Sumo-metric.

investigate rule combination by designing optimum sequences of rules. On the other hand, we may propose a recursive version, which iteratively applies the best scoring rule at each step of the algorithm. A combination of these two solutions can also be thought by maximizing a global quality criterion at each step of the algorithm. This is really an important issue to sentence reduction as we evidenced that, on average, three rules may fire for each sentence. As mentioned earlier, so far, we only tackled bubbles where the kernel implies an alignment with a void segment ($K \xrightarrow{transf} \emptyset$). However, more general transformations may be investigated. Indeed, any transformation ($K \xrightarrow{transf} Y$), where $Y \neq \emptyset$, respecting $Sz_{(K)} > Sz_{(Y)}$, may be a relevant compression example. Finally, in order to reach better performance for paraphrase extraction in real-world environments, which only reached 61.8% in our experiments, we may propose to combine both the Sumo-metric and the asymmetric InfoSimba $AIS(.\|.)$ (see Equation 2.32) to combine surface-based and "semantic" similarity measures ■

Chapter 6

Construction of Lexical-Semantic Resources

Contents

6.1 Prototype-based Ontologies 139

6.1.1 Discovering Highly Related Words by Similarity 140

6.1.2 Discovering Highly Related Words by Interchangeability . . . 142

6.1.3 Discovering Word Generality 150

6.2 Terminological Ontologies 159

6.3 Future Work 168

Lexical-semantic structures play an essential role in information retrieval (IR) and natural language processing (NLP). By coding the semantic relationships between concepts of discourse, they can enrich the reasoning capabilities of applications in IR and NLP such as word sense disambiguation [Tanaka 2007], text clustering [Hotho 2003], query-expansion [Bhogal 2007], personalized information retrieval [Mylonas 2008] to name but a few. However, their development is largely limited by the efforts required for their construction. In order to reduce the amount of work for engineering large lexical-semantic structures, also known as ontologies in their strongest sense, many researches have been appearing in the recent years to learn such structures from texts fostering new surveys about this field [Buitelaar 2005] [Biemann 2005] [Cimiano 2009]. Learning lexical-semantic resources from text instead of manually creating them has undeniable advantages. First, creating resources from texts within a domain may fit the semantic component neatly and directly, which will never be possible with general-purpose resources. Second, the cost per entry is greatly reduced, giving rise to much larger resources than an advocate of a manual approach could ever afford.

Although automatically creating semantic resources of any kind is a difficult task [Biemann 2005], different learning methods have been proposed. They can be grouped into three main classes: (1) the similarity-based methods [Hindle 1990] [Pereira 1993] [Caraballo 1999] [Paaß 2004] [Cimiano 2004] (2) the set-theoretical approaches [Petersen 2004] [Cimiano 2005] and (3) the associative frameworks [Sanderson 1999] [Sanderson 2000] [Dias 2008]. The first two methods adopt Harris' distributional hypothesis [Harris 1968] and represent terms as context feature

vectors but differ in the way these are processed. While similarity-based methods use word similarity measures to compute pairwise similarities between vectors in a numerical space (polythetic approach), set-theoretical approaches partially order objects according to the inclusion relationship between their feature sets in a symbolic space. Oppositely, the associative models do not follow Harris' hypothesis and hierarchically organize terms using asymmetric association measures, which evaluate the subsumption relation strength based on term distributions over documents (monothetic approach).

Learning conceptual hierarchies upon different paradigms mandatorily leads to different lexical-semantic structures. On the one hand, the similarity-based methods discover prototype-based ontologies [Biemann 2005], which are distinguished by typical instances rather than by definitions or axioms. Categories are formed by extensionally collecting instances rather than describing the set of all possible instances and selecting the most typical members. As a consequence, nodes in the lexical-semantic structures are usually labeled by relevant subsuming words present in the child nodes. An illustration of a prototype-based ontology is given in Figure 6.1. On the other hand, set-theoretical methodologies build semantic-based ontologies, which structure the data into units which are formal abstractions of concepts of human thought, thus allowing meaningful comprehensible interpretation. So, set-theoretical methodologies can be seen as conceptual clustering techniques, which provide intensional descriptions for the abstract concepts or data units they produce. An illustration of a semantic-based ontology is given in Figure 6.2.

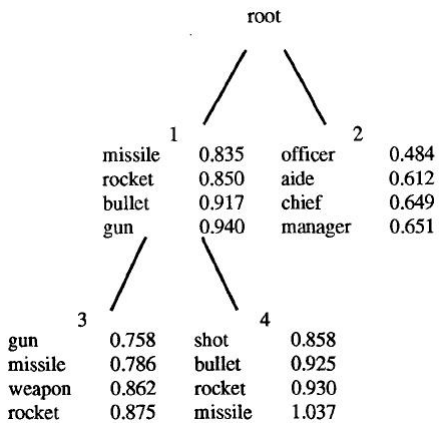

Figure 6.1: Prototype-based ontology from [Pereira 1993].

Finally, the associative models build terminological ontologies [Biemann 2005], which partially specify subtype-supertype relations and describe concepts by concept labels or synonyms rather than prototypical instances. Well-known examples of terminological ontologies are the general-purpose lexical-semantic database Word-

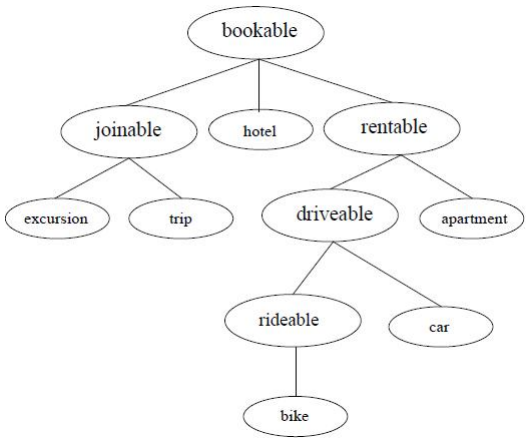

Figure 6.2: Semantic-based ontology from [Cimiano 2005].

Net [Miller 1990] and the UMLS[1] for the medical domain. An illustration of a terminological ontology is given in Figure 6.3.

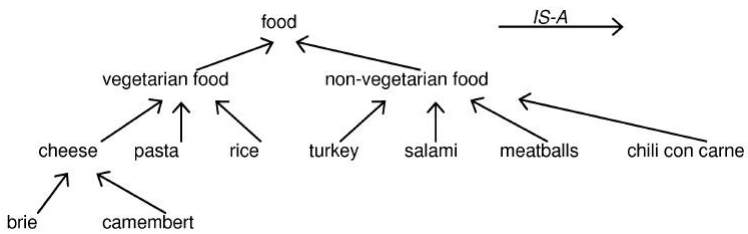

Figure 6.3: Terminological ontology from [Biemann 2005].

As knowledge-rich approaches may inherit the main shortcomings and limitations of man-made lexical resources (i.e. limited vocabulary size, unclear classification criteria and obviously considerable time and effort required to building semantic-lexical relations), we proposed to both study prototype-based and terminological ontologies based on shallow-linguistic resources or even on no resource at all but texts.

The similarity-based methods are best suited for discovering prototype-based ontologies. Within this context, attributional word similarity measures play an essential role. As mentioned in Chapter 2, different strategies have been proposed to extract synonyms or near synonyms [Hindle 1990] [Grefenstette 1993] [Lund 1995] [Landauer 1997] [Sahlgren 2001] [Terra 2003] [Ehlert 2003] [Weeds 2004] [Freitag 2005] [Heylen 2008]. The overlying idea is to find a word similarity measure capable of identifying highly related words so that clustering algorithms may successfully find clusters of synonyms, which then can be

[1] http://www.nlm.nih.gov/research/umls/ [17th *September*, 2010].

structured into a taxonomy. Within this scope, we proposed in [Moralyiski 2007] a new idea, which combines both local and global distributional representations of words to avoid as much as possible the bottleneck of polysemy [Freitag 2005].

However, later, we demonstrated in [Dias 2010] that although most metrics perform well on TOEFL-like[2] test cases, the results of attributional similarity measures decrease almost exponentially when more words form the TOEFL-like test cases. We conducted a simple experiment with a set of 1000 random test cases, created in the manner described in [Freitag 2005], with up to 10 decoy words and 1 synonym. We then solved the test cases using a contextual similarity measure i.e. the cosine similarity measure (see Equation 2.14) with features weighted with the PMI (see Equation 2.13). The increase of the number of decoys caused a rapid drop of the probability to rank first the synonym as shown in Figure 6.4.

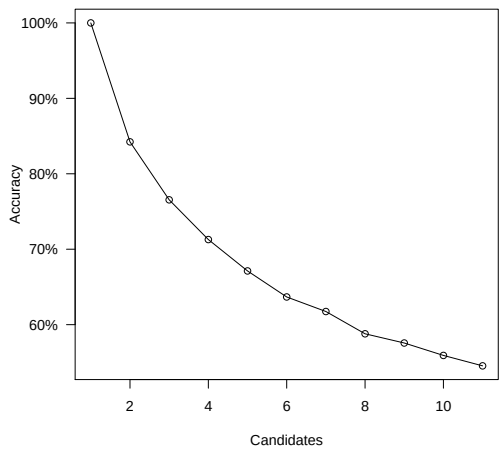

Figure 6.4: Accuracy by candidates counts.

As a consequence, as mentioned in [Heylen 2008], the exhaustive search is only capable to find the most salient semantic relations i.e. the ones that are established in the language and are frequent enough to be well represented, but also the ones that are usually included in manually built thesauri as well. In particular, [Heylen 2008] who used distributional similarity measures to unsupervisedly extract synonyms from shallow parsed corpora, reported that *the dependency-based model finds a tightly related neighbor for 50% of the target words and a true synonym for 14%*. In fact, the low accuracy of attributional similarity measures is a well-known problem and shows all the limitations of the similarity-based approach to build ontologies with high accuracy. As a consequence, so far, mainly prototype-based ontologies can be built based on the similarity-based paradigm. In order to improve the extraction

[2]Test of English as a Foreign Language.

of highly related words, we proposed to automatically extract nominal expressions within tight semantic relations (e.g. synonymy, hyperonymy/hyponymy, siblings) based on a new paradigm, which introduces paraphrase alignment [Dias 2010]. In particular, this terminological selection may lead to the construction of fine-grained prototype-based ontologies as we are capable of obtaining meaningful clusters with few noisy words inside.

The second main problem presented by similarity-based methods (which mostly build hierarchical structures based on hierarchical clustering) is the labeling of internal nodes. We already experimented this limitation in [Dias 2006b]. Indeed, selecting the best words in the child clusters to label the parent clusters is not an easy task. Many approaches have been proposed but all depend on language, such as the use of syntactical patterns [Caraballo 1999]. So, we proposed an unsupervised language-independent methodology to order words by levels of generality/specificity in [Dias 2008] so that nodes within a hierarchical structure can be populated with some degree of accuracy. In particular, we proposed to use asymmetric measures to characterize subsumption between words and finally obtain a complete order by running the TextRank algorithm [Mihalcea 2004] over a directed acyclic graph (DAG).

Interestingly, the results obtained in [Dias 2008] opened new insights for the automatic unsupervised construction of terminological ontologies. Indeed, just by looking at the asymmetry between words based on document contexts, we were capable of building meaningful lexical-semantic structures for a well-behaved set of words[3]. As a consequence, following the work proposed by [Sanderson 1999] [Sanderson 2000] and our intuition in [Dias 2008], we proposed a new framework to automatically build terminological ontologies based on asymmetric measures and the pretopology theory. This work is still under development but we will show its main ideas based on our paper [Cleuziou 2010].

6.1 Prototype-based Ontologies

We first started to tackle the construction of prototype-based ontologies based on our earlier work on lexical chains [Dias 2006b] (see Chapter 5, section 5.2.2). In particular, two main problems appeared when constructing prototype-based ontologies: (1) clustering highly related words with high accuracy and (2) populating nodes of the hierarchy with relevant words. We deal with both problems in the following sections.

[3]By well-behaved, we mean that all words present at least and at most an hypernym/hyponym or synonym relation with some other word in the set.

6.1.1 Discovering Highly Related Words by Similarity

The task of recognizing synonyms can be defined as in [Turney 2001] i.e. *given a problem word and a set of alternative words, choose the member from the set of alternative words that is most similar in meaning to the problem word.* Based on this definition, many algorithms [Landauer 1997] [Turney 2001] [Turney 2003] [Sahlgren 2001] [Ehlert 2003] [Terra 2003] [Freitag 2005] have been proposed and evaluated using multiple-choice synonym questions taken from the TOEFL. However, most of the works proposed so far, independently of their categorization, have in common the fact that word representation is built on global corpus evidence. As a consequence, all the senses of a polysemous word share a single description. This fact is clearly a drawback for any word meaning analysis. Indeed, this would mean that, to be synonyms, two words should share, as many as possible of their senses, while they usually do share just one.

A first attempt to take into account local corpus evidence is proposed in [Rapp 2004] who separate corpus evidences for distinct word occurrences in a corpus to build a matrix that is afterwards subjected to a single value decomposition and analyzed to discover the major word senses. However, they do not propose any evaluation and validation of their work, neither it is reproducible on a small scale i.e. single texts. Following the same idea, we proposed in [Moralyiski 2007] a method to measure syntactical oriented attributional similarity based on the *one sense per discourse* paradigm. So, instead of relying exclusively on global distributions, we build word representations and compare them within documents limits. In this way, we hopefully compare two specific senses of each word at a time and we argue that combining the local representation with the global approach may lead to improved results.

The common attributional similarity approach of gathering statistics from large corpora discards the information within single texts, which has shown promising results as in [Turney 2001]. Indeed, building context vectors of syntactical modifiers[4] based on the overall corpus by treating it as a single huge text implies the assumption that words are monosemous. Instead, the local attributional similarity approach aims at introducing the document dimension to the word meaning acquisition process. As a consequence, different noun meanings are not merged together into a single vector. The formal expression of the the local similarity $Lsim(.,.)$ is given in Equation 6.1 where D is the set of texts in the corpus where both nouns[5] n_1 and n_2 appear and $S(.,.)$ is any similarity measure described in Chapter 2.

$$Lsim(n_1, n_2) = \frac{\sum_{d \in D} S(n_1, n_2)}{card(D)}. \tag{6.1}$$

[4]In our case, we take into account adjectives, verbs and adverbs.
[5]At this stage, we are only interested in noun synonyms.

In fact, the global similarity may work as an indicator that words n_1 and n_2 are similar and the local similarity confirms that n_1 and n_2 are not just only similar, but instead good synonym candidates. For that purpose, we propose in Equation 6.2 a combination of both the global similarity $Gsim(.,.)$ (for which the similarity measure $S(.,.)$ is computed over the entire corpus) and the local similarity $Lsim(.,.)$ defined in Equation 6.1. In particular $\alpha \in [0,1]$.

$$Psim(n_1, n_2) = Gsim(n_1, n_2)^{\alpha} \times Lsim(n_1, n_2)^{1-\alpha}. \tag{6.2}$$

In order to test our assumption, we implemented the vector space model with the cosine similarity measure (see Equation 2.14) and different weighting schemes: term frequency (CosTf), term frequency weighted by inverse document frequency (CosTfIdf) (see Equation 2.19)[6], pointwise mutual information (CosPMI) (see Equation 2.13) as in [Terra 2003] and conditional probability (CosProb) (see Equation 2.13) as in [Weeds 2004]. We also implemented two probabilistic similarity measures: the Ehlert model (see Equation 2.18) as proposed in [Ehlert 2003] and the Lin model (see Equation 2.16). The evaluation was conducted on the subset of the 23 noun questions of a 50 multiple-choice synonym questions taken from the ESL (test for students of English as second language) provided by Peter Turney. Moreover, the statistics were evaluated based on a corpus automatically built from the Web so that sufficient counts were available to provide statistically relevant results. In particular, the corpus was shallow parsed using the MontyLingua software [Liu 2004b] and consists of 39 million words and 122 thousand word types in nearly 16 thousand documents. Table 6.1 presents the results for the ESL test cases[7], where each target word is associated to four decoys.

Experiment	$Gsim(.,.)$	$Lsim(.,.)$	$Psim(.,.)$
CosTf	39.13%	73.91%	69.57%
CosTfIdf	73.91%	65.22%	65.22%
CosPMI [Terra 2003]	60.87%	78.26%	82.61%
CosProb [Weeds 2004]	65.22%	82.61%	91.30%
Ehlert [Ehlert 2003]	65.22%	60.87%	69.57%
Lin	56.52%	78.26%	69.57%

Table 6.1: Performance of the $Psim(.,.)$.

The overall best results were obtained by $Psim(.,.)$ based on the cosine similarity measure combined with conditional probability as weighting factor, confirming the recent work of [Weeds 2004]. In particular, 91.30% accuracy (21 correct answers over 23) was reached. In Table 6.2, we illustrate how the global similarity highlights related words yet the local similarity is the measure that selects the correct option.

[6] The classical version of the vector space model.

[7] In particular, α was settled to 0.5. We will see in the other section that α can automatically be tuned.

a) stem	$Gsim(.,.)$	$Lsim(.,.)$	$Psim(.,.)$
b) column	0.0066	0.0370	0.0002
c) bark	0.0230	0.0225	0.0005
d) stalk	0.0278	**0.0577**	**0.0016**
e) trunk	**0.0288**	0.0151	0.0004

Table 6.2: Global vs. Local similarity for the CosProb.

Although these experiments show an improvement by combining local and global representations, we can not reliably state that this new methodology will solve the problem of synonymy detection. First of all because, as shown in Table 6.3, we do not reach the levels obtained by [Turney 2003] on the overall TOEFL and not just the subset of 23 nouns, which is equal to 97.50%. Moreover, as all other measures, if more words are presented as decoys, the $Psim(.,.)$ accuracy will also automatically decrease as shown in [Dias 2010].

Work	Best result
[Landauer 1997]	64.40%
[Sahlgren 2001]	72.00%
[Turney 2001]	73.75%
[Terra 2003]	81.25%
[Ehlert 2003]	82.00%
[Freitag 2005]	84.20%
[Turney 2003]	97.50%

Table 6.3: Accuracy on the overall TOEFL question set.

In fact, although high results are presented, even by other authors, they must be carefully interpreted. Indeed, solving pre-existing test cases does not mean that we are capable of automatically extracting synonyms. Another interesting study is proposed by [Heylen 2008], who apply an exhaustive search based on attributional similarity measures and show that synonyms were well encountered in only 14% of the cases. These results limit somehow the application of attributional similarity measures to learn ontologies based on clustering. Although many relevant clusters may be found, they will unlikely contain only synonymous i.e. the optimum case. However, one important contribution of this study is the fact that it seems difficult to treat synonymy without analyzing local contexts. In fact, this result will guide our second contribution in the field.

6.1.2 Discovering Highly Related Words by Interchangeability

Most of the works on discovering highly semantically related words have mainly been dealing with the distributional analysis paradigm. However, the relative

success of the vector space model on synonymy tests presented in the literature is mainly due to the structure of test cases i.e. a pair of unequivocally synonymous words and a small set of mostly unrelated decoys as shown in [Dias 2010]. In fact, the exhaustive search which is the obvious way to verify all the possible semantic connections between words of a given vocabulary is known to lead to low precision due to frequency sensitivity and word polysemy. So, in order to discover pairs of semantically related words that may be used in figurative or rare sense, we need to have them highlighted by their local environment as in the information extraction strategy and evaluate their semantic similarity by looking at their local and global distributional representations as in [Moralyiski 2007]. Our method aims at creating TOEFL-like tests of one target word plus a list of words, as short as possible, that are predominantly in paradigmatic relations with the target. Eventually, a candidate word may be interchangeable with the target word in context. One of the early formulations of this idea is attributed to Gottfried Leibniz who states that two words are synonymous if they are interchangeable in statements without a change in the truth value of the statements where the substitution occurs[8]. As a consequence, our goal is to find words, which are interchangeable in a local environment and then elect the ones, which show higher attributional similarity to obtain high accuracy extraction[9] of tightly related words. For that purpose, we propose to align paraphrases from automatically crawled Web pages such as in [Cordeiro 2007b] and discover words that are possibly substitutable for one another in context. Then, we solve these automatically extracted TOEFL-like test cases by applying different attributional similarities. This new trend within the area shows that lists of synonyms, hyponyms/hypernyms, siblings or instance-of can be extracted as in [Dias 2010], or even more in [Grigonyté 2010] such as antonyms or associated words. This methodology is mainly unsupervised and language-independent[10], which allows to automatically extract highly semantically related words in real-world environments and as such defines a quality terminological lexicon to be clustered to obtain a fine-grained prototype-based ontology.

Words prove to be rather promiscuous with respect to the semantic frames in which they can fit into. This specific behavior has primary origin in polysemy and in the creative use of language. Globally, a single context may not be enough to select the sense of a word. Rather, [Kaplan 1950] proposes that the semantic relations between the words within a sentence select their meanings. Following the same idea, [Charles 2000] collected a number of sentences, removed a word from

[8] It is commonly accepted that synonymy can be valued over a continuous scale rather than being a dichotomous relation i.e. for all synonyms there are statements whose meanings are changed by the substitution of one word for another. Thus, a more realistic view requires to see the synonymy as a continuous scale.

[9] As opposed to identification.

[10] Indeed, we use a shallow-parser in order to reduce irrelevant statistical evidences and to minimize computational complexity. But, as shown in [Ehlert 2003], comparable results without linguistic resources can be obtained if large quantities of texts are available and processing time is not an issue.

each of them and asked two groups of human subjects to recover the missing words when presented with a list of sentences and with a list of words taken from the same sentences. He observed that sentences impose stronger lexical preference than disconnected words and thus were more reliable evidence for measuring semantic similarity of pairs of words. Therefore, following these ideas, we aim at finding pairs of sentences in which one word is substituted by another one and then making a confident decision whether both words share the same meaning or not based on attributional similarity measures. Detecting paraphrases provides an elegant solution to the first part of the problem.

Similarly to what we did for sentence reduction in Chapter 5, we first built a square matrix evaluating the Sumo-metric (see Equation 5.9) between all the sentences extracted from the corpus of 30 days Web news stories, the $\{WNS30\}$ corpus. Based on this matrix, we then applied the quality threshold clustering algorithm (QT) [Heyer 1999] to build clusters of paraphrases so that stronger test cases could be obtained[11].

The next step aims at aligning the paraphrases inside each cluster i.e. detecting their common parts so as to evidence what differentiates them. Although the combined approach of biology-based alignment algorithms proposed in [Cordeiro 2007b] proposes an elegant solution to learn sentence reduction rules, it revealed useless for the present case. Indeed, biology-based alignment algorithms allow to align pairs of sequences. However, in the case of clusters of paraphrases, we may have more than two sentences to align. For that purpose, we applied the multiple sequence alignment paradigm. Different algorithms exist as proposed in [Barzilay 2003c] [Notredame 2007]. But, we preferred to use the one proposed by [Ahonen-Myka 1999], which allows to avoid the negative effects of stop-words in alignment algorithms. In fact, similarly to a mutation matrix, stop-words are statistically kept off from the overall process. The algorithm first extracts maximal frequent sequence (MFS) sets of paraphrases. A frequent sequence (FS) is defined as a non contiguous sequence of words that must occur in the same order more often than a given sentence-frequency threshold and MFS are constructed by expanding a FS to the point where the frequency drops below the threshold. This expansion is done through a greedy algorithm for which neither stemming nor stop-word removal are necessary. This way, a set of MFS is assigned to each cluster of paraphrases. In order to illustrate the overall process, we present in Figure 6.6, the alignment obtained from the cluster of paraphrases given in Figure 6.5.

[11] At the time of our research, we had in mind the results of [Barzilay 2003c] who argued that clusters of paraphrases could lead to better learning of text-to-text rewriting rules compared to just pairs of sentences. Moreover, clustering algorithms might lead to better performance than stand-alone similarity measures as they might take advantage of the different structures of sentences in the cluster to detect a new similar sentence. In fact, we will discover later in [Cordeiro 2007b] and [Grigonyté 2010] that this statement does not stand so strongly.

1. *Kazakhs are outraged by the wildly anticipated mock documentary feature Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*

2. *The news follows controversy surrounding the comedy film Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan which cut so close to the funny bone.*

3. *Meanwhile Borat is leaping to the big screen in the mockumentary Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*

Figure 6.5: A cluster of 3 paraphrases.

[{1:*Kazakhs are outraged by the wildly anticipated mock documentary feature*} {2:*The news follows controversy surrounding the comedy film*} {3:*Meanwhile Borat is leaping to the big screen in the mockumentary*}] Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan [{2:*which cut so close to the funny bone*}]

Figure 6.6: The alignment corresponding to the cluster of Figure 6.5.

The final step is to form TOEFL-like test cases from the aligned segments in the clusters. The notion of test implies one word in a specific position, or target word, for which we are searching matches among a list of candidates. So far, we have clusters of sentences conveying nearly the same message, but slightly differing in expression and with the corresponding parts aligned. We now need to search for candidates among the words, which appear out of the MFS i.e. the different parts of the paraphrases. In order to extract lists of interchangeable words, we first lemmatize and assign part-of-speech tags to the aligned paraphrases with the MontyLingua software [Liu 2004b]. This step is necessary since we are interested in nominal semantic relations and only open class words with the same part-of-speech are eligible candidates[12]. So, those parts of the paraphrases that lie between two successive parts of a MFS are potential candidates for synonymy. As a consequence, we build test cases as in algorithm 11.

Algorithm 11 The test case algorithm.

```
for each aligned sub-segment (or kernel) do
  for each open class word do
    Create a list of candidates from the rest of the segments that share left and
    right MFS contexts.
  end for
end for
```

[12]It is clear that this process could be done without part-of-speech tagging, leaving the decision process to the test solving part. But, this would prejudice the accuracy of the task as more words form the test case, the less accurate the decision may be.

For example, from the words from the first aligned paraphrase in Figure 6.6, we extract two test cases for the target words *kazakh* and *feature* as shown in Figure 6.7[13].

1. *kazakh* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*
2. *feature* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*

Figure 6.7: Two created TOEFL-like test cases.

Once TOEFL-like test cases have been created, we propose to extract the correct candidate synonyms by applying attributional similarity measures as proposed in section 6.1.1 i.e. by combining local and global word representations. For that purpose, we built a corpus as proposed in [Moralyiski 2007] with 500 million words in 110 thousand documents in which each sentence is a predicate structure based on the shallow-parsing structure given by MontyLingua. In order to keep the evaluation manageable, we only retained at random 1000 clusters of sentences and from them extracted 1058 noun test cases. Few clusters yielded more than one test. Then, we manually classified them into 5 classes with respect to whether the test contained a pair of words in one of the following relations: synonymy, siblings (or co-hyponymy), hypernymy/hyponymy (or is-a) or instance-of. Otherwise, we labeled it as None. This situation is illustrated in Table 6.4. It is interesting to observe, that the cases of synonymy together with co-hyponymy are more populous than the other two categories. It is no surprise that words with the same level of generality are preferred substitutes for the sake of paraphrasing[14].

Synonyms	Siblings	Is-a	Instance-of	None	Overall
117	108	61	86	686	1058

Table 6.4: Classification of the Test Cases.

It is also interesting to notice that out of the 117 pairs that we found to be in synonymy roles only 29 were present in WordNet as such. An excerpt of the annotated tests is given in Table 6.5. They all contain a pair of words that could be regarded in a given semantic relation in a given context.

In order to quantify the feasibility of the methodology, we retained the 372 test cases labeled with a specific semantic relation and performed a comparative study. For all the similarity measures and the respective weighting schemes (i.e. CosTfIdf, CosPMI, CosProb, Ehlert and Lin), we solved each test using the global, local

[13] Of course, six more test cases would be extracted from this paraphrase cluster for the nouns *news*, *controversy*, *film*, *Borat*, *screen* and *mockumentary*.

[14] We will see that this issue is very important for the sake of prototype-based ontology construction. In particular, we will see in the last section of this chapter that the work we developed in [Bastos 2009] is based on both the level of generality and the level of semantic closeness.

Semantic relation	TOEFL-like test cases
Synonyms	*body* | *panel*
	michael | *mike*
	administration | *government*
	condition | *disease* | *treatment*
	seat | *place* | *american* | *congress* | *election*
Siblings	*idea* | *plan*
	amazon | *ebay*
	journalist | *videographer*
	blaze | *wildfire* | *santa*
	reality | *point* | *campaign*
Is-a	*conspiracy* | *obstruction*
	capability | *repair*
	status | *fame* | *fortune*
	game | *play* | *room* | *sideline*
	allegation | *statement* | *admission* | *family* | *friday*
Instance-of	*july* | *month*
	community | *un*
	patriot | *team* | *right*
	fedex | *company* | *order*
	schwarzenegger | *star* | *film* | *terminator*

Table 6.5: Manually annotated tests. The respective relations hold between the first and the second words of each test.

and product similarity measures. In particular, the α parameter from Equation 6.2 was trained using well-known synonymy tests i.e. the TOEFL [Landauer 1997], the ESL [Turney 2003], the Reader Digest [Turney 2003] and the Freitag test set [Freitag 2005]. The results are summarized in Tables 6.6, 6.7 and 6.8.

	Lin	Ehlert	CosProb	CosPMI	CosTfIdf
Synonyms	42%	58%	42%	**75%**	50%
Siblings	**65%**	29%	53%	**65%**	47%
Is-a	42%	29%	46%	**58%**	54%
Instance-of	14%	26%	23%	**30%**	26%
Overall	33%	34%	36%	**49%**	39%

Table 6.6: Accuracy of $Gsim(.,.)$ on 372 tests.

The first observation we can make from Table 6.6 is that the combination cosine with PMI is nearly sufficient to extract the closest semantic relations[15]. However, most global measures achieve results near random guessing for the category instance-of. This is no surprise since, in order to be solved, most of the cases in this category

[15] Although, the best results were obtained by the CosProb for the ESL test in section 6.1.1.

	Lin	Ehlert	CosProb	CosPMI	CosTfIdf
Synonyms	54%	58%	**71%**	50%	58%
Siblings	**59%**	47%	47%	53%	41%
Is-a	38%	42%	42%	42%	**46%**
Instance-of	40%	42%	35%	40%	**49%**
Overall	43%	46%	46%	44%	**48%**

Table 6.7: Accuracy of $Lsim(.,.)$ on 372 tests.

	Lin	Ehlert	CosProb	CosPMI	CosTfIdf
Synonyms	50%	**63%**	46%	58%	58%
Siblings	**71%**	41%	41%	59%	64%
Is-a	46%	42%	50%	**58%**	46%
Instance-of	33%	37%	35%	34%	**42%**
Overall	43%	44%	40%	46%	**48%**

Table 6.8: Accuracy of $Psim(.,.)$ on 372 tests.

reduce to a problem of finding the most salient property associated to a proper name. For example, the pair ⟨*President, Luiz*⟩ refers to *Luiz Inácio Lula da Silva*. However, *Luiz* is a common name and as such is a very polysemous word[16]. Here is where the local similarity comes into play. Since, it always compares monosemous representations, it is bound to associate *President* with *Luiz* in those documents where the president of Brazil is the subject. As a result, the performance of the local similarities show statistically significant improvements over the global similarities for the instance-of test cases. Moreover, the results evidence that a single measure can not solve the entire problem. The synonym relation is best treated by global values, the instance-of relation is best treated by local values, while the Lin model deals best with the siblings for the product values. We summarize all these results in Table 6.9.

