Efficient corpus selection for statistical machine translation

Sadaf Abdul Rauf

Université du Maine - Ecole Doctorale STIM
Laboratoire d'Informatique, equipe reconnaissance de la parole et traduction automatique

Sélection de Corpus en Traduction automatique statistique
Efficient corpus selection for Statistical Machine Translation

Thése

présentée et soutenu publiquement le 17 Janvier 2012 à Le Mans

pour l'obtention du

Doctorat de l'Université du Maine
(spécialité Informatique)

par

Sadaf Abdul Rauf

Membres du jury: Laurent Besacier (rapporteur)
Hélène Maynard (rapporteur)
Holger Schwenk (directeur)
Marcello Federico (examinateur)
Yannick Estève (examinateur)

Dedicated to Pakistan.

Acknowledgements

Most humbly I am grateful to Allah Almighty, The One Who, to say the least, gave me the understanding, the strength and the perseverance to carry out this task and Who has helped me all through my life.

The theme of this thesis was suggested by my research supervisor, Holger Schwenk, to whom I would like to express my deepest gratitude for his continuous guidance, suggestions and encouragement. Starting with small ideas, the various facets of research explored over the period of four years evolved into a nicely knit topic. Had it not been the supervision and vision of Holger, this work might have lacked quality, consistency and strength, which he emphasised on and taught wonderfully while keeping me on track.

I would like to pay my gratitude to Dr. Laurent Besacier and Dr. Hélène Maynard for conferring me the honor by being referees for my PhD research work. I am extremely thankful to Dr. Marcello Federico and Dr. Yannick Estève for agreeing to be the members of my thesis evaluation committee. I am truly indebted to all my jury members for the wonderful defense experience they gave me.

I am grateful to the higher education commission of Pakistan (HEC) for providing research grant under HEC Overseas Scholarship 2005, the French government under the projects INSTAR (ANR JCJC06 143038) and COSMAT (ANR 2009 CORD 004 01), and the European project FP7 Euromatrix Plus.

Many thanks to all my friends and colleagues who have helped me at various stages of my thesis. Special thanks to Patrik Lambert for his help and guidance in various phases of my work and for his admirable cooperation whenever I approached him with a problem. I owe sincere thanks to Haithem Afli and Kashif Shah for the kind help in thesis administrative details. My visits to Le Mans have always been source of motivation and

pleasure. This I owe to the friendly colleagues at Le Mans: Loic Barrault, Yannick Esteve, Patrik Lambert, Haithem Afli, Kashif Shah, Walid Aransa, Anthony Rousseau, Frederic Blain.

I am indebted to my brother in law Rizwan-ur-Rehman and my friends for giving their precious time to look after the kids so that I could keep my appointments. Thanks are also due to all my friends and contacts whose company kept me sane.

I would like to record my gratitude to the academic office and support, Université du Maine, especially Mme Martine Turmeau, Etienne Micoulaut, Teva Merlin and Bruno Richard for their assistance and cooperation.

Last but by no means the least, I am deeply grateful to each member of my family, especially my parents-in-law, parents and husband, for their full support and constant encouragement all the way through my thesis. Special thanks to my sons Shahzain and Umar, who suffered neglect for so long in order to make this possible.

Contents

List of Figures

List of Tables

LIST OF TABLES

Chapter 1

Introduction

Machine translation (MT) has been a subject of interest since a long time. It involves the use of computers to translate text or speech from one natural language into another one. Like many other computational problems it has been largely reinvented in the last two decades by the incorporation of machine learning methods. Statistical machine translation (SMT) is the machine learning approach to the problem and thrives on data. Parallel and monolingual corpora are the raw material for SMT systems. *Parallel corpora* are sentence aligned multilingual corpora, where each sentence in one language has a translation in the other language. But their shortage is a bottleneck for many languages (and domains). The high expense of parallel corpus creation has motivated research in techniques to produce parallel corpora.

Parallel corpora have proved to be an indispensable resource for statistical machine translation [Brown et al., 1990; Och and Ney, 2002] as well as many other natural language processing applications, including building and extending bilingual lexicons and terminologies. However, beyond a few language pairs such Arabic-English, Chinese-English or a couple of European languages, and a few domains such as parliamentary debates or legal texts (proceedings of the Canadian or European Parliament [Koehn, 2006], or of the United Nations[1]), they remain a sparse resource, often due to the huge expense (human as well as monetary) required for their creation. Also the language jargon used in such corpora is not very well suited for everyday life translations or translations of some other domain, thus a dire need arises for more parallel corpora which are well suited for domain specific translations.

State-of-the-art machine translation based on the statistical approach is a data driven process. Treating translation as a machine learning problem, the SMT algorithm

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T4A

learns rules of translation from previously translated texts [Lopez, 2008a]. It enables rapid construction of systems for any language pair if sufficient training material is available. The performance of the system thus heavily depends on the quality and quantity of the corpus used for training. Generally, more appropriate bitexts lead to better performance. SMT systems use parallel texts as training material for the translation model and monolingual corpora for target language modeling. Though enough monolingual data is available for most languages, it is the parallel corpus that is insufficient with respect to its need.

