HAL
open science

MEASURING AND IMPROVING COMPARABLE CORPUS QUALITY

Li Bo

To cite this version:

Li Bo. MEASURING AND IMPROVING COMPARABLE CORPUS QUALITY. Information Retrieval [cs.IR]. Université de Grenoble, 2012. English. NNT : . tel-00997769

HAL Id: tel-00997769

https://theses.hal.science/tel-00997769v1

Submitted on 2 Jun 2014

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.

MEASURING AND IMPROVING COMPARABLE CORPUS QUALITY

A Dissertation

Submitted to the Faculty

of

Université de Grenoble

by

Bo Li

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

June 2012

Université de Grenoble

Grenoble, France

Dedicated to my parents.

ACKNOWLEDGMENTS

It is hard to express here my thanks in a few words. There are many people who have been helping and encouraging me throughout the PhD study. Without their precious help, I could not have arrived at this point.

Especially, I would like to give the sincere thanks to my supervisor Prof. Eric Gaussier who led me to the areas of comparable corpora exploitation and CLIR. With his help to my work and my personal life, I have experienced a very smooth and enjoyable life in France. His concentration and patience on research work has inspired me to understand the role I should play in the complex work. He is both an excellent researcher and a great friend for me. I also want to thank a lot my co-supervisor Dr. Jean-Pierre Chevallet who has continuously supported me to advance in my research.

By the way, I feel very lucky to have been able to work with all the kind and talent colleges in the AMA team of the LIG lab. Lastly, I would like to give my acknowledges to the French government which has funded me to do the interesting work.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES viii

ABSTRACT ix

1 Introduction 1
- 1.1 Bilingual Corpora 1
- 1.2 Quality of Comparable Corpora 5
- 1.3 Work in the Thesis 6

2 State-of-the-art Review 9
- 2.1 Comparable Corpus Quality 9
  - 2.1.1 Comparing Different Corpora 10
  - 2.1.2 Parallel Subpart Extraction 12
- 2.2 Bilingual Lexicon Extraction 15
  - 2.2.1 Extraction from Parallel Corpora 15
  - 2.2.2 Extraction from Comparable Corpora 17
- 2.3 Cross-Language Information Retrieval 22
  - 2.3.1 Monolingual IR Models 23
  - 2.3.2 CLIR Strategies 29
  - 2.3.3 Resources for IR Experiments 31
- 2.4 Conclusion 33

3 Comparability Measures 35
- 3.1 The Notion of Comparability 35
- 3.2 Developing Comparability Measures 37
  - 3.2.1 Measures Based on Vocabulary Overlapping 38
  - 3.2.2 Baseline Measures 43
- 3.3 Experimental Validation 45
  - 3.3.1 Resources in the Experiments 45
  - 3.3.2 Evaluating Comparability Measures 46
- 3.4 Conclusion 62

4 Enhancing Comparable Corpus Quality 65
- 4.1 General Methodology 65
- 4.2 Approaches to Enhancing Corpus Quality 67
  - 4.2.1 The Greedy Approach 67

Page
4.2.2 The Clustering Approach 70
4.3 Experimental Validation 77
4.3.1 Improving Corpus Quality 78
4.3.2 Bilingual Lexicon Extraction 80
4.4 Conclusion 84

5 CLIR Models and Comparable Corpora 87
5.1 Information-based CLIR Models 87
5.1.1 Monolingual Models 88
5.1.2 Cross-Language Extensions 90
5.2 Validation of CLIR Models 93
5.2.1 Theoretical Validation 94
5.2.2 Experimental Validation 97
5.2.3 Embedding Extracted Lexicons 101
5.3 Conclusion 104

6 Conclusion and Discussion 107

A An Example of Structured and Unstructured Queries 111
A.1 The Original Query 111
A.2 Structured and Unstructured Queries 111

MY PUBLICATIONS 113

LIST OF REFERENCES 114

LIST OF TABLES

Table Page

3.1 The information of the corpora used in the experiments in section 3.3. (k=1000, m=1000k) 46

3.2 Correlation scores of the vocabulary overlapping measures with the gold standard. The rows TF-IDF correspond to the TF-IDF weighting schema, and the rows P/A correspond to the Presence/Absence weighting method. 52

3.3 Correlation scores of the baseline comparability measures with the gold standard 55

3.4 The degree of robustness with the measure M and different dictionary coverages. The coverage range is set to $[r, r+\epsilon]$ as in definition 3.3.1 with ϵ being fixed to 0.02. 60

3.5 The degree of robustness with the measure M^c and different dictionary coverages. The coverage range is set to $[r, r+\epsilon]$ as in definition 3.3.1 with ϵ being fixed to 0.02. The table is designed in the same format as Table 3.4. 61

4.1 The size of the original and external corpora (k=1000) 78

4.2 Information of the resulting corpora enhanced by different methods. The row "Coverage" identifies the coverage of the resulting corpus vocabulary w.r.t. the original corpus $\mathcal{C}^0$. The row "M" gives, for each comparable corpus, the comparability score measured by M. 79

4.3 Performance of bilingual lexicon extraction from different corpora . . . 82

4.4 Comparison of the precision for words of different frequencies 83

5.1 Notations used in IR models in chapter 5 88

5.2 Characteristics of different CLEF collections 98

5.3 Comparison of different cross-language extensions of LL and SPL in terms of MAP scores, where "00-02", "03" and "04" respectively correspond to the data sets "CLEF 2000-2002", "CLEF 2003" and "CLEF 2004". A † indicates, for each model, that the difference with the best performing extension (in bold) is significant. 99

Table Page

5.4 Comparison of different CLIR strategies (SYN, QT, DT) for language models in terms of MAP scores, where "00-02", "03" and "04" have the same meaning as in Table 5.3. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant. . 100

5.5 Comparison of different CLIR systems in terms of MAP scores on all language pairs and collections, where "00-02", "03" and "04" have the same meaning as in Table 5.3. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant. . 102

5.6 MAP scores of the CLIR experiment using the lexicons extracted from comparable corpora (with JV_{LL} model). The significant differences against the baseline are marked with the †. 103

LIST OF FIGURES

Figure Page

1.1 The English article "History" has a cross-language link to the French article "Histoire". The two articles are highly related but not parallel. . . 4

3.1 Constructing the test corpus group G_a with gold-standard comparability levels 49

3.2 Evolution of the measures M_{ef}, M_{fe}, M, M^v_{ef}, M^v_{fe} and M^v w.r.t. gold standard on the corpus group G_c (x-axis: gold-standard comparability scores, y-axis: comparability scores from the measures) 53

3.3 Evolution of M and M^c w.r.t. different dictionary coverages on comparable corpora $\mathcal{P}_1$, $\mathcal{P}_1^{0.7}$, $\mathcal{P}_1^{0.4}$ and $\mathcal{P}_1^{0.1}$ in G_a (x-axis: different dictionary coverages; y-axis: comparability scores from M or M^c). 57

3.4 The frequency histogram of the comparability scores (from M) on $\mathcal{P}_1^{0.1}$ between the coverage 0.56 and 0.58 (x-axis: comparability scores; y-axis: frequency). 59

4.1 The general methodology for enhancing comparable corpus quality. The resulting corpus consists of three parts: Part 1 corresponds to the high-quality subpart of the original comparable corpus; Part 2/3 is the enriched version of the French/English low-quality subpart of the original corpus. 66

4.2 Performance of bilingual lexicon extraction from different corpora with varied N values from 1 to 300. The five lines from the top down in each subfigure are corresponding to the results for $\mathcal{C}^2$, $\mathcal{C}^{2'}$, $\mathcal{C}^1$, $\mathcal{C}^{1'}$ and $\mathcal{C}^0$ respectively. 83

ABSTRACT

Bilingual corpora are an essential resource used to cross the language barrier in multilingual Natural Language Processing (NLP) tasks. Most of the current work makes use of parallel corpora that are mainly available for major languages and constrained areas. Comparable corpora, text collections comprised of documents covering overlapping information, are however less expensive to obtain in high volume. Previous work has shown that using comparable corpora is beneficent for several NLP tasks. Apart from those studies, we will try in this thesis to improve the quality of comparable corpora so as to improve the performance of applications exploiting them. The idea is advantageous since it can work with any existing method making use of comparable corpora.

We first discuss in the thesis the notion of *comparability* inspired from the usage experience of bilingual corpora. The notion motivates several implementations of the comparability measure under the probabilistic framework, as well as a methodology to evaluate the ability of comparability measures to capture gold-standard comparability levels. The comparability measures are also examined in terms of robustness to dictionary changes. The experiments show that a symmetric measure relying on vocabulary overlapping can correlate very well with gold-standard comparability levels and is robust to dictionary changes.

Based on the comparability measure, two methods, namely the *greedy approach* and the *clustering approach*, are then developed to improve the quality of any given comparable corpus. The general idea of these two methods is to choose the high-quality subpart from the original corpus and to enrich the low-quality subpart with external resources. The experiments show that one can improve the quality, in terms

of comparability scores, of the given comparable corpus by these two methods, with the clustering approach being more efficient than the greedy approach. The enhanced comparable corpus further results in better bilingual lexicons extracted with the standard extraction algorithm.

Lastly, we investigate the task of Cross-Language Information Retrieval (CLIR) and the application of comparable corpora in CLIR. We develop novel CLIR models extending the recently proposed information-based models in monolingual IR. The information-based CLIR model is shown to give the best performance overall. Bilingual lexicons extracted from comparable corpora are then combined with the existing bilingual dictionary and used in CLIR experiments, which results in significant improvement of the CLIR system.

1 INTRODUCTION

Comparable corpora[1] have been shown to be useful in several multilingual natural language processing (NLP) tasks. Since parallel corpora of high quality are considerably more expensive to obtain, it makes sense to resort to comparable corpora in applications where parallel resources are not readily available. Compared to the usage of parallel corpora, there are several problems not well considered or solved, and specific to comparable corpora, which motivates the work in this thesis. In order to facilitate the description of the thesis work, we will introduction in this chapter the necessary background of comparable corpora. To be exact, we will first introduce in section 1.1 the concepts related with comparable corpora. The quality of comparable corpora will then be discussed in section 1.2. Lastly we will discuss in section 1.3 problems that remain unsolved in previous work prior to presenting a summary of the thesis work.

1.1 Bilingual Corpora

The bilingual corpus, a text collection consisting of content in two languages, is a commonly used resource in many multilingual NLP tasks. In previous work, two types of bilingual corpora have been broadly used: parallel corpora and comparable corpora. Parallel corpora are comprised of text that is the translation of each other. This kind of resource has been used for a long period in applications such as bilingual lexicon extraction (Kay and Roscheisen, 1993), statistical machine translation (SMT) (Och and Ney, 2003) and cross-language information retrieval (CLIR) (Ballesteros and Croft, 1997). It is however expensive to construct parallel corpora of high

[1]In corpus linguistics, comparable corpora refer to both monolingual, bilingual and multilingual comparable corpora. We will only consider bilingual ones in this thesis.

quality, especially the ones between minor languages or the ones covering broad topics (Markantonatou et al., 2006). For instance, one can construct parallel corpora by referring to resources such as books translated into different languages, international laws, movie subtitles, and software manuals in various languages. To alleviate the intensive human labor involved in building parallel corpora by hand, researchers have tried to exploit the web for automatic construction. Existing work (e.g. (Ma and Liberman, 1999; Resnik and Smith, 2003; Zhang et al., 2006)) makes use of such features as URL, web page structure and page content to pick web pages containing parallel content. However, as one can find, the parallel content on the web usually appears on the websites of international companies, organizations and news agencies, which again falls into constrained topics (e.g. government regulations and news) and limited language pairs (i.e. the major languages in the world). For an overview of parallel corpora that are publicly available, one can refer to the UN proceedings corpus, the Hansard corpus[2] and the Europarl corpus (Koehn, 2005), which are commonly used resources for SMT research.

Comparable corpora, as another kind of bilingual corppora, are cheaper to obtain in high volume, even for minor languages, since it only demands that the text collection covers related content which can be easily found on the web. For example, the newspapers published by various news agencies in different countries and in the same period usually report the same hot affairs in the world, which thus covers similar topics and can be used as a comparable corpus. In (Skadina et al., 2010), they list several examples of comparable corpora which are:

- The comparable part of the Multilingual Corpora for Cooperation (MLCC) (Armstrong et al., 1998). This corpus includes financial newspaper articles from the early 1990s in six European languages: Dutch (8.5 million words), English (30 million words), French (10 million words), German (33 million words), Italian (1.88 million words), and Spanish (10 million words).

[2]Both the UN proceedings corpus and the Hansard corpus are available from: http://www.ldc.upenn.edu

- The Bulgarian-Croatian comparable corpus (Bekavac et al., 2004) in news domain. This corpus is comprised of 3.5 million tokens (393 thousand Bulgarian words and 3.1 million Croatian words) and was built from subsets of two large newspaper corpora of respective languages. The subsets were chosen according to the same criteria (e.g. identical year, same domain, etc.).

- The English-Swedish comparable corpus in news domain (Talvensaari et al., 2007). This corpus was built by an automatic approach from articles by a Swedish news agency and a US newspaper.

- The English-French-Norwegian comparable corpus in science domain (Flottum, 2003). This corpus contains 450 reviewed scientific papers in three disciplines (economics, linguistics and medicine) amounting to 3.2 million words.

- The INTERA Multilingual Corpus (Gavrilidou et al., 2006). This resource contains content in various domains such as law, health, education, tourism, environment, politics and finance. The comparable corpus is in 4 language pairs: Bulgarian - English (2 million words), Greek - English (4 million words), Serbian - English (2 million words), and Slovene - English (4 million words).

More recently, the content of Wikipedia[3] has been used as a comparable corpus in several NLP applications such as (Ni et al., 2009; Otero and Lopez, 2010; Smith et al., 2010). Each Wikipedia article, corresponding to a *concept*, usually has a cross-language link to its versions in other languages. Then two versions of the same concept are highly related but, in most cases, are not the translation of each other (refer to an example in Figure 1.1). In addition, one can note that, the Wikipedia category structure clusters all the articles related to a topic under a category. Articles in different languages and in the same category thus talk about similar topics. Those facts listed above draw a clear picture that Wikipedia can be used as a good comparable corpus.

[3]http://www.wikipedia.org

History

From Wikipedia, the free encyclopedia

History (from Greek *ἱστορία - historia*, meaning "inquiry, knowledge acquired by investigation"[2]) is the discovery, collection, organization, and presentation of information about past events. History can also mean the period of time after writing was invented. Scholars who write about history are called historians. It is a field of research which uses a narrative to examine and analyse the sequence of events, and it sometimes attempts to investigate objectively the patterns of cause and effect that determine events.[3][4] Historians debate the nature of history and its usefulness. This includes discussing the study of the discipline as an end in itself and as a way of providing "perspective" on the problems of the present.[3][5][6][7] The stories common to a particular culture, but not supported by external sources (such as the legends surrounding King Arthur) are usually classified as cultural heritage rather than the "disinterested investigation" needed by the discipline of history.[8][9] Events of the past prior to written record are considered prehistory.

Historia
by Nikolaos Gysis (1892)

Amongst scholars, the fifth century BC Greek historian Herodotus is considered to be the "father of history", and, along with his contemporary Thucydides, forms the foundations for the modern study of history. Their influence, along with other historical traditions in other parts of their world, have spawned many different interpretations of the nature of history which has evolved over the centuries and are continuing to change. The modern study of history has many different fields including those that focus on certain regions and those which focus on certain topical or thematical elements of historical investigation. Often history is taught as part of primary and secondary education, and the academic study of history is a major discipline in University studies.

> Those who cannot remember the past are condemned to repeat it.[1]
>
> —George Santayana

(a) A segment of the English article

Histoire

L'**histoire** est à la fois l'étude des faits, des événements du passé et, par synecdoque, leur ensemble. L'histoire est un récit, elle est la construction d'une image du passé par des hommes et des femmes (les historiens et historiennes) qui tentent de décrire, d'expliquer ou de faire revivre des temps révolus. Ce récit historique n'est pas construit par intuition intellectuelle, mais à partir de sources. L'histoire s'attache avec ces sources à reconstruire plusieurs pans du passé. Au cours des siècles, les historiens ont fortement fait évoluer leurs champs d'intervention et ont aussi réévalué leurs sources, ainsi que la manière de les traiter.

Historia, allégorie de l'histoire
Peinture de Nikolaos Gysis (1892).

L'histoire, qui n'est pas seulement une réflexion sur le passé, se construit aussi selon une méthode. Celle-ci a évolué au cours du temps, évolution qu'on appelle l'historiographie. La méthode historique s'appuie sur un ensemble de sciences auxiliaires qui aident l'historien à construire son récit. Par delà les époques et les méthodes, et quel que soit le but sous-jacent du travail de l'historien, l'histoire est toujours une construction humaine, inscrite dans l'époque où elle est écrite. Elle joue un rôle social et elle est convoquée pour soutenir, accompagner ou juger les actions des Hommes.

(b) A segment of the French article

Figure 1.1. The English article "History" has a cross-language link to the French article "Histoire". The two articles are highly related but not parallel.

Previous work has shown that comparable corpora can be successfully applied to tasks such as bilingual lexicon extraction (for an example, refer to (Rapp, 1999), more studies will be cited in section 2.2), the enhancement of SMT systems (Munteanu et al., 2004; AbduI-Rauf and Schwenk, 2009), the enhancement of CLIR systems (Talvensaari et al., 2007), and the modeling of topics across languages (Ni et al., 2009; Boyd-Graber and Blei, 2009). These findings bring hope that comparable corpus can be used to bridge the language barrier in some applications in the absence of parallel resources. In this thesis, to exploit comparable corpora, we will make use of the task of bilingual lexicon extraction as a testbed, since it is the mostly investigated task using comparable corpora. The extracted lexicons will further be used to enhance the CLIR system. The representative work of bilingual lexicon extraction and CLIR will be reviewed in chapter 2.

1.2 Quality of Comparable Corpora

The definition for comparable corpus is rather vague compared to that of parallel corpus. For example, (Ji, 2009) defines comparable corpora as document collections describing similar topics. In (Munteanu et al., 2004; Hewavitharana and Vogel, 2008), comparable corpus is defined as a text collection covering overlapping information. In most of the other studies (e.g. (Fung and Yee, 1998)), they directly construct and use comparable corpora according to their intuitions regardless what constitutes a *real* comparable corpus. We will give in section 3.1 more discussions regarding the notion of *comparability*. According to those existing definitions and various comparable corpora used in previous work, one can find that different comparable corpora can be quite different from each other, depending on the extend to which two monolingual parts of comparable corpus are related to each other. Intuitively, parallel corpora can be seen as a special case of comparable corpus, i.e. the one with the highest comparability level, since there is no bilingual corpus better than parallel corpus in terms of usability in applications. If one does not consider the corpus size, the quality

of different parallel corpora should be the same, which is not the case for comparable corpora.

Data driven NLP tasks depend a lot on the quality of the text resource used, and the experience is that the better the corpus is, the better one algorithm can perform. The theoretical principle under the assertion is that the corpus of higher quality can provide more information useful for the corresponding NLP tasks. We conjecture here that comparable corpora of higher quality can lead to applications of better performance, a fact that will actually be validated in the following sections and that has been ignored in previous work. Among comparable corpora of different quality, we would like to pick the ones of good quality or we can try to enhance the low-quality ones. However, as one can find, there has not been any practical measure on which one can rely to judge the quality of comparable corpora. We will thus try to establish in the thesis a measure to capture different comparability levels. With this measure, we will be able to develop methods to enhance the low-quality comparable corpus so as to improve the performance of applications relying on comparable corpora.

1.3 Work in the Thesis

As we have discussed in section 1.2, the quality of comparable corpora is an important feature affecting their usability in NLP applications. However, this fact has never been systematically investigated by previous work. Most of previous studies did not pay attention to the corpus quality prior to the usage of comparable corpora. We will focus in this thesis on topics related with the quality of comparable corpora and their applications to NLP tasks. To exploit comparable corpora, we choose here the tasks of bilingual lexicon extraction and CLIR.

The structure of the thesis is as follows:

- We first review in chapter 2 existing work for several NLP tasks that will be discussed in the thesis. In section 2.1, we introduce previous work trying to compare two corpora and to extract the parallel subparts from a comparable

corpus. The approaches for bilingual lexicon extraction from comparable corpora and CLIR are respectively detailed in section 2.2 and section 2.3.

- We then develop and evaluate in chapter 3 the comparability measures that can be used to judge comparable corpus quality. The notion of comparability is discussed in section 3.1 according to the the intuition one can have from the usage experiences of various bilingual corpora. Following this notion, several candidate measures are developed in section 3.2 and evaluated in section 3.3.

- Based on the comparability measure, two approaches are developed in chapter 4 to improve the quality of comparable corpora. The *greedy approach* is introduced in section 4.2.1, prior to the presentation of the *clustering approach* having improved performance in section 4.2.2. These two approaches are validated in section 4.3 to show that one can improve the quality of the given comparable corpus and enhanced comparable corpora lead to better lexicons extracted.

- The application of comparable corpora in CLIR is discussed in chapter 5. We first propose in section 5.1 novel CLIR models extending the information-based models recently introduced in (Clinchant and Gaussier, 2010). The proposed CLIR models are then validated in section 5.2.2 to show their best performance overall. The information-based CLIR model can be further improved with lexicons extracted from comparable corpora.

- Lastly, the thesis is concluded in chapter 6.

2 STATE-OF-THE-ART REVIEW

We will focus in the thesis on the quality of comparable corpora and their applications in bilingual lexicon extraction and CLIR. This chapter will be devoted to the description of existing techniques in areas related with the thesis work. We will first introduce in section 2.1 the work related with corpus quality, namely the methods for comparing different corpora and for extracting parallel subparts from comparable corpora. The approaches for bilingual lexicon extraction from bilingual corpora, both parallel and comparable corpora, will then be reviewed in section 2.2. Lastly, we would like to introduce in section 2.3 the CLIR models and standard resources used in CLIR experiments. This chapter will be concluded in section 2.4.

2.1 Comparable Corpus Quality

Different from parallel corpora, the quality of comparable corpora is an important characteristic featuring its usability, which has been discussed before in section 1.2. There have been only a few studies trying to investigate the formal quantification of how similar two corpora are. We will first review in section 2.1.1 previous attempts to measure the similarity of two corpora. These approaches either deal with two corpora in the same language, or are computationally infeasible in the case of bilingual corpora. We will then discuss in section 2.1.2 the work trying to extract parallel subparts from comparable corpora. These methods resemble our thesis work in that both try to extract a high-quality subpart from the original comparable corpus. There is however a significant difference as one will find from section 4.3.1.

2.1.1 Comparing Different Corpora

The task of comparing corpora is relevant to the *genres* of corpora. Genre is a concept often used in the literature on language variation. A genre can be defined as a category of text assigned on the basis of external criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognized grouping of texts (Lee, 2001). According to the user's philosophy, different sets of genres can be established for the text collection. As an example used in (Sharoff, 2010), one can find that there are 15 genres defined for the Brown Corpus, 70 genres used in David Lee's classification of the British National Corpus (BNC), and 120 genre labels in the Russian National Corpus (RNC). Let us take for an example 4 out of the 15 genres defined for the Brown Corpus:

- B. PRESS: Editorial (including *Institutional Daily*, *Personal*, *Letters to the Editor*)
- D. RELIGION (including *Books*, *Periodicals*, *Tracts*)
- J. LEARNED (including *Natural Sciences*, *Medicine*, *Mathematics*, *Social and Behavioral Sciences*, *Political Science, Law, Education*, *Humanities*, *Technology and Engineering*)
- N. FICTION: Adventure and Western (including *Novels*, *Short Stories*)

One can find from the four genres that texts in different genres (e.g. B. PRESS: Editorial and N. FICTION) cover more different topics that texts in the same genre (e.g. B. PRESS: Editorial) which fall into narrow topics related with editorial information. The task of genre assignment to large set of documents is usually accomplished through standard classification or clustering approaches (Sharoff, 2007; Sharoff, 2010), which resembles corpus comparison in that corpora in the same genre are similar in terms of content and text styles. The impact of corpus genre on the methods proposed in this thesis will be discussed in section 4.4. Corpus genres are not a major problem considered in our work (1) since they are a qualitative but not a quantitative

characteristic of the corpora, and (2) since comparable corpora are mostly built from texts belonging to the same or similar genres. We will now turn to the approaches to comparing two corpora.

In corpus linguistics, there have been some studies trying to compare two corpora in order to answer questions such as *"what sort of a corpus is this?"* and *"how does this corpus compare to that?"*. The studies on this topic such as (Rayson and Garside, 2000) and (Kilgarriff, 2001) make use of similar ideas to compare two corpora. The idea of these approaches is that key words can be identified for each corpus in a statistical way, which are used to differentiate two corpora. These methods mostly focus on a qualitative analysis and do not investigate into a precisely quantitative measure for corpus similarity. Moreover, they try to compare two corpora in the same language, which is different from the work in this thesis dealing with bilingual corpora.

As for bilingual corpora, the work (Saralegi et al., 2008) is the only one we can find to measure the degree of comparability of two corpora in different languages. They infer a global comparability score from the similarity of all cross-language document pairs. Let the corpus C_1 in the language L_1 have m documents $d_1^i (i = 1, 2, \ldots, m)$ and let C_2 be a corpus in the language L_2 consisting of n documents $d_2^j (j = 1, 2, \ldots, n)$. First, the document similarity between each two documents in different languages is computed using the tool *Dokusare* (Saralegi Urizar and Alegria Loinaz, 2007). One can thus obtain a $n * m$ matrix DM, where each element s_{ij}, corresponding to the element on row i and column j, is the similarity between d_1^i and d_2^j. They then define a process called EMD to estimate from DM a global score for the bilingual corpus. Since the number of documents in comparable corpora is usually very large, the computation of similarity over all cross-language document pairs makes this measure not practical, especially when the comparability measure needs to be called frequently. Recently, under the European 7th framework[1] project ACCURAT (Skadina et al., 2010), researchers plan to study and investigate existing measures and

[1]http://cordis.europa.eu/fp7/

metrics for assessing corpus comparability and document parallelism in the context of under-resourced comparable corpora. Despite the existing work, there has not been any study developing and evaluating the comparability measure in a systematic and quantitative way.

2.1.2 Parallel Subpart Extraction

Due to the lack of parallel resource, there have been some work trying to extract the parallel subparts, in most cases sentences, from comparable corpora. These studies are similar with our thesis work trying to improve the quality of the given comparable corpus. There is however a significant difference that we will try in our work to preserve most of the original vocabulary.