	Lin	Ehlert	CosProb	CosPMI	CosTfIdf
Synonyms	-	-	-	$Gsim$	-
Siblings	$Psim$	-	-	-	-
Is-a	-	-	-	$Gsim$ or $Psim$	-
Instance-of	-	-	-	-	$Lsim$
Overall	-	-	-	$Gsim$	-

Table 6.9: Best methodology by semantic category.

The results with this new methodology show that synonymy extraction as compared to synonymy identification is possible and may lead to the creation of highly semantically related clusters for the sake of prototype-based ontology learning. How-

[16]It is clear that MWU identification would lead to better results. We propose this integration in [Grigonyté 2010].

ever, not only synonyms are extracted but a larger set of semantic relations such as co-hyponyms, hypernyms/hyponyms and instance-of. Although different measures seem to be more suited to some specific semantic relations, this result must be confirmed on a larger scale. Indeed, still many automatically created test cases are erroneous, which need to be mechanically withdrawn. Different reasons stand for this situation. First of all, clusters of paraphrases seem to induce more noise than they produce interesting alternatives. Second, the alignment process is more difficult when dealing with more than two sentences, thus leading to incorrect test cases due to misaligned sequences. Third, wrong part-of-speech tagging is also responsible for inaccurate test cases. In order to deal with all these problems, we recently proposed a similar study in [Grigonyté 2010], which first normalizes two domain specific corpora so that MWU are first identified. In particular, the Sumo-metric is adapted to take into account MWU and shows improved results for the extraction of paraphrases. Second, the alignment is performed by the methodology proposed in [Cordeiro 2007b] (see Chapter 5, section 5.3.3). Finally, only one-to-one word alignments are aligned instead of looking at more complicated cases. Within this context and without the application of attributional similarity measure, but just by extracting one-to-one word alignments, we were able to automatically extract synonyms or near synonyms with 71.29% precision for the computer security domain and 66.06% for the cancer research domain. Moreover, in order to assess the quality of these results, we calculated the similarity between all extracted pairs of synonyms following the distributional analysis paradigm using the cosine similarity measure and the log-likelihood ratio association measure [Dunning 1993] (see Equation 2.9) as the weighting scheme of the context features[17]. The distribution of the similarity measure for all noun synonyms is shown in Figure 6.8.

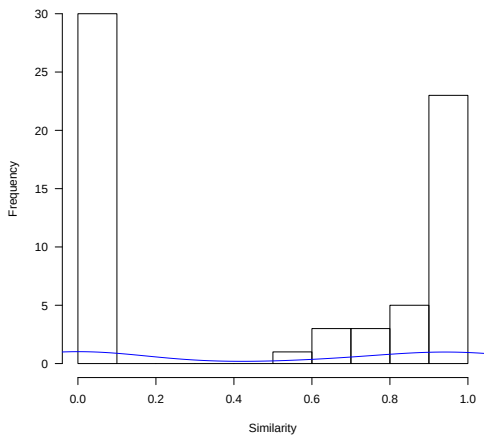

Figure 6.8: Synonym pairs similarity distribution.

[17]Only the global similarity was used.

The results clearly show that all extracted synonyms are highly correlated in terms of context. Indeed, most of the values are encountered in the last quantile and none have similarities lower than 0.5[18]. These results confirm our initial hypotheses that combining both local contexts and distributional representation studies may lead to the identification of highly semantically related words, which may form tightly accurate clusters to initiate prototype-based ontology learning.

6.1.3 Discovering Word Generality

One of the main problems to build prototype-based ontologies following the similarity-based method is to find ways to populate internal nodes of a hierarchical structures. [Pereira 1993] is the first to propose a divisive hierarchical clustering methodology based on a deterministic annealing procedure, which selects at each level of the hierarchy the best divisive clusters. However, this methodology implies that all the words in upper levels of the hierarchy are incrementally spread over the lower levels as illustrated in Figure 6.1. As a consequence, it is unclear whether the words in the parent nodes really subsume the ones in their children. For instance, in Figure 6.1, *weapon* clearly seems misclassified with regard to its parent node. Later, [Paaß 2004] present an interesting approach based on bayesian modeling of word occurrence by probabilistic hierarchical latent semantic classes. In particular, they do not assume a fixed number of classes and a fixed hierarchy, but allow the algorithm to select an appropriate topology. However, the results are not as satisfactory as expected. In fact, as in [Pereira 1993], most internal nodes contain words already embodied in the leaves. Moreover, their experiments show many odd classifications such as *club* (in leaf [22] in their paper [Paaß 2004]), which is subsumed by the node [65] which contains a list of football clubs *Hertha*, *Dortmund*, *Borussia*.

Instead, [Caraballo 1999] proposes a bottom-up clustering strategy, which automatically builds a hierarchical structure of nouns and then labels the internal nodes of the resulting tree with hypernyms from the nouns clustered underneath. For that purpose, he proposes the pattern-based approach already presented in Chapter 2. However, the pattern-based approach is language-dependent and sensitive to polysemy.

Another methodology is proposed in [Petersen 2004] based on the set-theoretical approach who state that *a good* [ontology] *representation avoids redundancy by capturing generalizations* where *a representation is said to be redundancy-free if every attribute and every object is stated exactly once*. This is indeed the optimal case. For that purpose, they propose a natural way of structuring these data by taking for every object the corresponding set of attributes and ordering these sets with respect to the superset relation. Thus, a top element can be added to

[18]The left-most bar shows that there were no sufficient statistics for 30 pairs of synonyms based on the existing corpora.

get a connected partial order if its attributes subsume all the attributes in the lower levels in terms of set inclusion. While [Petersen 2004] applies this paradigm through formal concept analysis (FCA) [Ganter 1999] based on noun feature vectors of their inflectional paradigms, [Cimiano 2005] propose to represent nouns as feature vectors of verb/prepositional phrase (PP)-complement, verb/object and verb/subject dependencies. In Figure 6.2, we illustrate the kind of acquired ontology. The main problem with the set-theoretical approach, as used so far, is that the nodes of the hierarchy contain attributes, which are not nouns and are definitely based on deep linguistic representation of words.

Based on our earlier work in [Dias 2006b], we propose to follow the two-step process as proposed in [Caraballo 1999] and focus on the second task[19] i.e. the extraction of the hyperonym/hyponym relations. To overcome the drawbacks proposed by previous works [Hearst 1992] [Riloff 1997] [Caraballo 1999] [Snow 2005] [Sang 2007] [Bollegala 2007] [Ohshima 2009], we present an unsupervised language-independent methodology, which takes into account asymmetry in language. Our method is based on the simple assumption that specific words tend to attract general words with more strength than the opposite. This idea is shared by [Michelbacher 2007] who state that *there is a tendency for a strong forward association from a specific term like adenocarcinoma to the more general term cancer, whereas the association from cancer to adenocarcinoma is weak.* In particular, we propose to automatically induce general/specific noun relationships from Web corpora frequency counts by running the TextRank algorithm [Mihalcea 2004] over a directed acyclic graph, where words are vertices and edges represent the asymmetric relation between vertices.

In [Michelbacher 2007], the authors clearly point at the importance of asymmetry in NLP. In particular, we deeply believe that asymmetry is a key factor to discover the degree of generality of terms. It is cognitively sensible to state that when someone hears about *mango*, he may induce the properties of a *fruit*. But, when hearing *fruit*, more common fruits will be likely to come into mind such as *apple* or *banana*. In this case, there exists an oriented association between *mango* and *fruit* (*mango* $\rightarrow$ *fruit*), which indicates that *mango* entails *fruit* while *fruit* does not attract *mango*. As a consequence, *fruit* is more likely to be a more general term than *mango*. Based on this assumption, asymmetric association measures are necessary to induce these associations. In particular, [Tan 2004] and [Pecina 2006] propose exhaustive lists of association measures. But the most important ones are reported in Chapter 2. Within this scope, we will compare the Braun-blanket (see Equation 2.25), the J-measure (see Equation 2.26), the Laplace (see Equation 2.27), the conviction (see Equation 2.28), the certainty factor (see Equation 2.29), the added value (see Equation 2.30) and the conditional probability (see Equation 2.17). So, the idea is that based on a given set of words and Web frequency counts

[19] Assuming that word clustering must be dealt apart from node labeling.

to evaluate their asymmetric relations, we are able to build a DAG by keeping the maximum asymmetric value between two words. We illustrate this situation in Figure 6.9 for the following is-a chain randomly extracted from WordNet: [level 1] [*mental disorder*, *mental disturbance*, *disturbance*, *psychological disorder*, *folie*], [level 2] [*conversion disorder*, *conversion reaction*, *conversion hysteria*] and [level 3] [*glove anesthesia*]. To obtain the DAG, we applied the conditional probability calculated from document hits retrieved by the Yahoo!^{TM} search[20] engine over the entire Web.

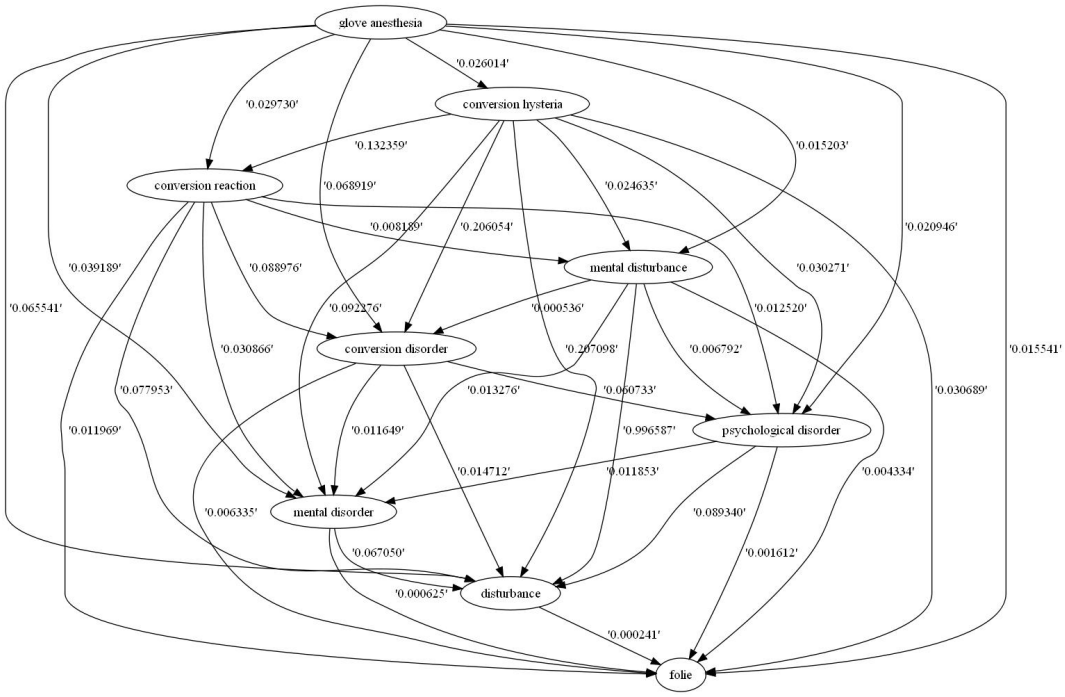

Figure 6.9: DAG built from the asymmetric matrix.

The second step of our methodology aims at defining a total order based on the created DAG. For that purpose, we propose to use a graph-based ranking algorithm, the TextRank [Mihalcea 2004]. Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph. The basic idea implemented by a graph-based ranking model is that of voting or recommendation. When one vertex links to another one, it is basically casting a vote for that other vertex. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex. Moreover, the importance of the vertex casting the vote determines how important the vote itself is, and this information is also taken into account by the ranking model. Hence, the score associated with a vertex is determined based on the votes that are cast for it, and the score of the vertices casting these votes.

[20] http://www.yahoo.com [19th *September*, 2010].

Based on these ideas, our intuition is that more general words will be more likely to have incoming associations as they will be associated to many specific words as opposed to specific words, which will have few incoming associations as illustrated in Figure 6.9. In particular, we experimented both the unweighted and weighted versions of the TextRank algorithm. For instance, the application of the weighted version of the TextRank algorithm for the DAG presented in Figure 6.9 would lead to the following total order by decreasing level of generality: [*folie*, *disturbance*, *mental disorder*, *psychological disorder*, *conversion disorder*, *mental disturbance*, *conversion reaction*, *conversion hysteria*, *glove anesthesia*].

In order to evaluate our methodology, we randomly extracted 800 seed synsets from WordNet for which we retrieved all their direct hypernym and hyponym synsets. For each seed synset, we then built the associated weighted and unweighted graphs DAG based on the asymmetric association measures referred to in the previous paragraphs and ran the TextRank algorithm to produce a general-specific ordered lists of terms. We first proposed an evaluation based on the number of order constraints that were respected by the generated order. WordNet can be defined as applying a set of constraints to words. Indeed, if word w is the hypernym of word x, we may represent this relation by the following constraint $w > x$, where $>$ is the order operator stating that w is more general than x. As a consequence, for each set of three synsets (the hypernym synset, the seed synset and the hyponym synset), a list of constraints can be established i.e. all words of the hypernym synset must be more general than all the words of the seed synset and the hyponym synset, and all the words of the seed synset must be more general than all the words in the hyponym synset. In order to evaluate our list of words ranked by the level of generality against the WordNet categorization, we just need to measure the proportion of constraints, which are respected as shown in Equation 6.3. This measure is called correctness.

$$correctness = \frac{\text{number of respected constraints}}{\text{number of constraints}}. \tag{6.3}$$

In Table 6.10, we present the results of the correctness for all seven asymmetric measures, both for the unweighted and weighted graphs.

Best results are obtained by the simple conditional probability and the Laplace measures reaching 65.69% correctness. However, the Braun-blanquet, the certainty factor and the added Value give results near the best ones. Only the J-measure and the conviction metrics seem to perform worst. It is also important to note that the difference between unweighted and weighted graphs is marginal, which clearly points at the fact that the topology of the graph is more important than its weighting. This is also confirmed by the fact that most of the asymmetric measures perform alike.

The evaluation can also be seen as a rank test between two ordered lists.

Measure	Type	Correctness
Braun-blanquet	unweighted	65.68%
	weighted	65.52%
J-measure	unweighted	60.00%
	weighted	60.34%
Conditional probability	unweighted	65.69%
	weighted	65.40%
Laplace	unweighted	65.69%
	weighted	65.69%
Conviction	unweighted	61.81%
	weighted	63.39%
Certainty factor	unweighted	65.59%
	weighted	63.76%
Added value	unweighted	65.61%
	weighted	64.90%
Frequency (baseline)	None	55.68%

Table 6.10: Average correctness.

Indeed, one way to evaluate the results is to compare the list of general/specific relationships encountered by the TextRank algorithm and the original list given by WordNet. However, we face one problem. WordNet does not give an order of generality inside synsets. In order to avoid this problem, we can order words in each synset by their estimated frequency given by WordNet as well as their frequency calculated by Web search hits. For that purpose, we propose to use the Spearman's rank correlation coefficient (ρ) [Spearman 1904]. The Spearman's ρ is a statistical coefficient that shows how much two random variables are correlated. It is defined in Equation 6.4 where d is the distance between every pair of words in the list ordered with TextRank and the reference list which is ordered according to WordNet or the Web and n is the number of pairs of ranked words. In particular, the Spearman's rank correlation coefficient is a number between -1 (no correlation at all) and 1 (very strong correlation) and results are given in Table 6.11.

$$\rho = \frac{6 \times \sum_{i=1}^{n} d_i^2}{n \times (n^2 - 1)}. \tag{6.4}$$

Similarly to what we evidenced for correctness, the J-measure and the conviction metrics are the measures, which less seem to map the correct order by evidencing low correlation scores. On the other hand, the conditional probability still gives the best results equally with the Laplace and Braun-blanquet asymmetric measures. It is interesting to note that in the case of the Web estimated list, the weighted graphs evidence much better results than the unweighted ones, although they do not show improved results compared to the WordNet list. On the one hand, these results show that our methodology is capable to map to WordNet lists as easily as Web lists. On the other hand, the fact that weighted graphs perform best shows

Measure	Type	ρ with WordNet	ρ with Web
Braun-blanquet	unweighted	0.38	0.30
	weighted	0.39	0.39
J-measure	unweighted	0.23	0.19
	weighted	0.27	0.27
Conditional probability	unweighted	0.38	0.30
	weighted	0.39	0.39
Laplace	unweighted	0.38	0.30
	weighted	0.38	0.38
Conviction	unweighted	0.30	0.22
	weighted	0.33	0.33
Certainty factor	unweighted	0.38	0.29
	weighted	0.35	0.35
Added value	unweighted	0.37	0.29
	weighted	0.38	0.38
Frequency (baseline)	None	0.14	0.14

Table 6.11: ρ value.

that the topology of the graph lacks in accuracy and needs the application of weights to counterpoint this lack. This conclusion could not be drawn from the first evaluation. Finally, it is clear that our methodology outperforms the baseline, which is built by only taking into account frequency of word occurrences.

Another way to evaluate the quality of the ordering of words is to apply hard clustering to the words weighted by their level of generality. By evidencing the quality of the mapping between three hard clusters generated automatically and the hypernym synset, the seed synset and the hyponym synset, we are able to measure the quality of our ranking. As a consequence, we propose to (1) perform a 3-means clustering over the list of ranked words, (2) classify the clusters by level of generality and (3) measure the precision, recall and F-measure of each cluster sorted by level of generality with the hypernym synset, the seed synset and the hyponym synset. For the first task, we used the implementation of the K-means algorithm of the NLTK toolkit[21]. For the second task the level of generality of each cluster is evaluated by the average level of generality of words inside the cluster. Finally, for the third task, the most general cluster and the hypernym synset are compared in terms of precision, recall and F-measure as respectively shown in Equation 6.5, 6.6 and 6.7. The same process is applied to the second most general cluster and the seed synset, and the third cluster and the hyponym synset.

$$precision = \frac{|\text{synset} \bigcap \text{cluster}|}{|\text{cluster}|}. \tag{6.5}$$

[21] http://nltk.sourceforge.net/ [19th *September*, 2010].

$$recall = \frac{|\text{synset} \bigcap \text{cluster}|}{|\text{synset}|}. \tag{6.6}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}. \tag{6.7}$$

In Table 6.12, we present the results of precision, recall and F-measure for both weighted and unweighted graphs for all the seven asymmetric measures. The best precision is obtained for the weighted graph with the conditional probability evidencing 47.62% and the best recall is also obtained by the conditional probability also for the weighted graph reaching 47.68%. Once again, the J-measure and the conviction metrics perform worst. These results also show that the weighting of the graph plays an important issue in our methodology. Indeed, most metrics perform better with weighted graphs in terms of F-measure.

Measure	Type	precision	recall	F-measure
Braun-blanquet	unweighted	46.61%	46.06%	46.33%
	weighted	47.60%	47.67%	47.64%
J-measure	unweighted	40.91%	40.86%	40.89%
	weighted	42.61%	43.71%	43.15%
Conditional probability	unweighted	46.54%	46.02%	46.28%
	weighted	47.62%	47.68%	47.65%
Laplace	unweighted	46.67%	46.11%	46.39%
	weighted	46.67%	46.11%	46.39%
Conviction	unweighted	42.13%	41.67%	41.90%
	weighted	43.62%	43.99%	43.80%
Certainty factor	unweighted	46.49%	46.52%	46.50%
	weighted	44.84%	45.85%	45.34%
Added value	unweighted	46.61%	46.59%	46.60%
	weighted	47.13%	47.27%	47.19%

Table 6.12: Clustering evaluation approach.

More experiments were made in [Dias 2008], which show that our approach performs better at higher levels of generality. For that purpose, we evaluated precision, recall and F-measure at each level of the hierarchy. We summarize these results in Table 6.13 for the conditional probability, which seems to be the best measure proposed so far. It is clear that while the precision drops with the level of generality, the recall inversely increases in both weighted and unweighted graphs.

This situation can easily be understood as most of the clusters created by the K-means present the same characteristics i.e. the upper level cluster usually has fewer words than the middle level cluster which in turn has fewer words than the last level cluster. We illustrate this situation based on the 3-means clustering of the word order obtained from Figure 6.9: [level 1] [*folie*], [level 2] [*mental*

Hierarchy level	Type	precision	recall	F-measure
Hypernym level	unweighted	59.20%	37.30%	45.77%
	weighted	58.71%	39.22%	47.03%
Seed level	unweighted	43.03%	37.67%	40.17%
	weighted	46.36%	33.02%	38.57%
Hyponym level	unweighted	37.38%	63.09%	46.95%
	weighted	37.79%	70.80%	49.27%

Table 6.13: Clustering evaluation approach by level of generality for the conditional probability.

disorder, *disturbance*] and [level 3] [*mental disturbance*, *psychological disorder*, *conversion disorder*, *conversion reaction*, *conversion hysteria*, *glove anesthesia*]. As a consequence, recall is artificially high for the hyponym level. But, on the opposite, the precision is high for higher levels of generality.

Although this situation does not match the structure of WordNet, it clearly evidences one of its lacuna. Indeed, the is-a relation in WordNet is not weighted. So, the level of generality/specificity between *transport* and *vehicle* is the same in WordNet as the one between *craft* and *hovercraft*. Of course, this situation does not reflect the reality between concepts. On the opposite, it seems that *mental disturbance* is a clear hyponym of *disturbance*, although they are treated as synonyms in WordNet. In fact, the order produced by our methodology seems to best embody linguistic phenomena reducing the level of generality/specificity when descending the taxonomy as illustrated in Figure 6.10.

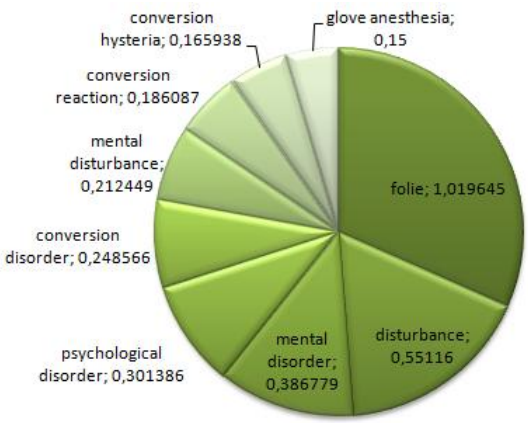

Figure 6.10: TextRank order values for the graph obtained in Figure 6.9.

However, this particular situation makes it difficult for clustering algorithms to reproduce the original clusters as hypernym level clusters will steadily group not much more than two words, while the others will gather more and more words as

the level of specificity increases. In fact, the main problem of this approach is the over-generation of edges towards the most general term. As a consequence, its rank is artificially high compared to the other words. Different strategies can be applied to overcome this situation. In particular, we proposed to log-transform the normalized order values, but with no significative impact on the final results as illustrated in Figure 6.11.

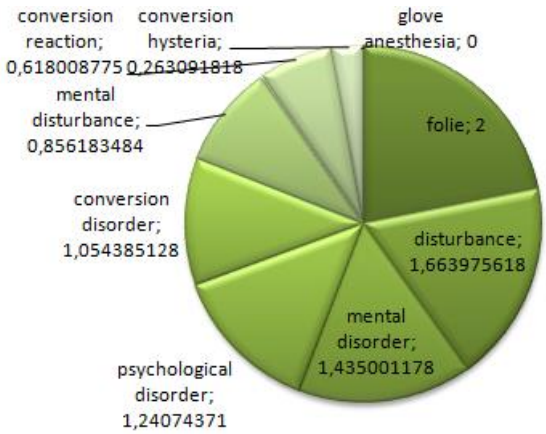

Figure 6.11: Log-transformation of normalized TextRank order values for the graph obtained in Figure 6.9.

Another solution is to compute recursively the TextRank algorithm by withdrawing at each iteration the hypernym level cluster, so that each level of analysis is evaluated as if it was at the (new) hypernym level, thus taking advantage of the good performance of our approach at upper levels of generality. This work is still under development evidencing some improvements, although less expressive than expected. Finally, we looked at the work proposed by [Sanderson 2000] who eliminate triangular transitive paths, while still keeping different paths between two words in the graph to obtain terminological ontologies. Although this technique seems to lead to better results, it is based on some assumption that longer paths should be kept. However, they only drop edges which form triangular equalities. We think that a more interesting idea can be proposed, which we explain in the following section.

Achieving the construction of prototype-based ontologies is still a long way from what we have achieved so far. However, we started to study the combination of semantic similarity with generality order to automatically build ontologies in [Bastos 2009] with some success within the medical domain. However, the complete process has not been reached so far. But, we will propose some of the ideas that we plan to carry out in the final section of this chapter, especially around multi-view clustering.

6.2 Terminological Ontologies

Based on our previous studies about the discovery of word order, it is evident to the light of Figure 6.9 that with a selected set of words sharing tight semantic relationships, it is possible to build terminological ontologies with some degree of accuracy. This idea was first evidenced in [Sanderson 1999] and then in [Sanderson 2000]. In particular, they proposed two associative methodologies, which present a document-based definition of subsumption according to which a certain word w_1 is more specific than a word w_2 if w_2 appears in most of the documents in which w_1 appears, but the opposite does not stand (i.e. $w_1 \rightarrow w_2$). [Sanderson 1999] are the first to extract salient words and phrases from documents and organize them hierarchically using the subsumption relation. They assume that a word w_2 subsumes a word w_1 if the documents in which w_1 occurs are a subset of the documents in which w_2 occurs constrained by $P(w_2|w_1) \geq 0.8$ and $P(w_1|w_2) < 1$. By gathering all remaining subsumption relations, they then build a lexical-semantic structure, which corresponds to a DAG. In [Sanderson 2000], the subsumption relation is relieved to the following expression $P(w_2|w_1) \geq P(w_1|t_2)$ and $P(w_2|w_1) > t$ where t[22] is a given threshold which must be tuned. Finally, all word pairs found to have a subsumption relationship are passed onto a transitivity module, which removes extraneous subsumption relationships in the way that transitivity is preferred over direct pathways, thus leading to a non-triangular DAG.

Although the associative models show interesting properties such as domain and language-independence as they only rely on word document distributions, they have been evidencing limitations so far. Indeed, all approaches tend to over-generate subsumption relations between words and as a consequence lead to the creation of unmanageable lexical-semantic structures when the number of terms increases as stated in [Fotzo 2004]. In order to overcome this drawback, we introduced in [Cleuziou 2010] a new approach based on the pretopology formalism, which builds a non-triangular DAG from pretopological operators over a proximity matrix. In particular, great emphasis is given to the topology of the object space compared to the works proposed by [Sanderson 1999] [Sanderson 2000] [Fotzo 2004]. Thus, from a given set of words from potentially different (sub-)domains and any related corpus (domain corpus or the Web), we assess the degree of generality/specificity of words as well as their semantic proximity using asymmetric similarity measures. So, from a proximity matrix reflecting the degree of attraction between words, we use the pretopology formalism to obtain, in a single step, an non-triangular DAG corresponding to the semantic structure of the (sub-)domains. A small example of an acquired terminological ontology is presented in Figure 6.12 based on the PubMed[23] corpus, where clustering of words (i.e. concepts) is allowed as proposed in [Fotzo 2004] but in a different manner.

[22] t is set to 0.8 in [Sanderson 2000].
[23] http://www.ncbi.nlm.nih.gov/pubmed [19th *September*, 2010].

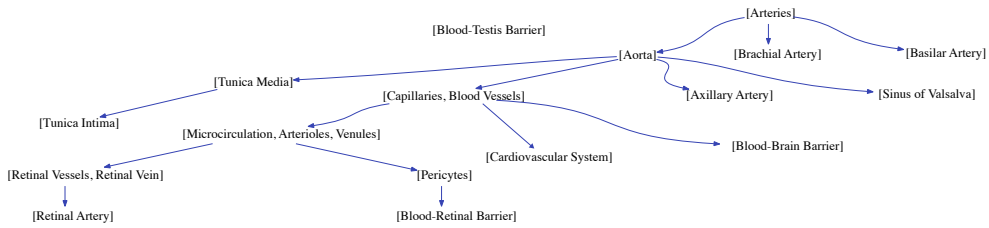

Figure 6.12: Terminological ontology acquired by pretopology.

The links between elements of a population can be modeled in several ways e.g. graph theory or a topological space. In topology, closure operators have been widely studied in algebra and computer science theory. However, the topological axioms and properties are too restrictive to model a space in concrete terms. Instead, pretopology models proximity in a more general way than topology and introduces interesting operations such as the pseudo-closure and significant mathematical objects like the maximal (minimal) closed subsets. We are aware that, according to the definition of pseudo-closure, we could work with graph theory. But in our case, using pretopology has several advantages. Our problematic can be viewed as a dynamic graph problem. However, when modeling dynamic graphs, one has to define a mathematical model and to develop algorithms that provide dynamic behaviors (e.g. adding or deleting edges or nodes). Designing algorithms to represent dynamics on a graph is a difficult task [Eppstein 1999]. Instead, pretopology generalizes graph and topology theories in one unified framework, where graphs, multigraphs, hypergraphs and topological spaces are particular cases of pretopological spaces [Belmandt 1993].

So, let's consider a non-empty set E and $\mathcal{P}(E)$, which designates all the subsets of E. A pretopological space is noted (E, a) where $a(.)$ is a pseudo-closure function denoted $a(.) : \mathcal{P}(E) \longrightarrow \mathcal{P}(E)$ which respects the two following conditions: $a(\emptyset) = \emptyset$ and $\forall A \in \mathcal{P}(E), A \subseteq a(A)$. It is important to notice that this function is not idempotent unlike in topology, where $a(a(A)) = a(A)$. As the definition of a pretopological space is very general, we can define such a space more precisely by a family of neighborhoods. First, let (E, a) be a pretopological space, we define the neighborhood of $x \in E$ as $\mathcal{N}(x)$ in Equation 6.8 where R is an asymmetric reflexive binary relation[24].

$$\mathcal{N}(x) = \{y \in E | xRy\}. \tag{6.8}$$

We then construct the pseudo-closure function based on a family of neighborhoods as defined in Equation 6.9.