Considerable effort has been put into the exploration of options to create parallel corpora. One obvious option to increase this insufficient resource could be to produce more human translations, but this is a very expensive option, in terms of both time and money. Germann [2001] report 140 translating hours to create a Tamil-English parallel corpus of about 1300 sentence pairs with 24,000 tokens on the Tamil side with an average translation rate of 170 words per hour. They concluded that translating a corpus of 100, 000 words in a month requires 4 to 5 full time translators. This of course forces to explore other scenarios for parallel corpus creation. Crowd sourcing could be another option [Ambati and Vogel, 2010; Bloodgood and Callison-Burch, 2010; Zaidan and Callison-Burch, 2011], but this has its own costs and thus is not very practical for all cases. The world wide web can also be crawled for potential "parallel sentences" [Fung et al., 2010; Hong et al., 2010; Ishisaka et al., 2009; Kilgarriff and Grefenstette, 2003; Nie et al., 1999; Resnik and Smith, 2003; Zhang et al., 2006]. Also, most of the found bilingual texts are often not direct translations of each other and not very easy to align. In recent works less expensive but very productive methods of creating such sentence aligned bilingual corpora were proposed. These are based on extracting "parallel" texts from already available "almost parallel" or "not much parallel" texts. The term "comparable corpus" is often used for such texts.

A comparable corpus is a collection of texts composed independently in the respective languages and combined on the basis of similarity of content. These are documents in one to many languages, that are comparable in content and form in various degrees and dimensions. Whereas a parallel corpus, also called bitext, consists in bilingual/multilingual texts aligned at the sentence level.

The raw material for comparable documents is often easy to obtain but the alignment of individual documents is a challenging task [Oard, 1997]. Potential sources of comparable corpora are multilingual news reporting agencies like AFP, Xinhua, Al-Jazeera, BBC etc, or multilingual encyclopedias like Wikipedia [Bharadwaj and Varma, 2011; Otero and Lopez, 2010; Smith et al., 2010], Encarta etc. Some of these compa-

rable corpora are widely available from LDC,[1] in particular the Gigaword corpora, or over the web for many languages and domains, e.g. Wikipedia. Comparable texts can also be found in large quantities over the world wide web, which is a continually growing resource. With the increasing diversity of the web, even low density languages will start having a significant presence. These comparable resources often contain many sentences that are reasonable translations of each other. Reliable identification of these pairs would enable the automatic creation of large and diverse parallel corpora, which in turn will open up many new and exciting research avenues.

However, identifying parallel sentences in comparable corpora is a hard problem. Even texts conveying the same information exhibit great differences at sentence level. Discovering the links of parallelism among these sentences requires isolated judgement of sentence pairs, independent of the context they appear in. Traditional sentence alignment algorithms [Fung and Church, 1994; Gale and Church, 1993; Wu, 1994] don't work here as they are designed to align sentences in parallel corpora and operate on the assumption that there are no re-orderings and that there are limited insertions and deletions between the two documents.

This dissertations presents two methods to improve SMT systems by exploiting the huge resources of comparable and monolingual corpora. The performance of an SMT system is heavily influenced by the domain of the training data. If there is a huge difference between the training and testing domains, translation quality decreases significantly. Thus, the need is not just of more parallel data but of more appropriate parallel corpora. The approaches we propose to help achieve this goal.

We propose methods to exploit comparable and monoligual corpora using information retrieval (IR) and SMT itself. The basic idea of our approaches is that we use SMT to get target language translations of the source texts. Information retrieval is then done using these automatic translations to find the matching sentences from the target language texts. In the realm of comparable corpora, we use the target language comparable corpora to find the matching sentences, and then select the most reliable parallel sentences pairs. On the other hand, when using monolingual corpora in the proposed framework, we use the development and test sets to find the matching sentences from the target language model data. These sentences are then translated back to the source language and used as additional bitexts.

Our approach to utilize comparable corpora is inspired by the work of Munteanu and Marcu [2005]. We devise a method to efficiently mine parallel corpora from comparable corpora. To do this, we translate the source language corpus to the target language

[1] http://www.ldc.upenn.edu/

using an SMT system. Using these translations we then find matching sentences from the target language corpus using IR. We choose the potential parallel sentence pairs by computing the similarity score between them. Our extracted parallel sentences have proved to be very helpful in improving state-of-the-art SMT systems. Though, using a similar scheme to address the problem as Munteanu and Marcu [2005], we differ from them in query creation where we use SMT translations as queries, emphasizing both precision and recall. Another significant difference is our post processing after IR, where for us the work is reduced to simple similarity score computation. Following similar lines [Do et al., 2010] and [Gahbiche-Braham et al., 2011] showed SMT improvements using their mined parallel sentences.

In our approach to employ monoligual corpora to prepare parallel texts, we take inspiration from the ideas of translation model adaptation and lightly-supervised training and provide an extension to the work of [Schwenk, 2008] by actively selecting the sentences to be used as additional bitexts. We use target language development and test data as queries and retrieve n-best matching sentences from the target language monolingual corpus. This gives us the sentences which are close to the development and test data. We then translate these sentences and add them as additional bitexts. Following this scheme, we were able to get considerable improvements in SMT scores.

1.1 Outline of the dissertation

This dissertation involves the research performed between 2008 and 2011 at the Laboratoire d'Informatique de l'Université du Maine (LIUM). This document is organized as follows:

Chapter 2 outlines the current state-of-art SMT technology. It describes the mathematical framework of the early word-based SMT systems and presents phrase-based systems as a natural evolution of the original approaches. A brief overview of the word alignment process and the different alignment models is also given. Language modeling concepts are also presented along with the concepts of smoothing and perplexity. The section on SMT concludes by discussing the difficulties of MT evaluation and presenting the most commonly used evaluation metrics. Our approach uses information retrieval techniques. These are presented in detail as we describe the underlying theory including the vector space model, the $tf.idf$ weighting system and the various steps of an IR system like indexing, query formulation and evaluation. The characteristics of parallel and comparable corpora are then described in detail along with an overview of the previous work which serves as an inspiration for our work.