In (Zhao and Vogel, 2002), to extract parallel sentences from bilingual news collection, they combine sentence length models and lexicon-based models under a maximum likelihood framework. Let S denote a news story in the source language consisting of sentences $s_1, s_2, \ldots, s_j, \ldots, s_J$ and T a news story comprised of sentences $t_1, t_2, \ldots, t_i, \ldots, t_I$ in the target language. The goal is to find an alignment A that gives maximum likelihood of aligning S and T. They then try to align the sentences (s_j, t_i) by using dynamic programming to find the Viterbi path between two sentence sequences in (S, T). The distance between (s_j, t_i) is computed based on both a translation lexicon model and a sentence length model. This alignment model considers *insertions* and *deletions* of sentences, which are the common phenomenon in noisy corpora.

The work in (Munteanu et al., 2004) relies on a maximum entropy classifier to identify parallel sentences from comparable corpora. They first select from the corpus candidate document pairs through an information retrieval approach. Then candidate sentence pairs are chosen from the document pairs according to two heuristics. A maximum entropy classifier is lastly used to judge whether a sentence pair is parallel or not. This classifier makes use of the general features as follows:

- Lengths of the sentences, as well as the length difference and length ratio;
- Percentage of words on each side that have a translation on the other side.

and alignment features based on IBM Model 1 (Brown et al., 1993):

- Percentage and number of words that have no connection;
- The top 3 largest fertilities;
- Length of the longest contiguous span;
- Alignment score.

In the later work (Munteanu and Marcu, 2006), they detect which segments of the source sentence are translated into segments in the target sentence by analyzing potential similar sentence pairs using a signal processing-inspired approach. The sub-sentential knowledge extracted in this manner is then used to enhance the machine translation system.

In (AbduI-Rauf and Schwenk, 2009), they first employ an SMT system trained with a small amount of parallel text to translate the source language part of comparable corpora. The translated text is then used as queries to retrieve related documents in the target language part of comparable corpora, which results in the highly related document pairs. Based on those documents pairs containing related content, parallel sentences are extracted according to several criteria such as:

- The length correlation between the parallel sentence length;
- The Levenshtein distance (the number of operations required to transform one sentence into the other. The operations include insertions, deletions and substitutions.) and translation error rate.

The system framework of the work (Sarikaya et al., 2009) resembles that of the work (AbduI-Rauf and Schwenk, 2009), which consists of such steps as document pairing, sentence alignments and a boosting step. The document pairing step in this

work makes use of the SMT system to map the source language part of comparable corpora to the target language part, prior to choosing the document pairs according to the standard vector space comparison. Sentence alignment is then accomplished by using the BLEU score (Papineni et al., 2002) as the similarity measure to compare the sentence pairs in document pairs. The alignment algorithm first tries to extract parallel sentence pairs with high confidence, which server as the anchor points. It then performs iterative context extrapolation around the anchor points to include more sentence pairs. This boosting procedure is developed to increase the amount of new sentence pairs that are correctly paired but fail to achieve a sufficiently high similarity score according to the BLEU measure used.

The work in (Tillmann and Xu, 2009) is more efficient than the methods mentioned above, because it does not need to pair the documents before one can choose candidate sentence pairs. The candidate sentence pairs can be picked from any two documents. A similarity score can directly be assigned to a sentence pair to judge if they are parallel. The similarity measure is inspired by phrase-based SMT (Koehn et al., 2003) which includes a feature that scores phrases pairs using lexical weights. The feature needs to be computed for two directions: source to target and target to source. A sentence pair is then scored as a phrase pair that covers all the source and target words.

The recent studies (Do et al., 2010) and (Do et al., 2011) have similar framework with the methods (AbduI-Rauf and Schwenk, 2009; Sarikaya et al., 2009). The difference is that they train the SMT system on a noisy parallel corpus but not on a parallel corpus as used in previous work. The trained SMT system is then used to translate the source language part of the comparable corpora. The parallel sentence extraction step is then the same as those used in (Sarikaya et al., 2009; AbduI-Rauf and Schwenk, 2009). Lastly, the SMT system can be trained again on the new corpus comprised of the original parallel corpus and parallel sentences extracted in this round. This whole process is iterated to increase the parallel sentences obtained and to enhance the performance of SMT system.

2.2 Bilingual Lexicon Extraction

Bilingual lexicons, word pairs in parallel translation, are an important resource in multilingual NLP tasks. It is however expensive to construct this kind of resource by hand. The existing dictionaries can be easily turned into machine-readable lexicons, which do not contain the newly appearing words and terminologies. This can be a problem for applications dealing with the web data containing many out-of-vocabulary words. To solve this problem, previous work has tried to automatically extract the bilingual lexicons from bilingual corpora, both parallel and comparable ones. The differences between two classes of approaches making use of two different corpora lie in the fact that the search space is much larger in the case of comparable corpora due to the lack of alignment information. We will review in this section approaches for bilingual lexicon extraction from bilingual corpus, focusing more on comparable corpora, which is one of the applications we will investigate to exploit comparable corpora in the thesis. To be exact, the extraction techniques based on parallel corpora are first reviewed in section 2.2.1. The methods for lexicon extraction from comparable corpora are then introduced in section 2.2.2

2.2.1 Extraction from Parallel Corpora

Bilingual lexicon extraction from parallel corpora is highly related with statistical machine translation (Brown et al., 1993) and conventionally named *word alignment*. The general idea of the word alignment process is that words in parallel translation usually appear in parallel sentences. We will thus first introduce in this section the approaches to aligning sentences in parallel corpora. Following that procedure, various methods for word alignment in parallel sentences are discussed.

The intuition under most studies on sentence alignment is that parallel sentences in different languages should posse similar features such as sentence length, corresponding words, and cognates. The methods presented in (Brown et al., 1991) and (Gale and Church, 1991) rely on the assumption that the length, in terms of number of

words or characters, of parallel sentences is highly correlated. For instance, the English translation of a long French sentence should also be long. The sentences are then aligned in order to maximize the overall length similarity. To solve the maximization problem, dynamic programming is employed in (Brown et al., 1991) and Hidden Markov Models is used in (Gale and Church, 1991). In addition to the length-based approaches above, one can also rely on some other features to judge if two sentences are similar. For instance, the work in (Chen, 1993) makes use of the lexical information by constructing a simple statistical word-to-word translation on the fly during alignment. Two alignment engines are combined in (Simard and Plamondon, 1998) to benefit from the robustness of the character-based methods and the accuracy of stochastic translation models.

Compared to sentence alignment, word alignment is a higher level task and thus more difficult. It is an important component in statistical machine translation (Brown et al., 1993; Och and Ney, 2000; Och and Ney, 2003). The task of word alignment is usually an unsupervised learning task given sentence-aligned parallel corpus, where one tries to learn the parameters for the unobserved model, for instance the IBM Models 1-5, which best explains the parallel corpus. It is also possible to combine sentence alignment and word alignment in a single procedure. In (Kay and Roscheisen, 1993), they propose an algorithm based on the notion that which word in one text corresponds to which word in the other text is essentially based on the similarity of their distributions. Then they iteratively make use of the intuition that better word alignments can lead to better sentence alignments which in turn refines the word alignments. The algorithm appears to coverage to the correct sentence alignment in only a few iterations. One can easily find some other studies dealing with the word alignment problem. The work (Gaussier, 1998) proposes the flow network models which allow one to estimate the parameters in a computationally efficient way. The study (Goutte et al., 2004) views word alignment as a problem of orthogonality non-negative matrix factorization. Word alignment in parallel corpora is not the main focus of the thesis and we do not plan to review all the studies here. One

can refer to the book (Veronis, 2000) for more discussions on this topic. Such tools as GIZA++ (Och and Ney, 2000), the Berkeley Word Aligner [2] and NATools [3] are publicly available for word alignment in parallel corpora.

Recently, some researchers refer to supervised approaches for word alignment, which try to improve the alignment performance through the knowledge obtained from human alignments. In (Haghighi et al., 2009), the inversion transduction grammar (Wu, 1997) constraints are exploited in supervised word alignment methods. A discriminative model is presented in (DeNero and Klein, 2010) that directly predicts which set of phrasal translation rules should be extracted from a sentence pair. This model scores extraction sets: nested collections of all the overlapping phrase pairs consistent with an underlying word alignment. In (Setiawan et al., 2010), they address the modeling, parameter estimation and search challenges that arise from the introduction of reordering models which capture non-local reordering in alignment modeling. The work in (Xu and Chen, 2011) shows a surprising result that the gain of human alignment over a state of the art unsupervised method (GIZA++) is less than 1 point in BLEU score and that the benefit of improved alignment becomes smaller with more training data.

2.2.2 Extraction from Comparable Corpora

It is a costly task to build parallel corpora of high volume and researchers have resorted to comparable corpora for bilingual lexicon extraction, although the performance on comparable corpora is not as good as the methods relying on parallel corpora. The alignment information, for instance the sentence alignments in parallel corpora, is generally not available for comparable corpora, so the search space for finding translation candidates in comparable corpora is much larger than that of parallel corpora. The basic assumption behind most studies on bilingual lexicon extraction from comparable corpora is a distributional hypothesis, stating that words

[2] http://code.google.com/p/berkeleyaligner/
[3] http://linguateca.di.uminho.pt/natools/

which are translations of each other are likely to appear in similar contexts across languages. Under the same assumption, previous works differ in terms of context representation or strategies to compare the context. Bilingual lexicon extraction is an important task considered in the thesis in order to exploit comparable corpora and we will detail typical methods below.

The work (Rapp, 1995) suggests that the identification of word translations should be possible with non-parallel and even unrelated texts with the co-occurrence patterns. For example, if in a text of one language two words w_a and w_b co-occur more often than expected from chance, then in a text of another language those words which are translations of w_a and w_b should also co-occur more frequently than expected. He shows that the assumption holds even in the case of unrelated German/English text. When comparing an English and a German co-occurrence matrix of corresponding words, he finds a high correlation between the co-occurrence patterns of the two matrices when the rows and and columns of both matrices are in corresponding word order, and a low correlation when the rows and columns are in random order. The method proposed in (Rapp, 1999) can be seen as a simple case of the gradient descent method in (Rapp, 1995). The large number of permutations to be considered in (Rapp, 1995) is reduced to a much smaller number of vector comparisons in (Rapp, 1999), which provides a practical implementation based on the co-occurrence clue that yields good results. To be exact, for each of the source (resp. target) language word, a co-occurrence vector can be built by considering the source (resp. target) corpus. Then two vectors are mapped to the same space and compared with each other through the help of the seed dictionary. The candidate translations in the target vocabulary can be ranked according to its similarity with the source word.

The work (Fung, 1995) assumes the statistical correlations between words and their translations in non-parallel corpus, which says that words with productive context in one language translate into words with productive context in another language, and words with rigid context translate into words with rigid context. She proposes

the *context heterogeneity* to measure how productive the context of a word is in a given domain. Then words can be matched with their translations through the context heterogeneity measure. Their later work in (Fung and McKeown, 1997; Fung and Yee, 1998) relies on the co-occurrence assumption that is similar with the one used in the work (Rapp, 1999). The work (Fung and McKeown, 1997) makes use of a bilingual list of known translation pairs (i.e. seed word list) constructed from online dictionaries. For each word in one language, a vector is built consisting of its statistical word signature feature which is the correlation of this word and each other word occurring in the same segment as the source word. The segment size in their work is not fixed and changes according to the frequency of the source word in the corpus. Finally the similarity of two correlation vectors, bridged by the bilingual dictionary, is computed and the translation candidates with high similarity are chosen. The work (Fung and Yee, 1998) also uses the words surrounding the source word in a certain window as the context of the source word. The window size here is always taken as a single sentence and the weight of each context dimension is measured in the TF-IDF manner. Two vectors are then made comparable through the mapping based on a bilingual dictionary. This approach also takes into account the reliability of bilingual seed words.

The work (Déjean et al., 2002) tests several different models in bilingual lexicon extraction from parallel or comparable corpora in specialized domains. They show that the combination of the models significantly improves results, and that the use of the thesaurus UMLS/MeSH is of primary importance. The studies in (Chiao and Zweigenbaum, 2002) and (Chiao et al., 2004) also deal with comparable corpora in medicine domain. They test and compare in (Chiao and Zweigenbaum, 2002) several weighting factors and similarity measures for the strategy relying on the distributional context. In order to prune translation alternatives in the later work (Chiao et al., 2004), they reweigh translation candidates in the target language by applying the same translation algorithm but in the reverse direction. The latter method (Chiao et al., 2004) shows an improvement in the quality of top candidate translations. In

studies such as (Deléger and Zweigenbaum, 2008) and (Deléger and Zweigenbaum, 2009), they make use of comparable corpora consisting of lay and specialized medical documents to extract similar text segemnts.

In (Gaussier et al., 2004), they present a geometric view on bilingual lexicon extraction from comparable corpora, which can interpret all the context vector approaches in a uniform framework. With such a presentation, they extend the standard approaches to dealing with the problems of dictionary coverage and *polysemy/synonym*. They try to find a vector space in which synonyms dictionary entries are close to each other, while polysemous ones still select different neighbors. The context vectors of words are then mapped to the new vector space for comparison. They show that the combination of relatively simple methods helps to improve the performance of bilingual lexicon extraction significantly. The work (Robitaille et al., 2006) proposes a method for compiling bilingual terminologies of multi-word terms (MWTs) based on the seed terms. They first collect MWTs semantically similar to the seeds in each language. Then they work out the alignments between the MWTs in both sets. The intuition is that both seeds have the same related terms across languages, and they believe that this will simplify the alignment process. The alignment is done by generating a set of translation candidates using a compositional method, and by selecting the most probable translations from that set. The method (Morin et al., 2007) tries to extract French-Japanese terminologies from comparable corpora, considering both single and multi-word terms. They show the fact that the quality of comparable corpora is more important than the quantity and that this ensures the quality of acquired terminological resources. They also conclude that besides the domains and subdomains, the type of discourse is important as a characteristic of comparable corpora. The work (Morin and Daille, 2010) introduces a general framework for the lexical alignment of MWTs from comparable corpora, which includes a compositional translation process and the standard lexical context analysis. The compositional method bridges the gap between MWTs of different syntactic structures through morpholog-

ical links. In more recent work (Hazem and Morin, 2012), they review the task of lexicon extraction as a question-answering (QA) problem.

Different from above work making use of the lexical context consisting of words co-occurring with the source word in a certain window, the approach (Garera et al., 2009) uses the dependency structure as the context. They use contexts derived from head-words linked by dependency trees instead of the immediate adjacent lexical words. With the deep semantic information, they gain significant improvement compared to the approaches solely relying on the lexical context. The study (Yu and Tsujii, 2009) introduces the phenomenon that they called *dependency heterogeneity* which is if one prepares the corpora with a dependency syntactic analyzer, a word in the source language shares similar heads and modifiers with its translation in the target language, no matter whether the two words occur in similar context or not. The method is advantageous in that bilingual dictionary is not demanded so as to bridge the context vectors. Their approach shows satisfactory performance on a small test set.

The work (Shezaf and Rappoport, 2010) introduces a method to create a high quality bilingual dictionary (Hebrew-Spanish) given a noisy one built from two pivot language lexicons. The profiling process is based on two monolingual corpora. An essential part of the algorithm is called the non-aligned signatures (NAS) context, which consists of words that co-occur with the source word most strongly in the corpus. NAS is used as a cross-language similarity score to remove from generated lexicons incorrect translations using cross-lingual co-occurrences. NAS is computed from the number of headword signature words that may be translated using the input noisy lexicons into words in the signature of a candidate translation.

There are some other studies investigating specific problems in the task of bilingual lexicon extraction from comparable corpora. For example, since the rare words, appearing with low frequency in the corpus, do not have sufficient context information, it is difficult to find their correct translations from the corpus. In (Pekar et al., 2006), prior to performing the mapping between vector spaces of different languages, the

method models context vectors of rare words using their distributional similarity to words of the same language to predict unseen co-occurrences as well as to smooth rare, unreliable ones. In (Prochasson and Fung, 2011), to solve the rare word problem, they incorporate two features, namely a context-vector similarity and a co-occurrence model between words, in aligned documents in a machine learning approach.

The exploitation of comparable corpora is still a rapidly evolving area. The Building and Using Comparable Corpora (BUCC) workshop, initiated in 2008, is an annual event to investigate the techniques in this area. For the most recent advance, one may refer to the latest proceedings of the workshop (Zweigenbaum et al., 2011).

2.3 Cross-Language Information Retrieval

Cross-Language Information Retrieval (CLIR) is the task of finding documents written in a language different from that of the query. If attempts to model multilinguality in information retrieval date back to the early seventies (Salton, 1969), a renewed interest was brought to the field by the rise of the Web in the mid-nineties, as pages written in many different languages were all of a sudden availability. International organizations, governments of multilingual countries, to name the most important, have been traditional users of CLIR systems, but the need for such systems in everyday life, even though less ascertained, becomes more and more clear, with the development of travels, tourism and multilinguality, at all levels. The recent book by J.-Y. Nie on CLIR (Nie, 2010) exposes in detail the need for cross-language and multilingual IR. In this thesis, we will investigate both novel CLIR models and the enhancement of CLIR performance with lexicons extracted from comparable corpora, and we would like to review in this section the classic CLIR strategies. Since most of the precious CLIR models are the extensions of monolingual IR models, we will introduce below the monolingual IR models in section 2.3.1 and their extensions to CLIR settings in section 2.3.2.

2.3.1 Monolingual IR Models

The monolingual IR models, aiming at weighting a query towards the documents in a document collection, have been studied broadly since the notion of *information retrieval* was firstly proposed in (Mooers, 1950). Such IR models as boolean retrieval models, vector space models, probabilistic models and language modeling approaches have been investigated in order to satisfy the users' information needs (Manning et al., 2008). The monolingual models are the basis of CLIR models and we will introduce the representative monolingual IR models below.

Boolean Models

The boolean models try to find exact matches according to the logic operations such as the logical product *AND*, the logic sum *OR* and the logical difference *NOT*. These operations can be combined to constitute complex queries. For example, with the query *"chemistry AND physics"*, one will retrieve the documents containing both the words *"chemistry"* and *"physics"*. With the query *"chemistry OR physics"*, one will fetch the documents containing at least one of the two words *"chemistry"* and *"physics"*. Given the query *"chemistry NOT physics"*, one will get the documents containing the word *"chemistry"* and meanwhile not containing *"physics"*. The basic boolean models have two problems: it is not easy for untrained users to compose an efficient query to accurately represent their information need; the retrieved result set of the boolean model is not ranked, which is however important in many applications. Additional operators such as the *proximity operator* can be integrated into the basic boolean models to produce more flexible output.

Vector Space Models

The vector space models try to represent queries and documents as vectors prior to comparing the vectors. The early work in (Salton, 1971) represents the query and the document as vectors in a Euclidean space containing dimensions corresponding to the different terms. The similarity of two vectors is then measured by the cosine of the angle between two vectors.

A crucial part of the vector space model is the method to weight each term in the vectors. The TF-IDF approach is a broadly used approach for this purpose. The simple TF-IDF approach does not take into account the *synonyms* and *polysemys* which affect the performance of the vector space representation. The techniques such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been proposed to solve the problem.

Probabilistic Models

In the vector space models for IR reviewed above, matching is done in a formally defined but semantically imprecise calculus of index terms. The probabilistic approaches for IR try to model the retrieval tasks in the framework of probability theory, which is thus easier to explain and extend in a mathematical way. The random variable R is used to denote if the document in consideration is relevant to the query. The values for R are thus picked from $\{0, 1\}$, corresponding to *relevant* and *irrelevant* respectively.

The work (Robertson and Jones, 1976) proposes to rank the documents by the probability $P(R = 1|d, q)$ which is the probability of relevance given a query q and a document d. In their work, both q and d are represented as vectors, respectively corresponding to $\vec{q}$ and $\vec{d}$ consisting of the term weights in the query/document. The vectors $\vec{q}$ and $\vec{d}$ consist of only 0-1 values identifying whether a term appears in the query/document. Under this setting, one vector representation, either $\vec{q}$ or $\vec{d}$, may correspond to several queries/documents with the same vocabulary. The model is called Binary Independence Model (BIM) due to this fact. The probability $P(R = 1|\vec{d}, \vec{q})$ can then be interpreted as the probability of finding the relevant documents in the document set represented by the same $\vec{d}$ vector. To realize the model in practice, they make use of $\frac{P(R=1|\vec{d},\vec{q})}{P(R=0|\vec{d},\vec{q})}$ instead of $P(R = 1|\vec{d}, \vec{q})$ to rank the documents, where $R = 0$ denotes the *irrelevance* of q and d, the opposite of $R = 1$.

As they show, the new formula in use can result in the same ranking of the documents but is easier to compute. With the Bayes rule, they have:

$$\frac{P(R=1|\vec{d},\vec{q})}{P(R=0|\vec{d},\vec{q})} = \frac{P(\vec{d}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{d}|R=0,\vec{q})P(R=0|\vec{q})} \tag{2.1}$$

The parts $P(R=1|\vec{q})$ and $P(R=0|\vec{q})$ only depend on $\vec{q}$ and not on $\vec{d}$, having no effects on the ranking result, so they can be ignored in equation 2.1 without changing the ranking function of the original model. To further reduce the computational cost, they assume that the terms are independent given the relevance information and the query (i.e. a naive Bayes conditional independence assumption), and obtain:

$$\frac{P(R=1|\vec{d},\vec{q})}{P(R=0|\vec{d},\vec{q})} \triangleq \frac{\prod_t P(d_t|R=1,\vec{q})}{\prod_t P(d_t|R=0,\vec{q})}$$

where d_t is the t-th component (i.e. a term) of the document vector. By assuming that terms not occurring in the query are equally likely to occur in relevant and irrelevant documents, and using the logarithm instead of the products, they can get the resulting function used for ranking the documents w.r.t. the query q, which is:

$$RSV(q,d) = \sum_{t:d_t=q_t=1} \log \frac{P(d_t=1|R=1,\vec{q})P(d_t=0|R=0,\vec{q})}{P(d_t=1|R=0,\vec{q})P(d_t=0|R=1,\vec{q})} \tag{2.2}$$

where RSV denotes Retrieval Status Value. The computation of the probabilities in equation 2.2 can be implemented in various ways. For a detailed discussion of related techniques, one can refer to the book (Manning et al., 2008). The BIM model has some obvious problems. For example, one loses such information as the document length by using the boolean representations of the documents and queries. Meanwhile, different terms are treated independently in the BIM model, which is different from the real situation.

The simple binary representation does not consider the term frequency and document length, potentially harming the retrieval performance. The BM25 weighting scheme (Jones et al., 2000), also known as Okapi weighting, is a probabilistic model considering those features ignored in BIM and has shown satisfactory performance across many collections and search tasks. The work (Turtle and Croft, 1990) makes

use of Bayesian networks to model the complex dependencies between the documents and the queries to facilitate the retrieval tasks. Two networks are built: one for the document collection and the other one for the query. A new network will be built for each new query, which is then attached to the document network. This results in a full network that can generalize simpler boolean and probabilistic models. This model has motivated the InQuery system (Callan et al., 1992) which has been used broadly in many IR experiments.

Language Modeling Approaches

The technique of language modeling tries to model a language by assigning a probability to any language string. It was originally proposed for the speech recognition task to pick a real sentence from several candidates. The idea of using language modeling in IR relies on the assumption that a document fits the topic of the query if the query is likely to be produced by the language model of the document (i.e. document model). Intuitively, it means that if the document contains the frequent occurrences of the words in the query, the document is likely to be relative to the query.

Language modeling approaches are a very general framework which can be implementated in various ways. The basic language modeling approach used in IR is the *query likelihood model* (Manning et al., 2008). This model tries to score the document d w.r.t. a query q based on the probability $P(d|q)$ that will be computed from the language modeling approach. With the Bayes rule, one has:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)}$$

Since $P(q)$ is the same for all the documents and $P(d)$ can be treated as uniform across all the documents in text retrieval, we can choose to only consider the part $P(q|d)$ without changing the ranking of documents. $P(q|d)$ is the probability of observing the query q given the document d. This probability can be obtained by assuming a generative process, which is the query q is generated from the language model M_d of the document d. To estimate the value $P(q|M_d)$, a simple idea is to assume the

independency of words w in the query q, which equals to a multinomial distribution assumption on the words in the query. With these considerations, we can have:

$$P(q|M_d) = Z_q \prod_{w \in V_d} P(t|M_d)^{n_w} \tag{2.3}$$

where Z_q is a normalization factor only dependent on the query q and can thus be ignored in the ranking function. The number of occurrences of the word w in the vocabulary V_d of d is denoted as n_w. One usually makes use of the log format of equation 2.3 to get a formula in the form of sum instead of multiply, which can make the implementation more efficient in real-world IR systems.

In order to compute the value $P(w|M_d)$ in equation 2.3, one can refer to maximum likelihood estimation (MLE) for M_d, which leads us to:

$$P_{ml}(w|M_d) = \frac{n_w}{l_d}$$

where l_d is the length, in terms of word count, of the document d. A problem of the MLE approach is that it assigns zero probability to the words that do not appear in the document, which will cause the value in equation 2.3 to 0. It means that a query with a single word not present in the document will get a score 0, even if the query is relative to the query. Meanwhile, solely relying on the original times of occurrences leads to the over estimation for the words occurring only once which might have appeared by chance.

The techniques to solve the zero probability problem in language models are called *smoothing*. In the smoothing process, one discounts the non-zero probabilities and allocates some probability to the words not present in the document. Several approaches have been developed to smooth the probability such as the one (Chen and Goodman, 1996) in the speech recognition task. We introduce here two approaches broadly used to smooth language models in IR, namely the Jelinek-Mercer (JM) method and the Bayesian smoothing with Dirichlet priors (DIR) (Zhai and Lafferty, 2004).

- The JM smoothing approach makes use of a linear interpolation to smooth the maximum likelihood model M_{ml} with the collection language model $P(w|C)$

where C is a text collection. Formally, the smoothed probability $P_\lambda(w|M_d)$ can be :

$$P_\lambda(w|M_d) = (1-\lambda)P_{ml}(w|d) + \lambda P(w|C)$$

where λ is a smoothing parameter controlling to what extent the maximum likelihood model is smoothed.

- The DIR smoothing strategy makes use of the Dirichlet distribution which is the conjugate prior for the multinomial distribution considered for the language model. The parameters of the Dirichlet distribution are:

$$(\mu P(w_1|C), \mu P(w_2|C), \ldots, \mu P(w_n|C))$$

and the smoothed probability $P_\mu(w|M_d)$ can be written as:

$$P_\mu(w|M_d) = \frac{c(w;d) + \mu P(w|C)}{\sum_w c(w;d) + \mu}$$

where μ is a smoothing parameter and $c(w;d)$ is the number of occurrences of w in d.