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E | A \cap \mathcal{N}(x) \neq \emptyset\}. \tag{6.9}$$

As the pseudo-closure function is not idempotent, its successive applications over

[24] Within the scope of our research.

any subset $A \in \mathcal{P}(E)$ generates a dilatation process. When this process reaches a fixed point (i.e. $a^{k+1}(A) = a^k(A)$), the generated subset is called a closed subset noted F_A, which is formally defined in Equation 6.10 where $a^k(.)$ stands for the kth pseudo-closure (e.g. $a^2(A) = a(a(A))$).

$$\forall A \in \mathcal{P}(E), \exists\ k \in \mathbb{N} \text{ tel que } F_A = a^k(A). \tag{6.10}$$

Within this scope, we introduce the notion of elementary closed subset F_x to designate the closure of a singleton of E as well as the family of elementary closed subsets $\mathcal{F}_e(E, a)$ over the pretopological space (E, a) such as $\mathcal{F}e(E, a) = \{F_x | x \in E\}$. Finally, we call maximal elementary closed subsets of (E, a), noted $\mathcal{F}_M(E, a)$, the elementary closed subsets, which are not included in any other elementary closed subset i.e. $F_x \in \mathcal{F}_M(E, a) \Leftrightarrow \forall\ F_{x'} \in \mathcal{F}_e(E, a),\ F_x \not\subset F_{x'}$. In particular, a structure is induced by the elementary closed subsets and the maximal closed subsets can be seen as the less homogenous groups of E. The nature of these particular subsets is very interesting in terms of analysis of the space as we can consider an inclusion relation between them, leading to a structural analysis algorithm, the core algorithm to build our ontology. Our algorithm is a top-down version of the algorithm proposed by [Largeron 2002]. Instead of considering minimal closed subsets, we consider maximal ones as defined in algorithm 12.

Algorithm 12 The core algorithm DIVIDE(E, $R_{\alpha,\varepsilon}$).

Input: E and $R_{\alpha,\varepsilon}$
Output: A non-triangular DAG
Build $W = \{w_i \in E | F_{w_i} \text{is maximal}\}$
Organize W in concepts C_1, ..., C_k such that $\forall w_i, w_j \in C_k \Rightarrow F_{w_i} = F_{w_j}$
for each C_k **do**
  if $C_k \neq F_{C_k}$ **then**
    DIVIDE($F_{C_k} \backslash C_k$, $R_{\alpha,\varepsilon}$).
  end if
end for

As an example, we apply the core algorithm to a small data set where elementary closed subsets are listed in Table 6.14. The final result is illustrated in Figure 6.13 which corresponds to a non-triangular DAG.

$x \in E$	F_x	$x \in E$	F_x
1	{1,2,3,4,5,6}	7	{7,8}
2	{1,2,3,4,5,6}	8	{7,8}
3	{1,2,3,4,5,6}	9	{4,5,6,7,8,9}
4	{4,5,6}	10	{10}
5	{4,5,6}	11	{1,2,3,4,5,6,7,8,9,10,11}
6	{4,5,6}		

Table 6.14: Example of elementary closed subsets.

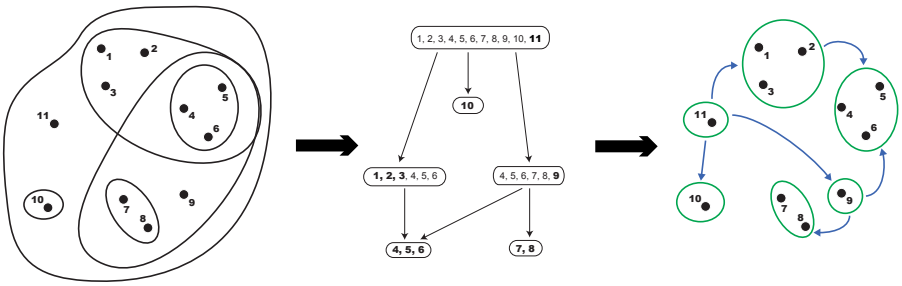

Figure 6.13: Resulting structure of the core algorithm based on Table 6.14.

For the construction of the lexical-semantic structure, we use the pretopological formalism, where E is the vocabulary combined with a matrix of asymmetric proximities P taking values in $[0, 1]$. This matrix reflects the semantic attractiveness of a word w_i oriented towards another word w_j. Thus, a high value of $P_{i,j}$ indicates that the semantic associated to word w_i strongly attracts the semantic of word w_j (i.e. $w_i \rightarrow w_j$), but the inverse is not necessarily true. In particular, we want to both treat the notion of semantic similarity and the degree of generality/specificity among all the words of the vocabulary. As such, we analyze the asymmetries existing between two terms ($P_{i,j}$ vs. $P_{j,i}$) when their values exceed ε, a threshold of interest fixed such as in [Sanderson 2000]. But unlike [Sanderson 2000], we deal with the notion of synonymy, also proposed to some extent in [Fotzo 2004]. So, a low asymmetry ($P_{i,j} \simeq P_{j,i}$) of a couple $\langle x_i, x_j \rangle$ may correspond to a relationship of synonymy, while a strong asymmetry ($P_{i,j} \gg P_{j,i}$) may symbolize a relationship of hyponymy (x_i hyponym of x_j). This observation leads us to propose, in Equation 6.11, the definition of an asymmetric binary reflexive relation R, which binds terms of E where $var(.)$ and $mean(.)$ respectively denote the variance and the mean of the values observed on the couple $\langle x_i, x_j \rangle$.

$$\begin{gathered} \forall w_i, w_j \in E, w_i R w_j \Leftrightarrow P_{i,j} \geq \varepsilon \wedge \\ \left(P_{i,j} \geq P_{j,i} \vee var(\{P_{i,j}, P_{j,i}\}) \leq \alpha . mean(\{P_{i,j}, P_{j,i}\})^2\right). \end{gathered} \tag{6.11}$$

Once the theoretical framework has been settled, we propose an evaluation of our methodology based on the well-known unified medical language system (UMLS) and the conditional probability[25] as primary information to build the asymmetric similarity matrix computed over PubMed as domain corpus and the entire Web as a huge general-purpose corpus. In particular, the evaluation environment is based on four distinct sub-domains of the UMLS, which were randomly selected i.e. the cardiovascular system (CS), the digestive system (DS), the respiratory system (RS) and the nervous system (NS). Each domain is represented by its own lexical-semantic structure present in the metathesaurus using the hypernym/hyponym relation as in Figure 6.14 for the referential CS sub-domain. As a consequence, given the domain vocabulary that appears in the UMLS structures, our objective is to evaluate

[25] Which proved to lead to improved results in [Dias 2008].

the ability of our methodology to automatically build from related texts a lexical-semantic structure that resembles to the UMLS one.

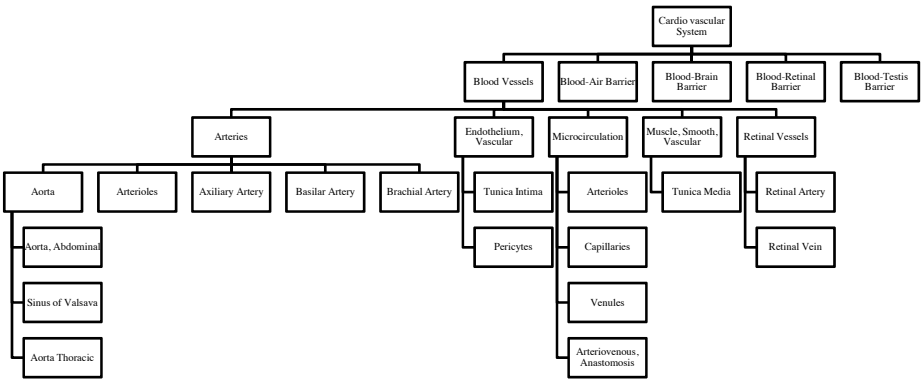

Figure 6.14: Baseline UMLS lexical structure of the CS sub-domain.

The results of ontology learning are complex structures that are intrinsically hard to evaluate [Dellschaft 2006]. Moreover, there are several possibilities of conceptualizations for one domain that might differ in their usefulness for different groups of people, but not in their soundness and justification. Based on these assumptions, we propose a general evaluation framework that consists in three different evaluations: (1) comparing lexical-semantic structures by means of structural measures (e.g. conceptual comparison level [Maedche 2002] and adapted metrics), (2) visualizing the general appearance of the obtained lexical-semantic structures as automatic metrics can be biased towards specific unsatisfiable topologies and (3) comparing the potential of the lexical-semantic structures in terms of word ontology-based similarity measures.

[Maedche 2002] proposed a way to compare ontologies at the conceptual level. Given a set of words X and two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ structuring X, the general principle is to compare for each entry $x \in X$ the matching between: the super/subconcepts of x in $\mathcal{O}_1$ and the super/subconcepts of x in $\mathcal{O}_2$. In our context, the acquired lexical structure $\mathcal{O}$ is a DAG $G(V, E)$ with V the set of nodes (concepts) and E the set of directed edges (relations): the subconcepts of $x \in V_x$ (V_x being the node/concept containing x), denoted as $Sub(x, \mathcal{O})$ are the nodes accessible from V_x. Similarly, the superconcepts of $x \in V_x$ ($Sup(x, \mathcal{O})$) are the nodes that can access to V_x. The comparison measure proposed in Equation 6.12 by [Maedche 2002] consists in using the Jaccard index J_1 as a matching quantifier where $S(x, \mathcal{O})$ denotes the conceptual environment (union of superconcepts and subconcepts) of x i.e. $S(x, \mathcal{O}) = Sub(x, \mathcal{O}) \cup Sup(x, \mathcal{O})$.

$$J_1(\mathcal{O}_1, \mathcal{O}_2) = \frac{\sum_{x \in X} \frac{|S(x,\mathcal{O}_1) \cap S(x,\mathcal{O}_2)|}{|S(x,\mathcal{O}_1) \cup S(x,\mathcal{O}_2)|}}{|X|}. \tag{6.12}$$

A perfect matching $J_1(\mathcal{O}_1, \mathcal{O}_2) = 1$ corresponds to structures where the term x has the same conceptual environment ($S(x, \mathcal{O}_1) = S(x, \mathcal{O}_2)$) in both structures. But, a perfect matching does not imply identical structures since the J_1 is not sensible to the hierarchical order of the concepts but only takes into account the unions of superconcepts and subconcepts. So, two structures with totally inverse orders have a perfect match. To avoid this inversion problem, we propose to consider separately subconcepts and superconcepts in the matching evaluation. For that purpose, we propose a new J_2 index in Equation 6.13, which is the (geometric) mean of two Jaccard indices for which a perfect match of 1 implies strictly identical structures.

$$J_2(\mathcal{O}_1, \mathcal{O}_2) = \frac{\sum_{x \in X} \left(\frac{|Sub(x,\mathcal{O}_1) \cap Sub(x,\mathcal{O}_2)|}{|Sub(x,\mathcal{O}_1) \cup Sub(x,\mathcal{O}_2)|} \cdot \frac{|Sup(x,\mathcal{O}_1) \cap Sup(x,\mathcal{O}_2)|}{|Sup(x,\mathcal{O}_1) \cup Sup(x,\mathcal{O}_2)|} \right)^{1/2}}{|X|}. \tag{6.13}$$

In order to perform our evaluation, we first analyzed the space of possible parameters $(\varepsilon, \alpha) \in [0, 1]^2$. In particular, we explored this space through discretization by deciles on the set of proximity values[26] and obtained best values for ε = 7th decile and α = 5th decile. Based on this parametrization, we performed our evaluation and summarize the results obtained by the pretopology model in Table 6.15.

Experiment	PubMed/Pretopology		Web/Pretopology	
Index	J_1	J_2	J_1	J_2
CS	0.20	0.18	0.29	0.00
DS	0.13	0.11	0.09	0.01
RS	0.14	0.11	0.09	0.00
NS	0.27	0.13	0.08	0.02
ALL	0.12	0.10	0.03	0.01

Table 6.15: Comparison between J1 and J2 indexes.

On the one hand, it is clear that J_1 always outperforms J_2. This can easily be explained as the proposed J_2 index restricts more the produced lexical-semantic structure than the J_1 index does. A clear example is illustrated with the value of 0.29 for J_1 compared to 0.00 for J_2 for the Web/Pretopology experiment. Indeed, by looking at the structure of the ontology built in this case, we had an inverted ontology such that the top concept of the ontology *cardiovascular system* was the deepest and only leaf of the ontology. As a consequence, J_2 is a more reliable metric, which takes into account such inversions.

On the other hand, the reader can legitimately ask himself what was our intention to produce an ontology directly from a non-domain corpus. The main reason to test whether we could obtain any reasonable results with the Web was based on the fact that in domain-specific corpora, experts usually do not mention

[26] Here, the conditional probability.

higher abstract levels in their texts. For instance, the term *cardiovascular system* is mentioned very few times in the PubMed corpus while more specific terms such as *aorta* are much more frequent. Based on this statement, the upper concept *cardiovascular system* was regularly misplaced in the taxonomy thus leading to ill-formed structures and lower J_2 indexes. But, counting word co-occurrences in corpora with different domains and different languages is dangerous as word sense disambiguation or language identification are not performed. For example, the term *colon* from the digestive system stands for *the part of the large intestine between the cecum and the rectum* but is also a MATLAB function and in French means a *colonist*. Moreover, when counting document hits over the Web, it is difficult to find two words, which do not show any relation. As a consequence, the proximity matrix is not sparse and finding good "general" ε and α may not be the best way to deal with the automatic learning of terminological ontologies[27]. This example clearly shows the incapacity of the Web to be a good base corpus to produce quality ontologies even with a pre-defined set of words, at least for the medical language.

Learning terminological ontologies is a goal of its own, but such structures are usually just a resource that should improve performance on NLP or IR tasks. Measuring improvements of ontology-supported approaches can best depict the gain for the application in focus. As a consequence, we proposed to evaluate the potential of the obtained terminological ontologies in the domain of word similarity based on the information-theoretic similarity measure proposed by [Lin 1998b] (see Equation 2.24). In particular, we compare both the similarities obtained with our structuring algorithm and the UMLS baseline structure, by means of the Kendall's correlation coefficient [Kendall 1938]. For two orders on the same set of values, the Kendall's correlation coefficient computes the number of concordant and discordant pairs and returns a correlation coefficient in $[-1, 1]$. We present the results of this evaluation in Table 6.16. We also performed the same test with the Spearman's rank correlation coefficient, which lead to similar results.

Experiment	PubMed/Pretopology	Web/Pretopology
CS	0.2670	0.3964
DS	0.2063	-0.0071
RS	0.2232	0.1478
NS	0.3843	0.0646
ALL	0.3700	0.0440

Table 6.16: Kendall's coefficient scores (ε = 7th decile, α = 5th decile).

The results obtained for the PubMed corpus are rather encouraging as high scores are obtained especially for the junction of the four sub-domains together with a Kendall's coefficient of 0.37 for the interval $[-1, 1]$. However, the results for the

[27] This will be discussed in the final section of this chapter.

Web corpus are rather low showing even negative results, except for the CS which surprisingly reaches a score of 0.39. In fact, although the structure of the ontology for the CS is ill-formed as confirmed by the J_2 index, it evidences an abnormal high statistical score, confirming the high value obtained by the J_1 index. A deeper analysis shows that terms can be erroneously subsumed by wrong upper concepts with high probability values. In this case, the Lin measure is wrongly over-evaluated. For example, in the generated lexical-semantic structure, *axillary arteries* and *arterioles* were subsumed by *retinal veins*, although *blood vessels* should be the most specific concept subsuming both sub-concepts.

Finally, we went on with this evaluation by performing clustering on Lin's similarity matrices in order to evaluate how the lexical-semantic structures would retrieve the expected reference knowledge. Figures 6.15 and 6.16 provide a visualization of the clustering obtained with Ward's agglomerative approach [Ward 1963] respectively with the UMLS and the pretopology model over the set of 128 terms.

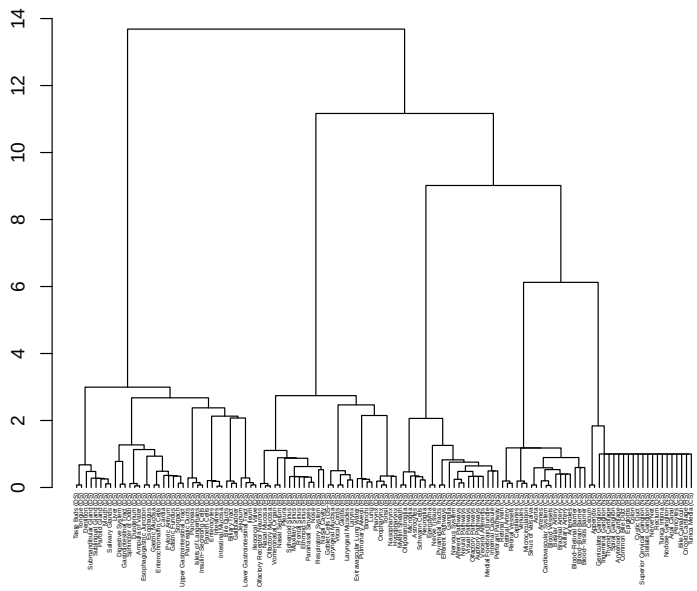

Figure 6.15: Dendrogram with the UMLS.

Furthermore, the confusion matrices in Table 6.17 confirm the strong matching between the clusters obtained with the pretopological-based similarities and the natural segmentation of the vocabulary i.e. the four sub-domains of the UMLS. In particular, the low purity observed for cluster 4 in the confusion matrices can

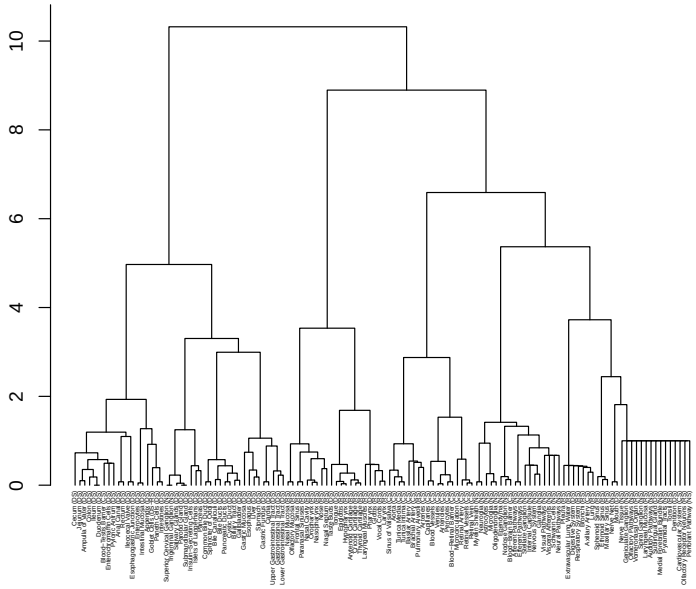

Figure 6.16: Dendrogram with the pretopology paradigm.

be explained by the disconnected terms that form a compact group (in terms of information loss) but heterogeneous in domain as shown in Figure 6.17 for the global generated terminological ontology.

Clusters	UMLS sub-domains			
	DS	RS	NS	CS
Cluster 1	38	0	0	0
Cluster 2	1	28	0	0
Cluster 3	0	0	21	0
Cluster 4	8	4	7	21

Clusters	UMLS sub-domains			
	DS	RS	NS	CS
Cluster 1	41	0	2	1
Cluster 2	3	19	0	0
Cluster 3	0	1	0	17
Cluster 4	4	12	26	3

Table 6.17: Confusion matrices obtained from the dendrograms with a 4-cluster cut: for the original UMLS structure (left) and for the new pretopological approach (right).

Although encouraging results have been achieved, this work is still in its beginning and many improvements are currently being studied. One of the main advantages of the pretopological framework is to look at the structure of the graph instead of working on the graph itself. As a consequence, recent experiments against the works

of [Sanderson 1999] and [Sanderson 2000] show that we are capable to leverage the problem of the over-generation of edges. On the one hand, great improvements are reached compared to [Sanderson 1999] for all evaluation indicators and not only for the number of generated edges. On the other hand, similar results are evidenced when compared to [Sanderson 2000] who apply the removal of transitive edges, although we are still decreasing the number of generated edges. Nevertheless, further evaluations need to be performed against [Sanderson 2000]. Indeed, the situation of the UMLS, as we used it in these experiments[28], is the ideal situation for [Sanderson 2000] as there are no synonyms in the vocabulary, but only words related by the is-a relation. This situation clearly limits the power of our model, which is capable to cluster similar concepts unlike [Sanderson 2000].

Besides, the pretopology formalism can easily model dynamic changes of parameters throughout the construction process. This is not the case with graph theory. As such, this may be a critical issue for the construction of terminological ontologies as we clearly understand that the (ε, α) space must be re-evaluated at each iteration of the core algorithm. Indeed, the deeper we go in the taxonomy, the higher ε should be and the lower α should be. In fact, at the beginning of the algorithm, we are more interested in highly connected words than in the confidence of the connections as we aim at encountering terms with highly populated closed subsets. However, when reaching the leaves of the structure, we are more interested in words connected with high confidence than ones with high "coverage" as we aim at encountering terms with high specificity. Moreover, the deeper we go in any taxonomy, the less likely we are able to find synonyms. These experiments are being carried out at the moment without results so far.

6.3 Future Work

As we mentioned earlier, we already started to study the automatic construction of prototype-based ontologies based on two paradigms in [Bastos 2009]: the level of semantic similarity between words and the level of generality between words. The idea is that there exist at least two dimensions in an ontology. In fact, this idea can be summarized as follows: two words are synonyms if they are semantically related as well as they share similar degree of generality. As a consequence, one can see the construction of an ontology in different ways. Most of the methodologies first deal with the semantic closeness and then encounter subsuming concepts. However, we can look at the problem the other way round. In [Bastos 2009], we proposed to build a prototype-based ontology by first partitioning the vocabulary by level of generality and then dealing with semantic closeness over the medical domain. Although the overall process has not been reached so far, it introduced a new trend in ontology learning. Moreover, the process is completely unsupervised, parameter-free and language-independent as the PoBOC algorithm [Cleuziou 2003]

[28] This data set was already available to initiate quick experiments.

is used as well as the evaluation of the level of generality as proposed in section 6.1.3.

However, it is clear that in order to obtain high accuracy ontologies, more reliable similarity measures are needed as well as a better understanding of the level of generality in texts. Within the context of semantic closeness, so far, we only presented models based on the vector space model as well as probabilistic models, although we know that the InfoSimba can lead to improved results. As a consequence, both the InfoSimba $IS(.,.)$ (see Equation 2.20) and the recursive InfoSimba $RIS_N(.,.)$ (see Equation 2.22) will be tested to evaluate to what extent they can improve accuracy of attributional similarity measures. Moreover, we aim at proposing a new attributional similarity measure based on the InfoSimba to capture the semantic closeness between words. One simple way to think at the problem is to say that two words x_i and x_j respectively represented by their feature vectors X_i and X_j are semantically related, within a given relation r (e.g. syntagmatic relation), if the co-occurring words within this relation are also semantically related with respect to the inverse relation r^{-1}. This situation is defined in Equation 6.14 where each W_{ij} corresponds to the attribute word at the jth position in the vector X_i.

$$IS_r(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.IS_{r^{-1}}(W_{ik}, W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.IS_{r^{-1}}(W_{ik}, W_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.IS_{r^{-1}}(W_{jk}, W_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.IS_{r^{-1}}(W_{ik}, W_{jl})\end{array}\right)}. \tag{6.14}$$

This situation can easily be illustrated for the case where nouns n_i and n_j are represented by their respective verb context vectors N_i and N_j i.e. $r = v$ and $r^{-1} = n$. In this case, Equation 6.14 would be represented as in Equation 6.15 where each V_{ij} corresponds to the verb attribute at the jth position in the feature vector N_i and to each V_{ij} is associated a noun feature vector $\vec{V}_{ij}$.

$$IS_v(N_i, N_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.IS_n(\vec{V}_{ik}, \vec{V}_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.IS_n(\vec{V}_{ik}, \vec{V}_{il})+\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.IS_n(\vec{V}_{jk}, \vec{V}_{jl})-\\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.IS_n(\vec{V}_{ik}, \vec{V}_{jl})\end{array}\right)}. \tag{6.15}$$

This way, we may be able to capture, at the same time, to what extent the context verbs attract related nouns as well as the context nouns attract related verbs. This idea can easily be adapted to the recursive version, which also may benefit the evaluation of semantic closeness.

With respect to the level of generality, further evaluations must also be carried out. In particular, we need to work on the topology of the generated DAG in the same way we did for the construction of terminological ontologies. Otherwise, clustering algorithms will endlessly produce hypernym clusters with few words and

hyponym clusters with many words. We may also test the asymmetric InfoSimba measure (see Equation 2.32) to see to what extent it may improve the correct word order. Finally, the bootstrapping methodology proposed in section 6.1.3 must deeply be studied to understand its capacity to retrieve reference clusters with high accuracy.

Once new similarity and generality measures will have been exhaustively tested, we will be able to propose a direct extension of the work initiated in [Bastos 2009]. The basic idea is to represent the semantic closeness and the level of generality/specificty as two different views and apply multi-view clustering algorithms such as the CoFKM proposed in [Cleuziou 2009]. Thus, we may extract highly relevant clusters, which may represent concepts as in WordNet (i.e. synsets) that we will then have to be linked together to build the final ontology.

The work proposed in [Dias 2010] introduced a new paradigm for the extraction of semantically related words by focusing on word interchangeability. We saw that this method is capable of selecting words within a tight semantic scope (e.g. synonymy, siblings, instance-of). However, these words still need to be grouped together following their semantic relationship. Within this context, we aim at studying different combinations of similarity and generality measures to see if we are able to isolate synonyms from hypernyms/hyponyms for instance. Another way to look at this problem is by applying multi-view clustering based on the semantic view and the generality view to understand to what extent highly relevant clusters can be found from TOEFL-like test cases. In this case, test cases will have to be constructed in a different manner i.e. joining all the interchangeable possibilities of a given word based on all paraphrase alignments. Indeed, at the moment, each paraphrase is treated independently from the other ones. As a consequence, we may fail to gather semantically related words in a single step.

Within the specific context of terminological ontologies, our pretopological space may be seen as an extension of a graph. However, we aimed at building a mathematically well-founded model, which allows to define different pretopological spaces that graph theory would not be able to cope with (or maybe with great difficulty). In particular, the notion of neighborhood $\mathcal{N}(x)$ may be operationally defined by different R relations, which may embody different types of graph structures. For instance, we may introduce dynamic changes of (ε, α) pairs along the structuralist process (dynamic graphs), repulsion and attraction restrictions (multi-graphs), dynamic inversions of subsumption relations (dynamic graphs) or multiple relations between words (multi-graphs). These are ideas for future works. But in a more recent future, we plan to extend our evaluation to further asymmetric similarity measures as defined in Chapter 2 and study the introduction of dynamic changes of (ε, α) pairs along the structuralist process ■

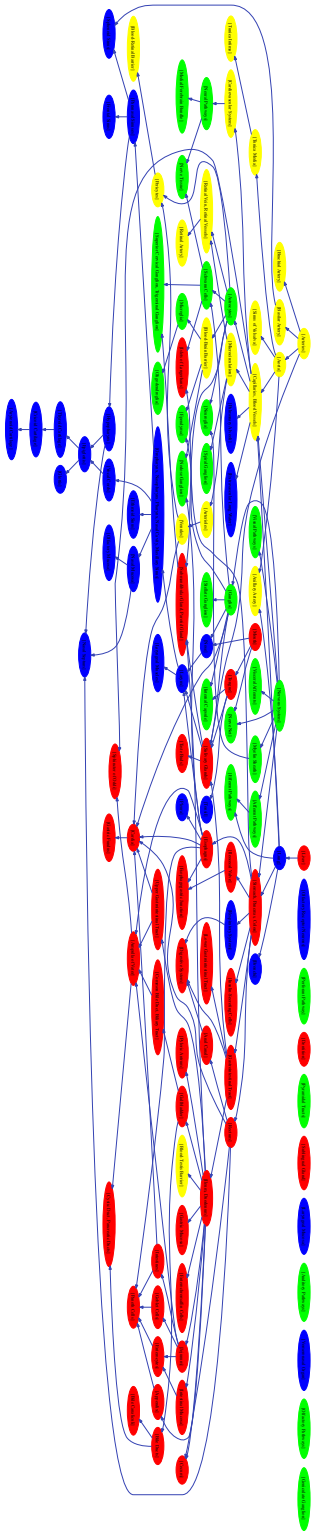

Figure 6.17: The terminological ontology generated by pretopology (128 terms).

Chapter 7

Subjectivity in Language

Contents

7.1 Related Work 176
7.2 Automatic Construction of Labeled Data Sets 179
7.3 Single-view Clustering 182
7.4 Multi-view Clustering 188
7.5 Future Work 191

The globalized access to the WWW via Weblogs or social networks raised new challenges to information access but also highlighted new problems. In particular, Web documents are no longer uniquely produced by legitimate sources (as if it was mostly the case in the beginning of the WWW) but can be by created by any well-intentioned or malicious users from earlier to older ages. The Web is clearly becoming the new support to express one's opinion without censorship. Although this is surely a great asset guaranteeing the freedom of expression of our modern societies, totally-free uncontrolled expression can lead to perverse effects such as the emission of false information or rumors, which can take enormous proportions as we can daily see in media such as Youtube^{TM}[1], Twitter^{TM}[2] or Facebook^{TM}[3]. In fact, gathering reliable information is becoming more and more difficult as more subjective information is produced everyday. Within the context of our research, we are especially interested in proposing objective information according to the users' needs, unlike most of other proposals, which focus on the evaluation of the public opinion. Indeed, learning subjectivity in text is an exciting research field, also called opinion mining, which offers enormous opportunities for various applications. In particular, it is likely to provide powerful functionalities for competitive analysis and marketing analysis through topic tracking and detection of unfavorable rumors. Although this issue is not to be put apart, it does not fulfil our principal goal. To our point of view, the user may have access to subjective contents but, in this case, he must be alerted that some of the contents he may read can be subjective. Indeed, it is important to guarantee the quality of information when this one is ruling most of our opinion. This way we present Information Digestion as a content selection process, where objective and subjective contents should be clearly highlighted.

[1] http://www.youtube.com [19th *September*, 2010].
[2] http://www.twitter.com [19th *September*, 2010].
[3] http://www.facebook.com [19th *September*, 2010].

Most works up-to-date tackle the simpler task of detecting polarity i.e. if a document is treating a given topic in a positive or negative way. Indeed, polarity is an important issue for topic tracking or opinion mining as companies, politics or even very important people (vip) search for positive recognition from the public opinion. However, our concern is different as we aim at offering reliable information to users rather than focusing on rumors, gossips or malicious contents. Of course, our discourse is purposely tendentious as negative and positive contents can be helpful for many tasks. One of the main difficulties in learning subjectivity in text is that the border between subjectivity and objectivity is fuzzy. Indeed, by definition, a negative or a positive statement is necessarily subjective. Only facts can be objective. In fact, this definition leaves very little space for objective contents. Let's take the famous problem of badly designed suspensions of the new launched Mercedes Class A in the nineteens. Saying that the Class A had a negative connotation at that time would certainly not be a subjective statement but rather a negative statement based on objective facts. To our point of view, polarity can only express subjectivity if not supported by facts. For example, public opinion may be subjective as it is usually based on misinformation, rumors or even manipulation. In this case, a negative statement can be subjective. So, saying that everything, which is negative is compulsory subjective does not stand to our point of view and negative and positive statements can be objective. Unfortunately, the definitions of the notions of subjectivity or sentiments in texts have received little attention, may be to the exception of [Boiy 2007]. As a consequence, we base our research on the most general paradigm, which aims at classifying texts whether they are objective or subjective, and let polarity be judged at a second stage of the process of opinion/sentiment learning. It is important to notice that less works have been proposed within this scope.