Chapter 3 describes our approach to extract parallel sentences from comparable corpora. We start by presenting the resources used, followed by the general architecture and then describe the individual components in detail. We then present detailed results for the translation from Arabic and French into English. The chapter concludes by presenting a theoretical and empirical comparison of our approach with one of the existing approaches .

Chapter 4 presents our work towards exploiting the huge monolingual target language reservoirs to 'make' parallel corpora close to the development and test sets. We first report the resources used in our scheme, followed by a detailed presentation of our proposed approach. We then present the experimental evaluation and show that we are able to improve already very competitive SMT systems.

Chapter 5 concludes the thesis by presenting a brief summary of the work and discussing various prospects for future research.

1.2 A brief history of Machine Translation

Though the technological foundation of Machine Translation (MT) traces back to 1949, its historical roots can be linked back to the 19th century when the text carved on the ancient Rosetta stone was translated [Budge, 1922] . The stone having same writings on it in two languages (Egyptian and Greek), using three scripts (hieroglyphic, demotic and Greek) was translated/decrypted by Jean-Francois Champollion who could read two of these scripts. This approach served as the starting point for translating languages based on information learned from available parallel texts.

The inceptive ideas of statistical machine translation were closely related to approaches from which information theory [Shannon, 1948] and cryptography [Shannon, 1949, 1951] arose. The famous paper by Weaver [1955], often marked as the starting point for SMT, viewed the problem of translating from one language to the other as decoding an encrypted version of it. The research on SMT gained its true momentum in the early 1990s, inspired by the progress made in speech recognition using statistical methods. The researchers at IBM developed purely statistical machine translation models [Brown et al., 1993, 1990]. The ease of availability of written translated texts and the increase in computation power paved the way for the development of statistical and corpus based MT approaches. Hutchins [2007] gives detailed and comprehensive account of various stages of development of MT since its inception, Dorr et al. [1999] also provides a survey documenting the history as well as the paradigms of MT.

1.2.1 Classification of MT approaches

Machine translation systems can be divided into two generations, the direct and the indirect systems. The first generation systems consists of ***direct systems***. These systems do word level or phrase level translation, usually using a minimal linguistic analysis of source side text [Hutchins and Somers, 1992]. The basic idea is to analyze the input text to the extent that some transformational rules can be applied, this could be part of speech analysis or some phrasal level information. Source language words are then replaced with target language words using a bilingual dictionary and some rearrangement rules to modify the word order according to target language [Arnold et al., 1993]. These direct systems are no longer in use.

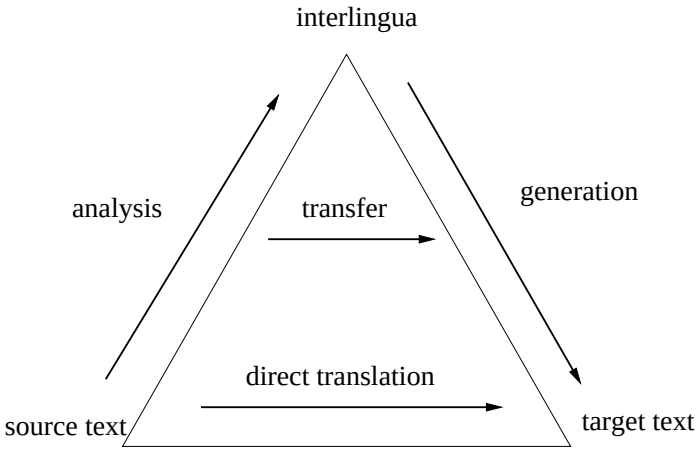

Figure 1.1: Machine Translation triangle (after Vauquois)

The second generation MT systems are classified as the ***indirect systems***. In such systems source language structure is analyzed and text is transformed into a logical form. The target language translation is then generated from the logical form of the text [Hutchins and Somers, 1992]. The transition from direct systems to indirect systems is illustrated in Figure 1.1 (the Vauquois triangle). Indirect systems can be further divided into ***interlingua*** and ***transfer based*** systems.

- In the *transfer based systems*, the source language is analyzed to an abstract level. A transfer module then transfers this abstract form to the corresponding abstract form in the target language through which the target translation text is generated by the generation module. Such systems require independent grammars for source

and target languages. Moreover, they need a comparative grammar or transfer rules to relate source structures to target structures [Arnold et al., 1993].

- The *interlingua approach* involves the use of an intermediate language, with the source language text translated to interlingua and interlingua translated to the target language text. As suggested by Hutchins and Somers [1992], interlingua is an intermediate meaning representation which encompasses all the information required for the generation of the target text without looking at the source text. This representation thus acts as a source text projection as well as the basis for the generation of the target text.

Though the translation quality achieved by the classical transfer based approaches is quite reasonable, it takes many person years to create a system that can well handle a wide variety of sentence constructions and perform well on unseen texts. Such systems need to be rebuild for each language pair. This bottleneck can be avoided by learning translation rules automatically from bilingual texts [Probst, 2005]. Corpus based approaches work on this principle.

Corpus-based approaches, as the name suggests, make use of already available corpora. These work by extracting translation related information from parallel corpora which have already been translated by a human translator. The availability of textual data in huge quantities in the early 1990s resulted in the popularity of corpus-based (a.k.a data-driven) approaches. These approaches have the advantage of reusability, once the required techniques have been defined for a language pair, they can be reused for any other language pair provided that enough parallel data is available [Ramis, 2006].