In the language modeling approaches above, one always considers that the query q is generated from the language model M_d of the document. The reverse direction, generating the document d from a query model M_q, is also feasible. Since the document is often longer than the query, giving a better estimation for its language model, it is more convenient to rely on $P(q|M_d)$ than on $P(d|M_q)$ to rank the documents. An important extension of the basic language modeling approach is to consider the language models of both the query and the document. The two language models can then be compared using the Kullback-Lebler (KL) divergence (Zhai and Lafferty, 2001). Formally, the relevance value of d with respect to q can be measured by the negative KL-divergence function:

$$RSV(q,d) = -KL(M_q||M_d) = -\sum_{w \in V} P(w|M_q) \log \frac{P(w|M_q)}{P(w|M_d)} \tag{2.4}$$

which is the one we will actually make use of in the experiment sections.

2.3.2 CLIR Strategies

There are several ways to cross the language barrier in CLIR models: through mapping the document representation into the query representation space (an approach known as document translation), through mapping the query representation into the document representation (an approach known as query translation) or through mapping both representations into a third space (interlingua approach). As for implementation, existing CLIR models fall into two categories:

- Model-independent approaches. Model-independent approaches treat translation and retrieval as two separate processes. The queries or the documents are first translated into the corresponding language of the documents or the queries. Monolingual IR models are then applied directly. A typical and also broadly used approach of this type is the machine translation (MT) approach (e.g. (Kraaij et al., 2003; Braschler, 2004)) which employs MT systems to translate the queries or documents before the monolingual retrieval process. The biggest problem of this approach is that MT systems are not readily available for all the language pairs. As one can see, the model-independent approaches are more like a monolingual IR task plus a translation process which is independent from the monolingual IR model.

- Model-dependent approaches. These methods integrate the translation and retrieval processes in a uniform framework. The first model-dependent approach commonly used in CLIR is the structure query technique (Pirkola, 1998; Darwish and Oard, 2003) which counts the occurrences of the translation candidates as the occurrence of the original word. This approach is realized as the SYN operator in the InQuery system. We list in Appendix A an example of the structured query technique. The other model-dependent approach, developed in (Federico and Bertoldi, 2002; Kraaij et al., 2003) in the context of language models, have the advantage of accounting better for the uncertainty of translation during retrieval. We will briefly introduce here the Query Translation

(QT) model and the Document Translation (DT) model (Kraaij et al., 2003). In order to explain how QT and DT work, we make use here of the language modeling approach based on KL-divergence defined in equation 2.4.

We denote d_t as the document in the target language, q_s as the query in the source language, V_t as the vocabulary of the target language, and V_s as the vocabulary of the source language. The QT model tries to estimate the query model M_{q_s} of the query q_s in the target language. Under these settings, the retrieval function in equation 2.4 can be rewritten as:

$$RSV(q_s, d_t) = -KL(M_{q_s}||M_{d_t}) = -\sum_{w_t \in V_t} P(w_t|M_{q_s}) \log \frac{P(w_t|M_{q_s})}{P(w_t|M_{d_t})} \tag{2.5}$$

and $P(w_t|M_{q_s})$ can be estimated as:

$$\begin{aligned} P(w_t|M_{q_s}) &= \sum_{w_s \in V_s} P(w_s, w_t|M_{q_s}) = \sum_{w_s \in V_s} P(w_t|w_s, M_{q_s}) P(w_s|M_{q_s}) \\ &\approx \sum_{w_s \in V_s} P(w_t|w_s) P(w_s|M_{q_s}) \end{aligned}$$

where the approximation is made by assuming the dependence of w_t on w_s is independent from M_{q_s}. The probability $P(w_t|w_s)$ can be estimated from the bilingual dictionaries or corpora, and $P(w_s|M_{q_s})$ can be simply estimated from the language model used for the query. Different from QT, the DT model tries to estimate the document model M_{d_t} in the source language. In this case, the retrieval function in equation 2.4 can be rewritten as:

$$RSV(q_s, d_t) = -KL(M_{q_s}||M_{d_t}) = -\sum_{w_s \in V_s} P(w_s|M_{q_s}) \log \frac{P(w_s|M_{q_s})}{P(w_s|M_{d_t})} \tag{2.6}$$

and $P(w_s|M_{d_t})$ can be estimated as:

$$\begin{aligned} P(w_s|M_{d_t}) &= \sum_{w_t \in V_t} P(w_s, w_t|M_{d_t}) = \sum_{w_t \in V_t} P(w_s|w_t, M_{d_t}) P(w_t|M_{d_t}) \\ &\approx \sum_{w_t \in V_t} P(w_s|w_t) P(w_t|M_{d_t}) \end{aligned}$$

where the approximation is made by assuming the dependence of w_s on w_t is independent from M_{q_t}. Similar with the QT model, the probability $P(w_s|w_t)$

should be estimated from the bilingual dictionaries or corpora, and $P(w_t|M_{d_t})$ is to be computed from the language model of the document.

2.3.3 Resources for IR Experiments

For academic research, there have been standard resources that can be employed to examine the performance of IR systems. We list here several of them which have been used broadly. We will introduce both the test collections and publicly available IR systems.

Several test collections with human judgements have been developed to help researchers evaluate their IR models. We list below the details of these data sets.

- Text Retrieval Conference (TREC). The TREC test set was initialized by the American National Institute of Standards and Technology (NIST), aiming at evaluating IR performance in different domains and environments, and was divided into various tracks. For example, the *Chemical IR Track* addresses challenges in building testbeds to evaluate the state of the art in chemical information retrieval. The *Web Track* investigates the web related retrieval tasks, where a huge amount of data is involved. The *Cross-Language Track* provides data sets adjusted for the CLIR tasks, with the queries in many languages. The *Question Answering Track* deals with deep semantics understanding in the retrieval tasks. The track that is mostly related with the IR task discussed in the thesis is the *Ad Hoc Track* and the *Cross-Language Track*.

- Cross Language Evaluation Forum (CLEF). The CLEF test set is specifically designed for the CLIR tasks. Most of the CLEF data set covers the main European languages such as English, French and Germany. Similar with the TREC test set, the CLEF test set is also divided into several tracks. For example, the *Multilingual Information Retrieval Track* provides the possibility of evaluating the search task towards a multilingual document collection using a query in several possible languages. The *Bilingual Information Retrieval Track*

is provided where the query can be chosen from one language and the target documents can be chosen from one of several languages. Different from the multilingual track, the documents in the bilingual track belong to only one language once chosen. The bilingual track is the one we will actually make use of in the thesis. There are also some other tracks such as the *GIRT Track* supporting the CLIR task in specific domains.

- NII Test Collections for IR Systems (NTCIR). Similar with CLEF, the NTCIR workshop is a series of evaluation workshops designed to evaluate relative research work. It is currently supported by Japan Society for Promotion of Science (JSPS) and National Institute of Informatics (NII). Such tracks as information retrieval, question answering and text summarization have been designed to evaluate various tasks. In addition to general IR tasks, the NTCIR forum pays special attention to Asian language (e.g. Japanese and Chinese) IR and CLIR.

Besides the publicly available test sets introduced above, one also needs an IR system with which one can easily test different IR models. We introduce here several IR systems which are open-source and have been used in many studies.

- **Lemur.** The Lemur system was developed by the University of Massachusetts, Amherst and Carnegie Mellon University. It provides search engines, browser toolbars, text analysis tools, and data resources for the research in information retrieval and text mining. The Indri search engine embedded in the Lemur project provides state-of-the-art retrieval performance and out-of-the-box user interface. The InQuery language coming with the Lemur system provides flexible structured queries such as the operator SYN. The project website is: http://www.lemurproject.org/.

- **Terrier.** The Terrier IR platform is a search engine developed by University of Glasgow. It is deployable on large-scale collections of documents. They have implemented in Terrier classic IR models such as TF-IDF, language modeling approaches, and DFR models. The system has been modularized for

easy alternation. Moreover, the system has provided customized interface for standard IR experiments such as CLEF and TREC. The project website is: http://terrier.org/.

- **Lucene.** The Apache Lucene is a high-performance text search engine, which is usually extended for commercial use. This system also provides support for query biased summarization. Regarding the IR models, Lucene only implements the boolean model and the basic vector space model. The original distribution of Lucene does not come with such classic IR models as language modeling approaches. The project website is: http://lucene.apache.org/.

2.4 Conclusion

As explained in section 1.3, we plan to focus in this thesis on comparable corpus quality and its application in NLP tasks. We have thus reviewed in this chapter existing work related to corpus quality and two NLP tasks, namely bilingual lexicon extraction and CLIR, relying on comparable corpora.

We have first discussed existing work concerning corpus quality. One class of work tries to compare the similarity of two corpora, in the same or different languages. The work in corpus linguistics tries to extract the key words that can differentiate two corpora in the same language, which mostly focus on a qualitative analysis but not a quantitive analysis. The work inferring a global comparability score for bilingual corpora is however computationally infeasible for large corpora. In a word, there has not been any practical measure on which one can rely to differentiate various comparability levels. The other class of work aims to extract parallel subparts from comparable corpora, which is similar with the thesis work trying to enhance comparable corpus quality. There is however a significant difference as we will show later.

We have then focused on two NLP tasks making use of comparable corpora. The first one is bilingual lexicon extraction which can be realized with both parallel corpora and comparable corpora. Bilingual lexicon extraction from comparable corpora relies

on a distributional hypothesis that similar words should appear in similar context in different languages. Previous studies on this topic differ in terms of either the representation for word context or strategies to compare two context sets. The second task we have reviewed is CLIR which extends the monolingual IR models. We have also discussed in this chapter standard resources for IR experiments.

3 COMPARABILITY MEASURES

As we have discussed in section 1.2, the quality of comparable corpora affects their usability in the NLP applications. However, the notion of *comparability* is rather vague in previous works, which constrains us from differentiating comparable corpora of different quality. In this chapter, we will try to develop efficient measures to quantify the quality of comparable corpora. We will first formalize in section 3.1 the concept of comparability following the usage experience from NLP practice. Then we will develop in section 3.2 several measures to approximate the notion of comparability. At last, in section 3.3, those comparability measures proposed above are validated by the experiments to show their ability of capturing different comparability levels as well as their robustness to dictionary changes. The chapter will be conclude in section 3.4.

3.1 The Notion of Comparability

Without a clear notion of *comparability*, it is hard to tell whether one comparable corpus can be seen as more comparable than the other. For the quality of comparable corpora, we make additional discussions here to complement the ones made in section 1.2. A comparability measure intends to capture the different comparability levels, i.e. to provide an estimation that if a comparable corpus has the quality good enough for relative NLP tasks. In this sense, the way we define comparability should depend on the target applications in consideration, since different applications might prefer comparable corpora of different genres. For example, in (Ni et al., 2009), one needs to get the documents with aligned topics in order to build the multilingual topic models. Comparable corpora containing overlapping information and without topic alignments are however sufficient for the task of bilingual lexicon extraction (refer to

the work cited in section 2.2.2). We will focus in this thesis on the task of bilingual lexicon extraction[1], which is one of the mostly investigated applications using comparable corpora. However, we believe that the notion of comparability argued here works with other applications as well.

The comparability measure can be defined on various levels such as sentences, documents and corpora. If one considers the comparability on the document or sentence level, it can directly be treated as the similarity of sentences/documents in two languages, where such classic strategies as vector space model can be applied. In this chapter, we aim to develop a measure that can work on all levels with specific attention to the corpus level. We intend to define comparability so as to reflect the usability of comparable corpora in NLP tasks and to direct the enhancement of existing comparable corpora. The intuition we are able to follow is that the following comparable corpora have decreasing comparability levels:

1. Parallel corpora;
2. Parallel corpora with noise;
3. Non-parallel corpora covering overlapping topics (i.e. strongly comparable);
4. Non-parallel corpora covering different topics (i.e. weakly comparable).

The bilingual corpus is used to bridge the language barrier in NLP tasks, so the more *bridge* information one can find from the corpus, the more one can rely on the corpus in multilingual NLP tasks, which is the other way of explaining our intuition above. It is thus no wonder the parallel corpus is of the best quality and the non-parallel corpus covering different topics is of the worst quality. Given these discussions, we would like to give additional comments on different text size levels. According to the above intuition, both the parallel corpus and the parallel document pair (i.e. only 2 documents) are of the highest comparability level, since there is not any other text of

[1]The CLIR task does not directly depend on comparable corpora. It depends on the bilingual lexicons extracted from comparable corpora, as we have discussed in section 1.1.

higher quality than *parallel* text. However, in terms of usability, the parallel corpus should be more useful in most NLP tasks such as statistical machine translation. Judged from the usability, one should say, compared to the parallel document pair, the parallel corpus is of higher quality and thus more comparable. The conflict here can be avoided by assuming that we are only interested in comparing the text resources belonging to similar scales.

The comparability measure we will develop needs to correlate with the intuition above, meaning that the measure can capture different comparability levels reflected by different comparable corpora. Moreover, we prefer to have a measure as simple in computational complexity as possible because it will be used intensively in other algorithms, which will be shown in the approaches to improving the corpus quality in section 4.2.

3.2 Developing Comparability Measures

Following the discussions on the notion of comparability in section 3.1, several implementations for comparability are proposed and discussed in this section which are:

- *Vocabulary overlapping approaches.* These methods try to measure how much the vocabularies of two corpora overlap with each other. It is realized through mapping the vocabulary of the source language corpus to a vocabulary in the target language, prior to the comparison of two vocabularies. We will make use in this thesis of the bilingual dictionary to perform the mapping process. When the *polysemy* of the words is not taken into account, we develop several context-free comparability measures. Otherwise, we can develop the corresponding context-based comparability measures considering word sense disambiguation. These measures will be presented in section 3.2.1.

- *Vector space approach.* This method makes use of the classic vector representation of the documents. The two monolingual corpora can be represented as two

vectors in two languages. Then the bilingual vector similarity can be computed through mapping two vectors into the same language space using approaches such as the one in (Gaussier et al., 2004). This approach will be introduced in section 3.2.2, severing as one baseline measure.

- *Machine translation approaches.* It is intuitive that the source language corpus can be translated into the target language, resulting in two corpora in the same language that can directly be compared relying on the matured techniques to evaluate machine translation systems. These measures will be introduced in section 3.2.2 and server as the second baseline measure.

For convenience, the following discussions will be made in the context of French-English comparable corpora. The conclusions however hold for any language pair.

3.2.1 Measures Based on Vocabulary Overlapping

In order to develop comparability measures satisfying the notion in section 3.1, we make use here of the intuition that it is easier to find the translation for each word in comparable corpora of higher quality. In practice, the *mathematical expectation* is employed to quantify if it is easy to find the translation candidates in the corpus. We will investigate in this section two strategies to implement the intuition: one is to consider the translation pairs without context information and the other one relies on context-based disambiguation.

Measures without Context

The measures introduced in this part are the extensions of the ones proposed in our former work (Li and Gaussier, 2010) and solely depend on the words themselves, ignoring any context information. It is easier to find the translation pairs between documents that are more comparable to each other, since authors tend to use similar words to depict similar topics in different languages (see (Morin et al., 2007) for a related analysis). The intuition can also be supported by the list of four types of

comparable corpora with decreasing comparability levels listed in section 3.1. It is easy to find many translation pairs from the parallel corpus, which is harder in the context of non-parallel corpora covering different topics. This intuition provides us with the possibility of measuring the degree of comparability based on if it is easy to find the translations for the words in the corpora.

As a natural choice, we make use of the *mathematical expectation* to estimate the difficulty in finding the translation pairs in the corpus. Let us assume that we have a French-English comparable corpus $\mathcal{C}$ consisting of a French part $\mathcal{C}_f$ and an English part $\mathcal{C}_e$. If we consider the translation process from the English part to the French part, the comparability measure M_{ef} can be defined as the expectation of finding, for each English word w_e in the vocabulary $\mathcal{C}_e^v$ of $\mathcal{C}_e$, its translation in the vocabulary $\mathcal{C}_f^v$ of $\mathcal{C}_f$. The definition for M_{ef} directly reflects our intuition. As one can note, a general English-French bilingual dictionary $\mathcal{D}$, independent from the corpus $\mathcal{C}$, is required to judge if two words are the translation of each other. Let σ be a function which indicates whether a translation from the translation set $\mathcal{T}_w$ of a word w is found in the vocabulary $\mathcal{C}^v$ of a corpus $\mathcal{C}$, i.e.:

$$\sigma(w, \mathcal{C}^v) = \begin{cases} 1 & \text{iff } \mathcal{T}_w \cap \mathcal{C}^v \neq \emptyset \\ 0 & \text{else} \end{cases} \tag{3.1}$$

M_{ef} is then defined as:

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \mathbb{E}(\sigma(w, \mathcal{C}_f^v) | w \in \mathcal{C}_e^v) = \sum_{w \in \mathcal{C}_e^v} \sigma(w, \mathcal{C}_f^v) \cdot \Pr(w \in \mathcal{C}_e^v) \tag{3.2}$$

where $\mathcal{D}_e^v$ is the English vocabulary of the given bilingual dictionary $\mathcal{D}$. As assumed above, comparable corpora and the general bilingual dictionary are independent from one another. It is thus natural to assume that the dictionary covers a substantial part of $\mathcal{C}_e^v$ and this substantial part can well represent the whole vocabulary. It means the expectation of finding the translation in $\mathcal{C}_f^v$ of a word w is the same for w in $\mathcal{C}_e^v$ and in $\mathcal{C}_e^v \cap \mathcal{D}_e^v$. This assumption amounts to a practical version of M_{ef} in equation 3.3:

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) \cdot \Pr(w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v) \tag{3.3}$$

The assumption is not always reliable while, for example, a remarkable part of the corpus vocabulary is not covered by the bilingual dictionary. This can happen when the bilingual dictionary is small or when the corpus contains content in the language other than the language of the corpus itself. The latter case is more like a technical problem where a language identifier (Lins and Gonçalves, 2004) can be employed to filter out the noisy text. We will thus only consider the first problem, regarding the dictionary coverage, in the following sections.

There are several possibilities to estimate $\Pr(w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v)$ in equation 3.3. However, the presence of common words suggests that one should not solely rely on the number of occurrences (i.e. term frequency), which is a broadly used approach in other fields like unigram language models, since the high-frequency words will dominate the final results. For example, in the *Europarl* corpus, the English word *Europe* and the French word `Europe` are very common words. It means that even if one piece of English text and one piece of French text are randomly picked from the *Europarl* corpus, one can still expect to find many translation pairs *Europe*-`Europe`. To avoid the bias common words can introduce in the comparability measure, one can weight each word w as ρ_w through TF-IDF or through the simple Presence/Absence (P/A) criterion. In this case, the part $\Pr(w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v)$ can be estimated as:

$$\Pr(w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v) = \frac{\rho_w}{\sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \rho_w} \tag{3.4}$$

With the P/A criterion, the weight ρ_w is 1 if and only if $w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v$, otherwise the value is 0. Alternatively, considering the TF-IDF weight for each word w, the weighting function ρ_w can be defined as:

$$\rho_w = tf_w * \log(1 + \frac{|\mathcal{C}_e|}{1 + df_w}) \tag{3.5}$$

where tf_w is the term frequency of the word w in the corpus $\mathcal{C}_e$, $|\mathcal{C}_e|$ is the total number of douments in the corpus $\mathcal{C}_e$ and df_w is the number of documents containing the word w.

Similarly, when considering the translation process from the French part to the English part, the counterpart of M_{ef}, M_{fe}, can be written as:

$$M_{fe}(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v) \cdot \Pr(w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v) \tag{3.6}$$

where $\Pr(w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v)$ is defined in the same way as in equation 3.4.

The two asymmetric measures M_{ef} and M_{fe} above reflect the degree of comparability when considering the translation process in only one direction. They can be combined to form a comprehensive measure M by reviewing the two directions as a whole. The difference between M_{ef} in equation 3.3 and M_{fe} in equation 3.6 comes from the part which estimates the probability of a word in the corpus vocabulary. We denote $\mathcal{C}^v$ as the whole vocabulary of $\mathcal{C}$ and $\mathcal{D}^v$ as the whole vocabulary of $\mathcal{D}$. Following the same idea, considering the English and French vocabularies as a whole, one can directly obtain $\Pr(w \in \mathcal{C}^v \cap \mathcal{D}^v)$ by replacing $\mathcal{C}_e^v \cap \mathcal{D}_e^v$ with $\mathcal{C}^v \cap \mathcal{D}^v$ in equation 3.4. The combined measure M, considering the translation in both directions, can then be written as:

$$M(\mathcal{C}_e, \mathcal{C}_f) = \sum_{w \in \mathcal{C}^v \cap \mathcal{D}^v} \sigma(w, \mathcal{C}^v) \cdot \Pr(w \in \mathcal{C}^v \cap \mathcal{D}^v) \tag{3.7}$$

We give here additional comments on the P/A criterion which is the one we will finally make use of in other algorithms in section 4.2. With this criterion, $\Pr(w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v)$ is directly $\frac{1}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v|}$ and $\Pr(w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v)$ is $\frac{1}{|\mathcal{C}_f^v \cap \mathcal{D}_f^v|}$. One can then obtain:

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \frac{1}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v|} \sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) \tag{3.8}$$

$$M_{fe}(\mathcal{C}_e, \mathcal{C}_f) = \frac{1}{|\mathcal{C}_f^v \cap \mathcal{D}_f^v|} \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v) \tag{3.9}$$

From equations 3.8 and 3.9, one can find that M_{ef} and M_{fe} can be interpreted as the proportion of words in one language (English or French) translated into the other language in the corpus. At last, according to equation 3.7, the combined measure M can be written as:

$$M(\mathcal{C}_e, \mathcal{C}_f) = \frac{\sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) + \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v)}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v| + |\mathcal{C}_f^v \cap \mathcal{D}_f^v|} \tag{3.10}$$

which corresponds to the overall proportion of words for which a translation can be found in comparable corpus. One can notice from equation 3.10 that M is a symmetric measure.

Measures with Context

In the measures without context information defined above, two words are treated as a translation pair, as in equation 3.1, if they appear as an entity in the bilingual dictionary. Due to the fact of *polysemy*, two words can be treated as the translation of each other according to the dictionary but might hold different senses in the corpus. We can make use of the simple assumption that words in parallel translation usually appear in similar contexts to disambiguate the translation candidates in the dictionary. This same assumption has been broadly exploited in the bilingual lexicon extraction tasks. We will embed the assumption in the function σ in equation 3.1 to build the context version of all comparability measures without context. Let us assume that the English word w_e (resp. French word w_f) appears in the context word set $\mathcal{S}_e$ (resp. $\mathcal{S}_f$) consisting of the words surrounding w_e (resp. w_f) in a certain window in the corpora. Then the similarity of the two context sets is measured by the overlap of the two sets which is directly the proportion of words of which the translation can be found in the counterpart set. Formally, the similarity of w_e and w_f, in terms of their context similarity, can be written as:

$$sim(w_e, w_f) = \frac{\sum_{w \in \mathcal{S}_e \cap \mathcal{D}_e^v} \sigma(w, \mathcal{S}_f) + \sum_{w \in \mathcal{S}_f \cap \mathcal{D}_f^v} \sigma(w, \mathcal{S}_e)}{|\mathcal{S}_e \cap \mathcal{D}_e^v| + |\mathcal{S}_f \cap \mathcal{D}_f^v|}$$

The enhanced version of the function σ in equation 3.1 is then defined as:

$$\sigma_c(w, \mathcal{C}^v) = \begin{cases} 1 & \text{iff } \exists w' \in \mathcal{T}_w \cap \mathcal{C}^v, \ \ \text{sim}(w, w') > \delta \\ 0 & \text{else} \end{cases}$$

where δ, empirically set to 0.3 in our experiments, is the threshold for the similarity. A word w is deemed to be translated, according to the function σ_c, if at least one of its translations w' identified by the function σ in the corpus, is similar to w based on the context similarity measure $\text{sim}(w, w')$. Replacing σ with σ_c in equation 3.1 will lead

to the context versions of the comparability measures above, which are respectively denoted as M^c_{ef}, M^c_{fe} and M^c.

3.2.2 Baseline Measures

In this section, we develop two categories of measures as baselines respectively based on the classic vector space model and the machine translation system. The first category, only relying on the bilingual dictionary, tries to map the word vectors from one language to the other and then compares two vectors in the same language. The second category makes use of machine translation systems to translate one corpus to the language of the other corpus, and then compares the two corpora in the same language. The second category of measures is not a practical method due to the lack of machine translation systems between minor language pairs.

The Measure Based on Vector Space Model

It is common practice to represent documents as vectors consisting of words occurring in the documents. The weight of each dimension, i.e. a word, is determined by methods such as TF-IDF. In the cross-language settings, one needs to compare two vectors in different languages, i.e. one vector in the source language needs to be mapped to a vector in the target language. We make use here a strategy similar with the one summarized in (Gaussier et al., 2004). Let us assume one has a vector $\vec{v_e}$ for the English corpus and a vector $\vec{v_f}$ for the French corpus. $\vec{v_f}$ is mapped to $\vec{v_e}$ by accumulating the contributions from words in $\vec{v_f}$ that yield identical translations. Whether one uses Dice, Jaccard or Cosine coefficient, the dot-product usually plays a central role. Let $f(w_e)$ (resp. $f(w_f)$) denote the weight of the word w_e (resp. w_f) in the vector $\vec{v_e}$ (resp. $\vec{v_f}$). In this paper, the weight is defined in the TF-IDF style as in equation 3.5. The dot product between the vectors $\vec{v_e}$ and $\vec{v_f}$ is obtained from:

$$< \vec{v_e}, \vec{v_f} >= \sum_{w_e \in \vec{v_e}} f(w_e) \sum_{w_f \in \mathcal{T}_{w_f} \cap \vec{v_f}} f(w_f) \tag{3.11}$$

where $\mathcal{T}_{w_f}$ is the translation set of w_f in the bilingual dictionary. Based on the dot product defined in equation 3.11, different measures can be deviated directly. We will make use of the cosine similarity in this paper and obtain the comparability measure M^v based on vector space model as:

$$M^v = cos(\vec{v_e}, \vec{v_f}) = \frac{< \vec{v_e}, \vec{v_f} >}{\sqrt{\sum_{w_e \in \vec{v_e}} (f(w_e)^2 + (\sum_{w_f \in \mathcal{T}_{w_f} \cap \vec{v_f}} f(w_f))^2)}}$$

Measures Based on Machine Translation

In addition to the measure based on vector space model, we propose here a direct approach for measuring the comparability based on machine translation systems. These approaches are not feasible solutions for comparability measure and just described here for the purpose of comparison, since machine translation systems are not readily available for all the languages. The idea is to translate the corpora from the source language into the target language prior to comparing two corpora in the same language where various methods can be employed. For the translation task, we will make use here of the google machine translation tool[2] which is one of the best systems as we can find.