Another particularity of the studies proposed so far about sentiment analysis is the fact that most of them mainly propose in-domain supervised classifiers. However, within the context of real-world environments, sentiment analysis must be dealt across domains. Within this context, we have been working on the development of cross-domain subjectivity/objectivity classifiers based on low-level features (e.g. unigrams and digrams) and high-level features (e.g. level of abstraction of words) following two different paradigms: (1) single-view classification and (2) multi-view classification. It is important to notice that although efforts have been made to propose language-independent classifiers, such has not been reached yet, although on-going work is being developed to turn our classifiers as language-independent as possible as we discuss in the end of this section.

A few researches dealt with cross-domain subjectivity classification and all argued that it is hard to learn a domain-independent classifier. One possible approach is to train the classifier on a domain-mixed set of data instead of training it on one specific domain [Aue 2005]. Another possibility is to propose high-level features

which do not depend so much on topics such as part-of-speech statistics [Finn 2006]. Within this context, we proposed in [Lambov 2009] a methodology, which aims at classifying texts at the subjectivity level (i.e. subjective vs. objective) based on high-level semantic features, which can apply to different domains. In particular, we propose a new feature based on the level of abstraction of nouns, which improved accuracy across domains using both support vector machines (SVM) and linear discriminant analysis (LDA) classifiers. Another important contribution in this area (in order to reach language-independency) is the automatic construction of labeled data sets. Indeed, supervised classification techniques require large amounts of labeled training data. However, the acquisition of these data can be time-consuming and expensive. From that assumption, we proposed in [Pais 2007] to automatically produce learning data from Web resources. In particular, we compare Wikipedia and Weblogs texts to reference objective and subjective corpora and concluded that Wikipedia texts convey objective messages while Weblogs evoke subjective contents.

One may see high-level features and low-level features as two different views to look at the same issue. Based on this assumption, we proposed a multi-view environment based on high-level features and low-level features in [Lambov 2010]. In particular, we apply the co-training algorithm without agreement to obtain maximum performance over two views (high-level and low-level features). Experimental results showed that our approach outperforms the stochastic agreement regularization (SAR) algorithm, which is the reference algorithm in the domain [Ganchev 2008]. However, while we use linguistic resources to build our high-level feature vectors, the SAR algorithm only uses unigrams and digrams as possible views. So, it is likely that SAR may achieve comparable results to the co-training algorithm with the same views. For that purpose, we are actually working on the adaptation of the SAR to more complex features and more than two views in order to present an exhaustive and impartial evaluation.

One of the drawbacks of our approach is that it relies too much on existing linguistic resources. As such, we do not comply to our initial ideas of language independence. As a consequence, we have been working lately on some ideas to build linguistic resources automatically. In particular, we proposed in [Rodrigues 2009] a new language-independent methodology to extract subjective lexicons. Moreover, as we will see in this chapter, evaluating the level of abstraction of texts leads to improved results for classification. So far, we used WordNet to understand the level of generality of texts. However, we plan to test our unsupervised methodology proposed in Chapter 6 in section 6.1.3. All these issues will be discussed in the final section of this chapter.

7.1 Related Work

Subjectivity and polarity of language have been investigated at some length. In particular, many features have been used to characterize opinionated texts at different levels of analysis: word, sentence and document. At word level, [Hatzivassiloglou 1997] were the first to distinguish between positive and negative adjectives through the hypothesis that adjectives joined by the conjunction *and* tend to be similar and dissimilar if joined by *but*. For that purpose, they used a four-step supervised learning algorithm to infer the semantic orientation of adjectives from constraints on these conjunctions. Their algorithm classified adjectives with accuracies ranging from 78% to 92% depending on the amount of available training data. In particular, they also based their study on the hypothesis that positive adjectives are more frequent than negative ones. Later, [Turney 2002] calculated the semantic orientation of a phrase based on the mutual information (see Equation 2.1) between the given phrase and the word *excellent* minus the mutual information with the word *poor*. Finally, [Esuli 2005] presented a semi-supervised methodology to identify the semantic orientation of words using their gloss definitions from online dictionaries. In particular, they provided a manually-composed set of words with positive and negative connotation and expanded it with the synonyms of the polar words. Finally, they used this expanded data set to predict the polarity of words on the basis of their glosses.

At sentence level, [Wiebe 1999] are the first to perform a statistical analysis of positive/negative sentences, finding that adjectives are statistically significantly and positively correlated with subjective sentences on the basis of the log-likelihood ratio test statistic (see Equation 2.9). Indeed, they found that the probability of a sentence being subjective was 56% by simply knowing that the sentence contained at least one adjective even though there were more objective than subjective sentences in the corpus. Later, [Yu 2003] proposed a technique based on the subjectivity scoring proposed by [Turney 2002] being a sentence positive (resp. negative) if most of the adjectives, adverbs, nouns and verbs in the sentence were positive (resp. negative). This work is closely related to the one proposed by [Turney 2002] in which a review is classified as recommended if the average semantic orientation of its phrases is positive. His algorithm achieved an average accuracy of 74% when evaluated on 410 reviews from Epinions[4], sampled from four different domains (reviews of automobiles, banks, movies and travel destinations). In particular, the accuracy ranged from 84% for automobile reviews to 66% for movie reviews.

At document level, [Pang 2002] first showed that the unigram model with SVM reached best results compared to more complex models in the domain of movie reviews. [Yu 2003] proposed a similar experiment based on multiple naive bayes classifiers, but more linguistically motivated i.e. with a significantly larger

[4] http://www.epinions.com/ [19th *September*, 2010].

set of seed words and orientation words from different syntactic classes other than adjectives (nouns, verbs and adverbs). Results from a large collection of news stories and a human evaluation of 400 sentences reported respectable performance in detecting opinions and classifying them at the sentence level as positive, negative, or neutral up to 91% accuracy. Later, [Pang 2004] proposed a unigram SVM classifier, which was only applied to the subjective portions of a document. In particular, a sentence cut-based classifier is used to identify subjective parts in texts, which are then used for text classification. Their results show that the created subjectivity extracts accurately represent the sentiment information of the originating documents in a much more compact form. In fact, depending on the choice of the polarity classifier, they were able to achieve highly statistically significant improvements (from 82.8% to 86.4%) for the polarity classification task while retaining only 60% of the words of the reviews. In parallel, [Wiebe 2004] is certainly the first work to propose a study of subjectivity in text and not just polarity. specifically, they derived a variety of subjectivity cues (frequencies of unique words in subjective-element data, collocations with one or more positions filled in by a unique word and distributional similarity of adjectives and verbs) from corpora and demonstrated their effectiveness on classification tasks. Moreover, they determined a relationship between low frequency terms and subjectivity and found that their method to extract subjective n-grams was enhanced by examining those that occur with unique terms. Finally, [Chesley 2006] presented a method using verb class information, and the Wikipedia dictionary to determine the polarity of adjectives. They used verb-class information in the sentiment classification task, since exploiting lexical information contained in verbs showed to be a successful technique for classifying documents, unlike previous research.

Unfortunately, all studies presented so far learn domain-dependent classifiers or study subjectivity in a single domain, maybe to the exception of [Wiebe 2004]. However, within the context of real-world environments, especially the Web, sentiment analysis must be dealt across domains. But, as mentioned in [Boiy 2007] [Lambov 2009], most research show difficulties in crossing domains. In particular, we showed in [Lambov 2009] that accuracy losses of 35% can be reached when evaluating a unigram domain-dependent SVM classifier against a cross-domain data set. Indeed, sentiment is orthogonal to topic and sentiment classification is clearly more difficult than topic classification. One possible approach to tackle cross-domain classification is to train a classifier on a domain-mixed data set instead of one specific domain. This idea is proposed by [Aue 2005] to learn a polarity classifier. Another possibility is to propose high-level features, which do not depend so much on topics such as part-of-speech statistics as proposed by [Finn 2006]. In this case, the part-of-speech representation does not reflect the topic of the document, but rather the type of text used in the document. So, just by looking at part-of-speech statistics, improved results were obtained comparatively to unigram models (low-level models) when trying to cross domains for polarity classifiers. Recently, an interesting language-independent methodology has been proposed to

leverage the problem of cross-domain classifiers. [Blitzer 2007] proposed to find anchor terms, which cross domains and evaluated the correlation between those words and words, which were specific to the domain. In this case, pivot features were discovered based on domain mutual information to relate training and target domains. Then, the overall approach extended the structural correspondence learning algorithm (SCL) to polarity classification. As a consequence, they identified a measure of domain similarity that correlated well with the potential for adaptation of a classifier from one domain to another. Within that context, best polarity results across domains reached an excellent score of 82.1% accuracy[5].

Finally, over the past few years, semi-supervised and multi-view learning proposals have emerged. [Ganchev 2008] proposed a co-regularization framework to learn across multiple related tasks with different output spaces. They presented a new algorithm for probabilistic multi-view learning, which uses the idea of stochastic agreement between views as regularization. Their algorithm called SAR works on structured and unstructured problems and generalizes to partial agreement scenarios. For the full agreement case, their algorithm minimizes the Bhattacharyya distance between the models of each of two views. For the specific case of cross-domain polarity classification, they obtained a maximum accuracy of 82.8% over the same data set used by [Blitzer 2007] combining the maximum entropy classifier and two randomly created views of unigrams. More recently, [Wan 2009] proposed a co-training approach to improve the classification accuracy of polarity identification of Chinese product reviews. First, machine translation services are used to translate English training reviews into Chinese reviews and also translate Chinese test reviews and additional unlabeled reviews into English reviews. Then, the classification problem can be viewed as two independent views: Chinese view with only Chinese features and English view with only English features. They then used the co-training approach to make full use of the two redundant views of features. A SVM classifier was adopted as a basic classifier in the proposed approach. Experimental results showed that their methodology outperforms baseline inductive classifiers and more advanced transductive classifiers.

Since sentiment in different domains can be expressed in different ways [Boiy 2007], supervised classification techniques require large amounts of labeled training data. However, the acquisition of these labeled data is time-consuming and expensive. Moreover, most of the studies proposed in this section are based on relatively small data sets, which do not guarantee reliability. For that purpose, we first proposed to automatically produce learning data from Web resources. To do so, we proposed to compare Wikipedia texts and Weblogs to reference objective and subjective corpora and managed to prove to some extent that texts from Wikipedia embody objectivity while Weblogs mainly convey subjective contents [Pais 2007]. Once huge quantities of training examples were available, we proposed in [Lambov 2009]

[5] Although the text data set is based on reviews for each domain, which may bias the results.

a single-view learning approach to cross-domain sentiment[6] classification based on state-of-the-art high-level characteristics that have been used to classify opinionated texts. Within this context, we proposed a new feature to classify sentiment texts based on the level of abstraction of nouns. An exhaustive evaluation showed that (1) the level of abstraction of nouns is a strong clue to identify subjective texts across domains, (2) high-level features allow to learn enhanced cross-domain models and (3) automatically labeled data sets extracted from Wikipedia and Weblogs give rise, on average, to the best cross-domain classifiers reaching accuracy levels of 74.5% with the LDA classifier. Finally, in [Lambov 2010], we proposed to combine high-level features (e.g. level of affective words, level of abstraction of nouns) and low-level features (e.g. unigrams and digrams) to learn models of subjectivity, which may apply to different domains. For that purpose, we proposed a new scheme based on the classical co-training algorithm [Blum 1998] over two views and joined two different classifiers LDA and SVM to maximize the optimality of the approach. Results showed that accuracy of 88% can be reached.

7.2 Automatic Construction of Labeled Data Sets

Up to now, the supervised and semi-supervised methodologies, which have been proposed to learn subjectivity in texts are based on a limited set of learning examples almost exclusively for the English[7] language. In fact, as most learning data sets are manually gathered and labeled, they are usually small and do not cover most of language subjectivity phenomena. To leverage this task, some authors [Wiebe 2004] [Chesley 2006] proposed to analyze texts, which should describe opinions (e.g. letters to the editor, news columns, reviews, political Weblogs) and facts (e.g. world, national and local news) by definition. As such, only manual post-editing is necessary. Although this issue is interesting, it clearly lacks theoretical background. Indeed, the characterization of subjectivity and objectivity is defined upon common sense beliefs, which should be at least experimentally supported. So, we followed this idea, but based on a more theoretical background. As such, we proposed in [Pais 2007] the assumption, which states that texts from Wikipedia should embody objectivity while Weblogs should convey subjective contents. But, somehow, this needs to be proved. As a consequence, we proposed to compare Wikipedia texts and Weblogs to a reference sentiment (subjective/objective) corpus, the subjectivity v1.0 corpus[8] built by [Pang 2004], to confirm our assumptions. For that purpose, we proposed an exhaustive evaluation based on (1) the Rocchio classification method [Rocchio 1971] for different part-of-speech tag levels and (2) language modeling.

In order to have a more complete view about subjectivity and objectivity in language, we first gathered large quantities of texts from Weblogs and Wikipedia.

[6] And not polarity as in the majority of the works proposed so far.

[7] With a few exceptions as proposed in [Mihalcea 2007] [Banea 2008] for Romanian and Spanish, and [Wan 2009] for Chinese.

[8] http://www.cs.cornell.edu/people/Pabo/movie-review-data [11th *September*, 2010].

On the one hand, we downloaded the English static version of Wikipedia in XML format[9] and extracted all the sentences giving rise to a corpus of 40 Gb. On the other hand, we crawled Weblogs domains from different themes that can be found in [Pais 2007] and gathered 12 Gb of Weblogs text sentences. Comparatively, the subjectivity v1.0 corpus contains 5.000 objective sentences collected from movie reviews and 5.000 subjective sentences gathered from customer review snippets. So, in order to afford an impartial comparison, we used a random sample of both Wikipedia and Weblogs corpora, maintaining statistical significance. The corpora are summarized in Table 7.1.

Corpora	Wikipedia	Weblogs	Objectivity	Subjectivity
Unique Sentences	411.293	984.682	5.000	5.000
Unique Words	224.112	79.680	15.065	14.146

Table 7.1: Dimensions of the corpora.

In order to test our initial assumptions, we first applied the simple Rocchio relevance feedback algorithm adapted for text classification [Rocchio 1971]. Rocchio relevance feedback algorithm is one of the most popular and widely used learning methods in information retrieval. It uses standard $tf.idf$ weighted vectors to represent text documents. For each category, it computes a prototype vector by summing the vectors of the training documents of the same category. Finally, the closest prototype vector in terms of cosine similarity measure (see Equation 2.14) with any given text classifies the text. Within the context of our work, instead of the $tf.idf$, we used the $tf.isf$ (already mentioned in Chapter 5, section 5.1.2), as we deal with classified sentences and not documents and performed an evaluation at different part-of-speech tag levels. In Table 7.2, we present the results where the test vector is the set of Wikipedia sentences and the trained vectors are the subjective and objective sentences from the subjectivity v1.0 corpus. The results confirmed our initial assumption that texts from Wikipedia convey objective contents, although the role of verbs seems less clear with respect to subjectivity as opposed to what is exposed in [Chesley 2006]. Similarly, we performed the same experiment where the test vector is the set of Weblogs sentences and the trained vectors are the subjective and objective sentences from the subjectivity v1.0 corpus. The results are presented in Table 7.3 and clearly show that at any part-of-speech level, Weblogs embody subjectivity. Finally, in order to confirm the assumptions formulated in [Wiebe 2004] [Chesley 2006] without any support, we proposed to test to what extent news article convey an objective language. For that purpose, we extracted a statistically significant random sample of the Reuters corpus[10] and performed classification with the Rocchio algorithm. The results are shown in Table 7.4 and confirm common sense judgments made by [Wiebe 2004] and [Chesley 2006], although verbs still seem to cause some confusions.

[9] http://download.wikimedia.org/enwiki/20071018 [6th *September*, 2007].
[10] http://trec.nist.gov/data/reuters/reuters.html [11th *September*, 2010].

Part-of-speech level	Subjective	Objective	Class
All Words	0.76	0.79	Objective
All ADJ	0.54	0.61	Objective
All V	0.71	0.67	Subjective
All N	0.66	0.69	Objective
All ADJ + All V	0.65	0.66	Objective
All ADJ + All N	0.65	0.68	Objective
All N + All V	0.70	0.69	Subjective
All ADJ + All N + All V	0.68	0.69	Objective

Table 7.2: Results with the Wikipedia test data set.

Part-of-speech level	Subjective	Objective	Class
All Words	0.60	0.56	Subjective
All ADJ	0.52	0.49	Subjective
All V	0.53	0.48	Subjective
All N	0.47	0.43	Subjective
All ADJ + All V	0.49	0.48	Subjective
All ADJ + All N	0.48	0.44	Subjective
All N + All V	0.50	0.45	Subjective
All ADJ + All N + All V	0.47	0.46	Subjective

Table 7.3: Results with the Weblogs test data set.

Part-of-speech level	Subjective	Objective	Class
All Words	0.64	0.68	Objective
All ADJ	0.30	0.40	Objective
All V	0.38	0.37	Subjective
All N	0.34	0.47	Objective
All ADJ + All V	0.36	0.38	Objective
All ADJ + All N	0.35	0.49	Objective
All N + All V	0.36	0.47	Objective
All ADJ + All N + All V	0.37	0.47	Objective

Table 7.4: Results with the Reuters test data set.

In spite of encouraging classifications, the values of the cosine similarity measure within the same morphological level between the trained and the test vectors are usually very close. This does not give much confidence about the results. For that purpose, we proposed another methodology based on language modeling. The basic idea is that objective and subjective languages are intrinsically different. As a consequence, if we build a language model based on Weblogs, the subjective part of the subjectivity v1.0 corpus should be more probable than the objective part, and vice and versa[11]. This probability is transformed into perplexity (Px) and entropy

[11] As language models need large quantities of texts, they are built from the Wikipedia, Weblogs

(H) measures within the CMU-Toolkit. The results of this experiment are given in Table 7.5 for a trigram language model.

	Wikipedia	Weblogs	Reuters
Objective	$Px = 691.27$	$Px = 2027.06$	$Px = 1104.03$
	$H = 9.43$	$H = 10.99$	$H = 10.11$
Subjective	$Px = 880.67$	$Px = 1991.09$	$Px = 1226.34$
	$H = 9.75$	$H = 10.96$	$H = 10.26$

Table 7.5: Results obtained with the Language Model.

To summarize the results in Table 7.5, the trained model Wikipedia shows lower perplexity and entropy for the objective sentences than for the subjective sentences. The opposite happens when using the trained model Weblogs. In that case, lower perplexity and entropy are shown for the subjective sentences than for the objective sentences. Once again, our assumptions are confirmed as objective (resp. subjective) sentences are intrinsically closer to the Wikipedia (resp. Weblogs) model than the subjective (resp. objective) sentences are. Indeed, the lower the perplexity and the entropy are, the closer to the model the sentences are. Finally, the trained model Reuters shows similar behavior as the Wikipedia model, thus validating the common sense judgments made by [Wiebe 2004] and [Chesley 2006].

Thanks to this analysis, we are now able to automatically build large data sets of learning examples based on common sense judgments. Moreover, Wikipedia texts exist in many languages as well as Weblogs, which emerge everyday all over the world. As a consequence, multilingual sentiment data sets can easily be compiled and new studies may rise to reach multilingual sentiment analysis based on corpus analysis as we show in [Rodrigues 2009] where we learn a subjective lexicon for the Portuguese language, without linguistic tools or resources as in [Mihalcea 2007] and [Banea 2008].

In the following sections, in which we introduce new single-view and multi-view supervised strategies for sentiment classification, we will use the new automatically labeled data set based on Wikipedia and Weblogs texts and compare results with manually tagged corpora.

7.3 Single-view Clustering

Following the ideas proposed in [Finn 2006] to reach cross-domain sentiment analysis, we proposed in [Lambov 2009] to first study the characteristics of language, which convey subjective contents in documents. According to [Boiy 2007], subjectivity can be expressed in different ways: evaluation (positive or negative), potency (powerful or unpowerful), proximity (near or far), specificity (clear or

and Reuters corpora in our experiments. In fact, evaluation is done the other way round.

vague), certainty (confident or doubtful) and identifiers (more or less) as well as direct expressions, elements of actions and remarks. Based on these assumptions, our methodology aims at classifying texts at the subjectivity level (i.e. subjective vs. objective) based on high-level features as opposed to low-level features (e.g. unigrams, digrams), which can apply to different domains. Within this scope, we identified eight characteristics based on well-known linguistic resources.

First, sentiment expressions mainly depend on some words, which can express subjective sentiment orientation. Within this scope, [Strapparava 2008] used a set of words extracted from the WordNet Affect lexicon proposed in [Strapparava 2004]. For example, the WordNet Affect lexicon encodes the fact that both *horror* and *hysteria* express negative fear as well as *enthusiastic* denotes positive emotion and *glad* refers to joy. As they are based on the classical WordNet [Miller 1990], most of these words are occurrences of general language. As such, the level of affective words in texts may successfully express subjectivity across domains.

Second, [Hatzivassiloglou 2000] considered two features for the identification of opinionated sentences in texts: (1) semantic orientation adjectives, which represent an evaluative characterization of the deviation of a word from the norm of its semantic group and (2) dynamic adjectives, which characterize the ability of a word to express a property in varying degrees. In particular, they noted that all sets involving dynamic adjectives and adjectives with positive or negative polarity are better predictors of subjective sentences than the class of adjectives as a whole. As a consequence, we used the proportion of these adjectives in texts to characterize their subjectivity level. For that purpose, we utilized the set of all dynamic adjectives manually identified in [Hatzivassiloglou 2000] and the set of semantic orientation labels assigned as in [Hatzivassiloglou 1997].

Third, [Chesley 2006] presented a method using verb class information. Their verb classes express objectivity and polarity. To obtain relevant verb classes, they used InfoXtract [Srihari 2006], an automatic text analyzer which groups verbs according to classes that often correspond to their polarity. As InfoXtract is not freely available, we reproduced their methodology by using the classification of verbs available in Levin's English verb classes and alternations [Levin 1993]. So, we proposed to evaluate the proportion of each corresponding class of verbs (i.e. conjecture verbs, marvel verbs, see verbs and positive verbs) as an interesting clue to identify subjectivity in texts.

Finally, there exists linguistic evidence that the overall level of generality is a characteristic of opinionated texts, i.e. subjectivity is usually expressed in more abstract terms than objectivity [Osgood 1971] [Boiy 2007]. Indeed, descriptive texts tend to be more precise and more objective and as a consequence more specific. In particular, [Boiy 2007] define specificity as *the extent to which a conceptualized object is referred to by name in a direct and clear way; or is only*

implied, suggested, alluded to, generalized, or otherwise hinted at. In other words, a word is abstract when it has few distinctive features and few attributes that can be pictured in the mind [Osgood 1971]. One way of measuring the abstractness of a word is by using the hypernym relation in WordNet. In particular, an hypernym metric can easily be defined as the path length to the root via the hypernym relation. So, a word having more hypernym levels is more likely to be concrete than one with fewer levels. As a consequence, the average hypernym level of all nouns in a given text may provide a good clue for sentiment classification.

Before performing any classification task, it is useful to evaluate to what extent the given high-level features are discriminative and allow representing distinctively the data sets in the given space of characteristics. For that purpose, feature selection and visualization techniques are usually used. To perform these experiments, we used three manually annotated standard corpora and the corpus automatically built from Wikipedia texts and Weblogs mentioned in section 7.2. One of the first corpora to be compiled for the sake of sentiment learning is the multi-perspective question answering opinion corpus $\{MPQA\}$[12] [Stoyanov 2004]. It contains 10.657 sentences in 535 documents from the world press on a variety of topics. All documents in the collection are marked with expression-level opinion annotations. The documents are from 187 different news sources in a variety of countries and date from June 2001 to May 2002. The corpus was collected and manually annotated with respect to subjectivity as part of the summer 2002 NRRC workshop on multi-perspective question answering. Based on the work done by [Pang 2004] who proposed to classify texts based only on their subjective and objective parts, we built a corpus of 100 objective texts and 100 subjective texts by randomly selecting sentences containing only subjective or objective phrases. This represents the "ideal" case where all the sentences in texts are either subjective or objective. The second corpus is built from the well-known subjectivity v1.0 corpus $\{RIMDB\}$. Similarly to what we did for the $\{MPQA\}$, we built a corpus of 100 objective texts and 100 subjective texts containing randomly selected 50 sentences assigned as subjective or objective exclusively. The third corpus $\{CHES\}$ was developed by [Chesley 2006] who manually annotated a data set of objective and subjective documents. It contains 496 subjective and 580 objective documents and is the only corpus to be manually tagged at the text level for subjectivity and not just polarity. In particular, objective feeds are from sites providing content such as world, national or local news focusing on different topics such as health, science, business, and technology. Comparatively, the subjective feeds include contents from newspaper columns, letters to the editor, reviews and political Weblogs. Finally, the fourth corpus $\{WBLOG\}$ is a compilation of 100 texts randomly selected from Wikipedia texts extracted from the corpus presented in section 7.2 and its subjective counterpart is a set of 100 texts randomly selected from the Weblogs of the same original corpus.

[12] http://www.cs.pitt.edu/mpqa/ [12th *September*, 2010].

In order to perform feature selection, we proposed to apply the Wilcoxon rank-sum test [Wilcoxon 1945]. In particular, the two sample Wilcoxon rank-sum test with one-sided alternative is carried out for all experiments. The samples contain 200 values for each one of the sets (100 objective texts and 100 subjective documents) and the exact p-value is computed. The exact 95% confidence interval for the difference of the location parameters of each of the sets is obtained by the algorithm described in [Bauer 1972] for which the Hodges-Lehmann estimator is employed. So, for each of the sets, we are 95% confident that the interval contains the actual difference between the feature values of subjective and objective texts. The results are shown in Table 7.6.

	$\{MPQA\}$	$\{RIMDB\}$	$\{CHES\}$	$\{WBLOG\}$
Affective words	$< 0,0001$	$< 0,0001$	$< 0,0001$	$< 0,0001$
Dynamic adjectives	$< 0,0001$	$< 0,0001$	$0,014$	$< 0,0001$
Semantic adjectives	$< 0,0001$	$< 0,0001$	$0,045$	$< 0,0001$
Conjecture verbs	$0,00024$	$< 0,0001$	$0,021$	$< 0,0001$
Marvel verbs	$< 0,0001$	$< 0,0001$	$0,44$	$< 0,0001$
See verbs	$< 0,0001$	$< 0,0001$	$0,006$	$< 0,0001$
Positive verbs	$< 0,0001$	$0,00011$	$0,075$	$0,00061$
Level of abstraction	$0,003$	$< 0,0001$	$< 0,0001$	$< 0,0001$

Table 7.6: Results of the Wilcoxon rank-sum test for 95% confidence.

The observed results are consistent with the hypothesis that most of the high-level features have good discriminative properties for subjectivity identification. As illustrated in Table 7.6, we can see that only the level of positive verbs does not significantly separate the objective sample from the subjective one over training corpora. This is mainly due to the fact that positive verbs do not occur frequently in texts thus biasing the statistical test. As a consequence, we discarded this feature from our classification task[13]. We can also see that the $\{CHES\}$ corpus shows an uncharacteristic behavior. This is mainly due to the fact that unlike the other corpora, it focuses on subjectivity exclusively based on polarity.

We also performed a visual analysis of the distribution of the data sets in the space of high-level features. The goal of this study is to give a visual interpretation of the data distribution in order to assess how well classification may perform. If objective and subjective texts can be represented in a distinct way in the reduced space of features, one may expect good classification results. To perform this study, we used multidimensional scaling [Kruskal 1977], which is a traditional data analysis technique. In particular, MDS allows to display the

[13]In fact, we also used the positive verbs in our classification tasks to confirm these results as combinations of features may lead to different results. But, indeed, positive verbs are seldom used in texts, which limits their impact and do not provide best results.

structure of distance-like data into an Euclidean space. In practice, the projection space, which is built with MDS from such a distance is sufficient to have an idea about whether data is organized into classes or not. For our purpose, we performed the MDS process on all corpora trying to visualize subjective texts from objective ones as illustrated in Figure 7.1.

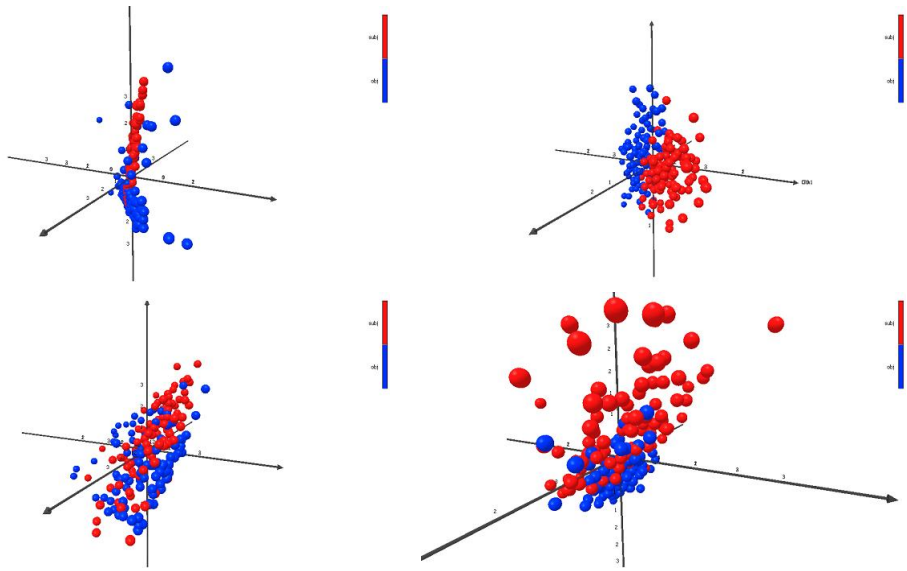

Figure 7.1: Multidimensional scaling: $\{MPQA\}$ (top-left), $\{RIMDB\}$ (top-right), $\{CHES\}$ (bottom-left) and $\{WBLOG\}$ (bottom-right).

The obtained visualizations show distinctly that a particular data organization can be drawn from the data. This visualization clearly shows that exclusively objective and subjective texts (i.e. $\{MPQA\}$ and $\{RIMDB\}$) may lead to improved results as data is well separated in the reduced 3-dimension space. In the case of the $\{CHES\}$ corpus and $\{WBLOG\}$ separating data in the space seems more difficult. Indeed, as these texts are not composed exclusively of subjective or objective sentences, the overlap in the space is inevitable. As [Wiebe 2004] state, 75% (resp. 56%) of the sentences in subjective (resp. objective) texts are subjective (resp. objective). However, a pattern in the space seems to emerge comforting us in the choice of our high-level features for our classification task.