Example based (EBMT) and *Statistical (SMT)* are among the most relevant examples of corpus based approaches. Example based MT usually works by making a database of translation examples using the available parallel corpus, these are then matched with input sentence based on a similarity measure. The target sequence is built by choosing and combining these examples in an appropriate way. Somers [1999] presents a survey of various EBMT techniques, whereas Hutchins [2005] gives an in depth historical review and commentary.

In SMT, the translation process is accomplished using purely statistical models, mainly a language and translation model trained from monolingual and parallel corpora respectively. Initially, SMT worked by translating word to word, current systems process sequence of words, called phrases. The research community is actively advancing and exploring new methods like incorporating linguistic knowledge etc. We give a

description of statistical approach to machine translation in chapter 3.

Corpus-based approaches need sentence-aligned parallel corpora and effective alignment is crucial for better translation output. A good description of algorithms that have achieved good results for sentence alignment is presented in [Gale and Church, 1993; Haruno and Yamazaki, 1996; Kay and Röscheisen, 1993].

1.3 Research Contributions

The main contributions of this Ph.D thesis dissertation are:

- A novel method to efficiently identify parallel sentences from comparable corpora. This is achieved by translating sentences from the comparable corpus in the source language to the target language, and then using these translations to search for potential parallel sentences in the comparable corpus. The search is done using IR techniques. Sentences are then selected using fast and efficient methods, i.e. simple WER/TER/TERp scores. The first publication of this work [Abdul-Rauf and Schwenk, 2009b] has **30** citations till date (14-06-2012). This work has also been published in a journal article [Abdul-Rauf and Schwenk, 2011].

- Another method enabling use of the target language monolingual corpora to improve state-of-the-art SMT systems. This method is an extension of previous works on unsupervised training by providing an active data selection technique using IR. We exploit the target language monolingual data by extracting the sentences most related to the development and test sets using IR. These retrieved sentences are then translated back to the source language and the automatic translations are used as additional parallel training data to build a new improved SMT system.

- The simplicity yet efficiency of our methods make them easily implementable for anyone. Both the proposed frameworks make use of SMT translations and open source tools.

- The parallel corpora 'generated' by our approaches systematically helped improve SMT systems considerably.

The research developed in this Ph.D. has been published in **8** papers in major conferences and **1** article in an international journal. These are referenced in the appropriate chapters and are given in Appendix A.

Chapter 2

State of the art

This chapter outlines the current state-of-art in SMT technology. The reader is referred to the literature for tutorials and textbooks that cover the whole field [Koehn, 2010; Lopez, 2008a]. We voluntarily limit the presentation to areas related to the research reported in this thesis.

It starts with the mathematical framework of the early word-based SMT systems. Section 2.2 then gives a brief overview of the word alignment process and the different alignment models used in SMT. Phrase based systems are then presented as a natural evolution of the original approaches (section 2.3). Language modeling concepts are presented in section 2.5 along with the concepts of smoothing and perplexity. The SMT section concludes by discussing the difficulties of MT evaluation and presenting the most commonly used evaluation metrics (section 3.2). Since we use information retrieval in our work, we present a brief overview of the IR process in section 2.7. We describe various steps in IR like query formulation, indexing and IR evaluation (section 2.8). We conclude this chapter by formally introducing the comparable corpora and presenting an account of previous works done in the direction of research presented in this thesis.

2.1 Mathematical basis

Let us consider the task of translating a sentence s in the source language S to a sentence t in the target language T. Treating sentences as sequences of words, we define them as $s = s_1, ..., s_i, ..., s_I$ and $t = t_1, ..., t_j, ..., t_J$, where t_i and s_j denote the words in position i of t and position j of s, respectively. The traditional way of viewing SMT is in the noisy channel (or source channel) metaphor, which regards the translation process as

a channel which distorts the target sentence and outputs the source sentence. The task of the translation decoder is, given the distorted sentence s, to find the sentence $\hat{t}$ which has the best probability to have been converted into s. Brown et al. [1993, 1990] suggested a statistical modeling of this distortion process.

Brown et al. [1993, 1990] took the view that every sentence in one language is a possible translation of any sentence in the other. Thus, to every pair of sentence (s,t) they assign a probability $\Pr(t|s)$, which is interpreted as the probability that the translator will produce t as a valid translation in the target language when presented with s in the source language. The chance of error is minimized by choosing the target string $\hat{t}$, for which $\Pr(t|s)$ is greatest, Mathematically,

$$\hat{t} = \arg\max_{t} \Pr(t|s) \tag{2.1}$$

Hence, the translation problem is formulated as choosing the best probability hypothesis among the set of all possible target sentences. Application of the Bayes rule gives the following expression:

$$\hat{t} = \arg\max_{t} \frac{\Pr(s|t)\Pr(t)}{\Pr(s)} \tag{2.2}$$

Since the source text s is constant across any alternative translation t, it can be disregarded, and therefore:

$$\hat{t} = \arg\max_{t} \Pr(s|t)\Pr(t) \tag{2.3}$$

This decomposition into two knowledge sources allows for separate modeling of two of the three fundamental components of the basic SMT system: the **translation model** $\Pr(s|t)$ and the **language model** $\Pr(t)$. Figure 2.1 shows the basic scheme of a typical SMT system under this framework. The rationale of decomposing the problem into two simpler components comes from the speech recognition paradigm and it facilitates computation in MT as well. The translation model (TM) gives the best translation according to the input text while the target language model (LM) ensures that the translation is syntactically correct, regardless of the input text. The **decoder** is the third major component which performs the search for the best translation $\hat{t}$ given the search space of all possible translations based on the probability estimates of the language and translation models.