In order to compare monolingual texts, one is the natural text and the other one is produced by the MT system, we will employ the transforms of the BLEU score (Papineni et al., 2002) initially developed to automatically evaluate MT systems with the reference translations. The idea is that a good translation candidate should share many n-grams with the reference translations. The computation of BLEU scores however depends on the statistics from the source-target sentence pairs which are not available in the context of comparable corpora. We will generalize the idea of BLEU and represent the corpus as a vector of n-grams. Each dimension of the vector is the weight of the corresponding n-gram computed in the same TF-IDF style as before. While setting n to 1, one will actually arrive at the standard vector space model of words. Two vectors of n-grams are then compared with each other with standard measures such as the cosine similarity used in this paper. Restricted by the

[2]http://translate.google.com

computational complexity, n will be set to 1 and 2 only in our work, corresponding to the comparability measures denoted as M^{g1} and M^{g2}.

3.3 Experimental Validation

In this section, we validate by the experiments the performance of the proposed comparability measures. We first introduce in section 3.3.1 resources used for the experiments. The procedure developed to validate the comparability measures is then detailed in section 3.3.2. To be exact, we design several comparable corpora with gold-standard comparability levels, prior to evaluating the comparability measures in terms of correlation scores with gold standard and robustness to dictionary changes.

3.3.1 Resources in the Experiments

For the experiments designed to compare and validate the comparability measures, several corpora are used: The parallel English-French *Europarl*[3] corpus, the TREC[4] *Associated Press* corpus and the corpora used in the multilingual track of CLEF[5] which includes the *Los Angeles Times*, *Glasgow Herald*, *Le Monde*, *SDA French 94* and *SDA French 95*. In addition to these existing corpora, two monolingual corpora from the Wikipedia dump[6] are built. For English, we construct the corpus *Wiki-En* by retrievling all the articles below the root category *Society*. For French, the corpus *Wiki-Fr* is built by getting all the articles below the category *Société*. The information of all the corpora used in the experiments is detailed in Table 3.1. Since the Europarl corpus we use has been aligned on the sentence level and stored as sentence pairs, the number of documents (Nr. docs) is not available in the table.

[3]http://www.statmt.org/europarl/

[4]http://trec.nist.gov/

[5]http://www.clef-campaign.org

[6]The Wikipedia dump files can be downloaded at http://download.wikimedia.org. In this paper, we use the English dump file on July 13, 2009 and the French dump file on July 7, 2009.

Table 3.1
The information of the corpora used in the experiments in section 3.3. (k=1000, m=1000k)

Name	Short name	Language	Nr. docs	Nr. words
Europarl	Europarl	English	-	51m
		French	-	55m
Associated Press	AP	English	243k	126m
Los Angeles Times	LAT94	English	113k	71m
Glasgow Herald	GH95	English	56k	27m
Le monde	MON94	French	44k	24m
SDA French 1994	SDA94	French	43k	13m
SDA French 1995	SDA95	French	43k	13m
Wiki-En	Wiki-En	English	368k	163m
Wiki-Fr	Wiki-Fr	French	378k	169m

The bilingual dictionary used in our experiments is constructed from an online dictionary. It consists of 33k distinct English words and 28k distinct French words, which constitutes 76k translation pairs. Standard preprocessing steps: tokenization, POS-tagging and lemmatization are performed on all the linguistic resources. We directly work on lemmatized forms of content words (nouns, verbs, adjectives, adverbs). This dictionary and some of the corpora will be also used in the process of enhancing the corpus quality in chapter 4.

3.3.2 Evaluating Comparability Measures

We try to evaluate the comparability measures in the following aspects:

1. Whether the designed comparability measures can capture the different comparability levels in the corpora;

2. Whether the proposed measures are robust to the dictionary coverage.

The first point has to be validated with corpora of gold comparability levels, and this kind of comparable corpora are not readily available as we can find. In this paper, the gold standards are built from the existing parallel corpora and monolingual corpora. The second point can be validated with various dictionaries of different coverages. Dictionaries of various sizes are obtained by randomly sampling the original bilingual dictionary. We will detail all the processes below.

Constructing Test Corpora

In order to test the comparability measures introduced before, one needs to have some corpora of known comparability levels. However, there has not been any work which tries to build the corpora with quantified comparability levels. In our work, the basic idea comes from the intuition established in section 3.1, which is that the comparability levels decrease from the *parallel corpus* to *non-parallel corpora covering different topics*. Following the intuition, we plan to develop gold-standard comparability scores from the parallel corpus *Europarl* and the monolingual corpus *AP*. We plan to build three groups of comparable corpora, following the different comparability levels listed in section 3.1:

- G_a: All the comparable corpora in G_a are built from the parallel corpus *Europarl*. One starts from the parallel corpus *Europarl*, of which the comparability level is the highest, and decreases the quality by exchanging some parallel parts with some non-parallel parts in *Europarl*. In another word, we bring some noise into the parallel corpus and the noise covers similar topics with the original parallel corpus.

- G_b: Similar with the construction process of G_a, one also starts from the parallel corpus *Europarl*. The difference is that one exchanges some parallel parts of *Europarl* with the content from the *AP* corpus. That is to say, the noise brought to the parallel corpus covers different topics from the original parallel corpus.

- G_c: One has tried to build two groups of corpora starting from the parallel corpus, producing parallel corpora with noise and corresponding to the first two classes (i.e. *parallel* and *parallel with noise*) of corpora mentioned in section 3.1. When all the parallel parts have been exchanged, as the case in G_a, one will get comparable corpora belonging to the class *non-parallel corpora covering similar topics*. In G_c, we start from the lowest comparability levels in G_a, i.e. the ones containing no parallel parts, and exchange certain parts containing similar topics with content from the AP corpus, meaning that the noise covering different topics is introduced into the *non-parallel corpora covering similar topics*. As a result, one obtains comparable corpora belonging to the class *non-parallel corpora covering different topics*.

For the three groups of corpora, we give more details of the construction process here. The first group G_a is built from the corpus *Europarl* only by following these steps:

Step 1: The English part of the *Europarl* corpus and its corresponding French part is split into 10 equal parts in terms of sentence number, leading to 10 parallel corpora denoted as $\mathcal{P}_1$, $\mathcal{P}_2$, ..., $\mathcal{P}_{10}$. The comparability level of the 10 parallel corpora are arbitrarily set to 1 (i.e. the highest level), since they are equally good and one can not find a corpus better than a parallel corpus in terms of quality.

Step 2: For each parallel corpus, e.g. $\mathcal{P}_i$ $(i = 1, 2, \ldots, 10)$, we replace a certain proportion p of the English part of $\mathcal{P}_i$ with content of the same size, again in terms of sentence number, from another parallel corpus $\mathcal{P}_j (j \neq i)$, producing the new corpus $\mathcal{P}'_i$ with less content in parallel and thus less comparable than $\mathcal{P}_i$. For each $\mathcal{P}_i$, as p increases, i.e. more noise is embedded in the parallel corpus, we obtain a series of comparable corpora with decreasing comparability scores. In our experiments, p is taken from 0 to 1 with the gap 0.01. All the $\mathcal{P}_i$ and their

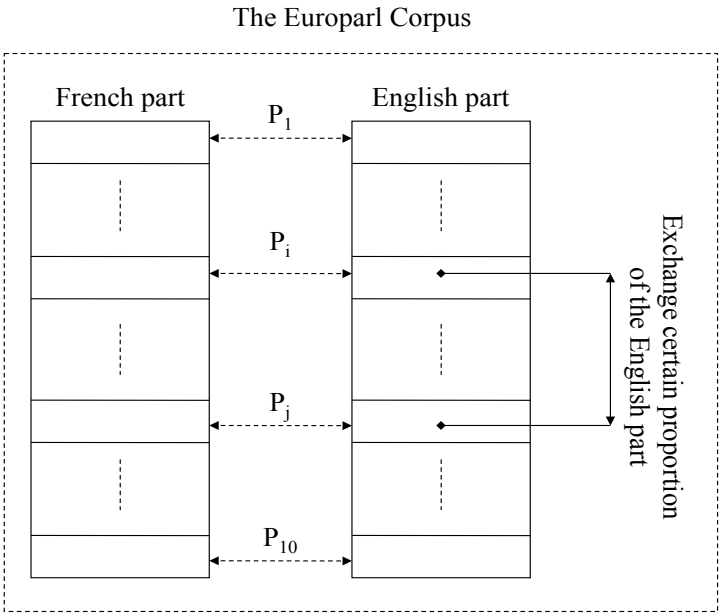


Figure 3.1. Constructing the test corpus group G_a with gold-standard comparability levels

respective descendant corpora according to different p constitute the group G_a. This process is illustrated in Figure 3.1.

The difference between building the corpora in G_b and in G_a is that, in G_b, the replacement in $\mathcal{P}_i$ is done with documents from the AP corpus and not from another parallel corpus $\mathcal{P}_j$ from *Europarl*. Compared with the corpora in G_a, we further degrade the parallel corpus $\mathcal{P}_i$ in G_b since the AP corpus covers different topics as in *Europarl*.

In G_c, we start with 10 comparable corpora $\mathcal{P}'_i$ from G_a instead of the parallel corpora. The 10 comparable corpora have the comparability scores of 0 in G_a, i.e. the least comparable ones in G_a. They thus contain documents from *Europarl* which are not the translation of each other. Each $\mathcal{P}'_i$ is further altered by replacing certain portions with documents from the AP corpus. Although $\mathcal{P}'_i$ itself is comparable and not parallel, its English part and French part cover similar topics embedded in the

Europarl corpus. Replacing certain part of $\mathcal{P}'_i$ with the content from AP will further degrade the comparability levels of $\mathcal{P}'_i$.

From the process of building the comparable corpora in G_a, G_b and G_c, one can note that the gold-standard comparability scores in different groups, e.g. G_a and G_c, can not be compared with each other directly, since the comparability scores are normalized to 0 and 1 in each group of corpora.

Correlation with Gold-standard Comparability Levels

The goal here is to assess whether comparability measures we have introduced can capture the differences in comparability introduced in the three different groups G_a, G_b and G_c. In order to quantify this, we use the Pearson correlation coefficient to measure the correlation between the proposed measures and the comparability scores of different corpora:

$$r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2}\sqrt{\sum_i (Y_i - \overline{Y})^2}}$$

where X_i denotes the comparability score provided by one measure on a given bilingual corpus and Y_i is the arbitrary comparability score (i.e. gold standard) assigned to this corpus in the construction process. $\overline{X}$ represents the average of X_is over all the bilingual corpora considered in G_a, G_b or G_c (and similarly for $\overline{Y}$).

Let us first recall the measures we have proposed in section 3.2:

- Measures based on vocabulary overlapping. These measures are defined as the mathematical expectation of finding the translation for each word in the corpus vocabulary. Corresponding to the English/French/whole vocabulary of the corpus, we have the measures M_{ef}, M_{fe}, M and their context version M^c_{ef}, M^c_{fe} and M^c. For the six measures, we will consider both the versions with the P/A weighting criterion and the versions with TF-IDF weighting schema, amounting to 12 measures in total.

- Baseline measures. The measures here, based on matured techniques in previous works, provide alternative choices for measuring the quality of comparable

corpora. We have the measure M^v based on the bilingual vector space model and the measurers M^{g1} and M^{g2} based on the machine translation system and n-grams representation.

We first make comparisons between measures based on vocabulary overlapping. The correlation scores are listed in Table 3.2. Each column in the table, e.g. M_{ef}, corresponds to the correlation scores between the specific comparability measure M_{ef} and the gold-standard comparability levels on different corpus groups, namely G_a, G_b and G_c. Let us first pay attention to the measures without context, i.e. M_{ef}, M_{fe} and M. One can find that M together with the P/A weighting schema performs the best and correlates very well with the gold standard on all the three groups of corpora, as the Pearson coefficient is close to 1. M_{fe} performs worst among the three measures with only one exception that, weighted by TF-IDF, M_{fe} performs better then M_{ef}. One can also conclude, for the best measure M, that using IDF to reduce the effects of frequent words does not help to solve the problem of too frequent words, with P/A playing better with M than TF-IDF. Although the weighting schema TF-IDF seems to be efficient for M_{fe} on G_b and G_c, the measure M_{fe} together with TF-IDF still performs far from the other two measures.

We then turn to the measures with context information, i.e. M^c_{ef}, M^c_{fe} and M^c. One can find from Table 3.2 that all the context-based measures, weighted by either TF-IDF or P/A, perform very well, which is also slightly better than the best performing measure M in the context-free family. Among the three measures with context, M^c performs slightly better than M^c_{ef} and M^c_{fe}. The measures M^c_{ef} and M^c_{fe}, with the context information, are better than their corresponding context-free version M_{ef} and M_{fe}. These findings are coincident with the assumption in the second part of section 3.2.1 that using context information can disambiguate between the translation candidates and thus leads to better performance of the measures.

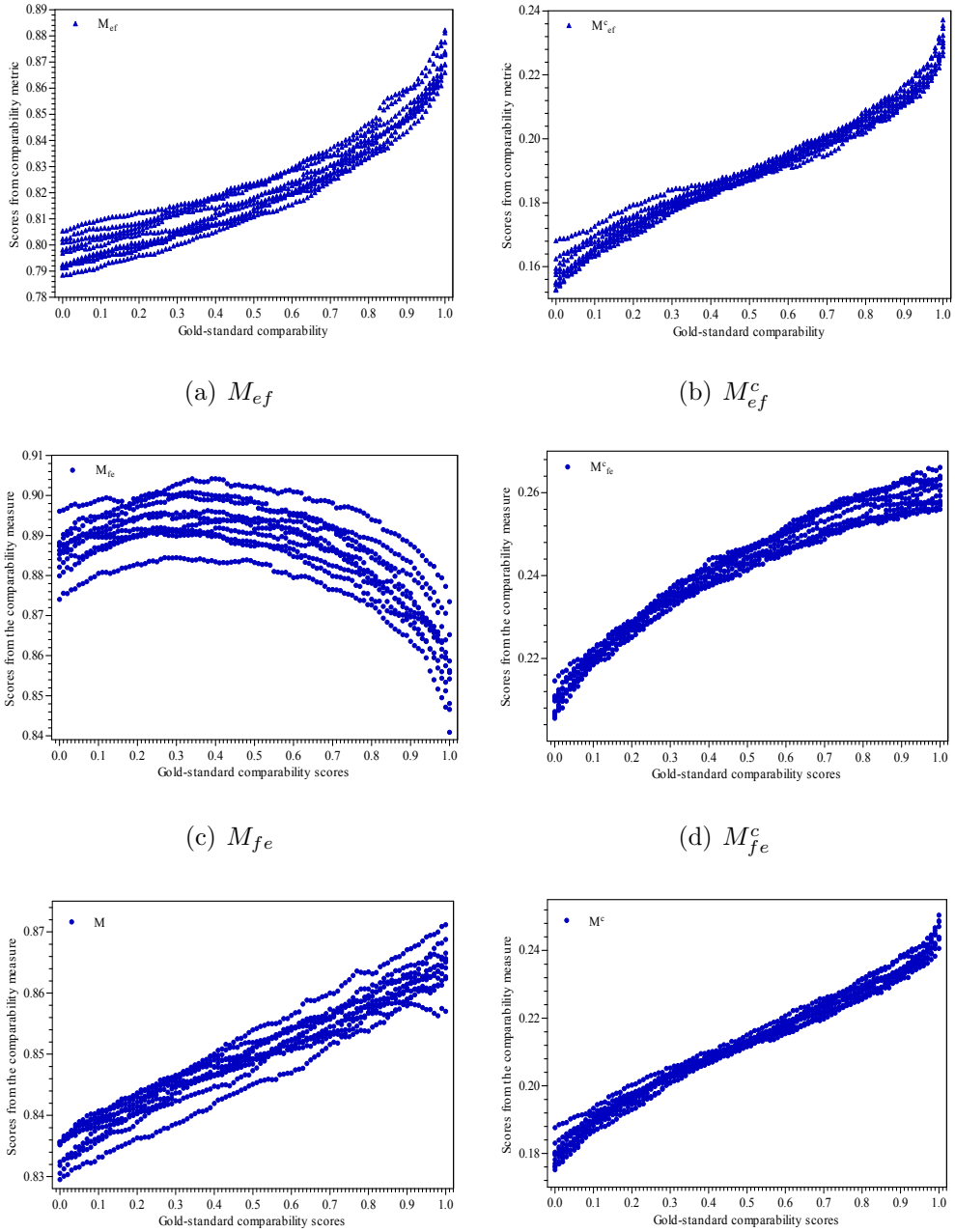
Deeper analysis is given here in order to show the performance of different measures. We will only consider here the P/A weighting schema so as to simplify the discussion. Figure 3.2 plots the measures M, M_{ef}, M_{fe}, M^c, M^c_{ef}, M^c_{fe} on 10 compa-

Table 3.2
Correlation scores of the vocabulary overlapping measures with the gold standard. The rows TF-IDF correspond to the TF-IDF weighting schema, and the rows P/A correspond to the Presence/Absence weighting method.

		M_{ef}	M_{fe}	M	M^c_{ef}	M^c_{fe}	M^c
G_a	TF-IDF	0.634	0.724	0.786	0.980	0.974	0.980
	P/A	0.897	0.770	0.936	0.976	0.966	0.972
G_b	TF-IDF	0.950	0.434	0.973	0.989	0.982	0.995
	P/A	0.955	0.190	0.979	0.975	0.977	0.978
G_c	TF-IDF	0.964	-0.292	0.962	0.980	0.934	0.991
	P/A	0.940	-0.595	0.960	0.984	0.968	0.990

rable corpora and their descendants in G_c with respect to their gold-standard comparability scores. We first compare here three context-free measures M_{ef}, M_{fe} and M. One can notice from Figure 3.2(c) that the comparability scores from M_{fe} even decrease at a certain point as the gold standard scores increase. The reason for the different performances is that asymmetric measures M_{ef} and M_{fe} are sensitive to the length of the corpus. Given a single English document and a large French document collection, it is very likely that we can find the translations for most of the English words according to the dictionary, although the two text sets are lowly comparable. In our case, since the average sentence length in AP is larger than that of *Europarl*, we increase the length of the English part of the test corpora remarkably when degrading the corpora in G_b and G_c, which leads to the poor performance of M_{fe}. The length related problem can be overcome by M which considers the translation in both directions.

We then consider the context-based versions M^c_{ef}, M^c_{fe} and M^c. All of these three measures perform very well on the three groups of corpora. Let us pay attention to the measure M_{fe} and its context version M^v_{fe}. The former one is sensitive to the corpus length and the latter one is not, resulting in different performance of two measures.



(a) M_{ef} (b) M^c_{ef}

(c) M_{fe} (d) M^c_{fe}

(e) M (f) M^c

Figure 3.2. Evolution of the measures M_{ef}, M_{fe}, M, M^v_{ef}, M^v_{fe} and M^v w.r.t. gold standard on the corpus group G_c (x-axis: gold-standard comparability scores, y-axis: comparability scores from the measures)

In order to explain the different performance, let us reuse the same example as above. Given a single English document and a large French document collection, although the probability of finding the translation for each word in the English part is high according the dictionary, the French translation candidate may hold a different sense from the English word, which can be disambiguated by the context version M^c_{fe}. As the results show here, M and all the context-based measures are able to capture all the differences in comparability artificially introduced in the degradation process we have considered above. Meanwhile, one can conclude from the results that it is easier to capture the different comparability levels in G_b than in G_a and G_c, which well coincides with our intuition, given the construction process considered.

We lastly list in Table 3.3 the results from the baseline measures. The results shown here are obtained in the same way as the ones for the vocabulary overlapping measures in Table 3.2. From the results one can find that the measure M^{g1} performs the best on all the three corpus groups and the measures M^{g2} and M^v perform worse. All these baseline measures do not perform as well as the measure M (weighed by P/A) or the three context-based measures only relying on vocabulary overlapping. The vector space model, being a standard approach, is broadly used to represent the text in previous work to capture the similarity on the sentence or document level. Due to the evaluation schema considered for the comparability measure, we however work here on the corpus level, which covers diverse topics compared to the document or sentence level. It is probably able to explain the fact that the baseline measures using vector space representation do not perform as well as the simple measures based on vocabulary overlapping. This finding is also partially supported by a recent report of the ACCURAT project[7].

Robustness of Comparability Measures

Since the two measures M and M^c perform the best in their respective class of measures, i.e. measures without context and measures with context, we choose to

[7]Related materials can be found on the deliverables of the project. The project website is: http://www.accurat-project.eu/.

Table 3.3
Correlation scores of the baseline comparability measures with the gold standard

	M^v	M^{g1}	M^{g2}
G_a	0.698	0.724	0.492
G_b	-0.611	0.479	0.228
G_c	-0.744	0.311	0.210

compare them in terms of the robustness w.r.t. the change of the dictionary. To simplify the discussion here, we follow the same consideration as above and only use the P/A weighting schema for this part of experiments. It is important that the comparability measure one retains remains consistent when the dictionary coverage of the corpus changes slightly, as this is a necessary condition to distinguish between different comparability levels. Indeed, if a slight change in the dictionary coverage entails an important change in the comparability score, it will become impossible in practice to compare different corpora, as they will likely have a different coverage with respect to the dictionary. We say, informally, that a comparability measure is robust at certain dictionary coverage range if the measure can distinguish between different comparability levels when the dictionary changes in this range. The experiments and analysis below try to validate the robustness of the measures.

In our experiments, several dictionaries of different sizes, corresponding to different coverages on the corpus vocabulary, are built by randomly choosing the subparts from the original dictionary. The coverage here is simply defined as the proportion of unique words in the corpus vocabulary that are covered by the dictionary. The definition is actually coincident with the definition of M and M^c, which corresponds to the proportion of words translated in the part of vocabulary covered by the dictionary in the whole vocabulary. In order to bridge the language barriers, a dictionary that is sufficient large is necessary. We thus choose to randomly pick certain proportions,

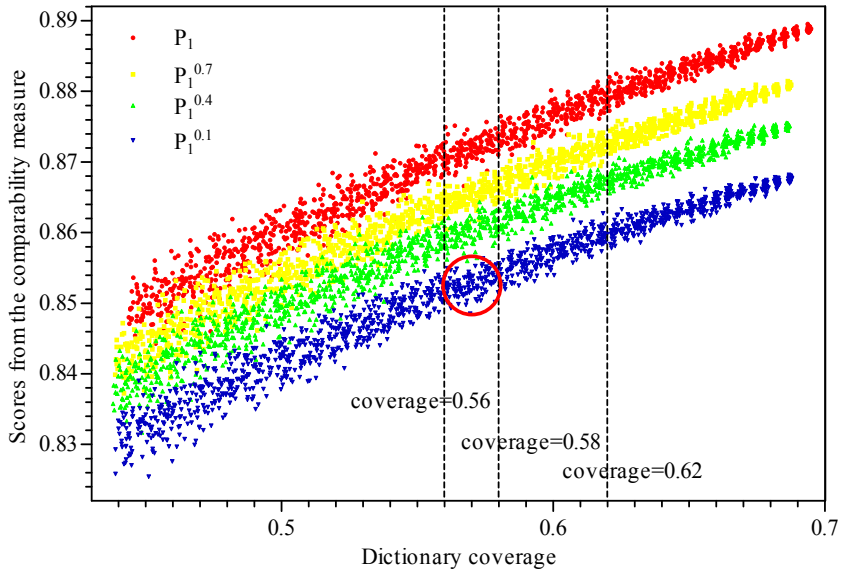
from 50% to 99% with a step of 1%, of the original dictionary listed in section 3.3.1. For each proportion, 30 different dictionaries are built by randomly sampling the original dictionary 30 times at this same proportion. These 1500 dictionaries (i.e. 50*30) are then used to compute M and M^c on different corpora with decreasing comparability scores in G_a, G_b and G_c.

As we have discussed in the corpus construction process above, in each of G_a, G_b and G_c, one tries to obtain a series of decreasing comparability levels starting from 10 parallel corpora (in G_a and G_b) or 10 high quality comparable corpora (in G_c). For the clarity of analysis, we only take here the first parallel corpus $\mathcal{P}_1$ from ($\mathcal{P}_1$, $\mathcal{P}_2$, ..., $\mathcal{P}_{10}$) together with its 10 descendant corpora which are built by setting the proportion p to 0.1, 0.2, ..., 1.0 in G_a. That is to say, we exchange 10%, 20%, ..., 100% percent of the content in the high quality corpora with noise from other corpora. In this case, we obtain 11 comparable corpora $\mathcal{P}_1$, $\mathcal{P}_1^{0.9}$, $\mathcal{P}_1^{0.8}$, ..., $\mathcal{P}_1^{0}$ with the gold comparability scores from 1 to 0, with a step of 0.1. Lastly, for readability reasons, we only plot in Figure 3.3 the comparability scores for some of the 11 comparable corpora, i.e. $\mathcal{P}_1$, $\mathcal{P}_1^{0.7}$, $\mathcal{P}_1^{0.4}$ and $\mathcal{P}_1^{0.1}$, w.r.t. the different coverages.

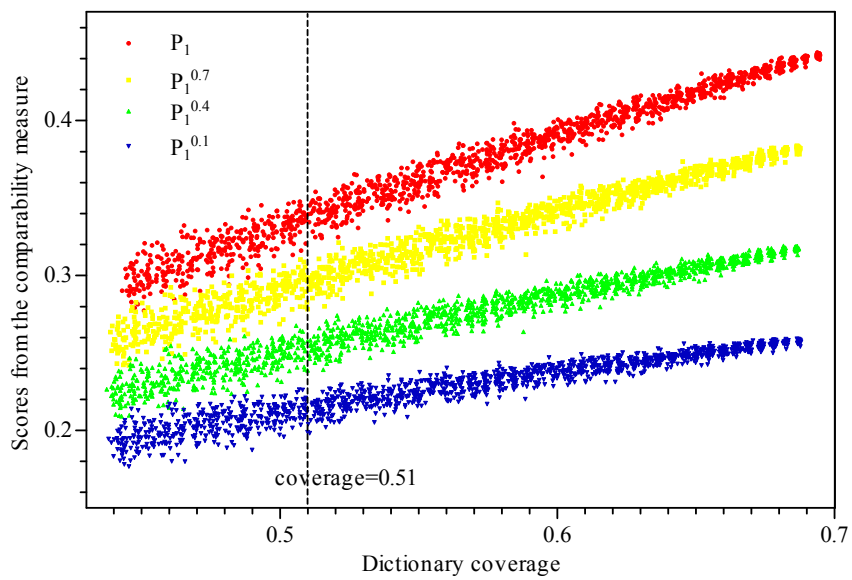
From Figure 3.3(a) one can find that when the dictionary coverage lies above a certain threshold (inspected from the figure, roughly set to 0.62), the differences between the 4 different comparability levels[8] can be captured very well, as the different data points are well separated. In another word, the different comparability levels can be captured very well by the measure M when the dictionary coverage is roughly above 0.62. The same conclusion can be drawn from the inspection of Figure 3.3(b) with another coverage threshold, roughly set to 0.51. One can thus conclude from the qualitative analysis that both M and M^v are robust to the changes of the dictionary at a certain point.