Finally, we performed classification both with SVM and LDA within and across domains. In particular, SVM [Joachims 2002] have consistently shown better performance than other classification algorithms for topical text classification in general [Joachims 1998], and for sentiment classification in particular [Pang 2002] [Pang 2004]. However, other algorithms have been proposed for sentiment classification, which perform well and even better in some cases than SVM e.g. K-nearest neighbors [Wiebe 2004], maximum entropy [Boiy 2007], C4.5 [Finn 2006], Naive Bayes [Yu 2003] [Aue 2005] and sequential minimal optimization [Aue 2005]. In

this study, we propose to use LDA as an alternative to SVM. Indeed, SVM proved to work better when the size of features is high, but in the context of our work, only seven features are used for classification. Within this context, LDA is particularly well-suited for our classification task as it constructs one or more linear combinations of the predictor variables such that the different groups differ as much as possible on these discriminant equations. So, the experiment framework was defined as follows. All experiments, both for SVM and LDA, were performed on a leave-one-out 5 cross validation basis using the SVMlight package[14] for SVM and the free software for statistical computing R[15] for LDA. As part-of-speech tagger, we used the MontyTagger module of MontyLingua[16] [Liu 2004b].

In order to evaluate the difference between high-level features with low-level ones, we first performed a comparative study within domains on our four data sets. For the high-level features, we took into account 7 features (affective words, dynamic and semantically oriented adjectives, conjuncture verbs, see verbs, marvel verbs and level of abstraction of nouns). For the unigram and digram feature representations, we used all the lemmas inside the corpora withdrawing their stop words and weighting them with the classical $tf.idf(.,.)$ measure [Salton 1975]. The results of these experiments are shown in Table 7.7, which respectively presents the accuracy levels for the SVM and the LDA classifiers within domains.

	$\{MPQA\}$	$\{RIMDB\}$	$\{CHES\}$	$\{WBLOG\}$
Unigrams (SVM)	80.6%	97.0%	72.6%	94.6%
Digrams (SVM)	67.8%	98.0%	70.6%	80.0%
7 features (SVM)	71.2%	86.8%	64.4%	76.2%
7 features (LDA)	93.5%	96.5%	71.0%	94.0%

Table 7.7: Accuracy results within domain.

The results evidence an important gain with low-level features compared to high-level features for the case of the SVM. However, the results obtained with the LDA classifier show similar results to the ones presented by the SVM classifier with low-level features. One of the main reasons of the success of the SVM classifier based on low-level features is that objective language and subjective language hardly intersect. In practice, this means that a word specific to the subjective (resp. objective) part of the corpus can easily distinguish between objective and subjective texts, although it does not necessarily carry any subjective content. As a consequence, we are in the middle of sentiment classification and topical classification as evidenced in Table 7.8, which respectively presents the accuracy levels for the SVM and the LDA classifiers across domains[17].

[14] http://svmlight.joachims.org/ [12th *September*, 2010].
[15] http://www.r-project.org/ [12th *September*, 2010].
[16] http://web.media.mit.edu/~hugo/montylingua/ [12th *September*, 2010].
[17] The 6 features line means that the level of abstraction of nouns was omitted from the seven original high-level features.

	$\{MPQA\}$	$\{RIMDB\}$	$\{CHES\}$	$\{WBLOG\}$
Unigrams (SVM)	53.8%	63.9%	59.9%	61.1%
Digrams (SVM)	54.4%	67.1%	55.0%	57.5%
7 features (SVM)	52.6%	69.5%	73.9%	71.0%
7 features (LDA)	67.6%	70.9%	73.6%	74.5%
6 features (LDA)	64.4%	67.7%	67.4%	68.7%

Table 7.8: Accuracy results across domain.

In order to test models across domains, we proposed to train different models based on one domain only at each time and test the classifiers over all domains together. So, each percentage in Table 7.8 can be expressed as the average results over all data sets. Within this context, best results overall were obtained for high-level features with the $\{WBLOG\}$ corpus as training data set and the LDA classifier with an average accuracy of 74.5%. Moreover, the results drop drastically when learning based on unigrams or digrams and the introduction of the level of abstraction of texts clearly improves the classification accuracy. Although these results are encouraging, new trends in sentiment classification recently appeared using multi-view learning such as [Ganchev 2008] [Wan 2009] who evidenced improved results to cross domains. Within this scope, we proposed in [Lambov 2010] that high-level and low-level features can be treated as different views.

7.4 Multi-view Clustering

Multi-view learning refers to a set of semi-supervised methods which exploit redundant views of the same input data [Blum 1998] [Collins 1999] [Brefeld 2005] [Sindhwani 2005]. However, although semi-supervised learning is usually associated to small labeled data sets and tries to automatically label new examples, multi-view learning aims at learning a compromise model of the different views. The most important work in multi-view sentiment classification is proposed by [Ganchev 2008], who presented a new algorithm called SAR, which outperformed the results proposed earlier by [Blitzer 2007] on the same data set. However, [Ganchev 2008] only use low-level features, which are randomly divided to form two "artificial" views. Instead, we aim at combining high-level features and low-level features to learn models of subjectivity, which may apply to different domains. For that purpose, we proposed in [Lambov 2010] a new scheme based on the classical co-training algorithm over two views [Blum 1998] and joined two different classifiers LDA and SVM to maximize the optimality of the approach. This work is at its initial stage and many future works must be endeavored. These issues will be discussed in the final section of this chapter.

In order to better understand the behavior of multi-view learning, we first applied SAR to our data sets. In particular, we used two views generated from a

	$\{MPQA\}$	$\{RIMDB\}$	$\{CHES\}$	$\{WBLOG\}$
Unigrams	65.3%	73.5%	72.2%	59.2%
Digrams	71.6%	75.2%	77.2%	65.1%

Table 7.9: SAR accuracy results for low-level features across domains.

random split of low-level features together with a maximum entropy classifier to learn a domain-independent model. For that purpose, we performed a leave-one-out 5 cross validation, where both labeled and unlabeled examples were provided for the learning process[18] and then new unlabeled examples were classified by the learnt model to evaluate accuracy. So, as we did previously in section 7.3, we proposed to train different models based on one domain only at each time and test the classifiers over all domains together. Thus, the accuracy results presented in Table 7.9 represent average values, which evaluate how well a model can cross different domains.

The results show indeed interesting properties. First, models built upon digrams constantly outperform models based on unigrams, thus highlighting the fact that combinations of words may embody interesting properties for sentiment classification. Based on these results, we recently studied in [Rodrigues 2009] the identification and extraction of sentiment MWU. Second, higher accuracy is reached compared to section 7.3 with less knowledge. Indeed, the baseline with single-view classification is 74.5% while 77.2% can be obtained with the SAR algorithm upon a random split of digrams. One great advantage of only using low-level features is the ability to reproduce such experiments on different languages without further resources than just texts. However, a good training data set will have to be produced as the best results are obtained from the manually annotated corpus $\{CHES\}$, while the automatically labeled corpus $\{WBLOG\}$ provides worst results.

These results were very encouraging and we tried to extend this experiment by introducing high-level features. However, the actual implementation of SAR[19] does not allow to test different types of views but only random subsets of unique views (i.e. unigrams are randomly divided into two subsets to "artificially" create two views). Moreover, it does not allow the implementation of different classifiers as new mathematically theoretical work would have to be demonstrated. Finally, [Ganchev 2008] only proved their algorithm for the existence of two views and not more. For that reason, we are already working on a new implementation of SAR with João Graça, co-author of [Ganchev 2008], Lionel Martin and Guillaume Cleuziou from the University of Orléans (France) to solve these limitations.

[18] It is important to note that the labeled examples are from one domain and the unlabeled ones are from a different one.

[19] Which was kindly provided by [Ganchev 2008].

Unfortunately, results are not available yet. As a consequence, we proposed to use the simpler co-training algorithm [Blum 1998], which is easily tunable both in terms of views and classifiers. The co-training algorithm is presented in algorithm 13.

Algorithm 13 The co-training algorithm.

Input: L a set of labeled examples from one domain, U a set of unlabeled examples from another domain
Output: Trained classifier $H2$
for k iterations **do**
  Train a classifier $H1$ on view $V1$ of L
  Train a classifier $H2$ on view $V2$ of L
  Allow $H1$ and $H2$ to label U
  Add the most confidently predicted P positive and N negative examples to L
end for

So, we proposed to combine a first view, which contains 7 high-level features (7F) and a second view, which contains low-level features (unigrams or digrams). Our expectations are that the low-level classifier will gain from the decisions of the high-level classifier and will self-adapt to different domains based on the high results of high-level features to cross domains. In Table 7.10, we show the results obtained using different combinations of classifiers for $N = P = 2$.

First view	Second view	$\{MPQA\}$	$\{RIMDB\}$	$\{CHES\}$	$\{WBLOG\}$
7F (SVM)	Unigr. (SVM)	61.0%	72.3%	78.8%	62.8%
7F (SVM)	Digr. (SVM)	66.4%	78.1%	75.3%	85.6%
7F (LDA)	Unigr. (SVM)	63.3%	74.9%	79.0%	63.5%
7F (LDA)	Digr. (SVM)	67.4%	78.1%	68.5%	86.4%

Table 7.10: Co-training accuracy across domains.

The benefit from the high-level features is clear. The best result is obtained by the combination of high-level features with the LDA classifier and digram low-level features with the SVM classifier trained over the automatically annotated corpus $\{WBLOG\}$. In this case, the average accuracy across domains is 86.4% outperforming SAR best performance 77.2%. It is important to notice that accuracy results were obtained from the second view classifier, i.e. the low-level classifier. Indeed, while the high-level classifier accuracy remains steady iteration after iteration, the low-level classifier steadily improves its accuracy based on the correct guesses of the high-level classifier. We illustrate the behavior of each classifier in Figure 7.2.
It is also interesting to notice that in almost all cases, digram low-level features provide better results than only unigrams. The only exception is the $\{CHES\}$ training set. But, it is especially evident for the $\{WBLOG\}$ training data set, where digrams drastically improve the performance of the co-training algorithm.

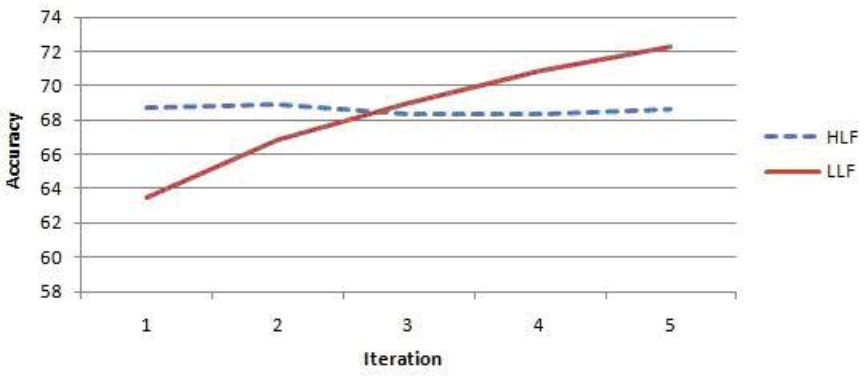

Figure 7.2: Low-level and high-level accuracies, iteration after iteration.

7.5 Future Work

The sentiment language can be thought as orthogonal to the topical language. Indeed, to some extent, one does not need to be an expert in a domain to understand whether a document is subjective or not. Of course, this assumption may not be as linear as we would like it to be. In fact, subjectivity is certainly a mixed between common language subjective words, expressions or even sentence structures, and domain-specific clues. As a consequence, sentiment classification is clearly a more difficult task than topic classification. A direct evidence of this assumption is that classifiers trained on a unique domain do not perform well in other domains, even when using deep linguistic resources of common language. Within this context, different studies have been emerging to tackle cross-domain sentiment classification. Indeed, sentiment classifiers need to be customizable to new domains in order to be useful in real-word environments such as the Web. Moreover, most of the studies have been focusing on polarity although subjectivity is a much more complex linguistic phenomenon as explained in [Boiy 2007]. For that purposes, we presented different experiments based on single-view and multi-view learning algorithms using high-level and low-level features.

However, many extensions to this work still need to be carried out. On the first hand, the SAR algorithm must be adapted to accept views with different feature types so that an impartial evaluation can be carried out. Indeed, the well-defined mathematical background of SAR (which makes it one of the reference multi-view learning algorithms in the field) together with results obtained just on low-level features makes us think that improved results can be obtained with little adaptations. We are actually working on this issue with João Graça, co-author of [Ganchev 2008].

On the second hand, recent experiments showed that using more than two views can lead to improved results over two views using the co-training algorithm

i.e. by dividing high-level features into different sets, thus providing new views. Within this context, the SAR algorithm is only defined for two views. So, we mathematically defined its formalism for 3 views and experiments will soon be carried out to verify its behavior. This work is being carried out in collaboration with Guillaume Cleuziou and Lionel Martin from the University of Orléans (France).

On the third hand, we just experimented the classical co-training algorithm, which is more adapted to semi-supervised learning rather than multi-view learning. As a consequence, we aim at proposing a new co-training algorithm based on agreement and not only on the increase of new training examples based on their classification confidence. As a consequence, new examples would be added to the training set if they are confidently classified with the same label by all classifiers. We also started this work but without results yet.

On the fourth hand, the experiments show that digrams usually provide best results than just unigrams. However, most digrams are not meaningful. As a consequence, we could identify relevant MWU within all corpora to reduce text representation space and at the same time increase the expressiveness of features.

Finally, and certainly the most important point of this research, is to leverage the necessity of pre-existing mostly manually built linguistic resources. Indeed, one of our main research directions focuses on the development of language- and domain-independent applications. However, the solutions proposed for sentiment classification are far from our ultimate goal. Nevertheless, this issue is already being studied. In fact, we need to turn all language-dependent features into language-independent features, or at least propose language-independent methodologies to automatically build linguistic resources. Within this scope, we already defined a statistical methodology to automatically extract sentiment words and MWU from randomly extracted Web pages from Wikipedia and Weblogs from any language in [Rodrigues 2009]. Moreover, the level of abstraction of words, which is one of the most relevant features, can be obtained by the automatic construction of terminological or prototype-based ontologies (see Chapter 6) or in the first place by our methodology to assess the level of generality/specifity between words (see Chapter 6, section 6.1.3) ■

Chapter 8

Conclusions

Contents

8.1 Conclusions **193**

8.2 Future Projects **197**

8.2.1 Temporal Information Retrieval 198

8.2.2 Personalized Information Retrieval 199

Reaching the conclusion of a thesis is always a peculiar moment as all the achievements of one's work are put forward. But at the same time, it seems that so little has been done compared to what remains to be done. This is the natural cycle of research. All along this thesis, we proposed to extract implicit knowledge about the language to understand the messages conveyed in texts within the context of Information Digestion. In particular, our challenge was to propose unsupervised language-independent methodologies based on complete raw texts. As a consequence, we will first focus on the main conclusions that can be drawn from what has been achieved so far and then present future projects, which can legitimately be integrated in the Information Digestion paradigm.

8.1 Conclusions

The main goal of this thesis was to show to what extent Information Digestion might be tackled by discovering implicit knowledge about the language. For that purpose, our constraints were high as we wanted to develop unsupervised language-independent methodologies from complete raw texts. To summarize what has been accomplished within this scope, we will report the main issues of each chapter in the remainder of this section.

Within the context of our ultimate goal, Chapter 3 is certainly the one, which best complied to our constraints. Indeed, the SENTA software takes any raw text as input and retrieves a list of MWU candidates. In particular, it is an unsupervised language-independent methodology, which can even be applied to character-based languages as shown in [Dias 2000d] [Dias 2000c] [Ribeiro 2001]. Furthermore, with the definition of the GenLocalMaxs algorithm, it does not rely on any threshold. Finally, recent studies showed that its bootstrapping application leads to better recall without precision loss. However, SENTA cannot be used as such to directly

populate a terminology. As a consequence, we proposed its hybrid version based on part-of-speech tagged corpora, the HELAS software, which gains in expressiveness of MWU although low precision results are obtained for digrams. As a matter of fact, the introduction of part-of-speech digrams raises frequency problems. It is clear that an optimized α parameter should be tuned for each level of extraction. However, by doing so, we would introduce an external parametric quantity, which would limit the universality of this solution. But, as strange as it can be, SENTA showed improved results when applied to normalize texts within the context of topic segmentation, Web snippet representation and cluster labeling. This opens an interesting debate. Would only terminologically "valid" MWU improve natural language processing? This is far from being a certainty. In fact, SENTA tends to thwart this statement.

In Chapter 4, we presented ephemeral clustering as the first way to reduce textual information from Web search engines. Although it may seem strange, it becomes clear when dealing with mobile information retrieval or adapted VIP interfaces. Within this scope, we presented different possible solutions. The most interesting solution is certainly the last one proposed in collaboration with Guillaume Cleuziou and David Machado, which takes into account the new informative similarity measure, the InfoSimba, proposed in Chapter 2 combined with a new learning strategy. As a matter of fact, the improved evaluation of Web snippets similarity allows to reach ephemeral clustering through query disambiguation. As a consequence, a few clusters are presented to the users, which likely embody the different meanings of the query. Compared to previous works, the introduction of the informative similarity measure is certainly the main contribution to the area. Indeed, the application of the vector space model clearly showed its limitations by over-generating clusters due to unsatisfactory evaluation of Web snippets similarity. Moreover, we showed that the introduction of linguistic knowledge can be avoided as it does not lead to any improved results, mainly due to the odd structure embodied by Web snippets. Once again, our main goals were attained as we propose an unsupervised language-independent methodology based on complete raw texts. Nevertheless, work still needs to be done to introduce text normalization and allow overlapping. Recent developments show how frequent itemsets computed by suffix-arrays can improve cluster labeling. These improvements can be accessed via our VIPACCESS[1] meta-search engine.

Chapter 5 was certainly the most challenging topic for Information Digestion. However, automatic text summarization is a wide area and only some of all its processes have been tackled. Within this scope, we proposed two different methodologies i.e. one which can be run "on the fly" and thus automatically plugged in our meta-search engine and one which needs to be processed off-line as it implies deep Web page understanding. The first approach is based on

[1] http://hultig.di.ubi.pt/vipaccess [27th *September*, 2010].

simple heuristics and do not propose cohesive nor coherent results. Although this methodology can evidence improvements with text normalization and advanced word weighting schemes [Dias 2006a] [Dias 2007b] [Dias 2009], it still suffers from lack of context and only provides low quality summaries. In particular, similarly to what we experienced with Web snippets, Web pages show odd text structures and extracting relevant textual information from Web pages is a hard problem. So, proposals such as the one of [Cai 2004], who designed a vision-based page segmentation algorithm should certainly be taken into account to deal with real-world heterogeneous texts. As a consequence, we focused on more complex solutions, which specifically take into account the discourse structure.

Within this scope, we first presented the ITOS algorithm, which sequentially segments any raw text based on text normalization and the InfoSimba informative similarity measure. Once again, the identification of MWU with the SENTA software and the evaluation of sentence and paragraph informative similarities showed to lead to improved results when compared to the state-of-the-art. Thus, we managed to deal with topic segmentation by proposing an unsupervised language-independent methodology based on complete raw texts. Nevertheless, the ITOS algorithm depends on tuning parameters that we propose to avoid in future work (see section 5.4 in Chapter 5).

We also proposed to automatically build a noun taxonomy based on the parameter-free PoBOC algorithm [Cleuziou 2003] combined with the InfoSimba over noun feature vectors to produce lexical chains. Although interesting results were obtained[2], the low quality of the taxonomy, limited by its topology, was one of the main factors for the obtention of over-weighted lexical chains in terms of words. But, if we are capable of constructing more structured ontologies, we may certainly obtain better results than the general-purpose taxonomy WordNet. In fact, building a taxonomy based on related specialized corpora may fit the semantic component neatly and directly, which will never be possible with general-purpose resources. That's why we endeavored researches within this scope in Chapter 6.

Finally, when tackling abstractive summarization, we evidenced some of the limitations of working based only on complete raw texts. Indeed, sentence reduction can benefit from shallow-parsing to the extent of 22% precision compared to a complete unsupervised language-independent methodology based on paraphrase alignment. However, this result is mainly due to the capacity of the learning algorithm to generalize more easily rather than because the shallow-linguistic process allows to better understand the content of texts. We are deeply convinced that new solutions can be proposed based on word clustering to attain comparable results without relying on shallow-parsing. In fact, by depending on clusters

[2] In particular, we compared the results to the state-of-the-art algorithms [Barzilay 1997] and [Hirst 1998b] and improved expressiveness was evidenced by our algorithm.

of words, which may embody conceptual information, instead of just words, generalization is likely to be facilitated. A clear example of this issue is that text normalization clearly improves paraphrase extraction and alignment as shown in [Grigonyté 2010]. As a consequence, in the near future, we aim at proposing a totally unsupervised methodology based on complete raw texts to reduce sentences based on the quality of kernels as well as the quality of kernel contexts. In a first step, this quality measure can evaluate the coverage of each word sequence.

Chapter 6 is certainly the most important to focus on for the rest of our research plans. Indeed, the automatic construction of lexical-semantic structures can be a great asset for many applications within the scope of Information Digestion. The first one is obviously the construction of high quality lexical chains. But, as we will see in the next section, the work proposed in this chapter may have a great impact on personalized Web search. In particular, we focused on two different approaches: prototype-based ontologies and terminological ontologies.

Within the context of prototype-based ontologies, we introduced an unsupervised language-independent methodology to evaluate the level of generality of words, which solves the problems evidenced by the pattern-based paradigm. We also introduced a new methodology based on the combination of unsupervised paraphrase extraction and alignment combined with a decision module based on attributional similarity measures. But, while the first method can slightly be improved by applying bootstrapping, the second one still needs deep analysis and evaluation. In particular, we clearly believe that the application of informative similarity measures may improve precision of the extraction process as all performed experiments are so far based on the vector space model. In this case, both the symmetric and the asymmetric InfoSimba may lead to improved results respectively for (1) the synonym and the co-hyponym relations and (2) the hypernym/hyponym and meronym/holonym relations. Moreover, the results presented in this chapter are based on shallow-parsed corpora. However, we recently assessed that better results can be obtained just with part-of-speech tagged corpora. Of course, we could only work based on raw texts but huge quantities (in terms of terabytes) would be necessary. Another way to deal with this problem would be to work based on clusters of words instead of single words as explained in the previous paragraph. Nevertheless, many studies show that comparative results can be obtained with terabytes of texts [Terra 2003]. As a consequence, the "one sense per discourse" proposals may also be tested on complete raw texts in a near future.

Finally, we proposed in [Bastos 2009] a new paradigm to build prototype-based ontology based on a two step process, which aims at first building generality clusters and then apply semantic clustering within generality clusters. So far, we did not reach the overall process, but all the methodology is unsupervised, language-independent and settles on complete raw texts. But, we do not think to stop here. In particular, we want to test multi-view clustering algorithms such

as the CoFKM proposed in [Cleuziou 2009] to take into account both levels of generality and semantic closeness between words in a single step.

Within the scope of terminological ontologies, we recently proposed with Guillaume Cleuziou and Vincent Levorato in [Cleuziou 2010] a new framework based on the pretopology theory. Recent results confirm the soundness of the unsupervised language-independent methodology, although a lot still needs to be proposed and evaluated. In particular, we expect that this formalism can lead to the creation of user personalized micro-worlds or micro-ontologies in the domain of adaptive information retrieval.

Finally, Chapter 7 is certainly the most atypical work as almost none of our initial constraints have been tested. To some extent, the only interesting idea is the automatic data labeling for learning purposes. In fact, all the interesting works based on our accumulated experience in the field are still to be tested. In particular, we proved that the level of abstraction of nouns is an important feature for subjectivity classification. Based on this conclusion, it is easy to test our unsupervised methodology, which provides a generality level to each given word within a list of related words. Moreover, we recently proposed in [Rodrigues 2009] an unsupervised methodology based on different simple heuristics to extract a set of subjective words. Consequently, we could replace the WordNet Affect lexicon with this automatically compiled data set. But our principal objective is to combine these new characteristics within the SAR framework [Ganchev 2008], which offers a mathematically well-founded model and reaches high results only based on unigrams or digrams. In particular, we are interested in generalizing SAR to tackle multi views (i.e. more than two views), so that we can attain higher levels of accuracy across domain based on language-independent characteristics.

Although much still needs to be proposed, performed and evaluated within each chapter of this thesis, new research directions have already appeared as a logical progress towards enhanced Information Digestion. They are expressed in the next section.

8.2 Future Projects

Future projects will mainly involve temporal IR and personalized IR as a means to reach deep Information Digestion. Indeed, such research areas may surely benefit from the works developed so far, in particular around the automatic construction of lexical-semantic resources as well as ephemeral clustering. Some ideas already emerged but always keeping in mind that language can always be understood to some extent by analyzing word and text similarities without requiring huge linguistic knowledge bases or tools.

8.2.1 Temporal Information Retrieval

The WWW is a huge information network from which retrieving and organizing quality relevant contents remains an open question. This particularity is even more important for ambiguous queries. In particular, many queries have temporal implicit intents but current search engines do not cope with this dimension. Better said, they give more attention to present information than past information. As a consequence, inferring the temporal ambiguity of queries, may play an important role for the improvement of Web search interfaces and for Information Digestion as a whole. Let's take a simple example. The query *football world cup* has undeniably a temporal ambiguity, which may be difficult to catch by current ephemeral clustering algorithms (see Figure 8.1) as all Web pages deal with the same topic. In fact, one may be interested in the *football world cup* of 1998 in France.

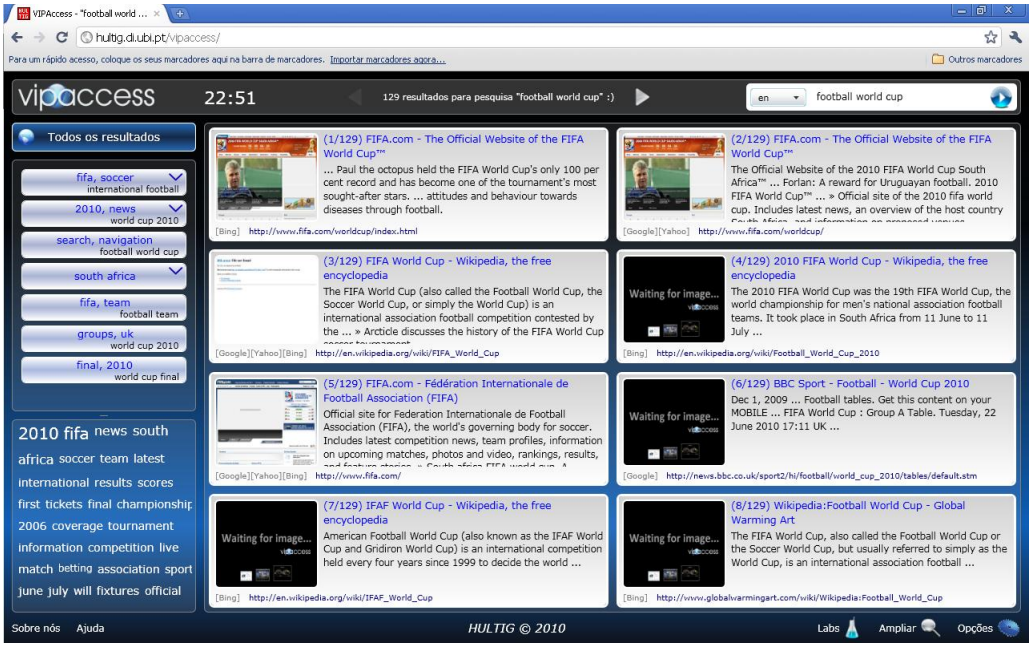

Figure 8.1: The HISGK-means algorithm for the query *football world cup* [27th *September*, 2010].

Although time is everywhere in the WWW, temporal information retrieval has not yet been involved in a systematic research process. In fact, few works have fully used temporal information for search purposes. [Jin 2008] and [Alonso 2009] are the exceptions. On the one hand, [Jin 2008] concentrate on the extraction on content time of Web pages and provide meaningful time-based search facilities. On the other hand, [Alonso 2009] propose to evidence the time dimension with time clusters. However, temporal ambiguity is much more difficult than exposed so far. Let's take another example. The query *world cup* is definitely ambiguous both in terms of concepts as well as in terms of time. This situation is illustrated in Figure 8.2. In

fact, only time clustering would not be sufficient to disambiguate from world cups of football and volleyball. Similarly, concept clustering may not be sufficient to disambiguate between the world cups of football, basketball and volleyball in the USA. In this last case, time clustering could help only to disambiguate between ⟨football, volleyball⟩ and basketball.

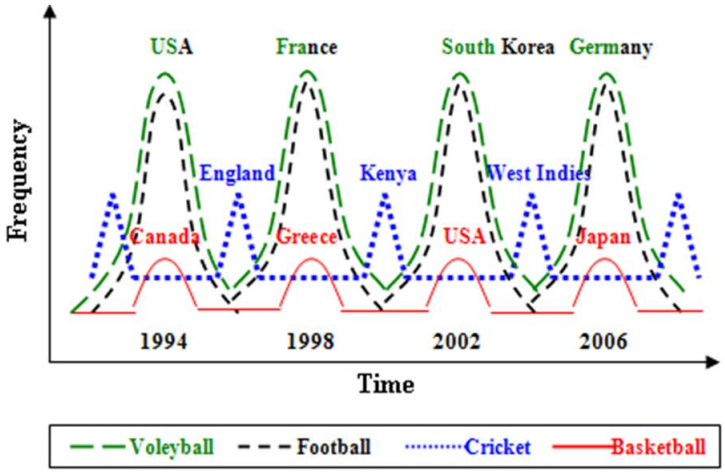

Figure 8.2: The query *world cup*.

As a consequence, we propose to develop a clustering framework tackling both conceptual and temporal dimensions in order to identify relevant documents both conceptually and timely.

8.2.2 Personalized Information Retrieval

Retrieving personalized information from the ever increasing information space of the Web is another way to look at Information Digestion. Due to the huge volume of the Web, searching for specific information has become tedious and time consuming. To alleviate this problem, efforts have been started to customize the view of the Web in a user specific way. This field is known as personalized information retrieval, which deals with the idea of retrieving specific information customized to the need of the user (i.e. building a user model). Within this scope, many proposals have recently been emerging such as [Golemati 2007] [Sieg 2007] [Challam 2007] [Tamine 2008] [Daoud 2010] [Bhowmick 2010].

User modeling can be described as the process of building the personal preferences of the users in terms of user's knowledge about the world, his behavioral aspects, goals, likes and dislikes i.e. his context. In particular, there are two kinds of contexts: (1) the short-term context, which is the surrounding information that emerges from the current user's information need in a single session and (2) the long-term context, which refers generally to the user's interests that have been inferred from past user sessions. While some works deal with both short-term and

long-term contexts in combination such as in [Tamine 2008], we intend to focus exclusively on long-term user profiles.