Early SMT systems used translation models working on a word-by-word basis. This was essentially done using a word alignment model (section 2.2) and word translation

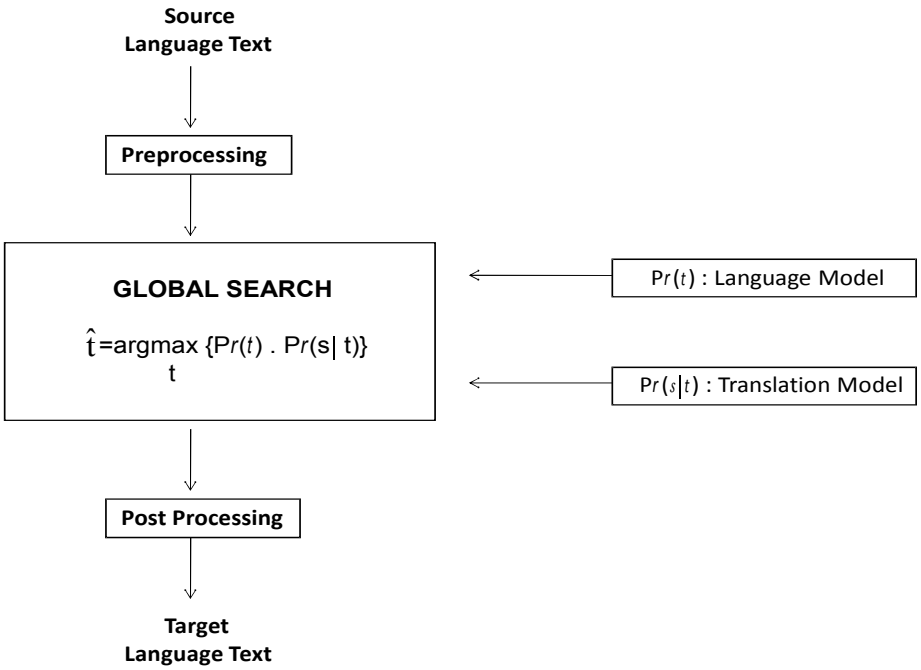

Figure 2.1: Architecture of the translation approach based on source channel framework.

probabilities. Almost all of the recent systems operate on sequences of words, the so-called phrase-based systems. Some phrase-based SMT approaches use non-contiguous phrases or phrases with gaps as in [Gimpel and Smith, 2011; Simard et al., 2005]. A phrase pair is a contiguous sequence of words, one phrase in each language, that are translation equivalents of each other. These phrase pairs are stored along with their frequency statistics and are then used as building blocks for new translations. All the probability distributions of the statistical models are automatically estimated from the sentence-aligned parallel corpus (section 2.3.1).

The language model assigns probabilities to the target language sequence based on the probabilities learned from monolingual target language corpora, hence the term target-side LM (section 2.5).

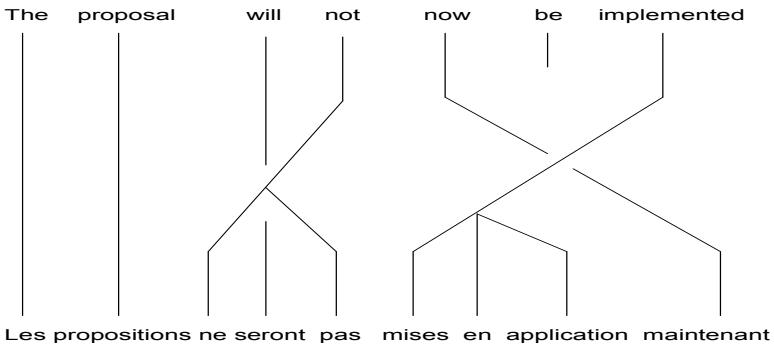

Figure 2.2: A word alignment example.

2.2 Word Alignment

Automatic word alignment is an important component of all SMT approaches. Given a bilingual sentence pair, the general definition of word alignment refers to any defined set of links between lexical units that are translations of each other. Brown et al. [1990] introduces the idea of alignment between a pair of strings as an object indicating for each word in the source string that word in the target string from which it arose.

A sentence pair can have many possible alignments. If the target sentence has length l and the source sentence has length m, then $l \times m$ different connections can be drawn between them since each target word can be aligned with any source word, thus there exist 2^{lm} possible alignments. A typical way to restrict the problem, proposed by Brown et al. [1993] is to assign each source word to exactly one target word. In this way, the number of possible alignments is limited to $(J+1)^l$. It is then the task of the translation model to assign significant probability to only some of the possible alignments by examining the alignments that it considers the most probable.

Figure 2.2 shows a graphical example of alignment between an English and French sentence. The lines, which are called *connections* indicate the origin in the English sentence of each of the words in the French sentence. The number of French words that an English word produces in an alignment is called its *fertility*, thus the word *implemented* producing the French string *mises en application* has a fertility of 3. Also, we see that the word *be* is not connected to any French word thus has a fertility of 0, such cases are handled by using NULL links in the alignment.