We have drawn for comparability measures the intuitive conclusion regarding robustness from Figure 3.3. In order to analyze the results quantitively, we will first try to define as below the *degree of robustness* of a comparability measure. The definition

[8]This is also true for all the 11 comparability levels although we only plot 4 in the figure.



(a) M



(b) M^c

Figure 3.3. Evolution of M and M^c w.r.t. different dictionary coverages on comparable corpora $\mathcal{P}_1$, $\mathcal{P}_1^{0.7}$, $\mathcal{P}_1^{0.4}$ and $\mathcal{P}_1^{0.1}$ in G_a (x-axis: different dictionary coverages; y-axis: comparability scores from M or M^c).

directly reflects the ability of a measure to capture different comparability levels at a certain coverage range. Then we depict the experiment results in Figure 3.3 in the framework of probability theory and estimate the robustness in terms of the probabilistic distribution. According to the qualitative analysis from Figure 3.3, we would like to define the degree of robustness as:

Definition 3.3.1 *Let us assume we have different comparable corpora* $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_k$, *with gold-standard comparability levels that are increasing, which can be written as:* $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \ldots \prec \mathcal{C}_k$. *The symbol* $\prec$ *is used here to denote the relation* less *in the gold-standard comparability levels. We further assume to have a bilingual dictionary* $\mathcal{D}$ *such that the coverage of* $\mathcal{D}$ *on all the* k *corpora* $\mathcal{C}_i (i = 1, 2, \ldots, k)$ *belongs to a range* $[r, r + \epsilon]$, *with* ϵ *being a small fixed value. Then, we define the degree of robustness of a comparability measure* M *w.r.t the dictionary coverage* r *as:*

$$\chi(M, r) = \min_{i \in 1,2,\ldots,k-1} \Pr(M(\mathcal{C}_{i+1}) > M(\mathcal{C}_i))$$

One can find from definition 3.3.1 that the degree of robustness states the guaranteed probability that a group of gold-standard comparability levels can be captured by the comparability measure M at a certain coverage range.

The experiments are then performed to estimate the degree of robustness of the comparability measures M and M^v. Let us first focus on the measure M. One can notice from Figure 3.3 that, for each of the test corpora, e.g. $\mathcal{P}_1^{0.1}$ in Figure 3.3(a), the comparability scores corresponding to a certain coverage range (e.g. from 0.56 to 0.58, identified by the circle in the figure) follow a normal distribution, according to the Shapiro-Wilk test (Shapiro and Wilk, 1965) at the significance level 0.05. This fact is also illustrated by the frequency distribution histogram in Figure 3.4. Thus, in a coverage range of size 0.02 (i.e. the value ϵ in the definition 3.3.1) here, the comparability scores of M for a specific corpus can be modeled as a normally distributed variable Z. Hence, on each span, the scores of M on two bilingual corpora, say $\mathcal{P}_1^{0.1}$ and $\mathcal{P}_1^{0.2}$, can be described as two normally distributed variables denoted as

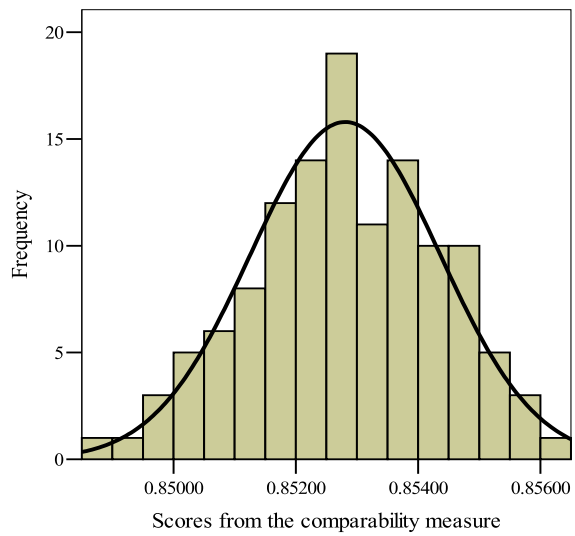


Figure 3.4. The frequency histogram of the comparability scores (from M) on $\mathcal{P}_1^{0.1}$ between the coverage 0.56 and 0.58 (x-axis: comparability scores; y-axis: frequency).

$Z_{0.1}$ and $Z_{0.2}$ of which the parameters (i.e. the mean μ and variance σ^2) can be estimated from the samples, i.e. all the comparability scores on $\mathcal{P}_1^{0.1}$ and $\mathcal{P}_1^{0.2}$ in the specific coverage range. With the estimated parameters, we can write:

$$Z_{0.1} \sim N(\mu_{0.1}, \sigma_{0.1}^2), \;\; Z_{0.2} \sim N(\mu_{0.2}, \sigma_{0.2}^2)$$

To computer the degree of robustness in definition 3.3.1, one needs to compute all the probability $\Pr(M(\mathcal{C}_{i+1}) > M(\mathcal{C}_i))$. Taking here the corpora $\mathcal{P}_1^{0.2}$ and $\mathcal{P}_1^{0.1}$ as an example, one needs to estimate $\Pr(Z_{0.2} > Z_{0.1})$. With the independence assumption between variables $Z_{0.1}$ and $Z_{0.2}$, the new variable $Z_{0.2} - Z_{0.1}$ should satisfy:

$$Z_{0.2} - Z_{0.1} \sim N(\mu_{0.2} - \mu_{0.1}, \sigma_{0.1}^2 + \sigma_{0.2}^2)$$

Then to compute $\Pr(Z_{0.2} > Z_{0.1})$ equals to the computation of $\Pr(Z_{0.2} - Z_{0.1} > 0)$ (in short, $\Pr(0.2, 0.1)$). In Table 3.4, we list, for different coverage ranges, the Pr values of different corpus pairs and also the degree of robustness at each coverage. One can find from the results that the higher the dictionary coverage is, the more reliable one

Table 3.4
The degree of robustness with the measure M and different dictionary coverages. The coverage range is set to $[r, r+\epsilon]$ as in definition 3.3.1 with ϵ being fixed to 0.02.

r	0.46	0.48	0.50	0.52	0.54	0.56
$\Pr(0.2, 0.1)$	0.69	0.70	0.73	0.74	0.76	0.79
$\Pr(0.4, 0.3)$	0.67	0.68	0.71	0.73	0.75	0.79
$\Pr(0.6, 0.5)$	0.59	0.60	0.61	0.63	0.65	0.69
$\Pr(0.8, 0.7)$	0.76	0.74	0.76	0.77	0.82	0.86
$\Pr(1.0, 0.9)$	0.76	0.68	0.75	0.78	0.82	0.82
$\chi(\mathbf{M}, \mathbf{r})$	**0.59**	**0.60**	**0.61**	**0.63**	**0.65**	**0.69**
r	0.58	0.60	0.62	0.64	0.66	0.68
$\Pr(0.2, 0.1)$	0.79	0.79	0.81	0.84	0.89	0.92
$\Pr(0.4, 0.3)$	0.84	0.83	0.88	0.91	0.95	0.98
$\Pr(0.6, 0.5)$	0.74	0.72	0.75	0.82	0.85	0.91
$\Pr(0.8, 0.7)$	0.89	0.86	0.91	0.96	0.96	0.99
$\Pr(1.0, 0.9)$	0.88	0.87	0.91	0.94	0.97	0.98
$\chi(\mathbf{M}, \mathbf{r})$	**0.74**	**0.72**	**0.75**	**0.82**	**0.85**	**0.91**

can rely on M to distinguish between different comparability levels. Furthermore, when the dictionary coverage is above a certain threshold (e.g. 0.64), we have a high confidence (≥ 0.82) that the different comparability levels between the corpus pairs can be reliably captured by M, so that the measure is robust in a given coverage range, here set to 0.02. The same conclusions can be drawn for the other comparable corpus pairs we have constructed (from G_b and G_c).

We also evaluate the robustness of the comparability measure M^c as above and the results are listed in Table 3.5. One can find from the results that, compared with M, it is much easier to achieve high confidence, and thus higher robustness, with M^c.

Table 3.5
The degree of robustness with the measure M^c and different dictionary coverages. The coverage range is set to $[r, r+\epsilon]$ as in definition 3.3.1 with ϵ being fixed to 0.02. The table is designed in the same format as Table 3.4.

r	0.46	0.48	0.50	0.52	0.54	0.56
$\Pr(0.2, 0.1)$	0.84	0.87	0.90	0.94	0.96	0.97
$\Pr(0.4, 0.3)$	0.79	0.85	0.87	0.90	0.94	0.93
$\Pr(0.6, 0.5)$	0.86	0.88	0.91	0.94	0.96	0.96
$\Pr(0.8, 0.7)$	0.81	0.80	0.85	0.88	0.92	0.90
$\Pr(1.0, 0.9)$	0.84	0.82	0.82	0.89	0.90	0.90
$\chi(\mathbf{M^c}, \mathbf{r})$	**0.79**	**0.80**	**0.82**	**0.88**	**0.90**	**0.90**
r	0.58	0.60	0.62	0.64	0.66	0.68
$\Pr(0.2, 0.1)$	0.98	1.00	1.00	1.00	1.00	1.00
$\Pr(0.4, 0.3)$	0.96	0.98	0.99	1.00	1.00	1.00
$\Pr(0.6, 0.5)$	0.98	0.99	1.00	1.00	1.00	1.00
$\Pr(0.8, 0.7)$	0.93	0.96	0.98	0.99	1.00	1.00
$\Pr(1.0, 0.9)$	0.96	0.95	0.98	0.99	1.00	1.00
$\chi(\mathbf{M^c}, \mathbf{r})$	**0.93**	**0.95**	**0.98**	**0.99**	**1.00**	**1.00**

Even when the dictionary coverage is only 0.54, the confidence is larger than 0.90 that M^c can capture the different comparability levels. However, the computational complexity of the context-based measures is usually very expensive. For this reason, we will still make use of the measure M weighted by the P/A criterion in the following experiments, as they require intensive calls to the comparability measure.

It could be useful to find a threshold for the coverage above which the comparability measure will work robustly on every comparable corpus. It is however difficult if one considers the experimentation. In order to build the gold-standard comparability scores, we can not think out another strategy to construct them. So all the

experiments must be performed on the artificial corpora we build. Different from the comparability measures, testing the robustness relies a lot on the style of the corpora. It is not easy to make a universal conclusion if the experiments are only done with these corpora. Having reviewed several comparability measures, we will turn to the problem of enhancing corpus comparability for bilingual lexicon extraction in the next chapter.

3.4 Conclusion

Data-driven NLP applications rely a lot on the quality of text resources used, a fact we conjecture here holding for the parallel corpus. However, there has not been any clear definition for the notion of comparability and it is hard to distinguish between different comparability levels. We have investigated in this section the intuitive notion and the measures to quantify the degree of comparability of comparable corpora. The experiments show that our proposed measure can correlate very well with gold-standard comparability levels and is robust to dictionary changes.

In oder to quantify the comparability levels, we have first discussed the notion of comparability according to the intuition one can have from the usage experience of bilingual corpora. The intuition is that the usability of bilingual corpora decreases from the *parallel corpus* to *non-parallel corpora covering different topics*, since the *bridge* information decreases in this case. This notion motivates several candidate implementations for the comparability measure as well as the methodology designed to evaluate the measures.

Based on the notion above, we have then developed several measures to approximate the notion of comparability. We make use here of the idea that it is easier to find the translation pairs from comparable corpora of higher quality. The mathematical expectation is then employed to quantify the difficulty in finding the translations for each word in the corpus vocabulary. One can judge the translation pairs based on a bilingual dictionary. The fact of *polysemy* however inspires one to use the context-

based disambiguation to filter out fake translation pairs holding different senses. We have thus developed two classes of measures: context-free measures only relying on the dictionary and context-based measures coming with word sense disambiguation. We have also designed several baseline measures extending previous work in related areas.

For the experiments, we have again followed the notion of comparability and developed comparable corpora with gold-standard comparability levels. The experiments show that a simple context-free measure M and all the context-based measures M^c_{ef}, M^c_{fe} and M^c can capture different comparability levels very well. Furthermore, the measures M and M^c are shown to be robust to dictionary changes. The context-based measure has more expensive computational cost and we will only rely on M in the following chapters.

4 ENHANCING COMPARABLE CORPUS QUALITY

We have proposed in chapter 3 the comparability measure based on vocabulary overlapping which can capture different comparability levels and is robust to dictionary changes. Based on the measure, we will develop in this chapter the approaches to improving the quality of any given comparable corpus. We will first discuss in section 4.1 the general methodology for enhancing corpus quality. Two such approaches, namely the *greedy approach* and the *clustering approach*, are introduced in section 4.2 to implement the general methodology. The efficiency of these approaches are then validated in the experiments in section 4.3, which shows that one can improve comparable corpus quality in terms of comparability scores, and that the enhanced comparable corpora lead to extracted lexicons of higher quality. Lastly, section 4.4 concludes this chapter.

4.1 General Methodology

We have reviewed in section 2.1.2 the studies trying to extract from comparable corpora the parallel content such as parallel sentences and sub-sentential segments. However, parallel sentences and segments do not exist in high volume in comparable corpora, especially in those low-quality ones. One can only extract a small amount of parallel segments from comparable corpora, which does not satisfy the need of such tasks as bilingual lexicon extraction aiming to extract the translations for each word in the original corpus vocabulary. This fact will be experimentally validated in section 4.3.1. we will consider in this chapter a methodology which can extract a high-quality subpart from the original corpus as well as enhance the low-quality subpart of the original corpus. With this methodology, one will be able to construct a resulting corpus resembling the original one without losing too much information

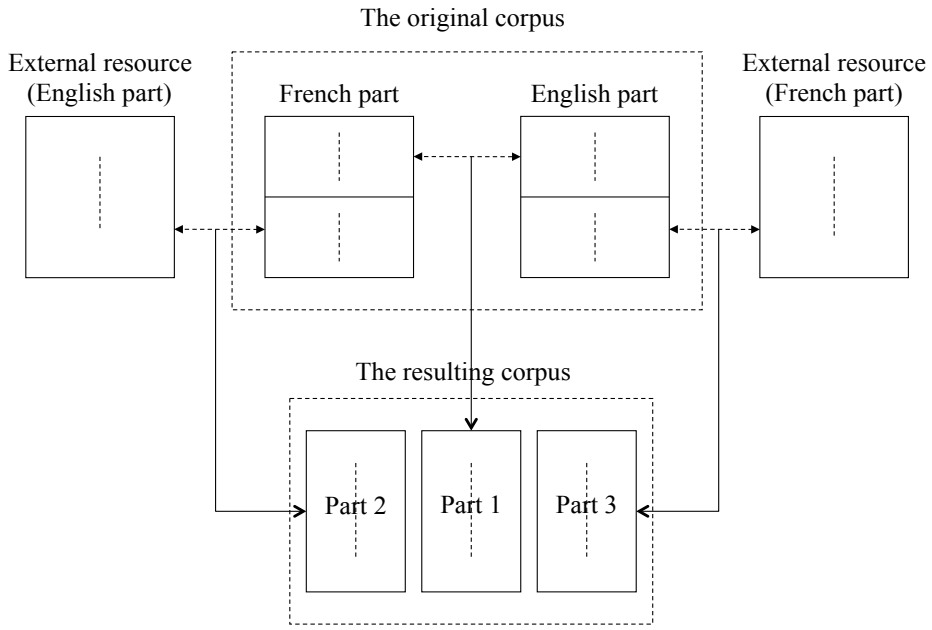


Figure 4.1. The general methodology for enhancing comparable corpus quality. The resulting corpus consists of three parts: Part 1 corresponds to the high-quality subpart of the original comparable corpus; Part 2/3 is the enriched version of the French/English low-quality subpart of the original corpus.

as in the methods for parallel subpart extraction. The methodology we will make use of in this chapter is illustrated in Figure 4.1.

The first step of our methodology is to extract a high-quality subpart ("Part 1" in Figure 4.1) from the original comparable corpus. Different from the work extracting parallel content, we aim here to extract from comparable corpora the subpart containing highly related content. The quality of extracted subpart should have a guaranteed quality, which is the degree of comparability in our work.

The second step is to enhance the low-quality subpart of the original corpus. The low-quality subpart is the content left in the original corpus after the high-quality subpart (i.e. "Part 1") being removed. The counterpart of low-quality content can not be found in the original corpus, and we thus need to refer to external resources in order

to complement the low-quality subpart. One can make use of either existing corpora covering overlapping topics with the original corpus or the web content containing a huge number of topics. As one can see from Figure 4.1, the French (resp. English) low-quality subpart of the original comparable corpus is enriched with the English (resp. French) external resource, resulting in the enhanced corpora "Part 2" and "Part 3". The genres of external resources actually affect the performance of the methodology presented here, which will be discussed in section 4.4.

4.2 Approaches to Enhancing Corpus Quality

We will develop in this section the strategies to improve the quality of any given comparable corpus in order to improve its usability. Two approaches are proposed following the general methodology in section 4.1: the first one, named the *greedy approach*, is introduced in section 4.2.1; the second one, named the *clustering approach*, overcomes some problems raising in the greedy approach and is detailed in section 4.2.2. The brief versions of these two approaches have been discussed in (Li and Gaussier, 2010) and (Li et al., 2011). In addition to the previous work, we will provide here sounder theoretical background and extensions for the two approaches.

4.2.1 The Greedy Approach

We here try to improve the quality of a given corpus $\mathcal{C}$, which we have referred to as the *original corpus* in section 4.1, by extracting the highly comparable subpart $\mathcal{C}_H$ which is above a certain degree of comparability (Step 1), and by enriching the lowly comparable part $\mathcal{C}_L$ with texts from other sources (Step 2). In the context of bilingual lexicon extraction, as we are interested in extracting information related to the vocabulary of the original corpus, we want the newly built corpus to contain a substantial part of the original corpus vocabulary. This can be achieved by preserving in Step 1 as many documents from the original corpus as possible, and by using in step 2 sources close to the original corpus.

Step 1: Extracting the High-quality Subpart $\mathcal{C}_H$

The strategy consisting of building all the possible sub-corpora of a given size from a given comparable corpora is not realistic as soon as the number of documents making up the corpora is larger than a few thousands. In such cases, better ways for extracting subparts have to be designed. The strategy we have adopted here aims at efficiently extracting a subpart of $\mathcal{C}$ above a certain degree of comparability and is based on the following property.

Property 1 *Let d_e^1 and d_e^2 (resp. d_f^1 and d_f^2) be two English (resp. French) documents from a bilingual corpus $\mathcal{C}$. We consider, as before, that the bilingual dictionary $\mathcal{D}$ is independent from $\mathcal{C}$. Let $(d_e^{1'}, d_f^{1'})$ be such that: $d_e^{1'} \subseteq d_e^1$, $d_f^{1'} \subseteq d_f^1$, which means $d_e^{1'}$ is a subpart of d_e^1 and $d_f^{1'}$ is a subpart of d_f^1.*
We assume:

$$\begin{array}{ll} (i) & \frac{|d_e^1 \cup d_e^2|}{|d_e^2|} = \frac{|d_f^1 \cup d_f^2|}{|d_f^2|} \\ (ii) & M_{ef}(d_e^{1'}, d_f^1) \geq M_{ef}(d_e^2, d_f^2) \\ & M_{fe}(d_e^1, d_f^{1'}) \geq M_{fe}(d_e^2, d_f^2) \end{array}$$

Then:

$$M(d_e^2, d_f^2) \leq M(d_e^1 \cup d_e^2, d_f^1 \cup d_f^2)$$

Proof[sketch]: Let $B = (d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v) \backslash (d_e^2 \cap \mathcal{D}_e^v)$. One can show, by exploiting condition (ii), that:

$$\sum_{w \in B} \sigma(w, d_f^1 \cup d_f^2) \geq |B| \cdot M_{ef}(d_e^2, d_f^2)$$

and similarly that:

$$\sum_{w \in d_e^2 \cap \mathcal{D}_e^v} \sigma(w, d_f^1 \cup d_f^2) \geq |d_e^2 \cap \mathcal{D}_e^v| \cdot M_{ef}(d_e^2, d_f^2)$$

Then exploiting condition (i), and the independence between the corpus and the dictionary, one arrives at:

$$\frac{\sum_{w \in (d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v} \sigma(w, d_f^1 \cup d_f^2)}{|(d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v| + |(d_f^1 \cup d_f^2) \cap \mathcal{D}_f^v|} \geq \frac{|d_e^2 \cap \mathcal{D}_e^v| \cdot M_{ef}(d_e^2, d_f^2)}{|d_e^2 \cap \mathcal{D}_e^v| + |d_f^2 \cap \mathcal{D}_f^v|}$$

The same development on M_{fe} completes the proof. ■

The property 1 shows that one can incrementally extract from a bilingual corpus a subpart with a guaranteed minimum degree of comparability η by iteratively adding new elements, provided (a) that the new elements have a degree of comparability of at least η and (b) that they are less comparable than the currently extracted subpart (conditions (ii)). This strategy is described in Algorithm 1. The code in line 3 tries to extract the document pair with the highest comparability score, which is then added to the resulting corpus by the code in line 6. The code in line 7 removes the current document pair from the corpus for the next circle.

Since the degree of comparability is always above a certain threshold and since the new documents selected (d_e^2, d_f^2) are the most comparable among the remaining documents, condition (i) is likely to be satisfied, as this condition states that the increase in the vocabulary from the second documents to the union of the two is the same in both languages. Similarly, considering new elements by decreasing comparability scores is a necessary step for the satisfaction of condition (ii), which states that the current subpart should be uniformly more comparable than the element to be added. Hence, the conditions for property 1 to hold are met in Algorithm 1, which finally yields a corpus with a degree of comparability of at least η.

Step 2: Enriching the Low-quality Subpart $\mathcal{C}_L$

This step tries to absorb knowledge from other resources, which will be called *external corpus*, to enrich the lowly comparable part $\mathcal{C}_L$ which is the left part in $\mathcal{C}$ during the creation of $\mathcal{C}_H$, i.e. $\mathcal{C}_L = \mathcal{C} \setminus \mathcal{C}_H$. One choice for obtaining the external corpus $\mathcal{C}_T$ is to fetch documents which are likely to be comparable from the Internet. In this case, we extract representative words for each document in $\mathcal{C}_L$, translate them using the bilingual dictionary and retrieve associated documents via a search engine (Baroni and Bernardini, 2004). An alternative approach is of course to use existing bilingual corpora. Once $\mathcal{C}_T$ has been constructed, the lowly comparable part $\mathcal{C}_L$ can be enriched in exactly the same way as in Step 1: First, Algorithm 1 is used on the English part of $\mathcal{C}_L$ and the French part of $\mathcal{C}_T$ to get the high-quality document

Algorithm 1: The Greedy Algorithm

Input:

English document set $\mathcal{C}_e^d$ of $\mathcal{C}$

French document set $\mathcal{C}_f^d$ of $\mathcal{C}$

The threshold η

Output:

$\mathcal{C}_H$, consisting of the English document set $\mathcal{W}_e$ and the French document set $\mathcal{W}_f$

1: Initialize $\mathcal{W}_e = \emptyset, \mathcal{W}_f = \emptyset, \text{temp} = 0$;
2: **repeat**
3: $(d_e, d_f) = \underset{d_e \in \mathcal{C}_e^d, d_f \in \mathcal{C}_f^d}{\arg\max}\ M(d_e, d_f)$;
4: $\text{temp} = \underset{d_e \in \mathcal{C}_e^d, d_f \in \mathcal{C}_f^d}{\max}\ M(d_e, d_f)$;
5: **if** temp $\geq \eta$ **then**
6: Add d_e into $\mathcal{W}_e$ and add d_f into $\mathcal{W}_f$;
7: $\mathcal{C}_e^d = \mathcal{C}_e^d \backslash d_e$, $\mathcal{C}_f^d = \mathcal{C}_f^d \backslash d_f$;
8: **end if**
9: **until** $\mathcal{C}_e^d = \emptyset$ or $\mathcal{C}_f^d = \emptyset$ or temp $< \eta$
10: **return** $\mathcal{C}_H$;

pairs. Then the French part of $\mathcal{C}_L$ is enriched with the English part of $\mathcal{C}_T$ by the same algorithm. All the high-quality document pairs are then added to $\mathcal{C}_H$ to constitute the final resulting corpus $\mathcal{C}_F$.

4.2.2 The Clustering Approach

We have proposed in section 4.2.1 a greedy approach to improving the quality of existing comparable corpus. A dramatic problem of the approach is that document pairs are considered separately and the relations between different document pairs are ignored. We will thus propose in this section a method that can take into account the relations between each two documents. If a comparable corpus covers a limited set of

topics, it is more likely to contain consistent information in different languages. The term *homogeneity* directly refers to this fact, and we will say, in an informal manner, that a corpus is homogeneous if it covers a limited set of topics. A homogeneous comparable corpus can be of higher quality and is more useful in the applications. For example, the distributional hypothesis underlying bilingual lexicon extraction method is more reliable when the documents in different languages describe the same or similar topics, since authors tend to use the same word combinations to describe similar topics (see (Morin et al., 2007) for a related analysis). In other words, the homogeneous comparable corpus can lead to improved bilingual lexicons extracted. The rationale for the algorithm we introduce here to enhance corpus comparability is precisely based on homogeneity.