Different representations of user profile exists. As explained in [Daoud 2010], the user profile can be seen as a bag of words [McGowan 2003], a graph of terms [Micarelli 2004], lists of concepts [Liu 2004a] or even a term/document matrix as in [Tamine 2008]. More recently, ontology-based user profiles have received great attention from the research community such as in [Golemati 2007] [Sieg 2007] [Challam 2007] [Bhowmick 2010] [Daoud 2010]. The idea is to build models of user context as ontological profiles by assigning implicitly derived interest scores to existing concepts in domain ontologies. However, all approaches are based on the idea that there exists a given "global" ontology to rely on. [Golemati 2007] even propose a standard ontology for modeling user profiles. This approach seems far from being applicable in real-world environments gathering heterogeneous collections of texts different in language, topic, genre, subjectivity, specificity. To overcome these drawbacks, we propose to automatically build user micro-ontologies or mini-worlds (i.e. taxonomic user profiles) based on the aforementioned methodologies of Chapter 6. Indeed, besides query terms, we may extract relevant categories as well as Web pages of interests and dynamically build and update the user's micro-ontology. In particular, the pretopology formalism associated to unsupervised language-independent methodologies for the extraction of meaningful terms from real-world heterogenous texts may lead to interesting results. A first approach has already been carried out in [Machado 2009a] just using query terms and subsuming visited cluster labels. Although interesting results have been obtained as illustrated in Figure 8.3, Web page interest clearly lacks.

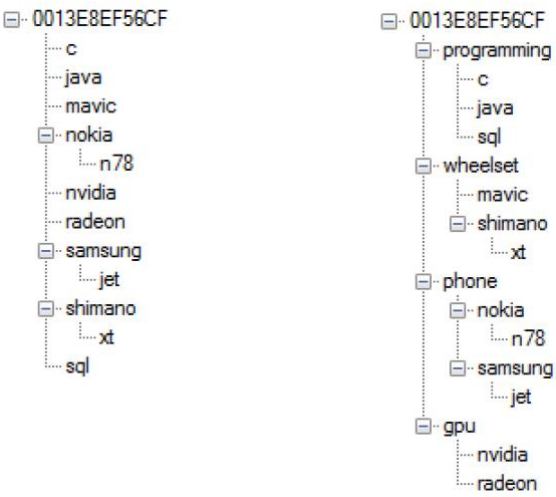

Figure 8.3: The user model based on queries (left) and the user model after cluster label subsumption (right).

As a consequence, we are currently working on a methodology, which applies formal concept analysis (FCA) [Ganter 1999] based on noun feature vectors to produce micro-ontologies. In particular, relevant concepts are extracted from input query terms, visited cluster labels and relevant words extracted from Web pages of interest for the user. Then, all words or concepts identified as relevant are represented as noun feature vectors. Finally, FCA is applied to build the mini-worlds. Although results are not available yet, many issues already appear to be tackled such as the update of the micro-ontologies and their reliability along the time. But, these are future extensions of a first process, which aims at building accurate hierarchical taxonomies from user interaction.

As we mentioned in the beginning of this chapter, concluding a thesis is always particular moment as it seems that so much has been done. But, at the same time, it appears that so little has been achieved compared to what still needs to be done. So, just to finish, I would like to mention that other future works can be visioned such as collaborative IR, social IR and context-aware IR. But, all these potential applications are likely to remain visions ■

Appendix A

Research Environment

Contents

A.1 Projects 203
A.2 Academic Staff 207
A.3 Collaborators and Partners 209

Most of the works presented in this thesis have been developed at the Center of Human Language Technology and Bioinformatics (HULTIG) of the University of Beira Interior (Portugal). In particular, many bachelor, master and PhD students endeavored hard work within funded national and bilateral international projects. Many collaborators also allowed to develop high quality research thanks to their experience in the field. All these contributors are listed in this annex.

A.1 Projects

Different projects have been accepted for funding, thus enabling the development of our research. In particular, we mainly competed to national projects and bilateral international projects. All funded projects can be found here.

VIPACCESS - Ubiquitous Web Access for Visually Impaired People: Visually impaired people are info-excluded due to the overwhelming task they face to read information on the Web. Indeed, unlike fully capacitated people, blind people can not read information by just scanning it quickly i.e. they can not read in the "diagonal". As a consequence, they have to come through all sentences of Web pages to understand if a document is interesting or not. This is obviously an overwhelming task, which clearly excludes visually impaired people from quick access to information. Although a lot has been done for blind people to access information with braille screens, braille keyboards, braille PDAs and text-to-speech interfaces, very little has been made to reduce the amount of information they have to deal with. For that purpose, we aim at proposing new human-computer interfaces that run on classical devices such as classical PCs, Pocket PCs or PDAs. As a consequence, blind people may have access to affordable technology. In particular, we aim at developing a wide range of technological solutions so that visually impaired people can take advantage of the new ubiquitous technologies that are fast growing in our every day lives. This project is funded by the National Portuguese Foundation for

Science and Technology with reference PTDC/PLP/72142/2006 from 2009 to 2011 and counts with the participation of the University of Porto (Portugal), the New University of Lisbon (Portugal), the MIT (United States of America) and the North Texas University (United States of America).

MEDON - Ontology for the Medical Domain: A vast amount of medical information exists in published documents and databases. However, clinicians usually use only a small subset of such information. Several obstacles may appear between the medical information and the clinician, such as: (1) the lack of knowledge on the existence of such information, (2) the lack of availability at the point-of-care, (3) the uncertainty on the precise applicability limits of the information and (4) the difficulty to retrieve the information, specially if the search interface requires a specialized technical formation. The use of natural language interfaces and ontologies are two promising approaches to overcome the above mentioned obstacles, conducing to an efficiency and quality increase on the clinician work. This project aims at creating a medical ontology allowing the representation of (1) medical procedures and algorithms, (2) clinical data and (3) time events. This project is funded by the National Portuguese Foundation for Science and Technology with reference PTDC/EIA/80772/2006 from 2009 to 2011 and counts with the participation of the University of Evora (Portugal) and DATA MEDICA, Inc. (Portugal).

PT-STAR - Speech Translation Advanced Research to and from Portuguese: S2SMT can be seen as a cascade of three major components: automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS). One of the main problems of this multidisciplinary area, however, is the still weak integration between the three components. The main goal of this project is to improve speech translation systems for Portuguese by strengthening this integration. Hence, the project encompasses three main tasks. The first one addresses the interface ASR/MT. The second one addresses the interface MT/TTS and the third one the MT module itself. A fourth task will build a proof of concept prototype. This project is funded by the National Portuguese Foundation for Science and Technology with reference CMU-PT/HuMach/0039/2008 from 2009 to 2012 and counts with the participation of the INESC-ID (Portugal), Carnegie Mellon University (United States of America) and the National Portuguese Foundation for Science and Technology itself.

SUMO - Automatic Text Summarization for Mobile Technologies: While Mobile Technologies (PDAs, Tablet PCs, Mobile Phones) are emerging in our every day life, their usage is still unmanageable for a variety of applications, even for the most basic ones like reading contents from the Web. Indeed, a very few techniques have been proposed for human-computer interaction with this kind of devices. The main objective of this project is to propose a summarization system that will enable easy reading of Web contents on mobile devices. Indeed, by summarizing

Web pages without loosing relevant contents and text coherence, people will quickly adopt these devices that otherwise will not enter the market as it has been the case for the WAP technology. Our other objective is to propose a language-independent summarization system that can be applied to any language without needing huge rich-knowledge databases that are only available for dominating languages. As a matter of fact, most of the applications in natural language processing are specialized for English and neglect all other languages thus restricting global access to information. This project was funded by the National Portuguese Foundation for Science and Technology with reference POSC/PLP/57438/2004 from 2005 to 2007 and counted with the participation of the University of Porto (Portugal).

SITE-O-MATIC - Web Automation and Adaptive Web: The Web currently poses a number of interesting research problems. From the user's point of view, the Web is becoming too large, too dynamic and increasingly unknown. From the editor's point of view, the Web is a constant demand for new information and timely updates. Moreover, the editor should not only maintain the contents of the site, but also permanently choose the site's navigational structure that best helps achieving the aims of the site's owner, user, or both. From the owner's point of view the need for such a constant labor intensive effort implies very high financial or personal costs. In this project we aim at developing a platform and a methodology to automate, as much as possible, the management activities of a Web site, taking into account the behavior of the users, and the aims of the owner. One of the effects of automation is the reduction of the editor's effort, and consequently of the costs for the owner. The other effect is that the site can more timely adapt to the behavior of the user, improving the browsing experience and helping the user in achieving his/her own goals when these are in accordance to the goals of the owner of the site. Our aims will be pursued by (1) defining a flexible Web site platform that allows the acquisition of quality Web data, as well as on line automatic transformation and customization of the site's structure and interface, (2) developing techniques for Web adaptation using data mining (association rules, collaborative filtering, bayesian approaches, graphical models) and (3) allowing the specification of topic focused content retrieval. This project was funded by the National Portuguese Foundation for Science and Technology with reference POSC/EIA/58367/2004 from 2005 to 2007 and counted with the participation of the University of Porto (Portugal), the Polytechnic Institute of Porto (Portugal), the Open University (Portugal) and the University of Minho (Portugal).

LEILA II - Learning Lexical Associations towards Ontology Construction: Basic natural language resources such as those in the UMLS specialist lexicon are a key asset for medical informatics. Beyond the specialist lexicon, medical lexicons have started to be generated for German and French. For the Portuguese language, some lexical resources do exist, but they are incomplete and scattered in multiple teams (Universities of Coimbra, Porto and Lisbon) or countries (Portugal

and Brazil). In this work, we propose to start the first study to build the future UMLS for Portuguese: the UMLP (Unified Medical Lexicon for Portuguese). This initiative is supported by the new School of Health Sciences of the University of Beira Interior, which proposes a new methodology for medical studies mainly based on e-learning. As a consequence, this lexicon is a key issue for the Faculty but also for all the Portuguese universities, which will sooner or later use this learning methodology. The first step of the construction of the UMLP will have to do with the construction of the compilation of existing medical lexicons in Portuguese. For that purpose, we will use traditional techniques to compute lexicons but with a new emphasis on phonetics, term frequency and term sense disambiguation. The second step of the construction of the UMLP will have to do with the construction of the metathesaurus. For that purpose, we will propose a new idea to automatically compute hierarchical structures based on soft clustering techniques from a similarity matrix of level-alike of generality similarity measures. This project was funded by the CRUP (The Council of the Deans of Portuguese Universities) with reference F - 48/07 in 2007 and counted with the participation of the University of Orléans (France).

LEILA - Learning Lexical Associations: Lexical associations include a large range of linguistic phenomena, such as compound nouns (e.g. interior designer), phrasal verbs (e.g. run through), adverbial locutions (e.g. on purpose), compound determinants (e.g. an amount of), prepositional locutions (e.g. in front of) and institutionalized phrases (e.g. con carne). In fact, lexical associations are frequently used in everyday language, usually to precisely express ideas and concepts that cannot be compressed into a single word. As a consequence, their identification is a crucial issue for applications that require some degree of semantic processing (e.g. information retrieval, machine translation, and automatic text summarization). In recent years, there has been a growing awareness in the natural language processing community of the problems that lexical associations pose and the need for their robust handling. For that purpose, syntactical, statistical and hybrid syntactic-statistical methodologies have been proposed. However, few works have attempted to tackle this problem through machine learning. This project aims at responding to this situation by introducing machine learning techniques in order to identify lexical associations from texts. This project was funded by the CRUP (The Council of the Deans of Portuguese Universities) with reference F - 20/05 between 2004 and 2005 and counted with the participation of the University of Orléans (France).

MULTILEXI - Multilingual Term Extraction for Lexicographic Purposes: Multilingual terminology resources are an indispensable aid for translators, domain experts and creators of terminological reference works, and especially for the development of multilingual language technologies such as machine translation, information retrieval systems and other language-based applications. In a time of rapid and global development of technological domains the creation of efficient and

up-to-date terminological resources is inevitably supported by appropriate computational resources and tools. These include the systematic archiving of multilingual documents and creation of multilingual corpora, and on the other hand tools and methodologies for automated term extraction from texts. The aim of this project is the development of robust tools for bilingual terminology extraction from domain-specific parallel corpora that could be applied to lexicographic and terminological purposes on a broad scale. The main goal is therefore to ensure the highest level of language and domain independence while producing a tool of adequate usability and portability to enable easy transfer of term extraction technology into practice. This project was funded by the GRICES (The Bureau of International Relations of Science and Superior Education) between 2004 and 2005 and counted with the participation of the University of Ljubljana (Slovenia).

A.2 Academic Staff

The work presented in this thesis is a global common efforts of bachelor, master and PhD students who joined funded project to endeavor innovative works within the scope of Information Digestion. All students who participated in these projects are listed below.

PhD students:

- David Machado (thesis due 2013)
- Dinko Lambov (thesis due 2010)
- Isabel Marcelino (thesis due 2010)
- João Paulo Cordeiro (thesis due 2010)
- Raycho Mulelov (thesis due 2010)
- Ricardo Campos (thesis due 2012)
- Rumen Moraliyski (thesis due 2010)
- Sebastião Pais (thesis due 2012)

Master students:

- Alexandre Gil (2003)
- Cláudia Santos (2006)
- Daniel Rodrigues (2009)
- David Machado (2009)
- Manuel Lourenço (2009)

- Nuno Guimarães (2009)
- Ricardo Campos (2005)
- Ruben Costa (thesis due 2010)
- Sebastião Pais (2007)
- Sónia Santos (2009)
- Tiago Barbosa (thesis due 2010)

Bachelor students:

- Bruno Conde (2006)
- Carmen Barroso (2007)
- Daniel Malaca (2008)
- Elsa Alves (2005)
- Fernando Cunha (2007)
- Hélio Santos (2007)
- Hugo Costa (2007)
- Hugo Veiga (2004)
- Luís Almeida (2008)
- Ruben Costa (2008)
- Sebastião Pais (2005)
- Sérgio Nunes (2001)
- Tiago Barbosa (2008)
- Victor Gonçalves (2007)

Research collaborators:

- Yuliya Dospatska (2010)
- Bono Nonchev (2007)

A.3 Collaborators and Partners

Developing high level research can only be reached by sharing one's ideas with other researchers of the field. For that purpose, we worked in straight collaborations with researchers from different universities and industrial partners in order to confront our ideas and as a consequence attain higher levels of exigence. These collaborators and partners are mentioned below.

Portuguese research collaborators:

- Artificial Intelligence and Computer Science Laboratory, University of Porto (Portugal)
- Department of Computer Science, Minho University (Portugal)
- Department of Computer Science, University of Évora (Portugal)
- Natural Language Processing Group, New University of Lisbon (Portugal)
- Spoken Language Systems Lab, Technical University of Lisbon (Portugal)

International research collaborators:

- Department of Computer Science, University of Helsinki (Finland)
- Department of Computer Science and Engineering, University of North Texas (USA)
- Department of Management Sciences, University of Waterloo (Canada)
- Department of Software and Computing Systems, University of Alicante (Spain)
- Ecole Supérieure d'Ingénieurs en Informatique et Génie des Télécommunications (France)
- Ecole des Mines de Paris (France)
- Faculty of Arts, University of Ljubljana (Slovenia)
- Faculty of Mathematics and Computer Science, University of Plovdiv (Bulgaria)
- Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (France)
- Institut de Recherche en Informatique et Systemes Aléatoires (France)
- Laboratoire d'Informatique Fondamentale d'Orléans (France)
- MIT Media Laboratory (USA)

- Research Group of Computer Linguistics, University of Tartu (Estonia)
- School of Humanities, Languages and Social Studies, University of Wolverhampton (UK)

Industrial research partners:

- Microsoft (Portugal)
- Imaxin.com (Spain)

Bibliography

[Agrawal 1996] R. Agrawal, H. Mannila, Srikant. R., H. Toivonen and A.I. Verkamo. *Fast Discovery of Association Rules.* Advances in Knowledge Discovery and Data Mining, 1996. 14

[Ahonen-Myka 1999] H. Ahonen-Myka. *Finding All Frequent Maximal Sequences in Text.* In Proceedings of the Workshop on Machine Learning in Text Data Analysis of the 16th International Conference of Machine Learning (ICML 1999), pages 11–17, 1999. 144

[Aitken 1926] A.C. Aitken. *On Bernoulli's Numerical Solution of Algebraic Equations.* Research Society Edinburgh, vol. 46, pages 289–305, 1926. 85

[Alonso 2009] O. Alonso, M. Gertz and R. Baeza-Yates. *Clustering and Exploring Search Results Using Timeline Constructions.* In Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), pages 97–106, 2009. 198

[Anderson 1998] A. Anderson and S. Nilsson. *Implementing Radixsort.* ACM Journal of Experimental Algorithmics, vol. 3, pages 156–165, 1998. 43

[Angheluta 2002] R. Angheluta, R. De Busser and M-F. Moens. *The Use of Topic Segmentation for Automatic Summarization.* In Proceedings of the Workshop on Text Summarization of the 40st Annual Meeting of the Association for Computational Linguistics (ACL 2002), pages 11–12, 2002. 94, 96, 101

[Anita 2002] H.M. Anita. Numerical methods for scientists and engineers. Birkhäuser Verlag, 2002. 117

[Aone 1999] C. Aone, M. E. Okurowski, J. Gorlinsky and B. Larsen. *A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques.* In Advances in Automatic Text Summarization, pages 71–80. 1999. 93

[Aue 2005] A. Aue and M. Gamon. *Customizing Sentiment Classifiers to New Domains: a Case Study.* In Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP 2005), pages 207–218, 2005. 174, 177, 186

[Banea 2008] C. Banea, R. Mihalcea, J. Wiebe and S. Hassan. *Multilingual Subjectivity Analysis Using Machine Translation.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pages 127–135, 2008. 179, 182

[Banerjee 2002] S. Banerjee and T. Pedersen. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*. In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2002), pages 136–145, 2002. 24

[Bannard 2003] C. Bannard, T. Baldwin and A. Lascarides. *A Statistical Approach to the Semantics of Verb Particles*. pages 65–72, 2003. 36

[Bannard 2007] C. Bannard. *A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora*. In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pages 1–8, 2007. 35, 36

[Barzilay 1997] R. Barzilay and M. Elhadad. *Using Lexical Chains for Text Summarization*. In Proceedings of the Workshop on Intelligent Scalable Text Summarization of the Joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL 1997), pages 10–17, 1997. 93, 94, 95, 96, 103, 105, 106, 108, 109, 110, 130, 131, 195

[Barzilay 1999] R. Barzilay, K. McKeown and M. Elhadad. *Information Fusion in the Context of Multi-document Summarization*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), pages 550–557, 1999. 93

[Barzilay 2003a] R. Barzilay and N. Elhadad. *Sentence Alignment for Monolingual Comparable Corpora*. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pages 25–33, 2003. 73

[Barzilay 2003b] R. Barzilay and N. Elhadad. *Sentence Alignment for Monolingual Comparable Corpora*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pages 25–33, 2003. 115

[Barzilay 2003c] R. Barzilay and L. Lee. *Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment*. In Proceedings of the 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, (NAACL/HLT 2003), pages 16–23, 2003. 115, 117, 119, 120, 121, 144

[Bastos 2009] S. Bastos. Unsupervised learning of ontology in the medical domain. Master's thesis, University of Beira Interior, 2009. 146, 158, 168, 170, 196

[Bauer 1972] D.F. Bauer. *Constructing Confidence Sets Using Rank Statistics*. Journal of the American Statistical Association, vol. 67, pages 687–690, 1972. 185

[Baxendale 1958] P. Baxendale. *Machine-made Index for Technical Literature - An Experiment.* IBM Journal of Research Development, vol. 2, no. 4, pages 354–361, 1958. 93

[Bécue 1993] M. Bécue and P. Peiro. *Les Quasi-segments pour une Classification Automatique des Réponses Ouvertes.* 1993. 16

[Beeferman 1997] D. Beeferman, A. Berger and J. Lafferty. *Text Segmentation using Exponential Models.* In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), pages 35–46, 1997. 94, 97, 100, 101, 131

[Beesley 1988] K.B. Beesley. *Language Identifier: a Computer Program for Automatic Natural-Language Identification of On-line Text.* In Proceedings of the 29th Annual Conference of the American Translators Association (ATA 1988), pages 47–54, 1988. 75

[Belmandt 1993] Z. Belmandt. Manuel de prétopologie et ses applications. Hermes Sciences Publications, 1993. 160

[Bentley 1993] J. Bentley and D. McIlroy. *Engineering a Sort Function.* Software - Practice and Experience, vol. 23, no. 11, pages 1249–1265, 1993. 43

[Bentley 1997] J. Bentley and R. Sedgewick. *Fast Algorithms for Sorting and Searching Strings.* In Proceedings of the 8th Annual ACM Symposium on Discrete Algorithms (SIAM 1997), pages 360–369, 1997. 43

[Bhogal 2007] J. Bhogal, A. Macfarlane and P. Smith. Information Processing and Management, vol. 43, no. 4, pages 866–886, 2007. 135

[Bhowmick 2010] P.K. Bhowmick, S. Sarkar and A. Basu. *Ontology based user modelling for Personalized Information Access.* International Journal of Computer Science and Applications, vol. 7, no. 1, pages 1–22, 2010. 199, 200

[Biemann 2005] C. Biemann. *Ontology Learning from Text: a Survey of Methods.* Journal for Language Technology and Computational Linguistics (LDV Forum), vol. 20, no. 2, pages 75–93, 2005. 135, 136, 137

[Bilal 2005] K. Bilal, U. Muhammad, A. Khan and M.N. Khan. *Extracting Multiword Expressions in Machine Translation from English to Urdu using Relational Data Approach.* World Academy of Science, Engineering and Technology, vol. 12, 2005. 31

[Blitzer 2007] J. Blitzer, M. Dredze and F. Pereira. *Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification.* In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), pages 187–205, 2007. 178, 188

[Blum 1998] A. Blum and T. Mitchell. *Combining Labeled and Unlabeled Data with Co-training.* In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT 1998), pages 92–100, 1998. 179, 188, 190

[Boguraev 2000] B.K. Boguraev and M.S. Neff. *Discourse Segmentation in Aid of Document Summarization.* In Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS 2000), page 3004, 2000. 94, 96

[Boiy 2007] E. Boiy, P. Hens, K. Deschacht and M-F. Moens. *Automatic Sentiment Analysis of On-line Text.* In Proceedings of the 11th International Conference on Electronic Publishing (ELPUB 2007), pages 349–360, 2007. 174, 177, 178, 182, 183, 186, 191

[Bollegala 2007] D. Bollegala, Y. Matsuo and M. Ishizuka. *Measuring Semantic Similarity Between Words Using Web Search Engines.* In roceedings of the 16th International Conference on World Wide Web (WWW 2007), pages 757–766, 2007. 12, 25, 151

[Bourigault 1993] D. Bourigault. *Analyse Syntaxique Locale pour le Repérage de Termes Complexes dans un Texte.* Traitement Automatique des Langues, vol. 34, no. 2, 1993. 34

[Brants 2000] T. Brants. *TnT - a Statistical Part-of-Speech Tagger.* In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000), pages 224–231, 2000. 104

[Brefeld 2005] U. Brefeld, C. Büscher and T. Scheffer. *Multi-view Discriminative Sequential Learning.* In Proceedings of the 16th European Conference on Machine Learning (ECML 2005), pages 60–71, 2005. 188

[Brockett 2005] C. Brockett and W.B. Dolan. *Support Vector Machines for Paraphrase Identification and Corpus Construction.* In Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005), 2005. 115

[Budanitsky 2006] A. Budanitsky and G. Hirst. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness.* Computational Linguistics, vol. 32, no. 1, pages 13–47, 2006. 109

[Buitelaar 2005] P. Buitelaar, P. Cimiano and B. Magnini, editeurs. Ontology learning from text: Methods, evaluation and applications, volume 123 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, 2005. 135

[Buyukkokten 2001] O. Buyukkokten, H. Garcia-Molina and A. Paepcke. *Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices.* In Proceedings of the 10th International Conference on World Wide Web (WWW 2001), pages 652–662, 2001. 129

[Cai 2004] S. Cai, S. Yu, J-R. Wen and W-Y Ma. *Block-based Web Search.* In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pages 456–463, 2004. 129, 195

[Camacho 1994] R. Camacho. *Learning Stage Transition Rules with Indlog.* Gesellschaft für Mathematik und Datenverarbeitung MBH, vol. 237, pages 273–290, 1994. 125

[Campos 2005] R. Campos and G. Dias. *Automatic Hierarchical Clustering of Web Pages.* In Proceedings of the ELECTRA Workshop of the 28th Annual International ACM SIGIR Conference (SIGIR 2005), pages 83–85, 2005. 33, 41, 44, 62, 63, 66, 89

[Campos 2008] R. Campos, G. Dias, C. Nunes and B. Nonchev. *Clustering of Web Page Search Results: A Full Text Based Approach.* International Journal of Computer and Information Science, vol. 9, no. 4, 2008. 63, 64, 72

[Caraballo 1999] S.A. Caraballo. *Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text.* In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pages 120–126, 1999. 11, 104, 135, 139, 150, 151

[Carenini 2008] G. Carenini and J.C.K. Cheung. *Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: the Effect of Corpus Controversiality.* In Proceedings of the 5th International Natural Language Generation Conference (INLG 2008), pages 33–41, 2008. 93

[Carpineto 2004] C. Carpineto and G. Romano. *Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO.* Journal of the Universal Computer Science, vol. 10, no. 8, pages 985–1013, 2004. 62, 70, 89

[Carpineto 2009] C. Carpineto, S. Osinski, G. Romano and D. Weiss. *A Survey of Web Clustering Engines.* ACM Computing Surveys, vol. 41, no. 3, pages 1–38, 2009. 61

[Challam 2007] V. Challam, S. Gauch and A. Chandramouli. *Contextual Information Retrieval Using Ontology Based User Profiles.* In Proceedings of the 9th triennial RIAO Conference: Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2007), pages 1–6, 2007. 199, 200

[Chandrasekar 1997] R. Chandrasekar and B. Srinivas. *Automatic Induction of Rules for Text Simplification.* Knowledge-Based Systems, vol. 10, pages 183–190, 1997. 111

[Charles 2000] W.G. Charles. *Contextual Correlates of Meaning.* Applied Psycholinguistics, vol. 21, no. 4, pages 505–524, 2000. 143

[Chartron 1988] G. Chartron. *Analyse des Corpus de Données Textuelles, Sondage de Flux d'Information.* PhD thesis, Université Paris 7, Paris, 1988. 16, 35

[Chesley 2006] P. Chesley, B. Vincent, L. Xu and R. Srihari. *Using Verbs and Adjectives to Automatically Classify Blog Sentiment.* In Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI/CAAW 2006), pages 27–29, 2006. 177, 179, 180, 182, 183, 184

[Choi 2000] F.Y.Y. Choi. *Advances in Domain Independent Linear Text Segmentation.* In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000), pages 26–33, 2000. 94, 97, 98, 101, 102

[Choueka 1983] Y. Choueka, T. Klein and E. Neuwitz. *Automatic Retrieval of Frequent Idiomatic and Collocation Expressions in a Large Corpus.* Journal for Literary and Linguistic Computing, vol. 4, pages 34–38, 1983. 34

[Church 1990] K. Church and P. Hanks. *Word Association Norms Mutual Information and Lexicography.* Computational Linguistics, vol. 16, no. 1, pages 23–29, 1990. 13, 14, 35

[Cicurel 2006] L. Cicurel, S. Bloehdorn and P. Cimiano. *Clustering of Polysemic Words.* In Proceedings of the 30th Annual Conference of the German Classification Society (GCS 2006), pages 595–602, 2006. 71, 74, 105

[Cilibrasi 2007] R.L. Cilibrasi and P.M.B. Vitanyi. *The Google Similarity Distance.* IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pages 370–383, 2007. 13

[Cimiano 2004] P. Cimiano, A. Hotho and S. Staab. *Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text.* In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), pages 435–439, 2004. 135

[Cimiano 2005] P. Cimiano, A. Hotho and S. Staab. *Learning Concept Hierarchies from Text Corpora using Formal Concept Anaylsis.* Journal of Artificial Intelligence Research, vol. 24, pages 305–339, 2005. 135, 137, 151

[Cimiano 2009] P. Cimiano, A. Mädche, S. Staab and J. Völker. *Ontology Learning.* In Handbook of Ontologies, pages 245–267. Springer Verlag, 2009. 135

[Clarke 2006] J. Clarke and M. Lapata. *Constraint-based Sentence Compression an Integer Programming Approach.* In Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006), pages 144–151, 2006. 111, 113, 128

[Cleuziou 2003] G. Cleuziou, L. Martin and C. Vrain. *PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data.* In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2003), pages 440–444, 2003. 63, 66, 71, 95, 168, 195

[Cleuziou 2008] G. Cleuziou and G. Dias. *Apprentissage de Mesures de Similarité Sémantiques: Etude d'une Variante de la Mesure InfoSimba.* In Proceedings of the 1st Joint Meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (SFC-CLADAG 2008), pages 233–236, 2008. 10, 22, 29

[Cleuziou 2009] G. Cleuziou, M. Exbrayat, L. Martin and J-H. Sublemontier. *CoFKM : a Centralized Method for Multiple-View Clustering.* In Proceedings of the 9th IEEE International Conference on Data Mining (ICDM 2009), pages 752–757, 2009. 170, 197

[Cleuziou 2010] G. Cleuziou, G. Dias and V. Levorato. *Modélisation Prétopologique pour la Structuration Sémantico-Lexicale.* In Proceedings of the 17èmes Rencontres de la Société Francophone de Classification (SFC 2010), 2010. 26, 132, 139, 159, 197

[Collins 1999] M. Collins and Y. Singer. *Unsupervised Models for Named Entity Classification.* In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP 1999), pages 100–110, 1999. 188

[Conroy 2001] J. M. Conroy and D. P. O'leary. *Text Summarization via Hidden Markov Models.* In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 406–407, 2001. 93

[Cooper 1970] L. Cooper and D. Steinberg. Introduction to methods of optimization. W.B. Saunders, 1970. 49

[Cordeiro 2007a] J.P. Cordeiro, G. Dias and P. Brazdil. *New Functions for Unsupervised Asymmetrical Paraphrase Detection.* Journal of Software, 2007. 65, 74, 96, 114, 115, 117, 118

[Cordeiro 2007b] J.P. Cordeiro, G. Dias, G. Cleuziou and P. Brazdil. *Biology Based Alignments of Paraphrases for Sentence Compression.* In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrase of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), 2007. 96, 119, 121, 122, 143, 144, 149

[Cordeiro 2009] J.P. Cordeiro, G. Dias and P. Brazdil. *Unsupervised Induction of Sentence Compression Rules.* In Proceedings of the Workshop on Language Generation and Summarisation of the Joint Conference of the 47th Annual

Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP 2009), 2009. 96, 121, 124, 126