2.2.1 Word alignment: mathematical ground

According to equation 2.3, we need to calculate the *inverted* translation probability $Pr(s|t)$ in order to translate the text s in the source language to a string t in the target language. The first challenging task is to establish the correspondences between the words in both sentences. Typically, the number of words and the order of the counterpart appearances in translated sentences is different. This modeling problem is addressed by using a hidden variable a which accounts for all possible pair-wise alignment links between the two sentences

$$\Pr(s|t) = \sum_{a} \Pr(s, a|t) \tag{2.4}$$

$\Pr(s, a|t))$ is generally expressed as:

$$\Pr(s, a|t) = \Pr(J|t) \sum_{j=1}^{J} \Pr(a_j|s_1^{j-1}, a_1^{j-1}, J, t) \cdot \Pr(s_j|a_1^{j}, s_1^{j-1}, J, t) \tag{2.5}$$

where,
J = length of the source sentence s
s_j = word in position j of source sentence s
a_j= hidden alignment of word s_j indicating the position at which s_j aligns in the target sentence

This equation is interpreted as: To generate a source sentence and an associated alignment from the target sentence, we can first choose the length J of the source sentence (given what we know about the target sentence). The choice of where to link the first source position (given the target sentence and the length of the source sentence) can then be made. Then we can choose the identity of the first source word (given the target sentence, the length of the source sentence and the target word linked to the first source position), and so on.

The alignment a_j can take up a zero value ($a_j = 0$) known as NULL word, which accounts for the cases when a source word is not aligned to any of the target words, these are represented by e_0. As explicitly introduced by IBM formulation, word alignment is a function from source positions j to target positions i, so that $a(j) = i$. This implies that the alignment solutions will never contain many-to-many links, but only many-to-one as only one function result is possible for a given source position j.

The right hand side of equation 2.4 sums over each such alignment in which s can be

a translation of t. The goal of the translation model is to maximize $\Pr(s|t)$ over all the sentences of the entire training corpus. Thus, it adjusts word translation probabilities so as to increase the probabilities of the translation pairs in the training corpus.

To calculate word translation probabilities, we need to know how many times a word is aligned with another word. But, each sentence pair can be aligned in many different ways, and each such alignment has some probability. This probability is given as :

$$\Pr(a|s,t) = \frac{\Pr(s,a|t)}{\Pr(s|t)} \tag{2.6}$$

Substituting equation 2.4 into 2.6, we get,

$$\Pr(a|s,t) = \frac{\Pr(s,a|t)}{\sum_a \Pr(s,a|t)} \tag{2.7}$$

Since both $\Pr(a|s,t)$ and $\Pr(s|t)$ are expressed in terms of $\Pr(s,a|t)$, we can get a relation between the word translation probabilities and the alignment probabilities by writing $\Pr(s,a|t)$ in terms of word translation probabilities and then maximizing $\Pr(s|t)$.

The model parameters or probabilities are learned by maximizing the likelihood of parallel data with the Expectation Maximization (EM) algorithm [Dempster et al., 1977]. The three fundamental models developed to calculate the probability in equation 2.5 are decomposed as follows :

- *Fertility model :* This defines the suggested number of source words that are generated by each target word, i.e. the probability that a target word e_i generates Φ_i words in the source sentence.

- *Lexicon model :* Represents the probability of producing a source word f_j given a target word e_i. It suggests strict dependencies between the source and target words.

- *Distortion model :* This model tries to define the reordering of the set of source words such that it best complies with the target language. It models the probability of placing a source word in the position j given that the target word is placed in position i in the target sentence.

Different combination of these models are known as the "IBM machine translation models" and are inspired by the generative process described in figure 2.3 which interprets the model decomposition of equation 2.5. Conceptually this process states that

for each target word, we first find how many source words will be generated (fertility model); then we find which source words are generated from each target word (lexicon model) and finally reorder the source word (distortion model) to obtain the source[1] sentence.

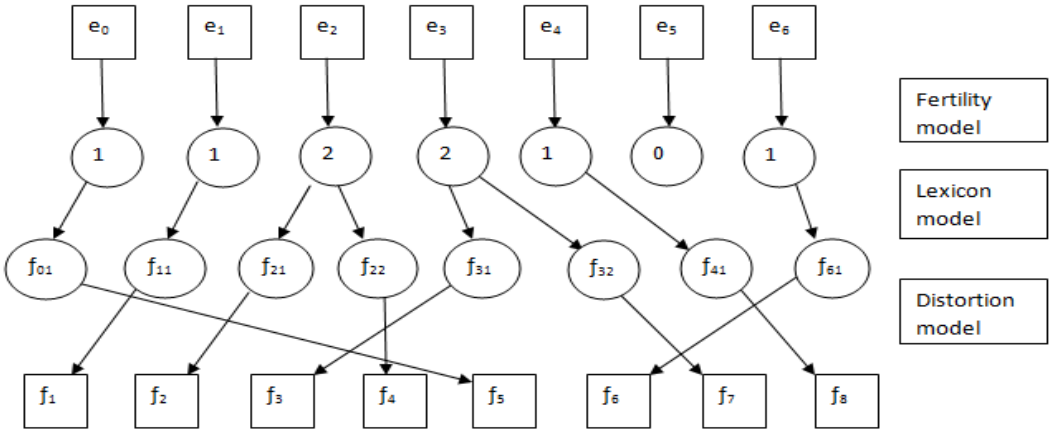

Figure 2.3: The IBM models.

The IBM models are :

- *IBM1:* This model is a simple word translation model, which makes use of co-occurrence of word pairs. It assigns a uniform distribution to the alignment probability, it assigns only lexicon probabilities.
- *IBM2:* This model adds local dependencies by introducing position parameters to the translation model i.e. lexicon plus absolute position.
- *Homogeneous HMM:* This model is a modification of the IBM2 model with the introduction of first-order dependencies in alignment probabilities [Vogel et al., 1996]. This deals with lexicon plus relative position.
- *IBM3:* This model introduces the choice of fertility Φ_i which depends only on e_i.
- *IBM4:* This model takes into account the relative movements conditioning the linking decision on the previous linking decisions, i.e. inverted relative position alignment.