Let us recall that our goal, in a first step, is to construct, from a given comparable corpus, an enhanced version of it which displays a higher degree of comparability and preserves most of the original vocabulary. We conjecture here that if one can guarantee a certain degree of homogeneity in addition to a certain degree of comparability, then the bilingual lexicons extracted from the obtained corpus will be of higher quality. As we will see, this conjecture will be fully validated in the experimental section. In order to find document sets which are similar with each other (i.e. homogeneous), it is natural to resort to clustering techniques. Furthermore, since we need homogeneous corpora for bilingual lexicon extraction, it will be convenient to rely on techniques which allows one to easily prune less relevant clusters. To perform all this, we use in this work a standard hierarchical clustering method but other clustering methods with associated pruning strategies can directly be used as well.

Bilingual Clustering Algorithm

The overall process retained to build high quality, homogeneous comparable corpora relies on the following steps:

Step 1: Using the bilingual similarity measure defined in equation 4.3 below, cluster English and French documents so as to get the bilingual dendrogram from the original corpus $\mathcal{C}$ by grouping documents with related content;

Step 2: Pick high quality sub-clusters by thresholding the obtained dendrogram according to the node depth;

Step 3: Combine all these sub-clusters to form a new comparable corpus $\mathcal{C}_H$, which thus contains homogeneous, high-quality subparts;

Step 4: Use again steps (1), (2) and (3) to enrich the remaining subpart of $\mathcal{C}$ (which will be denoted as $\mathcal{C}_L$ and $\mathcal{C}_L = \mathcal{C} \setminus \mathcal{C}_H$) with external resources $\mathcal{C}_T$.

The overall framework of this approach is similar with the greedy approach and the first three steps are summarized in Algorithm 2. As one can note, only $\mathcal{C}$ is used in order to build $\mathcal{C}_H$, through clustering and pruning of documents. As such, Algorithm 2 aims at extracting the most comparable and homogeneous subpart of $\mathcal{C}$. Once this has been done, i.e. once $\mathcal{C}$ has been exploited, one needs to resort to some other resources if one wants to build a homogeneous corpus with a high degree of comparability from $\mathcal{C}_L$ (which is the part of $\mathcal{C}$ left after removing $\mathcal{C}_H$). To do so, we simply replace, in step (4), the input corpus $\mathcal{C}$ with two comparable corpora: The first one consists of the English part of $\mathcal{C}_L$ and the French part of an external corpus $\mathcal{C}_T$; The second one consists of the French part of $\mathcal{C}_L$ and the English part of $\mathcal{C}_T$. The two high-quality subparts obtained from these two new comparable corpora in step (4) are then added to $\mathcal{C}_H$ to constitute the resulting comparable corpus $\mathcal{C}_F$ of higher quality.

Similarity Measure

The similarity measure between the clusters plays a central role in the hierarchal clustering algorithm. Let us assume that we have two document sets (i.e. clusters) $\mathcal{R}_1$ and $\mathcal{R}_2$. In the task of bilingual lexicon extraction, two document sets are similar to each other and should be clustered if the combination of the two can complement the content of each single set, which relates to the notion of homogeneity introduced

Algorithm 2: The Bilingual Clustering Algorithm

Input:

Set $\mathcal{U}$ of all English and French documents from $\mathcal{C}$

The depth threshold θ

Output:

$\mathcal{C}_H$, high quality, homogeneous subpart of $\mathcal{C}$

1: Set $\mathcal{C}_H$ as the null set $\emptyset$;

2: Cluster $\mathcal{U}$ in order to obtain a bilingual dendrogram $\mathcal{T}$;

3: Set m as the maximal depth of $\mathcal{T}$;

4: Remove from $\mathcal{T}$ the low-quality sub-clusters of which the depth is lower than the depth threshold computed from $m \cdot \theta$;

5: Add all the remaining documents in $\mathcal{T}$ to $\mathcal{C}_H$;

6: **return** $\mathcal{C}_H$;

before. In other words, both the English part $\mathcal{R}_1^e$ of $\mathcal{R}_1$ and the French part $\mathcal{R}_1^f$ of $\mathcal{R}_1$ should be comparable to their counterparts (respectively the French part $\mathcal{R}_2^f$ of $\mathcal{R}_2$ and the English part $\mathcal{R}_2^e$ of $\mathcal{R}_2$). This leads to the following similarity measure for $\mathcal{R}_1$ and $\mathcal{R}_2$:

$$sim(\mathcal{R}_1, \mathcal{R}_2) = \beta M(\mathcal{R}_1^e, \mathcal{R}_2^f) + (1 - \beta) M(\mathcal{R}_2^e, \mathcal{R}_1^f) \tag{4.1}$$

where $\beta\,(0 \leq \beta \leq 1)$ is a weight controlling the importance of the two subparts ($\mathcal{R}_1^e$, $\mathcal{R}_2^f$) and ($\mathcal{R}_2^e$, $\mathcal{R}_1^f$). Intuitively, one would like to give more weight in the combination to the larger subpart, as it contains more information. We use here the number of document pairs to represent the amount of information contained in a comparable sub-corpus. Thus, the weight β can be defined as the proportion of possible document pairs in the current comparable corpus ($\mathcal{R}_1^e$, $\mathcal{R}_2^f$) to all the possible document pairs, which is:

$$\beta = \frac{\#_d(\mathcal{R}_1^e) \cdot \#_d(\mathcal{R}_2^f)}{\#_d(\mathcal{R}_1^e) \cdot \#_d(\mathcal{R}_2^f) + \#_d(\mathcal{R}_2^e) \cdot \#_d(\mathcal{R}_1^f)}$$

where $\#_d(\mathcal{R})$ stands for the number of documents in $\mathcal{R}$. As the clusters are first formed from single document, in English or in French, one can see that the similarity measure corresponds to a *normalized* comparability score between the English and French sub-clusters making up the new cluster. However, this measure does not integrate the relative length of the two parts in different languages, which may actually impact the performance of bilingual lexicon extraction. If a 1-to-1 constraint is too strong (i.e. assuming that all clusters should contain the same number of documents in two languages), having completely unbalanced corpora is also not desirable. We thus introduce a penalty function ϕ aiming at penalizing corpora for which the number of documents in the English part and in the French part is too different:

$$\phi(\mathcal{R}) = \frac{1}{(1 + \log(1 + \frac{|\#_d(\mathcal{R}^e) - \#_d(\mathcal{R}^f)|}{\min(\#_d(\mathcal{R}^e), \#_d(\mathcal{R}^f))})} \tag{4.2}$$

The values of the function ϕ lie between 0 and 1 where small values identify strict penalty. The above penalty function leads us to a new similarity measure sim_l which is the one finally used in Algorithm 2:

$$sim_l(\mathcal{R}_1, \mathcal{R}_2) = sim(\mathcal{R}_1, \mathcal{R}_2) \cdot \phi(\mathcal{R}_1 \cup \mathcal{R}_2) \tag{4.3}$$

Theoretical Analysis

The clustering process used in Algorithm 2 guarantees that comparable documents covering similar content are grouped before documents which are comparable but cover different topics. Thus, the corpus $\mathcal{C}_H$ obtained through this algorithm will be homogeneous, i.e. its documents will belong to a small set of related topics. Furthermore, the fact that the comparable corpus $\mathcal{C}_F$ obtained through steps 1 to 4 above directly derives from the original corpus $\mathcal{C}$ is an indicator that most of the vocabulary of $\mathcal{C}$ will be preserved in $\mathcal{C}_F$. We will see in the experimental section that this is indeed the case. What is not clear yet is whether we can have any lowest guarantee concerning the degree of comparability of $\mathcal{C}_F$. The following development establishes such a guarantee.

Property 2 *As used before, let* $\mathcal{R}_1 = \mathcal{R}_1^e \cup \mathcal{R}_1^f$ *and* $\mathcal{R}_2 = \mathcal{R}_2^e \cup \mathcal{R}_2^f$ *be two document clusters to be combined in the clustering process. We will denote by* $\mathcal{R}$ *the union of* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *(i.e.* $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$*). We assume that:*

(i) The dictionary $\mathcal{D}$ *is independent of* $\mathcal{R}_1$ *and* $\mathcal{R}_2$*;*

(ii) $$\frac{|\mathcal{R}_1^e \cup \mathcal{R}_2^e|}{|\mathcal{R}_2^e|} = \frac{|\mathcal{R}_1^f \cup \mathcal{R}_2^f|}{|\mathcal{R}_2^f|}.$$

Then we have:

$$M(\mathcal{R}^e, \mathcal{R}^f) \geq min\{M(\mathcal{R}_1^e, \mathcal{R}_1^f), M(\mathcal{R}_2^e, \mathcal{R}_2^f), M(\mathcal{R}_1^e, \mathcal{R}_2^f), M(\mathcal{R}_2^e, \mathcal{R}_1^f)\} \tag{4.4}$$

Condition (i) states that the proportion of English (resp. French) words translated in the French (resp. English) corpus is homogeneous in the English (resp. French) corpus, which is reasonable to assume if the bilingual dictionary is independent from the corpus. Furthermore, condition (ii) is likely to be satisfied in our settings as (a) all corpora are preprocessed to remove documents too short or too long, and (b) the penalty used in the similarity measure in equation 4.3 guarantees clusters with a similar number of documents in different languages. We now give the sketch of the proof of this property.

Proof[sketch]: Let $Q = \mathcal{R}_1^e \cap \mathcal{R}_2^e$ be the intersection between the vocabularies of $\mathcal{R}_1^e$ and $\mathcal{R}_2^e$. Using the fact that $M_{ef}(\mathcal{R}_i^e, \mathcal{R}_i^f) \leq M_{ef}(\mathcal{R}_i^e, \mathcal{R}_i^{f'})$ for all $\mathcal{R}_i^{f'}$ such that $\mathcal{R}_i^f \subseteq \mathcal{R}_i^{f'}$ (and similarly for the French to English direction), we have, for $i = 1, 2$:

$$\sum_{w \in \mathcal{R}_i^e \setminus Q} \sigma(w, \mathcal{R}_1^f \cup \mathcal{R}_2^f)) \geq |\mathcal{R}_i^e \setminus Q| \cdot \max\{M_{ef}(\mathcal{R}_i^e, \mathcal{R}_1^f), M_{ef}(\mathcal{R}_i^e, \mathcal{R}_2^f)\}$$

and, for the words in Q:

$$\sum_{w \in Q} \sigma(w, \mathcal{R}_1^f \cup \mathcal{R}_2^f)) \geq$$
$$|Q| \cdot \max\{M_{ef}(\mathcal{R}_1^e, \mathcal{R}_1^f), M_{ef}(\mathcal{R}_2^e, \mathcal{R}_2^f), M_{ef}(\mathcal{R}_1^e, \mathcal{R}_2^f), M_{ef}(\mathcal{R}_2^e, \mathcal{R}_1^f)\} \tag{4.5}$$

Then, using the independence assumption between the corpus and the dictionary, one can arrive at:

$$\sum_{w \in (\mathcal{R}_1^e \cup \mathcal{R}_2^e) \cap \mathcal{D}^e} \sigma(w, \mathcal{R}_1^f \cup \mathcal{R}_2^f)) \geq$$
$$|(\mathcal{R}_1^e \cup \mathcal{R}_2^e) \cap \mathcal{D}^e| \cdot \min\{M_{ef}(\mathcal{R}_1^e, \mathcal{R}_1^f), M_{ef}(\mathcal{R}_2^e, \mathcal{R}_2^f), M_{ef}(\mathcal{R}_1^e, \mathcal{R}_2^f), M_{ef}(\mathcal{R}_2^e, \mathcal{R}_1^f)\}$$

The same derivation on M_{fe} and the application of condition (ii) completes the proof. ■

The inequality in 4.4 shows that, in order to maximize the comparability score of the newly formed cluster (i.e. $M(\mathcal{R}^e, \mathcal{R}^f)$), one should rely on those clusters $\mathcal{R}_1$ and $\mathcal{R}_2$ such that the right-hand side of inequality 4.4 is maximal. As $\mathcal{R}_1$ and $\mathcal{R}_2$ are existing clusters, originally formed by choosing sets of English and French documents with the highest comparability scores, the minimum is attained on either $M(\mathcal{R}_1^e, \mathcal{R}_2^f)$ or $M(\mathcal{R}_2^e, \mathcal{R}_1^f)$ (otherwise $\mathcal{R}_1^e$ and $\mathcal{R}_2^f$ would have been grouped before). Thus, in order to maximize the comparability score of the newly formed cluster, the clustering process should select those clusters with a high score for $M(\mathcal{R}_1^e, \mathcal{R}_2^f)$ and $M(\mathcal{R}_2^e, \mathcal{R}_1^f)$. This is exactly what the similarity measure defined in equations 4.1 and 4.3 and used in our clustering process aims at. Hence, the overall process we have defined leads to enhanced comparable corpora, which are both homogeneous and with a high degree of comparability.

Computational Considerations

As comparable corpora usually consist of a large number of documents, the agglomerative clustering algorithm may cost a lot of memory space and computational time. We address this problem (a) by providing a lower bound of the comparability measure which can be computed efficiently, (b) by filtering out document pairs with comparability scores less than a predefined threshold η, empirically set to 0.3 in the experiments, and (c) by updating the similarity matrix iteratively in an efficient way during the clustering process. As the clustering process involves at each iteration the merging of the two closest clusters, relying on a lower bound ensures that the clusters to be merged have a high comparability score.

In our implementation, the measure $M(\mathcal{C}_s, \mathcal{C}_t)$ in equation 4.1 is replaced by a lower-bound[1] $\frac{1}{\#_d(\mathcal{C}_s)\cdot\#_d(\mathcal{C}_t)} \sum_{d_e \in \mathcal{C}_s, d_f \in \mathcal{C}_t} M(d_e, d_f)$ which yields a similarity measure defined as the accumulative value of all the connections between two clusters. It is feasible, with this new measure, to update the similarity matrix iteratively in the clustering process. Assuming the clustering process merges, at some point, clusters $\mathcal{R}_1$ and $\mathcal{R}_2$ into $\mathcal{R}_{new}$, the similarity matrix between clusters should be updated and the similarity between $\mathcal{R}_{new}$ and any other cluster (e.g. $\mathcal{R}_3$) should be computed. According to equation 4.3 and the new similarity, the similarity between $\mathcal{R}_{new}$ and $\mathcal{R}_3$ can be written as:

$$sim_l(\mathcal{R}_{new}, \mathcal{R}_3) = \frac{(N_{\mathcal{R}_1} + N_{\mathcal{R}_2}) \cdot \phi(\mathcal{R}_1 \cup \mathcal{R}_2)}{\#_d(\mathcal{R}^s_{new}) \cdot \#_d(\mathcal{R}^t_3) + \#_d(\mathcal{R}^s_3) \cdot \#_d(\mathcal{R}^t_{new})}$$

where $(j = 1, 2)$ and:

$$N_{\mathcal{R}_j} = \frac{(\#_d(\mathcal{R}^s_j) \cdot \#_d(\mathcal{R}^t_3) + \#_d(\mathcal{R}^s_3) \cdot \#_d(\mathcal{R}^t_j)) \cdot sim_l(\mathcal{R}_j, \mathcal{R}_3)}{\phi(\mathcal{R}_j \cup \mathcal{R}_3)}$$

In the clustering process, since $sim_l(\mathcal{R}_1, \mathcal{R}_3)$ and $sim_l(\mathcal{R}_2, \mathcal{R}_3)$ are already known before the computation of $sim_l(\mathcal{R}_{new}, \mathcal{R}_3)$, one can directly update the similarity matrix at each iteration. Denoting N_c the number of clusters before a merge, the complexity of this update is $\mathcal{O}(N_c)$, whereas it reaches $\mathcal{O}(N_c \times \bar{C}^2)$ with the direct application of equations 4.1 and 4.3 (with $\bar{C}$ the average number of documents per cluster).

4.3 Experimental Validation

We design in this section two experiments to test the performance of the approaches for improving comparable corpus quality. The first experiment in section 4.3.1 tries to employ the two approaches to improving the quality of comparable corpora and evaluate the performance in terms of the comparability scores. The second experiment in section 4.3.2 tries to use comparable corpus of better quality in

[1]For space constraints, we do not show here that the new measure we introduce is indeed a lower bound of M.

the task of bilingual lexicon extraction to show that the NLP application can benefit from the enhanced quality of comparable corpora.

4.3.1 Improving Corpus Quality

The experiments we have designed here aim at assessing whether the algorithms we have introduced yield corpora of higher quality in terms of comparability scores. In our experiments, we use the two methods described in section 4.2, as well as the representative approaches trying to extract parallel sentences or sub-sentential fragments from comparable corpora such as (Munteanu et al., 2004; Munteanu and Marcu, 2006).

For the two methods introduced in this chapter, one needs to have the original corpus and the external corpus. All the corpora in use have been described in section 3.3.1. We will use in this part the corpora *GH95* and *SDA95* as the original corpora $\mathcal{C}^0$. Two classes of external corpora are considered to prove that the efficiency of our algorithm is not confined to a specific external resource. The first external corpus $\mathcal{C}_T^1$ consists of the corpora *LAT94*, *MON94* and *SDA94*. The second external corpus $\mathcal{C}_T^2$ consists of *Wiki-En* and *Wiki-Fr*. The size of the original and external corpora are given in Table 4.1.

Table 4.1
The size of the original and external corpora (k=1000)

	$\mathcal{C}^0$	$\mathcal{C}_T^1$	$\mathcal{C}_T^2$
English	56k	109k	368k
French	42k	87k	378k

For the clustering approach, we obtain the resulting corpora $\mathcal{C}^1$ (with the external corpus $\mathcal{C}_T^1$) and $\mathcal{C}^2$ (with the external corpus $\mathcal{C}_T^2$). For the greedy approach, we can obtain the resulting corpora $\mathcal{C}^{1'}$ (with $\mathcal{C}_T^1$) and $\mathcal{C}^{2'}$ (with $\mathcal{C}_T^2$) from $\mathcal{C}^0$. The results

are then listed in Table 4.2. In terms of lexical coverage, $\mathcal{C}^1$ covers 97.9% of the vocabulary of $\mathcal{C}^0$, while $\mathcal{C}^2$ covers 99.0% of the vocabulary of $\mathcal{C}^0$. Hence, most of the vocabulary of the original corpus has been preserved, which was one of the requirements behind our approach. Concerning comparability scores, the comparability of $\mathcal{C}^1$ reaches 0.924 and that of $\mathcal{C}^2$ is 0.939. Both corpora are more comparable than the original corpus $\mathcal{C}^0$ of which the comparability is 0.881. Furthermore, both $\mathcal{C}^1$ and $\mathcal{C}^2$ are more comparable than $\mathcal{C}^{1'}$ (comparability 0.912) and $\mathcal{C}^{2'}$ (comparability 0.915), which shows that homogeneity is crucial for comparability.

Table 4.2
Information of the resulting corpora enhanced by different methods. The row "Coverage" identifies the coverage of the resulting corpus vocabulary w.r.t. the original corpus $\mathcal{C}^0$. The row "M" gives, for each comparable corpus, the comparability score measured by M.

	$\mathcal{C}^0$	$\mathcal{C}^{1'}$	$\mathcal{C}^{2'}$	$\mathcal{C}^1$	$\mathcal{C}^2$
Coverage	100%	95.1%	98.0%	97.9%	99.0%
M	0.882	0.912	0.916	0.924	0.939

The work presented in (Munteanu et al., 2004) and (Munteanu and Marcu, 2006) try to extract parallel sentences and sub-sentential fragments from comparable corpora. Depending on the quality of comparable corpora, there might be only a small number of or absolutely no parallel sentences in them, which results in a low-coverage corpus. For instance, with the corpus $\mathcal{C}^0$ and $\mathcal{C}^1_T$, we can only fetch 5124 parallel sentences by the approach (Munteanu et al., 2004), which covers 23.8% of the original vocabulary. This result does not satisfy the needs of our work in the context of bilingual lexicon extraction which demands that most of the original vocabulary can be preserved in the resulting corpus. Such studies as (Zhao and Vogel, 2002), (AbduI-Rauf and Schwenk, 2009), (Sarikaya et al., 2009), (Tillmann and Xu, 2009) and (Do et al., 2010), trying to mine the parallel content from comparable corpora, have the same problem.

4.3.2 Bilingual Lexicon Extraction

Following the experiments in section 4.3.1, we will further show in this part that the bilingual lexicons extracted from the enhanced comparable corpora are of higher quality. As previous studies on bilingual lexicon extraction from comparable corpora radically differ on resources used and technical choices, it is very difficult to compare them in a unified framework (Laroche and Langlais, 2010). More importantly, our approach aims at enhancing corpus comparability, and can be coupled with any existing bilingual lexicon extraction method once the corpus has been enhanced. It is thus more interesting to directly asses whether such a coupling can lead to increased performance. To extract bilingual lexicons from comparable corpora, we directly use here the method proposed by Fung and Yee (Fung and Yee, 1998), and which has been referred to as the *standard approach* in more recent studies (Déjean et al., 2002; Gaussier et al., 2004; Yu and Tsujii, 2009). In this approach, each word w is represented as a context vector consisting of its surrounding words in the documents. Source (or target) context vectors are then translated with an existing bilingual dictionary. Finally, a translation score is given to any word pair based on the cosine of their respective context vectors.

Experimental Settings

In order to measure the performance of the lexicons extracted, we divide the bilingual dictionary mentioned in section 3.3.1 into 2 parts: 10% of the English words together with their translations are randomly chosen and used as the evaluation set, the remaining words being used to compute context vector similarity. Given the comparable corpus $\mathcal{C}$, English words not present in $\mathcal{C}_e^v$ or with no translation in $\mathcal{C}_f^v$ are excluded from the evaluation set. For each English word in the evaluation set, all the French words in $\mathcal{C}_f^v$ are then ranked according to their similarity with the English word. Precision, recall and the NMR measure are then computed on the first N translations.

The precision amounts in this case to the proportion of lists containing the correct translation (in case of multiple translations, a list is deemed to contain the correct translation as soon as one of the possible translations is present). The recall is the proportion of correct translations found in the lists to all the translations provided in the corpus. This evaluation procedure has been used in previous studies and is now standard. The precision or recall measure is not precise enough as it does not distinguish between candidate translations of different ranks. We thus use an additional measure NMR, previously discussed in (Voorhees, 1999; Yu and Tsujii, 2009) to show the ability of the algorithm to precisely rank the selected translation candidates. Assuming the total number of English words in the evaluation set is m, NMR is defined as $\frac{1}{m}\sum_{i=1}^{m}\frac{1}{r_i}$ where r_i is the rank of the first correct translation in the candidate translation list for the i-th word in the evaluation set. If the correct translation does not appear in the top N candidates, $\frac{1}{r_i}$ will be set to 0. In our experiments, N is set to 20.

Furthermore, several studies have shown that it is easier to find the correct translations for frequent words than for infrequent ones (Pekar et al., 2006), since frequent words are coupled with more context information. To take this fact into account, we distinguished different frequency ranges to assess the validity of our approach for all frequency ranges. Empirically, words with frequency less than 100 are defined as low-frequency words (W_l), whereas words with frequency larger than 400 are high-frequency words (W_h), and words with frequency in between are medium-frequency words (W_m).

Results and Analysis

In a first series of experiments, bilingual lexicons were extracted from the corpora obtained by the clustering approach ($\mathcal{C}^1$ and $\mathcal{C}^2$), the corpora obtained by the greedy approach ($\mathcal{C}^{1'}$ and $\mathcal{C}^{2'}$) and the original corpus $\mathcal{C}^0$. Table 4.3 displays the results obtained. Each of the last two columns "$\mathcal{C}^1 > \mathcal{C}^0$" and "$\mathcal{C}^2 > \mathcal{C}^0$" contains the absolute and the relative difference (in %) w.r.t. $\mathcal{C}^0$. As one can note, the best results are obtained from the corpora built with the clustering approach. The lexicons extracted

Table 4.3
Performance of bilingual lexicon extraction from different corpora

	$\mathcal{C}^0$	$\mathcal{C}^{1\prime}$	$\mathcal{C}^{2\prime}$	$\mathcal{C}^1$	$\mathcal{C}^2$	$\mathcal{C}^1 > \mathcal{C}^0$	$\mathcal{C}^2 > \mathcal{C}^0$
Precision	0.226	0.277	0.325	0.295	0.461	0.069, 30.5%	0.235, 104.0%
Recall	0.103	0.122	0.145	0.133	0.212	0.030, 29.1%	0.109, 105.8%
NMR	0.119	0.150	0.175	0.150	0.257	0.031, 26.1%	0.138, 116.0%

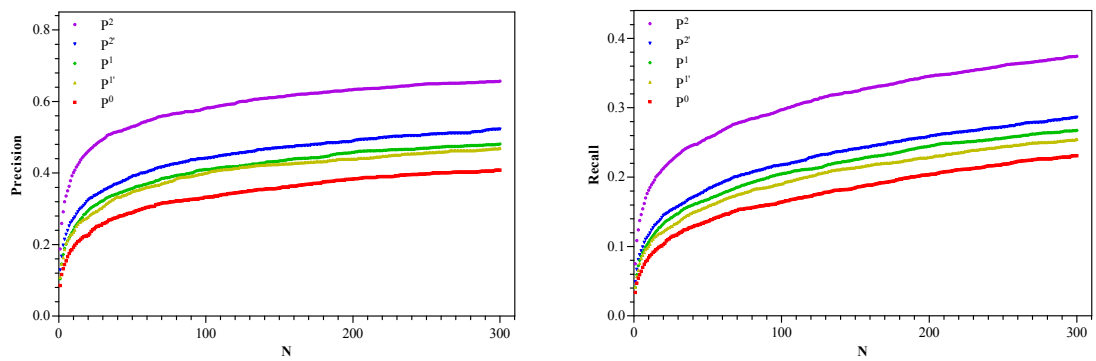
from the enhanced corpora through clustering are of much higher quality, in terms of precision, recall and NMR, than the ones obtained from the original corpus and from the corpora built according to the greedy methodology. The difference is more remarkable with $\mathcal{C}^2$, which is obtained from a large external corpus $\mathcal{C}^2_T$. Intuitively, one can expect to find, in larger corpora, more documents related to a given corpus, an intuition which seems to be confirmed by our results.

To assess the behavior of the methods w.r.t. word frequencies, we focus on the best results on $\mathcal{C}^{2\prime}$ from the greedy approach and the best results on $\mathcal{C}^2$ from the clustering approach. Table 4.4 summarizes the results obtained. As one can note, and not surprisingly, the results obtained with high frequency words are better than the ones obtained with low frequency words. Furthermore, our approach is superior for words in all the frequency ranges. The overall precision can be increased by 41.8% relatively, from 0.325 to 0.461. Comparing $\mathcal{C}^2$ with the original corpus $\mathcal{C}^0$, we note, for the overall precision, a relative increase of 104.0%, from 0.226 to 0.461, which is very satisfactory in the context of general, large evaluation sets. Lastly, the improvement for the low-frequency and medium-frequency ranges is more significant in $\mathcal{C}^2$, which demonstrates that our approach behaves much better on what is generally considered to be a hard problem (Pekar et al., 2006).