[Croft 1980] W.B. Croft. *A Model of Cluster Searching based on Classiffication.* Information System, vol. 5, pages 189–195, 1980. 59

[Daelemans 2004] W. Daelemans, A. Höthker and E. Tjong Kim Sang. *Automatic Sentence Simplification for Subtitling in Dutch and English.* In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 1045–1048, 2004. 110, 111, 113

[Dagan 1994] I. Dagan. *Termight: Identifying and Translating Technical Terminology.* pages 34–40, 1994. 34

[Daille 1996] B. Daille. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology.* The balancing act combining symbolic and statistical approaches to language, pages 49–66, 1996. 35, 75, 104

[Daoud 2010] M. Daoud, L. Tamine and M. Boughanem. *A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile.* In Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2010), pages 171–182, 2010. 199, 200

[David 1990] S. David and P. Plante. *Termino Version 1.0.* Rapport technique, Centre d'Analyse de Textes par Ordinateur, Université du Québec, Canada, 1990. 34

[De Jong 1975] K.A. De Jong. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems.* PhD thesis, 1975. 50

[Debusmann 2004] R. Debusmann. *Multiword Expressions as Dependency Subgraphs.* In Proceedings of the 2nd Workshop on Multiword Expressions of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 56–63, 2004. 34

[Deligne 1995] S. Deligne and F. Bimbot. *Language Modelling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams.* In Proceedings of the 20th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995), pages 169–172, 1995. 42

[Dellschaft 2006] K. Dellschaft and S. Staab. *On How to Perform a Gold Standard Based Evaluation of Ontology Learning.* In Proceedings of the 5th International Semantic Web Conference (ISWC 2006), pages 228–241, 2006. 163

[Demonet 1975] M. Demonet, A. Geoffroy, J. Gouaze, P. Lafon, M. Mouillaud and M. Tournier. Des tracts en mai 68. mesures de vocabulaire et de contenu. 1975. 13

[Dempster 1977] A.P. Dempster, N.M. Laird and D.B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, vol. 39, no. 1, pages 1–38, 1977. 119

[Dias 1999a] G. Dias, S. Guilloré and J.G.P. Lopes. *Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora.* In Proceedings of the 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999), pages 333–339, 1999. 10, 14, 17, 32, 35, 36, 38, 63, 70, 94

[Dias 1999b] G. Dias, S. Guilloré and J.G.P. Lopes. *Multilingual Aspects of Multiword Lexical Units.* In Proceedings of Workshop on Language Technologies of the 32th annual meeting fo the Societas Linguistica Europea (SLE 1999), pages 11–21, 1999. 39

[Dias 1999c] G. Dias, S. Guilloré, S. Vintar and J.G.P. Lopes. *Identifying and Integrating Terminologically Relevant Multiword Units in the IJS-ELAN Slovene-English Parallel Corpus.* In Selected Papers of 10th Computational Linguistics In the Netherlands (CLIN 1999), pages 29–40, 1999. 39

[Dias 2000a] G. Dias, S. Guilloré, J.C. Bassano and J.G.P. Lopes. *Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?* In Proceedings of the 6ème Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO 2000), pages 1–20, 2000. 53, 55, 65, 75, 95

[Dias 2000b] G. Dias, S. Guilloré and J.G.P. Lopes. *Benefiting from Multi-domain Corpora for Extracting Terminologically Relevant Multiword Lexical Units.* In Proceedings of the 9th EURALEX International Congress (EURALEX 2000), pages 339–350, 2000. 49

[Dias 2000c] G. Dias, S. Guilloré and J.G.P. Lopes. *Extraction Automatique d'Associations Textuelles à Partir de Corpora Non Traités.* In Proceedings of the 5èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2000), pages 213–221, 2000. 9, 33, 193

[Dias 2000d] G. Dias, S. Guilloré and J.G.P. Lopes. *Mining Textual Associations in Text Corpora.* In Proceedings of the Workshop on Text Mining associated to the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000), pages 92–95, 2000. 9, 33, 193

[Dias 2000e] G. Dias, S. Guilloré and J.G.P. Lopes. *Normalisation of Association Measures for Multiword Lexical Unit Extraction.* In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA 2000), pages 207–216, 2000. 10, 16, 35, 39

[Dias 2001a] G. Dias, H. Kaalep and K. Muischnek. *Automatic Extraction of Verb Phrases from Annotated Corpora: A Linguistic Evaluation for Estonian.* In

Proceedings of the Workshop on Collocation of the joint 39th Annual Meeting of the Association of Computational Linguistics and 10th Conference of the European Chapter of the Association of Computational Linguistics (EACL/ACL 2001), 2001. 35, 36, 39

[Dias 2001b] G. Dias and S. Nunes. *Combining Evolutionary Computing and Similarity Measures to Extract Collocations from Unrestricted Texts.* In Proceedings of the International Conference On Recent Advances in Natural Language Processing (RANLP 2001), pages 261–264, 2001. 33, 37, 47, 51

[Dias 2001c] G. Dias and S. Nunes. *Does Natural Selection Apply to Natural Language Processing? An Experiment for Multiword Unit Extraction.* In Proceedings of the Natural Language Processing Knowledge Engineering Workshop of the 2001 IEEE Systems, Man, and Cybernetics Conference (IEEE SMC 2001), pages 1–8, 2001. 38, 51

[Dias 2002] G. Dias. *Extraction Automatique d'Associations Lexicales à Partir de Corpora.* PhD thesis, Univeristy of Orléans and New University of Lisbon, 2002. 6, 15, 32, 36, 38, 39, 40, 41

[Dias 2003] G. Dias. *Multiword Unit Hybrid Extraction.* In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003), pages 41–49, 2003. 33, 36, 52, 53, 55

[Dias 2004a] G. Dias, S. Madeira and H. Veiga. *Webspy.* Rapport technique, University of Beira Interior, 2004. 66, 70

[Dias 2004b] G. Dias and S. Nunes. *Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment.* In Proceedings of the 4th International Conference On Languages Resources and Evaluation (LREC 2004), pages 1717–1721, 2004. 51

[Dias 2005a] G. Dias and E. Alves. *Discovering Topic Boundaries for Text Summarization based on Word Co-occurrence.* In Proceedings of the International Conference On Recent Advances in Natural Language Processing (RANLP 2005), pages 187–191, 2005. 10

[Dias 2005b] G. Dias and E. Alves. *Unsupervised Topic Segmentation Based on Word Co-occurrence and Multi-Word Units for Text Summarization.* In Proceedings of the ELECTRA Workshop of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pages 41–48, 2005. 101

[Dias 2005c] G. Dias and S. Vintar. *Unsupervised Learning of Multiword Units from Part-of-Speech Tagged Corpora: Does Quantity means Quality?* In Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005), pages 669–680, 2005. 35, 41, 53, 54

[Dias 2006a] G. Dias and B. Conde. *Efficient Text Summarization for Web Browsing On Mobile Devices.* In Proceedings of the Workshop on Ubiquitous User Modeling of the 17th European Conference on Artificial Intelligence (ECAI 2006), pages 9–12, 2006. 91, 129, 195

[Dias 2006b] G. Dias, C. Santos and G. Cleuziou. *Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining.* In Proceedings of the Workshop on Information Extraction Beyond the Document of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006), pages 36–47, 2006. 11, 17, 21, 29, 31, 36, 41, 45, 55, 94, 95, 104, 109, 139, 151

[Dias 2006c] G. Dias, S. Sousa and M. Crochemore. Special issue of the journal traitement automatique des langues on scaling natural language processing: Complexity, algorithms and architectures. Hermés Science Publications, Paris, 2006. 42

[Dias 2007a] G. Dias, E. Alves and J.G.P. Lopes. *Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation.* In Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007), pages 1334–1340, 2007. 10, 11, 29, 31, 41, 45, 55, 94, 96, 98, 101, 130

[Dias 2007b] G. Dias and B. Conde. *Accessing the Web on Handheld Devices for Visually Impaired People.* In Proceedings of the 5th Atlantic Web Intelligence Conference (AWIC 2007), pages 80–87, 2007. 91, 129, 195

[Dias 2008] G. Dias, R. Mukelov and G. Cleuziou. *Unsupervised Graph-Based Discovery of General-Specific Noun Relationships from Web Corpora Frequency Counts.* In Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008), 2008. 26, 27, 28, 132, 135, 139, 156, 162

[Dias 2009] G. Dias, S. Pais, F. Cunha, H. Costa, D. Machado, T. Barbosa and B. Martins. *Hierarchical Soft Clustering and Automatic Text Summarization for Accessing the Web on Mobile Devices for Visually Impaired People.* In Proceedings of the 22nd Florida Artificial Intelligence Research Society Conference (FLAIRS 2009), 2009. 31, 33, 41, 44, 55, 62, 64, 89, 91, 129, 195

[Dias 2010] G. Dias, R. Moraliyski, J.P. Cordeiro, A. Doucet and H. Ahonen-Myka. *Automatic Discovery of Word Semantic Relations using Paraphrase Alignment and Distributional Lexical Semantics Analysis.* Journal of Natural Language Engineering, 2010. 29, 32, 56, 120, 121, 132, 138, 139, 142, 143, 170

[Dice 1945] L. Dice. *Measures of the Amount of Ecologic Association between Species.* Journal of Ecology, 1945. 14

[Dolan 2004] W.B. Dolan, C. Quirk and C. Brockett. *Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources.* In Proceedings of 20th International Conference on Computational Linguistics (COLING 2004), 2004. 115, 116

[Dunning 1993] T Dunning. *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, vol. 19, no. 1, 1993. 14, 15, 16, 35, 149

[Edmundson 1969] H.P. Edmundson. *New Methods in Automatic Extracting.* Jornal of ACM, vol. 16, no. 2, pages 264–285, 1969. 93, 129

[Ehlert 2003] B. Ehlert. Making accurate lexical semantic similarity judgments using word-context cooccurrence statistics. Master's thesis, University of California, 2003. 20, 137, 140, 141, 142, 143

[Enguehard 1993] C. Enguehard. *Acquisition de Terminologie à partir de Gros Corpus.* Informatique et Langue Naturelle, pages 373–384, 1993. 35

[Eppstein 1999] D. Eppstein, Z. Galil and G.F. Italiano. *Dynamic Graph Algorithms.* In Algorithms and Theory of Computation Handbook, chapitre 8. CRC Press, 1999. 160

[Erjavec 2002] T. Erjavec. *The IJS-ELAN Slovene-English Parallel Corpus.* International Journal of Corpus Linguistics, vol. 7, no. 1, pages 1–20, 2002. 54

[Esuli 2005] A. Esuli and F. Sebastiani. *Determining the Semantic Orientation of Terms through Gloss Classification.* In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005), pages 617–624, 2005. 176

[Fano 1961] R. Fano. Transmission of information: A statistical theory of communications. 1961. 13

[Farzindar 2004] A. Farzindar and G. Lapalme. *Legal Text Summarization by Exploration of the Thematic Structures and Argumentative Roles.* In Proceedings of the Workshop on Text Summarization Branches Out of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 27–38, 2004. 94, 96

[Fazly 2007] A. Fazly and S. Stevenson. *Distinguishing Subtypes of Multiword Expressions using Linguistically-motivated Statistical Measures.* In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions of the 45st Annual Meeting of the Association of Computational Linguistics (ACL 2007), pages 9–16, 2007. 35

[Feldman 1998] R. Feldman, M. Fresko, Y. Kinar, Y Lindell, O. Liphstat, M. Rajman, Y. Schler and O. Zamir. *Text Mining at the Term Level.* pages 65–73, 1998. 35

[Fellbaum 1998] C.D. Fellbaum. Wordnet: An electronic lexical database. MIT Press, 1998. 104

[Ferragina 2008] P. Ferragina and A. Gulli. *A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering.* Software: Practice and Experience, vol. 38, no. 2, pages 189–225, 2008. 62, 65, 69, 75, 77, 89

[Ferret 2002] O. Ferret. *Using Collocations for Topic Segmentation and Link Detection.* In Proceedings of the 9th International Conference on Computational Linguistics (COLING 2002), pages 1–7, 2002. 98, 101

[Filatova 2004] E. Filatova and V. Hatzivassiloglou. *Event-Based Extractive Summarization.* In Proceedings of the Workshop on Text Summarization Branches Out of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 104–111, 2004. 93

[Finley 2002] S.H. Finley and S.M. Harabagiu. *Generating Single and Multi-Document Summaries with GISTEXTER.* In Proceedings of the Workshop on Automatic Summarization, pages 30–38, 2002. 93

[Finn 2006] A. Finn and N. Kushmerick. *Learning to Classify Documents According to Genre.* American Society for Information Science and Technology, Special issue on Computational Analysis of Style, vol. 57, no. 11, pages 1506–1518, 2006. 175, 177, 182, 186

[Firth 1957] J.R. Firth. *A Synopsis of Linguistic Theory.* Studies in Linguistic Analysis, pages 1–32, 1957. 17

[Fotzo 2004] H.N. Fotzo and P. Gallinari. *Learning "Generalization/Specialization" Relations between Concepts: Application for Automatically Building Thematic Document Hierarchies.* In Proceedings of the 7th triennial RIAO Conference: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval (RIAO 2004), pages 143–155, 2004. 159, 162

[Frantzi 1996] K. Frantzi and S. Ananiadou. *Extracting Nested Collocations.* pages 41–46, 1996. 16, 35, 49, 78

[Freitag 2005] D. Freitag, M. Blume, J. Byrnes, E. Chow, S. Kapadia, R. Rohwer and Z. Wang. *New Experiments in Distributional Representations of Synonymy.* In Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005), pages 25–32, 2005. 17, 20, 27, 28, 29, 137, 138, 140, 142, 147

[Fung 2003] B.C.M. Fung, K. Wang and M. Ester. *Hierarchical Document Clustering using Frequent Itemsets.* In Proceedings of the 3rd SIAM International Conference on Data Mining (SDM 2003), pages 59–70, 2003. 62, 77, 89

[Gale 1991] W. Gale. *Concordances for Parallel Texts.* In Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research, Using Corpora, 1991. 14, 15, 35

[Gale 1992] W.A. Gale, K.W. Church and D. Yarowsky. *One sense per discourse.* In Proceedings of the Workshop on Speech and Natural Language of the Human Language Technology Conference (HLT 1991), pages 233–237, 1992. 105

[Galley 2003] M. Galley and K. McKeown. *Improving Word Sense Disambiguation in Lexical Chaining.* In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), pages 1486–1488, 2003. 103, 106, 110

[Ganchev 2008] K. Ganchev, J. Graça, J. Blitzer and B. Taskar. *Multi-View Learning over Structured and Non-Identical Outputs.* In Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008), pages 204–211, 2008. 175, 178, 188, 189, 191, 197

[Ganesan 2010] K. Ganesan, C. Zhai and J. Han. *Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions.* In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 2010. 93

[Ganter 1999] B. Ganter and R. Wille. Formal concept analysis: Mathematical foundations. Springer Verlag, 1999. 151, 201

[Gil 2003a] A. Gil and G. Dias. *Efficient Mining of Textual Associations.* In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2003), pages 549–555, 2003. 42, 44, 70

[Gil 2003b] A. Gil and G. Dias. *Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora.* In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003), pages 25–33, 2003. 33, 42, 44, 70

[Goldberg 1989] D.E. Goldberg. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, 1989. 48

[Golemati 2007] M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras and C. Halatsis. *Creating an Ontology for the User Profile: Method and Applications.* In Proceedings of the 1st International Conference on Research Challenges in Information Science (RCIS 2007), pages 1–7, 2007. 199, 200

[Grefenstette 1992] G. Grefenstette. *Use of Syntactic Context to Produce term Association Lists for Text Retrieval.* In Proceedings of the 15th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 1992), pages 89–97, 1992. 18, 20

[Grefenstette 1993] G. Grefenstette. *Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques.* In Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research, 1993. 18, 32, 137

[Grigonyté 2010] G. Grigonyté, J.P. Cordeiro, R. Moraliyski, G. Dias and P. Brazdil. *A Paraphrase Alignment for Synonym Evidence Discovery.* In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 2010. 30, 31, 121, 132, 143, 144, 148, 149, 196

[Gross 1996] G. Gross. Les expressions figées en français. Ophrys, Paris, 1996. 32, 48, 52, 53, 56

[Habert 1993] B. Habert and C. Jacquemin. *Noms Composés, Termes, Dénominations Complexes: Problématiques Linguistiques et Traitements Automatiques.* Traitement Automatique des Langues, vol. 34, no. 2, pages 5–41, 1993. 33, 52

[Hahn 2000] U. Hahn and I. Mani. *The Challenges of Automatic Summarization.* IEEE Computer, vol. 33, no. 11, pages 29–36, 2000. 92

[Halliday 1976] M.A.K. Halliday and R. Hasan. Cohesion in english. Longman, 1976. 102

[Harris 1968] Z. Harris. *Mathematical Structures of Language.* Wiley, 1968. 10, 11, 17, 135

[Hatzivassiloglou 1997] V. Hatzivassiloglou and K.R. McKeown. *Predicting the Semantic Orientation of Adjectives.* In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics (EACL 1997), pages 174–181, 1997. 176, 183

[Hatzivassiloglou 1999] V. Hatzivassiloglou, J.L. Klavans and E. Eskin. *Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning.* In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 1999), pages 203–212, 1999. 115

[Hatzivassiloglou 2000] V. Hatzivassiloglou and J. Wiebe. *Effects of Adjective Orientation and Gradability on Sentence Subjectivity.* In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), pages 299–305, 2000. 183

[Hearst 1992] M.A. Hearst. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proceedings of the 14th Conference on Computational linguistics, pages 539–545, 1992. 11, 151

[Hearst 1994] M.A. Hearst. *Multi-paragraph Segmentation of Expository Text*. In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994), pages 9–16, 1994. 94, 97, 98, 100, 101, 102

[Hearst 1996] M.A. Hearst and J.O. Pedersen. *Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results*. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996), pages 76–84, 1996. 61, 62, 63, 65, 66, 71, 73, 89

[Heim 1983] I. Heim. *File Change Semantics and the Familiarity Theory of Definiteness*. In Meaning, Use and Interpretation of Language, pages 164–189. 1983. 102

[Heyer 1999] L.J. Heyer, S. Kruglyak and S. Yooseph. *Exploring Expression Data: Identification and Analysis of Coexpressed Genes*. ACM Computing Surveys, vol. 9, pages 1106–1115, 1999. 64, 119, 144

[Heylen 2008] K. Heylen, Y. Peirsman, D. Geeraerts and D. Speelman. *Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms*. In Proceedings of the 6th International Language Resources and Evaluation (LREC 2008), pages 3243–3249, 2008. 29, 137, 138, 142

[Hindle 1990] D. Hindle. *Noun Classification from Predicate-Argument Structures*. In Proceedings of the 28th Annual Meeting on Association for Computational Linguistics (ACL 1990), pages 268–275, 1990. 17, 18, 104, 135, 137

[Hirst 1998a] G. Hirst and D. St-Onge. *Lexical Chains as Representation of Context for the Detection and Correction of Malapropisms*. In WordNet: An Electronic Lexical Database and some of its Applications, pages 1–24. 1998. 102, 103, 105, 110

[Hirst 1998b] G. Hirst and D. St-Onge. *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*. WordNet: An Electronic Lexical Database and Some of its Applications, 1998. 24, 195

[Hoare 1962] C.A.R. Hoare. *Quicksort*. Computer Journal, vol. 5, no. 1, pages 10–15, 1962. 43

[Hotho 2003] A. Hotho, S. Staab and G. Stumme. *Ontologies Improve Text Document Clustering*. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), pages 541–544, 2003. 135

[Jain 1999] A. Jain, M. Murty and P. Flynn. *Data Clustering: a Review*. ACM Computing Surveys, vol. 31, pages 264–323, 1999. 64, 119

[Jelinek 1991] F. Jelinek, B. Merialdo, S. Roukos and M. Strauss. *A Dynamic Language Model for Speech Recognition.* In Proceedings of the Workshop on Speech and Natural Language of the Human Language Technology Conference(HLT 1991), pages 293–295, 1991. 42

[Jiang 1997] J. Jiang and D. Conrath. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.* In Proceedings of the International Conference on REsearch in Computational Linguistics (ROCLING 1997), pages 19–33, 1997. 24

[Jiang 2002] Z. Jiang and A. Joshi. *Retriever Improving Web Search Engine Results using Clustering.* Managing Business with Electronic Commerce: Issues and Trends, pages 59–81, 2002. 61, 62, 67, 73, 89

[Jin 2008] P. Jin, J. Lian, X. Zhao and S. Wan. *TISE: A Temporal Search Engine for Web Contents.* In Proceedings of the 2008 Second International Symposium on Intelligent Information Technology Application (IITA 2008), pages 220–224, 2008. 198

[Jing 2000a] H. Jing. *Sentence Reduction for Automatic Text Summarization.* In Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000), pages 310–315, 2000. 95, 111, 112, 113

[Jing 2000b] H. Jing and K.R. McKeown. *Cut and Paste Based Text Summarization.* In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics (NAACL 2000), pages 178–185, 2000. 93, 110, 112

[Joachims 1998] T. Joachims. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* In Proceedings of 10th the European Conference on Machine Learning (ECML 1998), pages 137–142, 1998. 186

[Joachims 2002] T. Joachims. Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers, 2002. 186

[Justeson 1993] J. Justeson and S. Katz. *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text.* Rapport technique, IBM, 1993. 35, 49, 52, 53, 56, 75, 104

[Kamp 1995] H. Kamp. *Discourse Representation Theory.* In Handbook of Pragmatics, pages 253–257. 1995. 102

[Kaplan 1950] A. Kaplan. *An Experimental Study of Ambiguity and Context.* Mechanical Translation, vol. 2, no. 2, pages 39–46, 1950. 143

[Kaszkiel 1997] M. Kaszkiel and J. Zobel. *Passage Retrieval Revisited.* In Proceedings of the 20th Annual International ACM SIGIR Conference on Research

and Development in Information Retrieval (SIGIR 1997), pages 178–185, 1997. 96

[Kaszkiel 1999] M. Kaszkiel, J. Zobel and R. Sacks-Davis. *Efficient Passage Ranking for Document Databases.* ACM Transactions on Information Systems, vol. 17, pages 406–439, 1999. 96

[Kendall 1938] M. G. Kendall. *A New Measure of Rank Correlation.* Biometrika, vol. 30, no. 1-2, pages 81–93, 1938. 165

[Kim 1994] P.K. Kim. *Indexing Compound Words from Korean Texts using Mutual Information.* Journal of KISS, vol. 21, no. 7, pages 1333–1340, 1994. 35

[Kit 1998] C. Kit and Y. Wilks. *The Virtual Approach to Deriving Ngram Statistics from Large Scale Corpora.* In Proceedings of the International Conference on Chinese Information Processing, pages 223–229, 1998. 42, 43

[Knight 2002] K. Knight and D. Marcu. *Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression.* Artificial Intelligence, vol. 139, no. 1, pages 91–107, 2002. 95, 110, 111, 112, 113, 114, 115, 116, 127, 128

[Kozima 1993] H. Kozima. *Text Segmentation Based on Similarity between Words.* In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993), pages 286–288, 1993. 94, 97, 100

[Krovetz 1998] R. Krovetz. *More than One Sense per Discourse.* Rapport technique, NEC Princeton NJ Labs, 1998. 105

[Kruskal 1977] J. B. Kruskal and Wish M. Multidimensional scaling. Sage Publications, 1977. 185

[Kuhn 1994] T. Kuhn, H. Nieman and E.G. Schukat-Talamazzini. *Ergodic Hidden Markov Models and Polygrams for Language Modelling.* In Proceedings of the 15th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994), pages 357–360, 1994. 42

[Kullback 1951] S. Kullback and R.A. Leibler. *On Information and Sufficiency.* Annals of Mathematical Statistics, vol. 22, no. 1, pages 79–86, 1951. 28

[Kummamuru 2001] R. Kummamuru and R. Krishnapuram. *A Clustering Algorithm for Asymmetrically Related Data with its Applications to Text Mining.* In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM 2001), pages 571–573, 2001. 77, 89

[Kummamuru 2004] R. Kummamuru, R. Lotlikar, S. Roy, K. Singal and R. Krishnapuram. *A Hierarchical Monothetic Document Clustering Algorithm for*

Summarization and Browsing Search Results. In Proceedings of the 13th International Conference on the World Wide Web (WWW 2004), pages 658–665, 2004. 61, 62, 64, 65, 66, 73, 75, 76, 77, 89

[Kupiec 1995] J. Kupiec, J.O. Pedersen and F. Chen. *A Trainable Document Summarizer.* In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995), pages 68–73, 1995. 93

[Kuroda 2008] M. Kuroda, M. Sakakihara and Z. Geng. *Acceleration of the EM and ECM Algorithms using the Aitken δ^2 Method for Log-linear Models with Partially Classified Data.* Statistics & Probability Letters, vol. 78, no. 15, pages 2332–2338, 2008. 85

[Labbé 1988] D. Labbé, P. Thoiron and D. Serant. Etude sur la recherche et la structure lexicale. 1988. 13

[Lambov 2009] D. Lambov, G. Dias and V. Noncheva. *High Level Features for Learning Subjective Language.* In Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009), 2009. 175, 177, 178, 182

[Lambov 2010] D. Lambov, G. Dias and J.V. Graça. *Multi-view Learning for Text Subjectivity Classification.* In Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis of the 19th European Conference on Artificial Intelligence (ECAI 2010), 2010. 175, 179, 188

[Lan 1992] Y. Lan and M.A. Mohamed. *Parallel Quicksort in Hypercubes.* In Proceedings of the 1992 ACM Symposium on Applied Computing: Technological Challenges of the 1990's (SIGAPP 1992), pages 740–746, 1992. 45

[Landauer 1997] T.K. Landauer and S.T. Dumais. *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction and Representation of Knowledge.* Psychological Review, vol. 104, no. 2, pages 211–240, 1997. 17, 19, 20, 137, 140, 142, 147

[Largeron 2002] C. Largeron and S. Bonnevay. *A Pretopological Approach for Structural Analysis.* Information Sciences, vol. 144, pages 169–185, 2002. 161

[Lawrie 2003] D. Lawrie and B. Croft. *Generating Hierarchical Summaries for Web Searches.* In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), pages 457–458, 2003. 61, 62, 70, 74, 75, 77, 79, 89

[Le Nguyen 2004] M. Le Nguyen, S. Horiguchi, A. Shimazu and B.T. Ho. *Example-based Sentence Reduction using the Hidden Markov Model.* ACM Transactions on Asian Language Information Processing, vol. 3, no. 2, pages 146–158, 2004. 95, 110, 111, 114, 128

[Leacock 1998] C. Leacock and M. Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification.* WordNet: An Electronic Lexical Database and Some of its Applications, pages 265–283, 1998. 24

[Levenshtein 1966] V. Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.* Soviet Physice-Doklady, vol. 10, pages 707–710, 1966. 73, 115, 117

[Levin 1993] B. Levin. English verb classes and alternations. University of Chicago Press, 1993. 183

[Li 2008] M Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer-Verlag, New York, 2008. 13

[Likasa 2003] A. Likasa, Vlassis. N. and J. Verbeek. *The Global K-means Clustering Algorithm.* Pattern Recognition, vol. 36, pages 451–461, 2003. 81, 84

[Lin 1998a] D. Lin. *Automatic Retrieval and Clustering of Similar Words.* In Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998), pages 768–774, 1998. 19

[Lin 1998b] D. Lin. *An Information-Theoretic Definition of Similarity.* In Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), pages 296–304, 1998. 24, 106, 165

[Lin 2001] D. Lin and P. Pantel. *Discovery of Inference Rules for Question Answering.* Natural Language Engineering, vol. 7, pages 343–360, 2001. 11

[Lin 2004] C.Y. Lin. *ROUGE: a Package for Automatic Evaluation of Summaries.* In Proceedings of the Workshop on Text Summarization Branches Out of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 74–81, 2004. 74

[Liu 2004a] F. Liu, C. Yu and W. Meng. *Personalized Web Search For Improving Retrieval Effectiveness.* IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pages 28–40, 2004. 200

[Liu 2004b] H Liu. *MontyLingua: An End-to-End Natural Language Processor with Common Sense*, 2004. 141, 145, 187

[Liu 2007] M. Liu, W. Li, M. Wu and H. Hu. *Event-Based Extractive Summarization Using Event Semantic Relevance from External Linguistic Resource.* In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), pages 117–122, 2007. 93

[Lloyd 1982] S.P. Lloyd. *Least Squares Quantization in PCM.* IEEE Transactions on Information Theory, vol. 28, no. 2, pages 129–137, 1982. 81

[Lourenço 2009] M. Lourenço. Extracção de palavras compostas por bootstrapping. Master's thesis, University of Beira Interior, 2009. 40, 41, 45

[Luhn 1958] H.P. Luhn. *The Automatic Creation of Literature Abstracts.* IBM Journal of Research Development, vol. 2, no. 2, pages 159–165, 1958. 93, 129

[Lund 1995] K. Lund, C. Burgess and R.A. Atchley. *Semantic and Associative Priming in High Dimensional Semantic Space.* Cognitive Science, pages 660–665, 1995. 17, 19, 27, 137

[Maarek 2000] Y. Maarek, R. Fagin, I. Ben-Shaul and D. Pelleg. *Ephemeral Document Clustering for Web Applications.* Rapport technique, IBM, 2000. 59, 61, 62, 63, 71, 89

[Machado 2009a] D. Machado. Procura estruturada de textos para perfis de utilizadores. Master's thesis, University of Beira Interior, 2009. 200

[Machado 2009b] D. Machado, T. Barbosa, S. Pais, B Martins and G. Dias. *Universal Mobile Information Retrieval.* In Proceedings of the 13th International Conference on Human Computer Interaction (HCII 2009), 2009. 62, 77

[Maedche 2002] A. Maedche and S. Staab. *Measuring Similarity between Ontologies.* In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002), pages 251–263, 2002. 163

[Manber 1990] U. Manber and G. Myers. *Suffix-arrays: A New Method for Online String Searches.* In Proceedings of the 1st Annual ACM Symposium on Discrete Algorithms (SIAM 1990), pages 319–327, 1990. 43

[Mani 1998] I. Mani and E. Bloedorn. *Multi-document Summarization by Graph Search and Matching.* In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI 1997), pages 622–628, 1998. 92

[Mani 2001] I. Mani. Automatic summarization. John Benjamins Publishing, 2001. 92, 93

[Mann 1988] W.C. Mann and S.A. Thompson. *Rhetorical Structure Theory: Toward a Functional Theory of Text Organization.* Text, vol. 8, no. 3, pages 243–281, 1988. 93

[Manning 2008] C.D. Manning, P. Raghavan and H. Schütze. Introduction to information retrieval. Cambridge University Press, 2008. 74

[Marcu 1997] D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD thesis, University of Toronto, 1997. 93

[Marques 2001] N. Marques and J.G.P. Lopes. *Tagging With Small Training Corpora.* In Proceedings of the 4th International Symposium on Intelligent Data Analysis (IDA 2001), pages 63–72, 2001. 95