[1]The process generates from target to source language due to application of Bayes rule equation 2.3

- *IBM5*: This model limits the waste of probability mass on impossible situations, this is non-deficient version of IBM model 4.

An inconvenience of the IBM models is that only one-to-many alignments are allowed, since the alignment mapping is restricted to source to target locations. This problem is generally tackled by performing alignments in source-to-target and target-to-source directions and then symmetrizing via the union, intersection or other methods [Och and Ney, 2003].

Och [2002] gives detailed information about the IBM models $1-5$ whereas [Och and Ney, 2003] presents a systematic performance comparison of various models.

2.3 Phrase-based models

In natural languages it is frequent for contiguous sequences of words to translate as a unit. To model this property, current SMT systems are not based on word to word translation but on translation of word phrases [Koehn et al., 2003]. This approach is a simplistic version of the alignment template approach [Och and Ney, 2004] and is known as phrase-based SMT [Zens et al., 2002].

Phrase-based SMT makes it possible to easily handle the collocational relations within the sentence. The translation process consists of grouping source words into phrases, which are contiguous sequence of words (not necessarily linguistically motivated). Some phrase-based SMT approaches use non-contiguous phrases or phrases with gaps as in [Gimpel and Smith, 2011; Simard et al., 2005]. The source phrases are then mapped into target phrases and are generatively inserted as lexically motivated by the word context. The decoding process in phrase-based translation is a very simple one, consisting of three main steps:

1. The source sentence is segmented into known phrases.

2. Each phrase is translated in isolation based on the probability estimates for the phrases from the translation table.

3. The translated phrases are permuted into final order to follow the natural order of the target language.

A source sentence can be split into many possible phrases. The choice of the phrase is based on the existing phrases in the *phrase translation table*. The mathematical for-

mulation for phrase-based systems is the same as for the word based systems, equation 2.3, however $\Pr(s|t)$ is further decomposed into:

$$\Pr(\bar{s}|\bar{t}) = \prod_{i=1}^{I} \phi(\bar{s_i}|\bar{t_i}) d(start_i - end_{i-1} - 1) \tag{2.8}$$

Here, the source sentence s is segmented into I phrases, $\bar{s_i}$, and each $\bar{s_i}$ is translated into a target phrase $\bar{t_i}$; Since the translation direction was mathematically inverted in the noisy channel model, the phrase translation probability $\phi(\bar{s_i}|\bar{t_i})$ is modeled from target to source. $\phi(\bar{s_i}|\bar{t_i})$ represents the phrase probability according to the translation table. The term $d(start_i - end_{i-1} - 1)$ in the formula represents the *distance based reordering model*. According to this model, the reordering of a phrase is relative to the previous phrase: $start_i$ and end_i denote the start and end words of the ith source phrase that translates into ith target phrase.

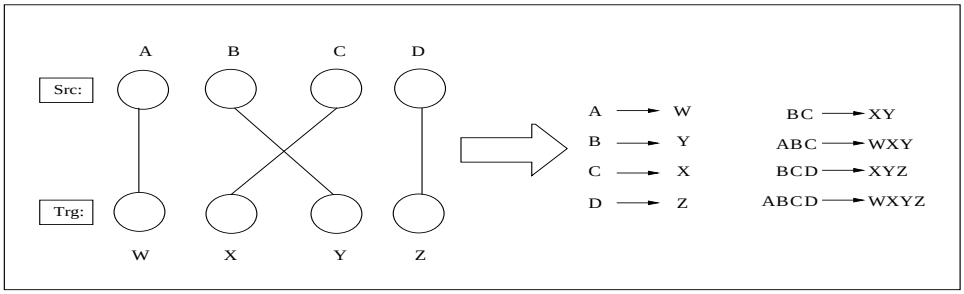

Figure 2.4: Phrase extraction from a certain word aligned pair of sentences.

2.3.1 Phrase translation table creation

The phrase translation probabilities are estimated over all bilingual phrases using the relative frequency of the target sequence given the source sequence. The phrase translation probabilities $\phi(\bar{s_i}|\bar{t_i})$ are learned using the Maximum Likelihood Estimation (MLE), that is, counts of the phrase pairs in the corpus:

$$\phi(\bar{s_i}|\bar{t_i}) = \frac{count(\bar{t_i}, \bar{s_i})}{\sum_{\bar{s_j}} count(\bar{t_i}, \bar{s_j})} \tag{2.9}$$

The phrases are extracted by applying a set of heuristics to the word aligned parallel corpora. According to a criterion any sequence of consecutive source words and consecutive target words which are aligned to each other and are not aligned to any

other token in the sentence become a phrase. Och et al. [1999b] and Zens et al. [2002] give details of the criterion. Figures 2.4 and 2.5 show phrase extraction examples.