In the above experiments, the value N is fixed to 20. Intuitively, the value N plays an important role in the above experiments. In a second series of experiments, we let N vary from 1 to 300 and plot the results obtained with different evaluation

Table 4.4
Comparison of the precision for words of different frequencies

	$\mathcal{C}^0$	$\mathcal{C}^{2'}$	$\mathcal{C}^2$	$\mathcal{C}^{2'} > \mathcal{C}^0$	$\mathcal{C}^2 > \mathcal{C}^0$	$\mathcal{C}^2 > \mathcal{C}^{2'}$
W_l	0.135	0.206	0.304	0.071, 52.6%	0.169, 125.2%	0.098, 47.6%
W_m	0.256	0.390	0.564	0.134, 52.3%	0.308, 120.3%	0.174, 44.6%
W_h	0.434	0.632	0.667	0.198, 45.6%	0.233, 53.7%	0.035, 5.5%
All	**0.226**	**0.325**	**0.461**	**0.099, 43.8%**	**0.235, 104.0%**	**0.136, 41.8%**



(a) Precision (b) Recall

Figure 4.2. Performance of bilingual lexicon extraction from different corpora with varied N values from 1 to 300. The five lines from the top down in each subfigure are corresponding to the results for $\mathcal{C}^2$, $\mathcal{C}^{2'}$, $\mathcal{C}^1$, $\mathcal{C}^{1'}$ and $\mathcal{C}^0$ respectively.

measure in Figure 4.2. In Figure 4.2(a) (resp. Figure 4.2(b)), the x-axis corresponds to the values taken by N, and the y-axis to the precision (resp. recall) scores for the lexicons extracted on each of the 5 corpora $\mathcal{C}^0$, $\mathcal{C}^{1'}$, $\mathcal{C}^{2'}$, $\mathcal{C}^1$ and $\mathcal{C}^2$. A clear fact from the figure is that both the precision and the recall scores increase according to the increase of the N values, which coincides with our intuition. As one can note, the clustering method consistently outperforms the greedy method and also the original corpus on all the values considered for N.

4.4 Conclusion

We have developed in this chapter two approaches to improving the quality of existing comparable corpus. These two approaches make use of the comparability measure developed in chapter 3 and try to extract the high-quality subpart, as well as to enhance the low-quality subpart, of the original corpus. The first approach, namely the *greedy approach*, tries to pick the high-quality document pairs greedily, which thus does not take into account the relations between different document pairs. The second method, namely the *clustering approach*, uses a clustering process to select high-quality sub-clusters, which partially relieves the problem of the greedy approach. The theoretical analysis has shown that both approaches are able to guarantee a least degree of comparability of the resulting comparable corpora.

The two approaches are then validated in the experiments which show that (1) The resulting comparable corpora are of higher quality, in terms of comparability scores, than the original corpus, with the clustering approach being more efficient than the greedy approach; and (2) Bilingual lexicons extracted from better comparable corpora are of higher quality. The extracted lexicons will be used to enhance the CLIR systems in the following chapter. Previous work trying to extract parallel subparts from comparable corpora resembles our work in this chapter. There is however a significant difference as we have shown in the experiments.

We have used in section 4.3.1 two types of external corpora in order to enhance the low-quality subpart of the original comparable corpus. One is the news collection comprised of *LAT94*, *MON94* and *SDA94*, which contains content of the same genre as the original corpus that is also news corpus. The other external resource is the corpora *Wiki-En* and *Wiki-Fr* built from the Wikipedia articles related to specific topics (i.e. categories) that are of similar but not the same genre as the news articles. Intuitively, using external resource of the same genre as that of the original corpus will lead to more significant enhancement of the original corpus, since one can expect to find more related content. However, we use here a much larger Wikipedia corpus

giving us the possibility of finding the content relative to the original corpus, and we have gotten the better results with Wikipedia corpus. Another choice for the external corpus is the content on the web. For a reference of genes on the web, one may refer to the work (Mehler et al., 2010). Intuitively, one can expect to get the best results through using the web corpus consisting of all possible genres in large volume. We do the same experiment as in section 4.3.2 with the greedy algorithm for simplicity, and get a precision of 0.30 compared to 0.28 on $\mathcal{C}^{1'}$ (i.e. resulting corpus with the news corpus as external resource) and 0.33 on $\mathcal{C}^{2'}$ (i.e. resulting corpus with the Wikipedia corpus as external resource). The results ontained with the web corpus do not comply with our expectation because the web corpus we have currently fetched consists of much noise which harms the performance of bilingual lexicon extraction relying on lexical context. The results also tell us that our approaches for enhancing corpus quality are robust to a certain degree to the external resource chosen.

5 CLIR MODELS AND COMPARABLE CORPORA

As one can find from the general strategies of CLIR in section 2.3, the model-dependent approaches rely on a cross-language extension of existing monolingual IR models. If most monolingual IR models have been extended to a cross-language setting, it is not true for all of them, and we will explore in this chapter the cross-language extension of the recently introduced information-based IR systems (Clinchant and Gaussier, 2010). The brief version of this part of work has appeared in our previous work (Li and Gaussier, 2012). In addition to the previous work, we will pay in this chapter additional attention to the enhancement of CLIR system with bilingual lexicons extracted from comparable corpora. The information-based IR models and their extensions to CLIR settings are first introduced in section 5.1. The performance of the proposed CLIR models is then validated in sections 5.2.1 and 5.2.2. In section 5.2.3, we make use of lexicons extracted from comparable corpora to enhance the information-based CLIR models. This chapter is finally concluded in section 5.3.

5.1 Information-based CLIR Models

The information-based models (Clinchant and Gaussier, 2010) have been shown to provide state-of-the-art performance in monolingual IR. The question of their possible extensions, and the quality of these extensions, to a cross-language setting remains unanswered until our work (Li and Gaussier, 2012). This section is devoted to the detailed introduction of CLIR extensions of this model. We will first introduce in section 5.1.1 the monolingual version of information-based models. Then several extensions are proposed to adapt the CLIR environment in section 5.1.2.

The notations we will use throughout this chapter are summarized in Table 5.1 (w represents a word).

Table 5.1
Notations used in IR models in chapter 5

Notation	Description
x_w^q	Number of occurrences of w in query q
x_w^d	Number of occurrences of w in document d
t_w^d	Normalized version of x_w^d
l_d	Length of document d
l_m	Average document length
L	Length of document collection
N	Number of documents in the collection
N_w	Number of documents containing w
$TS(w)$	Set of translations of w
$DS(w)$	Set of documents containing w (i.e. $N_w = \|DS(w)\|$)
RSV(q, d)	Retrieval Status Value of document d for query q

5.1.1 Monolingual Models

Information-based models for IR, recently introduced in (Clinchant and Gaussier, 2010), compute the similarity between queries and documents through the quantity of information brought by document terms on query words. Two such models, referred to as the Log-Logistic model (in short LL) and the Smoothed Power Law model (in short SPL), were shown in (Clinchant and Gaussier, 2010; Clinchant and Gaussier, 2011) to be either on par or to outperform other IR models on several collections and in different settings, as the one of pseudo-relevance feedback.

Information-based models are based on the following retrieval status value[1]:

$$\begin{aligned} RSV(q,d) &= \sum_{w \in q} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \\ &= \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \end{aligned} \tag{5.1}$$

where:

- t_w^d is a normalization function depending on the number of occurrences, x_w^d, of w in d, and on the length, l_d, of d, and satisfies:

  $$\frac{\partial t_w^d}{\partial x_w^d} > 0; \ \ \frac{\partial t_w^d}{\partial l_d} < 0; \frac{\partial^2 x_w^d}{\partial (t_w^d)^2} \geq 0$$

  In the thesis, we follow the settings of the original work and set: $t_w^d = x_w^d \log(1 + c\frac{l_m}{l_d})$ where c is a smoothing parameter;

- P is a probability distribution defined over a random variable, X_w, associated to each word w. This probability distribution must be:

  - Continuous, the random variable under consideration, t_w^d, being continuous;
  - Compatible with the domain of t_w^d, i.e. if t_{min} is the minimum value of t_w^d, then $P(X_w \geq t_{min} | \lambda_w) = 1$;
  - Bursty, i.e. it should be such that:
    $\forall \epsilon > 0, \ g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$ is strictly increasing in x;

- And λ_w is a collection-dependent parameter of P. As suggested in the original work, it is set as:

  $$\lambda_w = \frac{N_w}{N} \tag{5.2}$$

[1]We introduce a slight modification, namely the normalization by the query length, in the formula given in (Clinchant and Gaussier, 2010), in order to provide a more intuitive explanation of the models. This modification does not change the ranking of the documents.

As one can note, equation 5.1 computes the information brought by the document on each query word (i.e. $-\log P(X_w \geq t_w^d|\lambda_w)$) weighted by the importance of the word in the query (i.e. $\frac{x_w^q}{l_q}$). In order to define a proper IR model, one needs to choose a particular bursty distribution. As mentioned above, two such distributions have been proposed and studied, and we will rely on them here. These are the log-logistic and smoothed power law distributions, associated to the models referred to as LL and SPL and defined as (see (Clinchant and Gaussier, 2010)):

$$
\begin{aligned}
RSV_{LL}(q,d) &= \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w}{\lambda_w + t_w^d} \\
RSV_{SPL}(q,d) &= \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}
\end{aligned}
$$

We now turn to cross-language extensions of this family of models.

5.1.2 Cross-Language Extensions

We will present in this section well-founded cross-language extensions of the information-based models, namely the LL and SPL models. These extensions are based on the following considerations:

- A generalization of the notion of *information* used in the information-based family;
- A generalization of the random variables used in this family;
- The direct expansion of query terms with their translations.

First of all, one can note that the information brought by a document on a query term in equation 5.1 is restricted to the query word itself. It is however possible to adopt a more general view by considering the mean information brought by all words in the document related to a given query term. Let $\mathcal{F}(w)$ denote the set of all the words related, through a relation we leave unspecified for the moment, to a given

query word w. Let us furthermore introduce the normalized relation between w and a word w' in d, a quantity we will denote as $\mathcal{A}(w, w', d)$, as:

$$\mathcal{A}(w, w', d) = \begin{cases} \frac{I_{\mathcal{F}(w)}(w')}{\sum_{w'' \in d} I_{\mathcal{F}(w)}(w'')} & \text{if } \sum_{w'' \in d} I_{\mathcal{F}(w)}(w'') > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $I_{\mathcal{F}(w)}$ is the indicator function of the set $\mathcal{F}(w)$. The mean information brought by all words of the document d related to a given query term w can then be defined as: $-\sum_{w' \in d} \mathcal{A}(w, w', d) \log P(X_{w'} \geq t^d_{w'} | \lambda_{w'})$, leading to the overall retrieval function:

$$RSV(q, d) \quad = \quad -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log P(X_{w'} \geq t^d_{w'} | \lambda_{w'})$$

Equation 5.1 is just a special case of the above formulation, obtained by setting $\mathcal{F}(w) = \{w\}$, i.e. considering that words are only related to themselves. The application to a cross-language setting then simply amounts to using the translation relation, $\mathcal{F}(w) = TS(w)$, for computing $\mathcal{A}(w, w', d)$ ($TS(w)$ denotes the translation set of word w in our general notations). This leads, for the LL and SPL models, to:

$$RSV_{LL}(q, d) \quad = \quad -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log(\frac{\lambda_{w'}}{\lambda_{w'} + t^d_{w'}}) \tag{5.3}$$

$$RSV_{SPL}(q, d) \quad = \quad -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log(\frac{\lambda_{w'}^{\frac{t^d_{w'}}{t^d_{w'}+1}} - \lambda_{w'}}{1 - \lambda_{w'}}) \tag{5.4}$$

The above equations 5.3 and 5.4 define two new CLIR models, which we will refer to as MI_{LL} and MI_{SPL}, MI standing for *Mean Information*.

A second extension consists in considering that the random variable used in the information-based family is not associated to a single word w, but to a set of words $\mathcal{F}(w)$, namely the words related to w. This defines a new retrieval function of the form:

$$RSV(q, d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \log P(X_{\mathcal{F}(w)} \geq t^d_{\mathcal{F}(w)} | \lambda_{\mathcal{F}(w)})$$

As before, equation 5.1 is just a special case obtained by setting $\mathcal{F}(w) = \{w\}$, and a cross-language version can be obtained by setting $\mathcal{F}(w) = TS(w)$. One needs however

to define $t^d_{\mathcal{F}(w)}$ and $\lambda_{\mathcal{F}(w)}$. We simply set here the first quantity, $t^d_{\mathcal{F}(w)}$, to the sum of the corresponding quantities for the words in $\mathcal{F}(w)$, which corresponds to the fact that we have indeed observed that many (normalized) occurrences of w in d, through its related words. The second quantity, $\lambda_{\mathcal{F}(w)}$, is set in a similar fashion, by considering the normalized document frequency of all the words in $\mathcal{F}(w)$ (see equation 5.2). This leads to the following cross-language version of the LL and SPL models:

$$t^d_{\mathcal{F}(w)} = \sum_{w' \in \mathcal{F}(w)} t^d_{w'}$$

$$\lambda_{\mathcal{F}(w)} = \frac{|\cup_{w' \in \mathcal{F}(w)} DS(w')|}{N}$$

$$\text{RSV}_{LL}(q,d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \log\left(\frac{\lambda_{\mathcal{F}(w)}}{t^d_{\mathcal{F}(w)} + \lambda_{\mathcal{F}(w)}}\right) \tag{5.5}$$

$$\text{RSV}_{SPL}(q,d) = -\sum_{w_f \in q} \frac{x^q_w}{l_q} \log\left(\frac{(\lambda_{\mathcal{F}(w)})^{\frac{t^d_{\mathcal{F}(w)}}{t^d_{\mathcal{F}(w)}+1}} - \lambda_{\mathcal{F}(w)}}{1 - \lambda_{\mathcal{F}(w)}}\right) \tag{5.6}$$

The above extension bears strong similarities with the SYN operator of the InQuery system, developed for CLIR purposes in (Pirkola, 1998). Indeed, the above formulation can also be obtained by considering that all related words form a single word. We have shown here, however, that it also derives from a different perspective, through the use, in the information-based family, of a single random variable to account for all related words. The setting of the associated parameters (t^d_w and λ_w) then naturally follows from the general framework of the information-based family of IR models. For this reason, we will refer to the above CLIR models as JV_{LL} and JV_{SPL}, JV standing for *Joint random Variable*.

Lastly, a third cross-language extension can directly be obtained by expanding all query terms with their translations. As in standard bilingual dictionaries translations

are not weighted, we resort to the following, simple extension of equation 5.1, which could however be extended by taking translation weights into account:

$$RSV(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log P(X_{w'} \geq t_{w'}^d | \lambda_{w'})$$

This leads, for the LL and SPL models, to:

$$RSV_{LL}(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}}{\lambda_{w'} + t_{w'}^d}) \tag{5.7}$$

$$RSV_{SPL}(q,d) = -\sum_{w \in q} \frac{x_w^q}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}^{\frac{t_{w'}^d}{t_{w'}^d+1}} - \lambda_{w'}}{1 - \lambda_{w'}}) \tag{5.8}$$

We will refer to the above CLIR models as QE_{LL} and QE_{SPL}, QE standing for *Query Expansion*.

To summarize, we have defined, through the above developments, three new CLIR versions of the LL and SPL models, within the general framework of information-based models for IR:

1. MI_{LL} and MI_{SPL}, corresponding to equations 5.3 and 5.4;

2. JV_{LL} and JV_{SPL}, corresponding to equations 5.5 and 5.6;

3. QE_{LL} and QE_{SPL}, corresponding to equations 5.7 and 5.8.

5.2 Validation of CLIR Models

In this section, we will, both theoretically and experimentally, validate the performance of information-based CLIR models. Prior to the experimental validation of these models in section 5.2.2, we want to address the question of whether it is possible to validate them from a theoretical point of view. We do so in section 5.2.1 by resorting to the axiomatic theory of IR.

5.2.1 Theoretical Validation

Heuristic retrieval constraints were first fully developed in the seminal work of Fang et al. (Fang et al., 2004), and followed by many others since, including (Fang and Zhai, 2006; Cummins and O'Riordan, 2007; Clinchant and Gaussier, 2010; Clinchant and Gaussier, 2011; Zhai, 2011). Such constraints state conditions IR models should satisfy, and there is now a large corpus of empirical evidence showing that models failing on one condition do not yield an optimal performance. As shown in (Clinchant and Gaussier, 2010), the LL and SPL models we have considered comply with all the conditions for *ad hoc* information retrieval, and so do their cross-language extensions. However, the cross-language setting also relies on new elements, the translations, and the question remains as to whether these new elements can be regulated through a particular CLIR condition. We develop such a condition below.

Let us assume a collection of French documents about rivers and lakes, and the English query *bank*. In this context, the possible translations of *bank* in French are *rive, berge, banc*[2]. Now let us assume that, in one document d, the words *berge* and *banc* appear two times each, and that, in another document, d', the word *rive* appears four times. Let us also assume that d and d' have roughly the same length and that *berge* and *banc* only occur in d and *rive* only in d'. All these assumptions can be met, for example, on a collection containing formatted articles on water flows. In this context, there is absolutely no difference between d and d' with respect to their relevance according to the query, and one would like a good CLIR strategy to assign the same score to these two documents. The following condition formalizes this intuition.

[2]One can certainly think of other possible translations, but this does not change our argument.

Condition 1 *Let q be a source language query consisting of a single term w, d and d' two target language documents of equal length. Furthermore, let $\{w'_0, w'_1, \cdots, w'_k\}$ be equally likely and equally good translations of w such that:*

$$\begin{cases} x^d_{w'_i} = 1, N_{w'_i} = 1, \ 1 \leq i \leq k \\ x^{d'}_{w'_0} = k, N_{w'_0} = 1 \end{cases}$$

Then, a good CLIR strategy should satisfy:

$$RSV(q, d) = RSV(q, d')$$

Because the translation of w is either diluted, in d, on several words, or concentrated, in d', on a single word, we will refer to the above condition as the *DC condition*, where DC stands for Dilution/Concentration. We now review the different CLIR models we proposed in light of this condition, focusing here on the LL model, the reasoning and results being the same for SPL.

As all translations in d have the same number of occurrences, they also have the same normalized frequency, which will be denoted by τ: $\tau = t^d_{w'_i}$, $1 \leq i \leq k$. Furthermore, it is direct to see: $t^{d'}_{w'_0} = k\tau$. The DC assumptions furthermore imply that all translations have the same parameter λ: $\lambda_{w'_i} = \frac{1}{N}$, $0 \leq i \leq k$. Given these assumptions, one can have:

- For the QE extensions:

  $$RSV_{LL}(q, d) = k \log(\tau N + 1), \ RSV_{LL}(q, d') = \log(k\tau N + 1)$$

  Let us define the function $\Delta_{QE}(\tau)$ as the difference between $RSV_{LL}(q, d')$ and $RSV_{LL}(q, d)$:

  $$\Delta_{QE}(\tau) = RSV_{LL}(q, d') - RSV_{LL}(q, d) = \log(k\tau N + 1) - k \log(\tau N + 1)$$

  The derivative of Δ_{QE} w.r.t. τ is:

  $$\frac{\mathrm{d}(\Delta_{QE}(\tau))}{\mathrm{d}\tau} = kN \cdot (\frac{1}{k\tau N + 1} - \frac{1}{\tau N + 1}) \tag{5.9}$$

In our settings for condition 1, $k > 1$, $\tau > 0$ and $N > 1$, so the derivative in equation 5.9 is strictly negative, which implies that Δ_{QE} is strictly decreasing with τ. Noticing that Δ_{QE} equals 0 when $\tau = 0$, one can conclude:

$$RSV_{LL}(q, d') < RSV_{LL}(q, d)$$

The QE strategy thus does not fulfill the DC condition.

- For the MI extensions:

$$RSV_{LL}(q, d) = \log(\tau N + 1),\ RSV_{LL}(q, d') = \log(k\tau N + 1)$$

This time, let us define the function $\Delta_{MI}(\tau)$ such that:

$$\Delta_{MI}(\tau) = RSV_{LL}(q, d') - RSV_{LL}(q, d) = \log(k\tau N + 1) - \log(\tau N + 1)$$

The derivative of Δ_{MI} w.r.t. τ is:

$$\frac{\mathrm{d}(\Delta_{MI}(\tau))}{\mathrm{d}\tau} = kN \cdot (\frac{1}{k\tau N + 1} - \frac{1}{k\tau N + k}) \tag{5.10}$$

With the same settings as before for k, τ and N, the derivative in equation 5.10 is strictly positive, which implies that Δ_{QE} is strictly increasing with τ. Given that Δ_{MI} equals 0 when $\tau = 0$, one can conclude:

$$RSV_{LL}(q, d') > RSV_{LL}(q, d)$$

The MI strategy thus does not fulfill the DC condition. One can find however that in this extension $\mathrm{RSV}_{LL}(q, d)$ is closer to $\mathrm{RSV}_{LL}(q, d')$ than in the QE one. The absolute values of Δ_{QE} and Δ_{MI} are a measure to quantify the fitness of the corresponding CLIR model to the DC condition. In order to compare the fitness of QE and MI to the DC condition, let us define here a function $\Delta_{QE/MI}(\tau)$ as:

$$\begin{aligned} \Delta_{QE/MI}(\tau) &= |\Delta_{QE}| - |\Delta_{MI}| \\ &= k\log(\tau N + 1) + \log(\tau N + 1) - 2\log(k\tau N + 1) \end{aligned}$$

Its derivative w.r.t τ is:

$$\frac{\mathrm{d}(\Delta_{QE/MI}(\tau))}{\mathrm{d}\tau} = \frac{(k-1)\cdot(k\tau N-1)}{(\tau N+1)\cdot(k\tau N+1)}$$

which is a positive value as soon as $k\tau N > 1$, a fact being met in practice. The function $\Delta_{QE/MI}$ is thus increasing with τ. Together with the fact that $\Delta_{QE/MI}$ equals 0 when $\tau = 0$, one can arrive at:

$$|\Delta_{QE}| > |\Delta_{MI}|$$

which implies that MI strategy satisfies the DC condition better than QE does.

- For the JV extension: $RSV_{JV}(q,d) = \log(k\tau N+1) = RSV_{JV}(q,d')$. This extension is thus fully compliant with the DC condition.

The above theoretical development thus reveals that both the MI and QE extensions do not fulfill the DC condition, the violation of the condition being less remarkable in the MI extension. Furthermore, the JV extension does fulfill the DC condition. As we will see in section 5.2.2, our experiments are in agreement with these findings.

5.2.2 Experimental Validation

We use in our experiments the English text collections from the bilingual tasks of the CLEF campaign[3], with English, French, German and Italian queries, from the year 2000 to 2004. Table 5.2 lists the number of documents (N_d), number of distinct words (N_w), average document length (DL_{avg}) in the English document collections, as well as the number of queries, N_q, in each task (all the queries are available in all languages). As the queries from the year 2000 to 2002 have the same target collection, they are combined in a single task. In all our experiments, we use bilingual dictionaries comprised respectively of 70k entries for the French-English language pair, 58k entries for the German-English language pair, and 67k entries for the Italian-English language pair. For evaluation, we use the Mean Average Precision (MAP) scores to evaluate

[3]http://www.clef-campaign.org

the different models. Lastly, we rely on a paired t-test (at the level 0.05) to measure significance difference between the different systems.

Table 5.2
Characteristics of different CLEF collections

Collection	N_d	N_w	DL_{avg}	N_q
CLEF 2000-2002	113,005	173,228	310.85	140
CLEF 2003	169,477	232,685	284.09	60
CLEF 2004	56,472	119,548	230.52	50

Comparisons of Various CLIR Extensions

In a first series of experiments, we compare the different extensions (MI, JV and QE) proposed for both the LL and SPL models. Information-based models rely on one parameter, namely c, used in the normalization step. As this normalization step is identical to the one used in DFR models ((Amati and Van Rijsbergen, 2002)), we use the default setting provided in Terrier[4] for this parameter: $c = 1$. The results we obtained for MAP scores are displayed, for the three language pairs (i.e. French(Fr)-English(En), Italian(It)-English(En), and German(De)-English(En)), in Table 5.3. As one can note, and in accordance to the theoretical validation developed in section 5.2.1, for both LL and SPL, the JV extension is significantly better than both the MI and QE extensions, and meanwhile MI provides better results than QE. In the following experiments aiming at comparing different CLIR systems, we will thus only rely on the JV extension for the two models LL and SPL of the information-based family.

Comparisons with Standard CLIR Models

We then compare the cross-language versions of LL and SPL we have introduced with CLIR versions of standard systems, namely: (a) a vector space model based on

[4]terrier.org

Table 5.3
Comparison of different cross-language extensions of LL and SPL in terms of MAP scores, where "00-02", "03" and "04" respectively correspond to the data sets "CLEF 2000-2002", "CLEF 2003" and "CLEF 2004". A † indicates, for each model, that the difference with the best performing extension (in bold) is significant.

Collection		LL			SPL		
		JV	QE	MI	JV	QE	MI
00-02	Fr-En	**0.417**	0.204†	0.375†	0.401†	0.194†	0.370†
	It-En	**0.393**	0.212†	0.370†	0.373†	0.184†	0.342†
	De-En	**0.410**	0.212†	0.375†	0.390†	0.199†	0.357†
03	Fr-En	**0.480**	0.223†	0.417†	0.462†	0.204†	0.420†
	It-En	**0.434**	0.213†	0.382†	0.421†	0.199†	0.375†
	De-En	**0.463**	0.220	0.394†	0.444†	0.203†	0.328†
04	Fr-En	**0.520**	0.309†	0.417†	0.432†	0.232†	0.346†
	It-En	**0.491**	0.297†	0.406†	0.421†	0.209†	0.317†
	De-En	**0.492**	0.297†	0.406†	0.422†	0.217†	0.328†

Robertson's *tf* and Sparck Jones' *idf* ((Robertson and Sparck Jones, 1988)), referred to as TF-IDF, (b) BM25 with the default parameter setting given by the Terrier system, (c) INQUERY (in short INQ) with the default parameters of the Lemur system, and (d) the Jelinek-Mercer and Dirichlet versions of the language models, again with the default parameters of the Terrier system ($\lambda = 0.15$ and $\mu = 2500$), referred to as LM_{JM} and LM_{DIR}. For the first three models, we directly rely on the SYN strategy, which amounts to considering all the translations of a given query term in the documents as forming a single word. This strategy has been shown to outperform other ones in different studies ((Pirkola, 1998; Sperer and Oard, 2000; Pirkola et al., 2001; Ballesteros and Sanderson, 2003)). For LM_{JM} and LM_{DIR}, two additional strategies have been explored in previous studies (e.g. (Kraaij et al., 2003))

and introduced in section 2.3.2: integration of the translations within the query model (QT), or within the document model (DT), and we first compare them here.