[Maynard 1999] D. Maynard and S. Ananiadou. *Identifying Contextual Information for Multi-Word Term Extraction.* In Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE 1999), pages 212–221, 1999. 35

[McCallum 2005] A. McCallum. *Information Extraction: Distilling Structured Data from Unstructured Text.* Queue, vol. 3, no. 9, pages 48–57, 2005. 5

[McGowan 2003] J.P. McGowan. A multiple model approach to. personalised information access. Master's thesis, University College Dublin, 2003. 200

[McKeown 1995] K. McKeown, J. Robin and K. Kukich. *Generating Concise Natural Language Summaries.* Information Processing and Management, vol. 31, no. 5, pages 703–733, 1995. 92

[McKeown 1999] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin. *Towards Multidocument Summarization by Reformulation: Progress and Prospects.* In Proceedings of the 16th National Conference on Artificial Intelligence (AAAI 1999), pages 453–460, 1999. 93

[McKeown 2003] K. McKeown, R. Barzilay, J. Chen, D. Elson, D. Evans, J. Klavans, A. Nenkova, B. Schiffman and S. Sigelman. *Columbia's Newsblaster: New Features and Future Directions.* In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003), pages 15–16, 2003. 91

[Medin 1990] D. Medin, L. Robert, L. Goldstone and D. Gentner. *Similarity Involving Attributes and Relations: Judgments of Similarity and Difference are not Inverses.* Psychological Science, vol. 1, no. 1, pages 64–69, 1990. 24

[Menéndez 1997] M.L. Menéndez, J.A. Pardo, L. Pardo and M.C. Pardo. *The Jensen-Shannon Divergence.* Journal of the Franklin Institute, vol. 334, no. 2, pages 307–318, 1997. 28

[Meyer 1975] D. Meyer, R. Schvaneveldt and M. Ruddy. *Loci of Contextual Effects on Visual Word Recognition.* Attention and Performance V, 1975. 13

[Micarelli 2004] A. Micarelli and F. Sciarrone. *Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System.* User Modeling and User-Adapted Interaction, vol. 14, no. 233, pages 159–200, 2004. 200

[Michalewicz 1996] Z. Michalewicz. Genetic algorithms + data structures = evolution programs. Springer, 1996. 48

[Michelbacher 2007] L. Michelbacher, S. Evert and H. Schütze. *Asymmetric Association Measures.* In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007), pages 1–6, 2007. 25, 26, 151

[Mihalcea 2004] R. Mihalcea and P. Tarau. *TextRank: Bringing Order into Texts.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), 2004. 27, 139, 151, 152

[Mihalcea 2007] R. Mihalcea and C. Banea. *Learning Multilingual Subjective Language via Cross-Lingual Projections.* In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pages 976–983, 2007. 179, 182

[Miller 1990] G. A. Miller. *WordNet: an On-line Lexical Database.* International Journal of Lexicography, vol. 3, no. 4, 1990. 11, 25, 95, 103, 137, 183

[Milligan 1985] G.W. Milligan and M.C. Cooper. *An Examination of Procedures for Determining the Number of Clusters in a Data Set.* Psychometrika, vol. 50, no. 2, pages 159–179, 1985. 84

[Min-Yen 2002] K. Min-Yen and K. McKeown. *Corpus-trained Text Generation for Summarization.* In Proceedings of the 2nd International Natural Language Generation Conference (INLG 2002), pages 1–8, 2002. 93

[Mithun 2010] S. Mithun. *Exploiting Rhetorical Relations in Blog Summarization.* In Advances in Artificial Intelligence, volume 6085, pages 388–392. 2010. 93

[Moens 2001] M-F. Moens and R. De Busser. *Generic Topic Segmentation of Document Texts.* In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 418–419, 2001. 94, 101

[Moralyiski 2007] R. Moralyiski and G. Dias. *One Sense per Discourse for Synonym Detection.* In Proceedings of the International Conference On Recent Advances in Natural Language Processing (RANLP 2007), pages 383–387, 2007. 132, 138, 140, 143, 146

[Morris 1991] J. Morris and G. Hirst. *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text.* Computational Linguistics, vol. 17, no. 1, pages 21–43, 1991. 93, 94, 97, 102, 103

[Muggleton 1991] S. Muggleton. *Inductive logic programming.* New Generation Computing, vol. 8, no. 4, pages 295–318, 1991. 96, 124

[Muggleton 1996] S. Muggleton. *Learning From Positive Data.* pages 358–376, 1996. 124, 126

[Muggleton 1999] S. Muggleton. *Inductive Logic Programming: Issues, Results and the Challenge of Learning Language in Logic.* Artificial Intelligence, vol. 114, no. 1-2, pages 283–296, 1999. 125

[Muskens 1996] R. Muskens. *Combining Montague Semantics and Discourse Representation.* Linguistics and Philosophy, vol. 19, pages 143–186, 1996. 102

[Mylonas 2008] P. Mylonas, D. Vallet, P. Castells, M. Fernandez and Y. Avrithis. *Personalized Information Retrieval Based on Context and Ontological Knowledge.* Knowledge Engineering, vol. 23, no. 1, pages 73–100, 2008. 135

[Needleman 1970] S. Needleman and C. Wunsch. *A General Method Applicable to the Search for Similarities in Amino Acid Sequence of Two Proteins.* National Academy of Sciences, no. 48, pages 443–453, 1970. 96, 121

[Nivre 2004] J. Nivre and J. Nilsson. *Multiword Units in Syntactic Parsing.* In Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 39–46, 2004. 31

[Notredame 2007] C. Notredame. *Recent Evolutions of Multiple Sequence Alignment Algorithms.* PLoS Computational Biology, vol. 3, no. 8, pages 1405–1408, 2007. 144

[Ohshima 2009] H. Ohshima and K. Tanaka. *Real Time Extraction of Related Terms by Bi-directional Lexico-syntactic Patterns from the Web.* In Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC 2009), pages 441–449, 2009. 12, 26, 151

[Ono 1994] K. Ono, K. Sumita and S. Miike. *Abstract Generation based on Rhetorical Structure Extraction.* In Proceedings of the 13th International Conference on Computational Linguistics (COLING 1994), pages 344–348, 1994. 93

[Osgood 1971] C.E. Osgood, G.J. Suci and P.H. Tannebaum. The measurement of meaning. University of Illinois Press, 1971. 183, 184

[Osinski 2004] S. Osinski, J. Stefanowski and D. Weiss. *Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition.* In Proceedings of the Intelligent Information Systems Conference (IIPWM 2004), pages 369–378, 2004. 62, 65, 66, 77, 89

[Paaß 2004] G. Paaß, J. Kindermann and E. Leopold. *Learning Prototype Ontologies by Hierarchical Latent Semantic Analysis.* In Proceedings of the Workshop on Knowledge Discovery and Ontologies of the Joint 15th European Conference on Machine Learning and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECMl/PKDD 2004), pages 1–12, 2004. 104, 135, 150

[Pais 2007] S. Pais. Classification of opinionated texts by analogy. Master's thesis, University of Beira Interior, 2007. 175, 178, 179, 180

[Pang 2002] B. Pang, L. Lee and S. Vaithyanathan. *Thumbs Up?: Sentiment Classification Using Machine Learning Techniques.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 79–86, 2002. 176, 186

[Pang 2004] B. Pang and L. Lee. *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.* In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 271–278, 2004. 177, 179, 184, 186

[Pantel 2002] P. Pantel and D. Lin. *Discovering Word Senses from Text.* In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), pages 613–619, 2002. 11, 71, 105

[Papineni 2002] K. Papineni, S. Roukos, T. Ward and W.J. Zhu. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), pages 311–318, 2002. 74, 117

[Pecina 2006] P. Pecina and P. Schlesinger. *Combining Association Measures for Collocation Extraction.* In Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006), pages 651–658, 2006. 13, 15, 16, 25, 26, 28, 33, 37, 47, 56, 151

[Pereira 1993] F. Pereira, N. Tishby and L. Lee. *Distributional Clustering of English Words.* In Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics (ACL 1993), pages 183–190, 1993. 104, 135, 136, 150

[Pereira 2004] R. Pereira, P. Crocker and G. Dias. *A Parallel Multikey Quicksort Algorithm for Mining Multiword Units.* In Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications of the 4th International Conference On Languages Resources and Evaluation (LREC 2004), pages 17–24, 2004. 33, 45, 47

[Petersen 2004] W. Petersen. *A Set-Theoretical Approach for the Induction of Inheritance Hierarchies.* Electronic Notes in Theoretical Computer Science, vol. 53, pages 296–308, 2004. 135, 150, 151

[Pevzner 2002] L. Pevzner and M. Hearst. *A Critique and Improvement of an Evaluation Metric for Text Segmentation.* Computational Linguistics, vol. 28, no. 1, pages 19–36, 2002. 101

[Phillips 1985] M. Phillips. Aspects of text structure: An investigation of the lexical organisation of text. North Holland Linguistic Series, North Holland, 1985. 97

[Piao 2003] S.S.L. Piao, P. Rayson, D. Archer, A. Wilson and T. McEnery. *Extracting Multiword Expressions with a Semantic Tagger.* In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003), pages 49–56, 2003. 36, 37

[Pon 2007] R.K. Pon, A.F. Cardenas, D.J. Buttler and T. J. Critchlow. *iScore: Measuring the Interestingness of Articles in a Limited User Environment.* In Proceedings of the IEEE Symposium on In Computational Intelligence and Data Mining (CIDM 2007), pages 354–361, 2007. 37

[Ponte 1997] J.M. Ponte and W.B. Croft. *Text Segmentation by Topic.* In Proceedings of the 1st European Conference on Research and Advanced Technology for Digitial Libraries (ECDL 1997), pages 120–129, 1997. 97, 98, 99, 100, 131

[Quinlan 1990] J.R. Quinlan. *Learning Logical Deinitions from Relations.* Machine Learning, vol. 5, no. 3, pages 239–266, 1990. 125

[Quinlan 1996] J.R. Quinlan. *Improved Use of Continuous Attributes in C4.5.* Journal of Artificial Intelligence Research, vol. 4, pages 77–90, 1996. 70

[Rada 1989] R. Rada, H. Mili, E. Bicknell and M. Blettner. *Development and Application of a Metric on Semantic Nets.* IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, pages 17–30, 1989. 24

[Radev 1998] D.R. Radev and K. McKeown. *Generating Natural Language Summaries from Multiple On-line Sources.* Computational Linguistics, vol. 24, no. 3, pages 469–500, 1998. 92

[Radev 2004] D.R. Radev, H. Jing, M. Stys and D. Tam. *Centroid-based Summarization of Multiple Documents.* Information Processing and Management, vol. 40, pages 919–938, 2004. 93

[Ramisch 2008] C. Ramisch, P. Schreiner, M. Idiart and A. Villavicencio. *An Evaluation of Methods for the Extraction of Multiword Expressions.* In Proceedings of the Workshop on Towards a Shared Task for Multiword Expressions of the 8th International Conference on Language Resources and Evaluation (LREC 2008), 2008. 36

[Rapp 2004] R. Rapp. *Utilizing the One-Sense-per-Discourse Constraint for Fully Unsupervised Word Sense Induction and Disambiguation.* In Proceedings of 4th Language Resources and Evaluation Conference (LREC 2004), pages 1–4, 2004. 140

[Resnik 1992] P. Resnik. *WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery.* In Proceedings of the AAAI Symposium on Probabilistic Approaches to Natural Language (AAAI 1992), pages 48–56, 1992. 24

[Resnik 1995] P. Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy.* In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), pages 448–453, 1995. 24

[Reynar 1994] J.C. Reynar. *An Automatic Method of Finding Topic Boundaries.* In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994), pages 331–333, 1994. 94, 97, 98

[Ribeiro 2001] A. Ribeiro, G. Dias, J.G.P. Lopes and J. Mexia. *Cognates Alignment.* In Proceedings of the Machine Translation Summit VIII (MT 2001), pages 287–292, 2001. 9, 33, 193

[Richardson 1995] R. Richardson and A. F. Smeaton. *Using WordNet in a Knowledge-Based Approach to Information Retrieval.* Rapport technique, Dublin City University, 1995. 24

[Riloff 1997] E. Riloff and J. Shepherd. *A Corpus-Based Approach for Building Semantic Lexicons.* In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), pages 117–124, 1997. 11, 151

[Rocchio 1971] J. Rocchio. *Relevance Feedback in Information Retrieval.* In The SMART Retrieval System: Experiments in Automatic Document Processing, chapitre 14, pages 313–323. Prentice-Hall, 1971. 179, 180

[Rodrigues 2009] D. Rodrigues. Construção automática de um dicionário emocional para o português. Master's thesis, University of Beira Interior, 2009. 45, 175, 182, 189, 192, 197

[Roget 1852] P.M. Roget. Roget's thesaurus. 1852. 11, 95, 103

[Rohall 2004] S.L. Rohall, D. Gruen, P. Moody, M. Wattenberg, M. Stern, B. Kerr, B. Stachel, K. Dave, R. Armes and E. Wilcox. *ReMail: a Reinvented Email Prototype.* In Proceedings of 2004 ACM Human Factors in Computing Systems (CHI 2004), pages 791–792, 2004. 91

[Rosch 1973] E.H. Rosch. *Natural Categories.* Cognitive Psychology, vol. 4, pages 265–283, 1973. 25

[Rumelhart 1986] D. E. Rumelhart and J. L. McClelland. *On Learning Past Tenses of English Verbs.* In Parallel Distributed Processing, chapitre 2, pages 216–271. MIT Press, 1986. i, 1

[Sahlgren 2001] M. Sahlgren. *Vector-based Semantic Analysis: Representing Word Meanings based on Random Labels.* In Proceedings of the Workshop on Semantic Knowledge Acquisition and Categorisation at the European Summer School in Logic, Language and Information (ESSLLI 2001), 2001. 17, 20, 137, 140, 142

[Salem 1987] A. Salem. La pratique des segments répétés. Klincksieck, 1987. 16, 35

[Salton 1975] G. Salton, C.S. Yang and C.T. Yu. *A Theory of Term Importance in Automatic Text Analysis.* American Society of Information Science, vol. 26, no. 1, pages 33–44, 1975. 31, 63, 73, 98, 187

[Salton 1993] G. Salton, J. Allan and C. Buckley. *Approaches to Passage Retrieval in Full Text Information Systems.* In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), pages 49–58, 1993. 96

[Sanderson 1999] M. Sanderson and B. Croft. *Deriving Concept Hierarchies from Text.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1999), pages 206–213, 1999. 26, 28, 135, 139, 159, 168

[Sanderson 2000] M. Sanderson and D. Lawrie. *Building, Testing, and Applying Concept Hierarchies.* Advances in Information Retrieval, vol. 7, pages 235–266, 2000. 26, 135, 139, 158, 159, 162, 168

[Sang 2007] E.T.K. Sang. *Extracting Hypernym Pairs from the web.* In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions (ACL 2007), pages 165–168, 2007. 12, 151

[Santos 2006] C. Santos. Alexia - acquisition of lexical chains for text summarization. Master's thesis, University of Beira Interior, 2006. 109

[Sardinha 2002] T.B. Sardinha. *Segmenting Corpora of Texts.* DELTA, vol. 18, no. 2, pages 273–286, 2002. 97

[Saussure 1959] F. Saussure. Course in general linguistics. 1959. 12

[Schmid 1994] H. Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* In Proceedings of the 2nd International Conference on New Methods in Language Processing (ICNMLP 1994), pages 44–49, 1994. 75

[Schneider 2000] R. Schneider and I. Renz. *The Relevance of Frerquency Lists for Error Correction and Robust Lemmatization.* 2000. 16, 35

[Sharoff 2004] S. Sharoff. *What is at Stake: a Case Study of Russian Expressions Starting with a Preposition.* In Proceedings of the Workshop on Multiword Expressions of the 42nd Annual Meeting of the Association of Computational Linguistics (ACL 2004), 2004. 36

[Shi 1992] H. Shi and J. Schaeffer. *Parallel Sorting by Regular Sampling.* Journal of Parallel and Distributed Computing, vol. 14, no. 4, pages 361–372, 1992. 45, 46

[Shimohata 1997] S. Shimohata. *Retrieving Collocations by Co-occurrences and Word Order Constraints.* pages 476–481, 1997. 35

[Sieg 2007] A. Sieg, B. Mobasher and R. Burke. *Web Search Personalization with Ontological User Profiles*. In Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007), pages 525–534, 2007. 199, 200

[Silber 2000] H.G. Silber and K.F. McCoy. *Efficient Text Summarization Using Lexical Chains*. In Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI 2000), pages 252–255, 2000. 93, 94, 95, 96, 103, 106, 129, 131

[Silva 1999] J. Silva, G. Dias, S. Guilloré and J.G.P. Lopes. *Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units*. In Proceedings of the 9th Portuguese Conference in Artificial Intelligence (EPIA 1999), pages 21–24, 1999. 14, 16

[Sindhwani 2005] V. Sindhwani and P. Niyogi. *A Co-regularized Approach to Semi-supervised Learning With Multiple Views*. In Proceedings of the Workshop on Learning with Multiple Views of the 22nd International Conference (ICML 2005), pages 1–6, 2005. 188

[Sinha 2007] R. Sinha and R. Mihalcea. *Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity*. In Proceedings of the International Conference on Semantic Computing (ICSC 2007), pages 363–369, 2007. 11

[Smadja 1993] F. Smadja. *Retrieving Collocations From Text: XTRACT*. Computational Linguistics, vol. 19, no. 1, pages 143–117, 1993. 34, 35, 36, 37, 53, 56

[Smadja 1996] F. Smadja, K.R. McKeown and V. Hatzivassiloglou. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Computational Linguistics, vol. 22, no. 1, 1996. 14

[Smith 1981] T.F. Smith and M.S. Waterman. *Identification of Common Molecular Subsequences*. Journal of Molecular Biology, vol. 147, no. 1, pages 195–197, 1981. 96, 121

[Snow 2005] R. Snow, D. Jurafsky and A. Y. Ng. *Learning syntactic patterns for automatic hypernym discovery*. In Proceedings of the Neural Information Processing Systems Conference (NIPS 2005), 2005. 11, 151

[Song 2008] W. Song and S.C Park. *An Improved Genetic Algorithm for Document Clustering with Semantic Similarity Measure*. In Proceedings of the 4th International Conference on Natural Computation (ICNC 2008), pages 536–540, 2008. 11

[Spärck Jones 1972] K. Spärck Jones. *A Statistical Interpretation of Term Specificity and its Application in Retrieval.* Journal of Documentation, vol. 28, pages 11–21, 1972. 73

[Spearman 1904] C. Spearman. *The Proof and Measurement of Association Between Two Things.* The American Journal of Psychology, vol. 100, no. 3-4, pages 441–471, 1904. 154

[Specia 2010] L. Specia. *Translating from Complex to Simplified Sentences.* In Computational Processing of the Portuguese Language, volume 6001, pages 30–39. 2010. 95, 111, 114

[Srihari 2006] R. Srihari, W. Li, T. Cornell and C. Niu. *InfoXtract: A Customizable Intermediate Level Information Extraction Engine.* Natural Language Engineering, no. 14, pages 33–69, 2006. 183

[Srinivasan 2000] A. Srinivasan. *The Aleph Manual.* Rapport technique, Oxford University, 2000. 96, 124

[Steinberger 2005] J. Steinberger, M.A. Kabadjov, M. Poesio and O. Sanchez-Graillet. *Improving LSA-based Summarization with Anaphora Resolution.* In Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pages 1–8, 2005. 93

[Stevenson 2004] S. Stevenson, A. Fazly and R. North. *Statistical Measures of the Semi-productivity of Light Verb Constructions.* In Proceedings of the 2nd Workshop on Multiword Expressions of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 1–8, 2004. 36

[Stevenson 2006] M. Stevenson and M.A. Greenwood. *Comparing Information Extraction Pattern Models.* In Proceedings of the Workshop on Information Extraction Beyond The Document of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006), 2006. 11

[Stokes 2002] N. Stokes, J. Carthy and A.F. Smeaton. *Segmenting Broadcast News Streams Using Lexical Chains.* In Proceedings of the 1st Starting Artificial Intelligence Researchers Symposium (STAIRS 2002), pages 145–154, 2002. 100

[Stokes 2004] N. Stokes, E. Newman, J. Carthy and A.F. Smeaton. *Broadcast News Gisting Using Lexical Cohesion Analysis.* In Advances in Information Retrieval, volume 2997, pages 209–222. 2004. 102

[Stoyanov 2004] V. Stoyanov, C. Cardie, D. Litman and J. Wiebe. *Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus.* In Proceedings of the AAAI Spring Symposium on Exploring

Attitude and Affect in Text: Theories and Applications, pages 77–89, 2004. 184

[Strapparava 2004] C. Strapparava and A. Valitutti. *WordNet-Affect: An Affective Extension of WordNet.* In Proceedings of the 4th Language Resources and Evaluation International Conference (LREC 2004), pages 1083–1086, 2004. 183

[Strapparava 2008] C. Strapparava and R. Mihalcea. *Learning to Identify Emotions in Text.* In Proceedings of the 2008 ACM Symposium on Applied Computing (SAC 2008), pages 1556–1560, 2008. 183

[Svore 2007] K. Svore, L. Vanderwende and C. Burges. *Enhancing Single-document Summarization by Combining RankNet and Third-party Sources.* In Proceedings of the Joint Conference Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2007), pages 448–457, 2007. 93

[Tamine 2008] L. Tamine, M. Boughanem and W.N. Zemirli. *Personalized Document Ranking: Exploiting Evidence from Multiple User Interests for Profiling and Retrieval.* Journal of Digital Information Management, vol. 6, no. 5, pages 354–365, 2008. 199, 200

[Tan 2004] P-N. Tan, V. Kumar and J. Srivastava. *Selecting the Right Objective Measure for Association Analysis.* Information Systems, vol. 29, no. 4, pages 293–313, 2004. 25, 26, 151

[Tanaka 2007] T. Tanaka, F. Bond, T. Baldwin, S. Fujita and C. Hashimoto. *Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information.* In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2007), pages 477–485, 2007. 135

[Teich 2004] E. Teich and P. Fankhauser. *Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with WordNet.* In Proceedings of the 2nd International Global WordNet Conference (GWC 2004), pages 326–331, 2004. 109

[Terra 2003] E. Terra and C.L.A. Clarke. *Frequency Estimates for Statistical Word Similarity Measures.* In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003), pages 165–172, 2003. 17, 20, 21, 29, 137, 140, 141, 142, 196

[Terra 2005] E. Terra and C.L.A. Clarke. *Comparing Query Formulations and Lexical Affinity Replacements in Passage Retrieval.* In Proceedings of the ELECTRA Workshop of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pages 11–17, 2005. 102

[Tomokiyo 2003] T. Tomokiyo and M. Hurst. *A Language Model Approach to Keyphrase Extraction.* In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003), pages 33–40, 2003. 35

[Tsuchiya 2006] M. Tsuchiya, T. Shime, T. Takagi, T. Utsuro, K. Uchimoto, S. Matsuyoshi, S. Sato and S. Nakagawa. *Chunking Japanese Compound Functional Expressions by Machine Learning.* In Proceedings of the Workshop on Multiword Expressions in a Multilingual Context of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pages 25–32, 2006. 38

[Turner 2005] J. Turner and E. Charniak. *Supervised and Unsupervised Learning for Sentence Compression.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005), pages 290–297, 2005. 95, 111, 112

[Turney 2000] P. Turney. *Learning Algorithms for Keyphrase Extraction.* Information Retrieval, vol. 2, no. 4, pages 303–336, 2000. 70

[Turney 2001] P.D. Turney. *Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.* In Proceedings of the 12th European Conference on Machine Learning (EMCL 2001), pages 491–502, 2001. 140, 142

[Turney 2002] P.D. Turney. *Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews.* In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), pages 417–424, 2002. 176

[Turney 2003] P.D. Turney, M.L. Littman, J. Bigham and V. Shnayder. *Combining Independent Modules in Lexical Multiple-Choice Problems.* In Proceedings of the 2003 International Conference on Recent Advances in Natural Language Processing (RANLP 2003), pages 482–489, 2003. 140, 142, 147

[Turney 2005] P.D. Turney and M.L. Littman. *Corpus-based Learning of Analogies and Semantic Relations.* Machine Learning, vol. 60, pages 251–278, 2005. 25

[Turney 2006] Peter D. Turney. *Similarity of Semantic Relations.* Computational Linguistics, vol. 32, pages 379–416, 2006. 24, 25

[Turney 2008] P.D. Turney. *The Latent Relation Mapping Engine: Algorithm and Experiments.* Journal of Artificial Intelligence Research, vol. 33, pages 615–655, 2008. 25

[Unno 2006] Y. Unno, T. Ninomiya, Y. Miyao and J. Tsujii. *Trimming CFG Parse Trees for Sentence Compression using Machine Learning Approaches.* In Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006), pages 850–857, 2006. 111, 112

[Utsuro 2007] T. Utsuro, T. Shime, M. Tsuchiya, S. Matsuyoshi and S. Sato. *Learning Dependency Relations of Japanese Compound Functional Expressions.* In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions of the 45st Annual Meeting of the Association of Computational Linguistics (ACL 2007), pages 65–72, 2007. 36

[Uzêda 2010] V.R. Uzêda, T.A.S. Pardo and M.G.V. Nunes. *A Comprehensive Comparative Evaluation of RST-based Summarization Methods.* ACM Transactions on Speech Language Processing, vol. 6, no. 4, pages 1–20, 2010. 93

[Van de Cruys 2007] T. Van de Cruys and B.V. Moirón. *Semantics-based Multiword Expression Extraction.* In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions of the 45st Annual Meeting of the Association of Computational Linguistics (ACL 2007), pages 25–32, 2007. 36, 37, 56

[Vandeghinste 2002] V. Vandeghinste. *Lexicon Optimization: Maximizing Lexical Coverage in Speech Recognition through Automated Compounding.* In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2002), pages 1270–1276, 2002. 113

[Vandeghinste 2004] V. Vandeghinste and Y. Pan. *Sentence Compression for Automated Subtitling: A Hybrid Approach.* In Proceedings of the Workshop on Text Summarization Branches Out of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pages 89–95, 2004. 110, 111, 113

[Vechtomova 2006] O. Vechtomova, M. Karamuftuoglu and S.E. Robertson. *On Document Relevance and Lexical Cohesion Between Query Terms.* Information Processing and Management, vol. 42, no. 5, pages 1230–1247, 2006. 102

[Venkatsubramanyan 2004] S. Venkatsubramanyan and J. Perez-Carballo. *Multiword Expression Filtering for Building Knowledge Maps.* In Proceedings of the Workshop on Multiword Expressions of the 42nd Annual Meeting of the Association of Computational Linguistics (ACL 2004), pages 40–47, 2004. 36

[Wan 2009] X. Wan. *Co-training for Cross-lingual Sentiment Classification.* In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2009), pages 235–243, 2009. 178, 179, 188

[Ward 1963] J. Ward. *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, vol. 58, pages 236–244, 1963. 166

[Weeds 2004] J. Weeds, D. Weir and D. McCarthy. *Characterising Measures of Lexical Distributional Similarity.* In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), 2004. 18, 137, 141

[Wiebe 1999] J.M. Wiebe, R.F. Bruce and T.P. O'Hara. *Development and Use of a Gold-standard Data Set for Subjectivity Classifications.* In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999), pages 246–253, 1999. 176

[Wiebe 2004] J. Wiebe, T. Wilson, R. Bruce, M. Bell and M. Martin. *Learning Subjective Language.* Computational Linguistics, vol. 30, no. 3, pages 277–308, 2004. 177, 179, 180, 182, 186

[Wilcoxon 1945] F. Wilcoxon. *Individual Comparisons by Ranking Methods.* Biometrics, vol. 1, pages 80–83, 1945. 185

[Willett 1988] P. Willett. *Recent Trends in Hierarchical Document Clustering: A Critical Review.* Information Processing and Management, vol. 25, no. 5, pages 577–597, 1988. 59

[Wong 1985] S.K.M. Wong, W. Ziarko and P.C.N. Wong. *Generalized Vector Spaces Model in Information Retrieval.* In Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1985), pages 18–25, 1985. 21

[Wong 2008] K-F. Wong, M. Wu and W. Li. *Extractive Summarization using Supervised and Semi-supervised Learning.* In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pages 985–992, 2008. 93

[Wu 1994] Z. Wu and M. Palmer. *Verbs Semantics and Lexical Selection.* In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994), pages 133–138, 1994. 24

[Xia 2006] H. Xia, S. Wang and T Yoshida. *A Modified Ant-based Text Clustering Algorithm with Semantic Similarity Measure.* Journal of Systems Science and Systems Engineering, vol. 15, no. 4, pages 474–492, 2006. 11

[Xiang 2003] J. Xiang and Z. Hongyuan. *Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming.* In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), pages 322–329, 2003. 94, 101

[Xu 1996] J. Xu and W.B. Croft. *Query Expansion Using Local and Global Document Analysis.* In Proceedings of the 19th Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval (SIGIR 1996), pages 4–11, 1996. 97

[Xu 2006] R.F. Xu and Q. Lu. *A Multi-stage Chinese Collocation Extraction System.* Lecture Notes in Computer Science, vol. 3930, pages 740–749, 2006. 38

[Yamamoto 2001] M. Yamamoto and K. Church. *Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus.* Computational Linguistics, vol. 27, no. 1, pages 1–30, 2001. 42, 43

[Yang 2003a] C.C. Yang and F.L. Wang. *Fractal Summarization for Mobile Devices to Access Large Documents on the Web.* In Proceedings of the 12th International Conference on World Wide Web (WWW 2003), pages 215–224, 2003. 129

[Yang 2003b] S. Yang. *Machine Learning for Collocation Identification.* In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2003), pages 315–321, 2003. 37, 38, 47, 56

[Yarlett 2008] D. Yarlett. *Language Learning Through Similarity-Based Generalization.* PhD thesis, Stanford University, 2008. 17

[Yu 2003] H. Yu and V. Hatzivassiloglou. *Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pages 129–136, 2003. 176, 186

[Zamir 1998] O. Zamir and O. Etzioni. *Web Document Clustering: A Feasibility Demonstration.* In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), pages 46–54, 1998. 7, 62, 63, 64, 65, 66, 73, 77, 89

[Zeng 2004] Q. Zeng, Q. He, C. Zheng and J. Ma. *Learning to Cluster Web Search Results.* In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pages 210–217, 2004. 62, 63, 66, 70, 76, 77, 89

[Zhang 2001] D. Zhang and Y. Dong. *Semantic, Hierarchical, Online Clustering of Web Search Results.* In Proceedings of the 6th Asia Pacific Web Conference (APWEB 2004), pages 69–78, 2001. 62, 73, 77, 89

[Zhang 2003] Y.Z. Zhang, N. Zincir-Heywood and E. Milios. *Summarizing Web Sites Automatically.* In Proceedings of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence (AI 2003), pages 283–296, 2003. 129