The results of the comparison between three language model-related strategies (SYN, QT, DT) are given in Table 5.4, for the MAP scores on three language pairs. As one can note, the SYN strategy outperforms the other ones, the difference being always significant. Because of that, we will rely for LM_{JM} and LM_{DIR} on the SYN strategy in the following experiments.

Table 5.4
Comparison of different CLIR strategies (SYN, QT, DT) for language models in terms of MAP scores, where "00-02", "03" and "04" have the same meaning as in Table 5.3. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant.

Collection		DT		QT		SYN	
		JM	DIR	JM	DIR	JM	DIR
00-02	Fr-En	0.371†	0.392†	0.364†	0.349†	0.393†	**0.410**
	It-En	0.350†	0.366†	0.321†	0.314†	0.372†	**0.388**
	De-En	0.373†	0.380†	0.349†	0.350†	*0.393*	**0.398**
03	Fr-En	0.442†	0.404†	0.398†	0.378†	**0.472**	0.424†
	It-En	0.416†	0.421†	0.375†	0.380†	**0.436**	0.386†
	De-En	0.427†	0.371†	0.381†	0.334†	**0.455**	0.410†
04	Fr-En	0.422†	0.422†	0.386†	0.419†	**0.451**	0.442
	It-En	0.391†	0.382†	0.378†	0.381†	**0.422**	0.420
	De-En	0.399†	0.387†	0.381†	0.380†	**0.433**	0.431

It is also interesting to note that DT yields results consistently better than QT, which is the worst performing strategy based on language modeling approach. Interestingly, QT is the only strategy which does not fulfill the DC condition introduced in section 5.2.1. Indeed, for Jelinek-Mercer smoothing with the smoothing parameter

λ (the reasoning and the results are the same for Dirichlet smoothing), we obtain, under the setting of the DC condition[5]:

$$\mathrm{RSV}_{QT}(q, d') - \mathrm{RSV}_{QT}(q, d) \quad = \quad \alpha(\log k - (k-1)\log((1-\lambda)\frac{1}{l_d} + \lambda\frac{1}{L}))$$

where α corresponds to the translation probability between the query word and any of its translations. The above quantity is strictly positive for $k > 1$. In contrast, both the DT and SYN strategy are compliant with the DC condition. It is straightforward to see this for SYN: the different words in d are grouped into a single word with k occurrences, hence making the setting in d identical to the one in d'. For DT, we obtain:

$$\mathrm{RSV}_{DT}(q, d') = \log\left(k\alpha((1-\lambda)\frac{1}{l_d} + \lambda\frac{1}{L})\right) = \mathrm{RSV}_{DT}(q, d)$$

Lastly, Table 5.5 gives the results obtained with the different CLIR systems we have reviewed, on all language pairs and all collections. The performance of the monolingual version of the CLIR systems is given in the line MON. First of all, one can note that either the model JV_{LL} obtains the best score (9 times out of 12) , or the difference with the best system is not significant. Furthermore, when JV_{LL} obtains the best score, the difference with the other models is most of the time significant. Indeed, for the cross-language part, only LM_{DIR} is on a par with JV_{LL} on the CLEF2000-2002 collection, only LM_{JM} is on a par with JV_{LL} on the 2003 collection, and all models are significantly below JV_{LL} on the 2004 collection.

5.2.3 Embedding Extracted Lexicons

This section is devoted to assessing whether bilingual lexicons extracted from comparable corpora can be used to improve the performance of CLIR systems. As we have shown in the above experiments that the information-based model JV_{LL} works best in the CLIR environment. We will utilize this model for this part of experiments. Constrained by the corpora for bilingual lexicon extraction, CLIR experiments will be

[5]We omit the derivation here as it is similar with the ones given in section 5.2.1, which is purely technical.

Table 5.5
Comparison of different CLIR systems in terms of MAP scores on all language pairs and collections, where "00-02", "03" and "04" have the same meaning as in Table 5.3. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant.

Data	Model	TF-IDF	BM25	LM_{JM}	LM_{DIR}	INQ	JV_{LL}	JV_{SPL}
00-02	MON	$0.448^{\dagger}$	$0.474^{\dagger}$	$0.462^{\dagger}$	$0.478^{\dagger}$	$0.423^{\dagger}$	**0.487**	0.483
	Fr-En	$0.364^{\dagger}$	$0.389^{\dagger}$	$0.399^{\dagger}$	0.410	$0.353^{\dagger}$	**0.417**	$0.401^{\dagger}$
	It-En	$0.333^{\dagger}$	$0.358^{\dagger}$	$0.372^{\dagger}$	0.388	$0.322^{\dagger}$	**0.393**	$0.373^{\dagger}$
	Ge-En	$0.350^{\dagger}$	$0.367^{\dagger}$	0.393	0.398	$0.342^{\dagger}$	**0.410**	$0.390^{\dagger}$
03	MON	$0.476^{\dagger}$	**0.503**	$0.492^{\dagger}$	$0.475^{\dagger}$	$0.437^{\dagger}$	0.503	0.500
	Fr-En	$0.416^{\dagger}$	$0.441^{\dagger}$	0.472	$0.424^{\dagger}$	$0.408^{\dagger}$	**0.480**	$0.462^{\dagger}$
	It-En	$0.376^{\dagger}$	$0.400^{\dagger}$	**0.436**	$0.386^{\dagger}$	$0.373^{\dagger}$	0.434	$0.421^{\dagger}$
	Ge-En	$0.397^{\dagger}$	$0.420^{\dagger}$	0.455	$0.410^{\dagger}$	$0.384^{\dagger}$	**0.463**	$0.444^{\dagger}$
04	MON	$0.519^{\dagger}$	$0.523^{\dagger}$	$0.511^{\dagger}$	**0.539**	$0.426^{\dagger}$	0.538	$0.529^{\dagger}$
	Fr-En	$0.423^{\dagger}$	$0.420^{\dagger}$	$0.451^{\dagger}$	$0.442^{\dagger}$	$0.376^{\dagger}$	**0.520**	$0.432^{\dagger}$
	It-En	$0.392^{\dagger}$	$0.383^{\dagger}$	$0.422^{\dagger}$	$0.420^{\dagger}$	$0.343^{\dagger}$	**0.491**	$0.421^{\dagger}$
	Ge-En	$0.401^{\dagger}$	$0.395^{\dagger}$	$0.433^{\dagger}$	$0.431^{\dagger}$	$0.353^{\dagger}$	**0.492**	$0.422^{\dagger}$

only performed on the French-English language pair. The following bilingual lexicons will be used in the CLIR systems:

- *bd*: the original dictionary as used in above CLIR experiments in section 5.2.2, which is a gold-standard dictionary constructed from the online dictionary;
- bd_1: the bilingual lexicons extracted from the original comparable corpus $\mathcal{C}^0$ in section 4.3.2;
- bd_2: the bilingual lexicons extracted in section 4.3.2 from the resulting corpus $\mathcal{C}^{2'}$, which is the best corpus constructed by the greedy approach;

- bd_3: the bilingual lexicons extracted in section 4.3.2 from the resulting corpus $\mathcal{C}^2$, which is the best corpus constructed by the clustering approach.

The CLIR system only using the original dictionary bd is set as the baseline. As we have discussed in section 5.1.2, the JV extension of information-based models has the same function as that of the SYN strategy. We will then combine in each experiment the extracted lexicons, e.g. bd_1, with bd in a weighted SYN style (Darwish and Oard, 2003), a strategy which augments the SYN strategy with translation weights as follows: let $WT(w)$ denote the translation weight of w, then the number of occurrences of w in d is set to $x_w^d WT(w)$. The results obtained are displayed in Table 5.6.

Table 5.6
MAP scores of the CLIR experiment using the lexicons extracted from comparable corpora (with JV_{LL} model). The significant differences against the baseline are marked with the †.

Data	bd	$bd + bd_1$	$bd + bd_2$	$bd + bd_3$
CLEF 2000-2002	0.417	0.415	0.420	0.428†
CLEF 2003	0.480	0.482	0.490†	0.495†
CLEF 2004	0.520	0.517	0.533†	0.540†

One can find from the results that with $bd + bd_3$, one always observes a significant improvement over the baseline. With $bd + bd_2$, one can not notice a significant improvement on CLEF 2000-2002 over the baseline. No significant improvement can be observed if one rely on the dictionary $bd + bd_1$ that is extracted from the original comparable corpus. It demonstrates that using the lexicons extracted from the comparable corpus of the best quality (i.e. $\mathcal{C}^2$) can enhance the performance of the JV_{LL} model. However, with the lexicons extracted from comparable corpora of lower quality (i.e. $\mathcal{C}^0$ and and $\mathcal{C}^{2'}$), one can not always achieve the significant improvement. Finally, the best results, in terms of MAP scores, reach 0.428 on CLEF 2000-2002, 0.495 on CLEF 2003 and 0.540 on CLEF-2004, which respectively accounts for 87.9%,

98.4% and 100.4% of the corresponding monolingual baselines. This result is also significantly better than the performance of any other CLIR model.

5.3 Conclusion

There have been a lot of studies investigating the extensions of classic IR models to CLIR environment. None of them however addressed the problem of extending the recently introduced information-based models to a cross-language setting. We have presented in this chapter several strategies for such an extension, (a) through the generalization of the information used in information-based models, (b) through the generalization of the random variables also used in this family, and (c) through the expansion of query terms. The strategy based on the generalization of the random variables plays a role similar to the one of the SYN strategy reviewed in previous studies. The good behavior of this strategy, noticed in these previous studies, is confirmed here for the information-based family.

We have furthermore introduced a novel CLIR condition, thus extending the axiomatic approach to IR to the cross-language setting. This new condition, which we referred to as the Dilution/Concentration condition, helps us assess from a purely theoretical point-of-view the different cross-language extensions we have introduced. The results obtained from this theoretical assessment are confirmed in our experiments. We have also used this condition to assess the possible strategies for building cross-language extensions of the language modeling approach to IR, and again find that the theoretical results are in line with the experimental ones.

We have also shown that the cross-language extension of the log-logistic model (LL) based on the joint random variable (and equivalent to the SYN strategy) yields the best performance on three collections and three language pairs. This model is never significantly below any other model, always significantly above most of them if not all of them. We thus believe this model to be a state-of-the-art CLIR model. Its

simple form, given by equation 5.5, also makes it appealing from an implementation perspective.

Lastly, we have combined into the CLIR system bilingual lexicons extracted from comparable corpora. One can find that the combination of extracted lexicons contributes to the CLIR performance. Moreover, the better the quality of bilingual lexicons is, the more significant improvement one can obtain for the CLIR experiment.

6 CONCLUSION AND DISCUSSION

In this thesis, we have focused on topics related to comparable corpora and their application in bilingual lexicon extraction and CLIR. Previous studies mining comparable corpora paid most attention to the mining algorithms themselves, which have met with a bottleneck in terms of performance. Different from previous work, we have tried here to enhance comparable corpus quality in order to improve the performance of applications relying on comparable corpora. This general idea is advantageous since it can work with any existing algorithm exploiting comparable corpora. The deployment of this idea on real-world applications such as bilingual lexicon extraction and CLIR further validates its efficiency.

We have first proposed in chapter 3 comparability measures to estimate the quality of comparable corpora. The measure is developed according to a notion inspired from the usage experience of bilingual corpora, which is the usability of various bilingual corpora decreases from *parallel corpora* to *non-parallel corpora covering different topics* as listed in section 3.1. The comparability measure then quantifies the expectation of finding the translations for each word in the corpus vocabulary. We have developed in practice two classes of comparability measures: one is to consider context-based disambiguation and the other one is not. The same notion has also motivated the experimental design to validate the efficiency of various comparability measures, where decreasing gold-standard comparability levels are obtained by incrementally introducing noise to the high-quality comparable corpus. The results show that a symmetric measure M (refer to equation 3.10) based on vocabulary overlapping performs very well, meaning that the measure can capture gold-standard comparability levels, which is reflected by high correlation scores. This simple measure has cheap computational cost and is robust to dictionary coverage as we have shown in the experiments. By the way, all context-based measures M^c_{ef}, M^c_{fe} and M^c also perform very well. They

are however computationally too expensive, making them infeasible under intensive calls.

Based on the comparability measure above, we have developed in chapter 4 two strategies, namely the *greedy approach* and the *clustering approach*, to enhance the quality of any comparable corpus. The general methodology of two approaches is to extract the high-quality subpart and to enrich the low-quality subpart of the original corpus. The greedy approach tries to choose the high-quality document pairs from the original comparable corpus to construct a high-quality subpart. The low-quality subpart is then enriched in a similar way by referring to an external resource. In the greedy approach, document pairs are chosen independent from one another and relations among different documents are not taken into account. To solve the problem, the clustering approach makes use of the hierarchical clustering algorithm and a pruning strategy to select the high-quality sub-clusters which are *homogeneous*. Both approaches can be validated from a theoretical perspective that a least degree of comparability can be guaranteed for the resulting corpora. These two approaches are then used in the experiments to show their ability to improve the quality of comparable corpora, with the clustering approach being more efficient than the greedy one. The enhanced comparable corpora are lastly used in the task of bilingual lexicon extraction with a standard method to produce lexicons of better quality. We have used in the experiments two kinds of external corpora to enhance the low-quality subpart of the original corpus, which shows that our algorithms are robust to a certain degree to the external corpus chosen.

The last part of work in chapter 5 concerns about CLIR. The information-based models have shown satisfactory performance in monolingual IR tasks. We have first tried in our work to extend this model to CLIR settings and develop three candidate extensions, namely the one generalizing the notion of *information*, another one generalizing the random variables, and the third one directly expending the query terms. We have then developed a novel CLIR condition that can be used to assess different CLIR models from a theoretical point-of-view. The experimental compar-

isons of various CLIR models comply with the theoretical assessment using the CLIR condition. Based on both theoretical and experimental validation of CLIR versions of information-based models, we have lastly concluded that the Log-logistic model (LL) with the Joint random Variable strategy for extension gives us a best-performing CLIR model. This CLIR model can be further enhanced with bilingual lexicons extracted from comparable corpora. Moreover, the better one comparable corpus is, the more significant improvement one can obtain by embedding in CLIR experiments extracted lexicons with the existing bilingual dictionary.

We have proposed in this thesis a set of methods that have been well validated to be able to measure and enhance comparable corpus quality. The enhanced comparable corpus then results in better NLP performance. There are several directions we can follow in future work.

- The first direction concerns about the style of gold-standard comparability levels. We have introduced in section 3.3.2 a strategy to construct a group of comparable corpora with decreasing comparability levels. The idea is that one can decrease corpus quality by exchanging the high-quality part of the original corpus with noisy content. The more noise one introduces into the original corpus, the more one can degrade the original corpus quality. We have then followed the list of comparable corpora in section 3.1 and built three groups of comparable corpora. This is the only approach we are aware of to construct the quantitative comparability levels. However, the degradation process we have considered is not the only style which exists among different comparable corpus quality in real-world cases. The idea of introducing noise is intuitive and easy to use, but only reflects one type of difference among different comparable corpora. One can think out, for example, an approach that relies more on the content and topic similarity but not explicitly on the proportion of noisy content in comparable corpora. In this case, manual efforts might be needed in order to organize the content and topics according to certain criteria. We are also interested in testing our comparability measures on different comparability

styles. By the way, we currently focus on the French-English language pair. The comparable corpora of other language pairs, especially those pairs between different language families, can be considered as well.

- The second direction is regarding the comparability measures themselves. The measures we have proposed in this thesis rely on the simple information that is the overlapping of corpus vocabularies. As we have shown, the measure performs very well in our experimental settings. The comparability measure however suffers from problems such as the bilingual dictionary coverage. Although we have validated in section 3.3.2 that the comparability measures M and M^c are robust to dictionary changes in a certain range, it will be more interesting if one can have a measure which are efficient regardless of the dictionary chosen. More deeply semantic information, for instance the topics contained in the corpus, can be employed here to make the comparability measure more robust.

- The third direction is to extend the DC condition we have proposed in section 5.2.1. The DC condition provides us with the possibility of assessing CLIR models from a purely theoretical perspective. In this condition, we have considered a rather simple setting where there is only one word in the query against two target documents. Similar to the set of constraints developed to feature a good monolingual IR model, the DC condition can be extended by considering more query terms in a query against more complex document set, which provides comprehensive constraints to assess the CLIR model.

APPENDICES

A AN EXAMPLE OF STRUCTURED AND UNSTRUCTURED QUERIES

A.1 The Original Query

For the example here, we take the French topic C089 from the CLEF-2001 task. The parts we make use of as the query text in above CLIR experiments are *title* and *desc*, and the part *narr* is simply ignored. The original text of the topic is as follows:

> <num> C089 </num>
> <FR-title> Faillite de M. Schneider </FR-title>
> <FR-desc> Faillite de l'agent immobilier allemand Schneider </FR-desc>
> <FR-narr> Les documents pertinents donnent des informations sur la faillite de l'agent immobilier allemand Schneider et sur les raisons de cette faillite. Ils prennent aussi en considération les omissions, les erreurs et la responsabilité des banques allemandes dans cette affaire. </FR-narr>

Before the further process, stop words are removed and inflected forms of words are lemmatized.

A.2 Structured and Unstructured Queries

The structured query is constructed based on the SYN approach which treats the translation candidates of a query term as synonyms. The unstructured query is however constructed by replacing each query term with all its translations and then combining all the translation candidates as a long query. Based on a bilingual dictionary, the structured and unstructured English queries for the French query C089 are listed below. For simplification, the weight of different query terms are not considered in this example.

Structured:

#sum(#syn(bankruptcy failure insolvency collapse bust crash fall breaking wall smash) Mr. Schneider #syn(agent factor) immovable German)

Unstructured:

#sum(bankruptcy failure insolvency collapse bust crash fall breaking wall smash Mr. Schneider agent factor immovable German)

MY PUBLICATIONS

MY PUBLICATIONS

[Journal] **Bo Li** and Eric Gaussier. 2012. Measuring and Improving Bilingual Corpus Comparability. *Submitted* to *Computational Linguistics*.

[Book Chapter] **Bo Li** and Eric Gaussier. 2012. Exploiting comparable corpora for lexicon extraction: measuring and improving corpus quality. In *BUCC: Building and Using Comparable Corpora*. Serge Sharoff, et al. (editors), Springer Press. (*to appear*)

[Conference] **Bo Li** and Eric Gaussier. 2012. An information-based cross-language information retrieval model. In *Proceedings of the 34th European Conference on Information Retrieval (ECIR 2012)*. (*acceptance rare=21%*)

[Conference] **Bo Li** and Eric Gaussier. 2012. Modèles d'information pour la recherce multilingue. In *Proceedings of CORIA 2012*. (**Best paper award.** In French)

[Conference] **Bo Li**, Eric Gaussier and Akiko Aizawa. 2011. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. (*acceptance rate=25%*)

[Conference] **Bo Li**, Eric Gaussier, Emmanuel Morin and Amir Hazen. 2011. Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In *Proceedings of TALN 2011*. (in French)

[Conference] **Bo Li** and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. (*acceptance rate=19%*)

LIST OF REFERENCES

LIST OF REFERENCES

AbduI-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20:357–389.

Armstrong, S., Kempen, M., McKelvie, D., Petitpierre, D., Rapp, R., and Thompson, H. S. (1998). Multilingual corpora for cooperation. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th ACM SIGIR*, pages 84–91.

Ballesteros, L. and Sanderson, M. (2003). Addressing the lack of direct translation resources for cross-language retrieval. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 147–152.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*.

Bekavac, B., Osenova, P., Simov, K., and Tadic, M. (2004). Making monolingual corpora comparable: a case study of bulgarian & croatian. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 75–82.

Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Inf. Retr.*, 7:183–204.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311.

Callan, J. P., Croft, W. B., and Harding, S. M. (1992). The INQUERY retrieval system. In *In Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics (ACL)*, pages 9–16.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318.

Chiao, Y.-C., Sta, J.-D., and Zweigenbaum, P. (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7.

Clinchant, S. and Gaussier, E. (2010). Information-based models for ad hoc IR. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241.

Clinchant, S. and Gaussier, E. (2011). Is document frequency important for PRF? In *Proceedings of ICTIR*, pages 89–100.

Cummins, R. and O'Riordan, C. (2007). An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, 28:51–68.

Darwish, K. and Oard, D. W. (2003). Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 338–344.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society For information Science*, 41(6):391–407.

Déjean, H., Gaussier, E., and Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

Deléger, L. and Zweigenbaum, P. (2008). Aligning lay and specialized passages in comparable medical corpora. *Studies in Health Technology and Informatics*, 136:89–94.

Deléger, L. and Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 2–10.

DeNero, J. and Klein, D. (2010). Discriminative modeling of extraction sets for machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1453–1463.

Do, T.-N.-D., Besacier, L., and Castelli, E. (2010). A fully unsupervised approach for mining parallel data from comparable corpora. In *Proceedings of the 14th Conference of the European Association for Machine Translation*.

Do, T.-N.-D., Castelli, E., and Besacier, L. (2011). Mining parallel data from comparable corpora via triangulation. In *Proceedings of the International Conference on Asian Language Processing*, pages 185–188.

Fang, H., Tao, T., and Zhai, C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*.

Fang, H. and Zhai, C. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174.

Flottum, K. (2003). Personal english, indefinite french and plural norwegian scientific authors? pronominal author manifestation in research articles. *Norsk Lingvistisk Tidsskrift*, 21:21–55.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 173–183.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING)*, pages 414–420.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184.

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *CoNLL 09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 129–137.

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th international conference on Computational linguistics*, pages 444–450.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 526–533.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Giouli, V., Calzolari, N., Monachini, M., Soria, C., and Choukri, K. (2006). Language resources production models: the case of the intera multilingual corpus and terminology. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.

Goutte, C., Yamada, K., and Gaussier, E. (2004). Aligning words using matrix factorisation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

Haghighi, A., Blitzer, J., DeNero, J., and Klein, D. (2009). Better word alignments with supervised itg models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 923–931.

Hazem, A. and Morin, E. (2012). QAlign: A new method for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 83–96.

Hewavitharana, S. and Vogel, S. (2008). Enhancing a statistical machine translation system by using an automatically extracted parallel corpus from comparable sources. In *Proceedings of the LREC 2008 Workshop on Comparable Corpora*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Ji, H. (2009). Mining name translations from comparable corpora by creating bilingual information networks. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 34–37.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36:779–808.

Kay, M. and Roscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6:97–133.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Kraaij, W., Nie, J.-Y., and Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistic*, 29:381–419.

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5:37–72.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652.

Li, B. and Gaussier, E. (2012). An information-based cross-language information retrieval model. In *Proceedings of the 34th European Conference on Information Retrieval (ECIR)*, pages 281–292.

Li, B., Gaussier, E., and Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 473–478.

Lins, R. D. and Gonçalves, P. (2004). Automatic language identification of written texts. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 1128–1133.

Ma, X. and Liberman, M. (1999). Bits: a method for bilingual text search over the web. In *Machine translation summit VII*.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M., and Yannoutsou, O. (2006). Using patterns for machine translation (mt). In *Proceedings of the European Association for Machine Translation*, pages 239–246.

Mehler, A., Sharoff, S., and Santini, M. (2010). *Genres on the Web: computational models and empirical studies*. Springer.

Mooers, C. (1950). Coding, information retrieval, and the rapid selector. *American Documentation*, 1:225–229.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multiword terms. *Language Resources and Evaluation*, 44:79–95.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 664–671.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the HLT-NAACL 2004*, pages 265–272.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-COLING)*, pages 81–88.

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2009). Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156.

Nie, J.-Y. (2010). *Cross-Language Information Retrieval*. Morgan & Claypool, New York, NY, USA.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Otero, P. G. and Lopez, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr.*, 4:209–230.

Prochasson, E. and Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1327–1335.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL)*, pages 320–322.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 519–526.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the ACL workshop on Comparing corpora*, pages 1–6.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29:349–380.

Robertson, S. E. and Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

Robertson, S. E. and Sparck Jones, K. (1988). Relevance weighting of search terms. In Willett, P., editor, *Document retrieval systems*, pages 143–160. Taylor Graham Publishing.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., and Utsuro, T. (2006). Compiling French-Japanese terminologies from the web. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 225–232.

Salton, G. (1969). Automatic processing of foreign language documents. In *Proceedings of the 1969 conference on Computational linguistics*, pages 1–28.

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc.

Saralegi, X., San Vicente, I., and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *6th International Conference on Language Resources and Evaluations - Building and using Comparable Corpora workshop*.

Saralegi Urizar, X. and Alegria Loinaz, I. (2007). Similitud entre documentos multilingues de caracter cientifico-tecnico en un entorno web. In *Proceedings of the SEPLN*, pages 71–78.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., and Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, pages 432–435.

Setiawan, H., Dyer, C., and Resnik, P. (2010). Discriminative word alignment with a function word reordering model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 534–544.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 3(52).

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*.

Sharoff, S. (2010). In the garden and in the jungle. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer.

Shezaf, D. and Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 98–107.

Simard, M. and Plamondon, P. (1998). Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13:59–80.

Skadina, I., Vasiljevs, A., Skadins, R., Gaizauskas, R., Tufis, D., and Gornostay, T. (2010). Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 6–14.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*.

Sperer, R. and Oard, D. W. (2000). Structured translation for cross-language information retrieval. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1):4.

Tillmann, C. and Xu, J.-m. (2009). A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 93–96.

Turtle, H. and Croft, W. B. (1990). Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–24.

Veronis, J. (2000). *Parallel text processing: Alignment and use of translation corpora*. Kluwer Academic Publishers.

Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403.

Xu, J. and Chen, J. (2011). How much can we gain from supervised word alignment? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 165–169.

Yu, K. and Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of HLT-NAACL 2009*, pages 121–124.

Zhai, C. (2011). Axiomatic analysis and optimization of information retrieval models. In *Proceedings of ICTIR*.

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 403–410.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22:179–214.

Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of chinese-english parallel corpus from the web. In *Proceedings of 28th European Conference on Information Retrieval*, pages 420–431.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*.

Zweigenbaum, P., Rapp, R., and Sharoff, S., editors (2011). *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Association for Computational Linguistics.