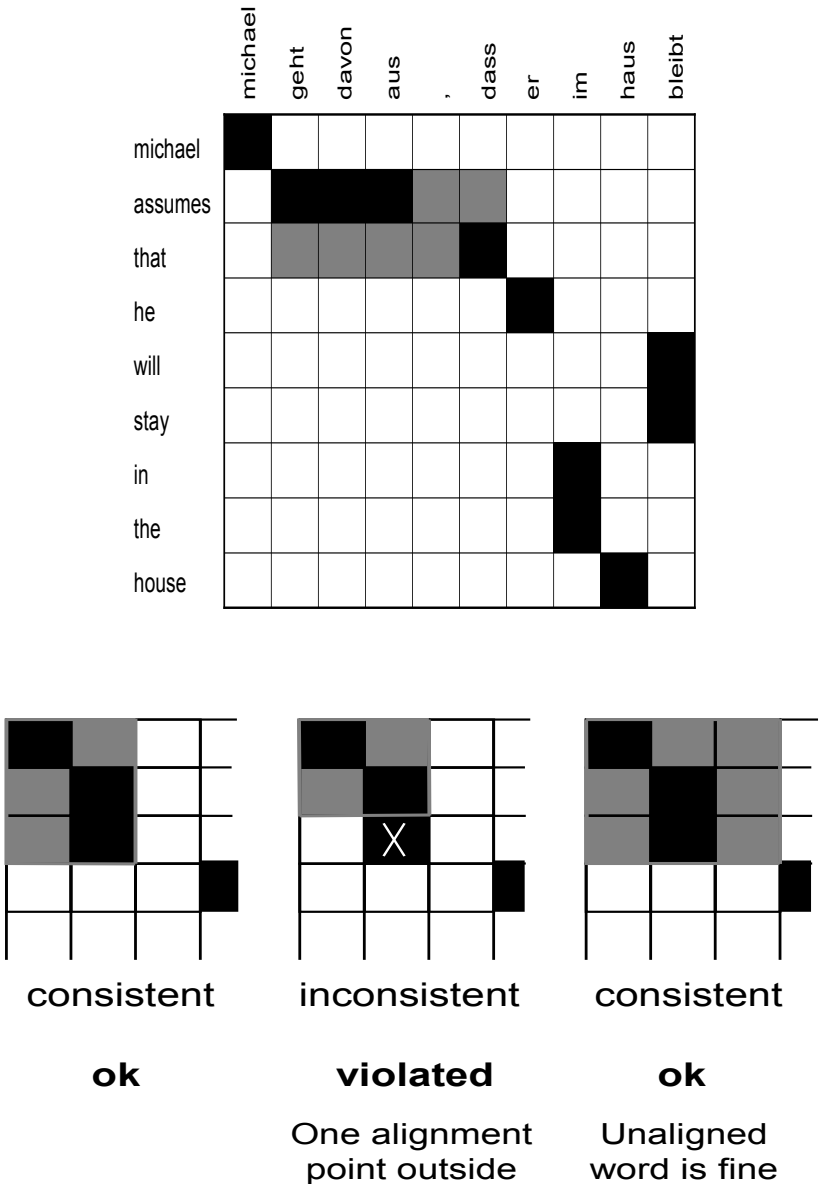

Figure 2.5: Extract phrase pair consistent with word alignment. Examples of consistent and inconsistent phrases. The grey part shows the probable phrases. Taken from slides of [Koehn, 2010].

Figure 2.4 shows an example where eight different phrases are extracted. Using alignment information, smaller phrases are extracted, and then using the criterion larger phrases are extracted, note that since $AB \rightarrow WY$ does not fulfill the criterion, it is not extracted as a phrase. Figure 2.5 shows another representation of word aligned sentences, with the black squares representing word alignments and the grey portions representing the possible phrases. According to the criterion *assumes that* $\Longleftrightarrow$ *geht davon aus , dass* is a consistent phrase to extract. We see in the graphical representation of the consistency criterion (bottom of Figure 2.5, middle figure with a cross) that since one alignment point is outside the suggested phrase pair, it defies the phrase extraction criterion, so this grey part is not a valid phrase, whereas the other two in the figure are.

The result of the phrase extraction process is pairs of source and target sentences which have consecutive words and are consistent with the word alignment matrix. These alignments are produced in both directions since alignment is asymmetric, the intersection (or other alignment methods) of these two alignments is then used. The word alignment can be produced using the IBM models (section 2.2).

A standard phrase-based SMT model is the product of three components, the phrase translation table $\phi(\bar{s_i}|\bar{t_i})$, the reordering or the distortion model d and the language model $\Pr(t)$, mathematically,

$$\arg\max_t \Pr(t|s) = \arg\max_t \prod_{i=1}^{I} \phi(\bar{s_i}|\bar{t_i}) d(start_i - end_{i-1} - 1) \prod_{i=1}^{|e|} \Pr(e_i|e_1...e_{i-1}) \quad (2.10)$$

Nowadays, many SMT systems follow a phrase-based approach, in that their translation unit is the bilingual phrase, such as [Bertoldi et al., 2006; Hewavitharana et al., 2005; Matusov et al., 2006] among many others. Most of these systems introduce a log-linear combination of models.

Among the most popular modern approaches to SMT include, *phrase-based models*, *factored translation model* [Koehn et al., 2007], *hierarchical approach* [Chiang, 2005, 2007; Wu, 1997], *N-gram based approach* [Casacuberta and Vidal, 2004; Casacuberta et al., 2002] and *syntax-based MT* [Och et al., 2003; Yamada and Knight, 2001].

2.4 The maximum entropy approach

The best estimated translation as given by equation 2.3 in the noisy channel approach (section 2.1) is optimal if the true probability distributions $\Pr(s|t)$ and $\Pr(t)$ are used. However, the available models and methods only provide poor approximations of the true probability distributions. In such case, a different combination of language and translation model might yield better results. Another limitation of the noisy channel approach is the difficulty to extend the baseline model to include features other than the language and translation model.

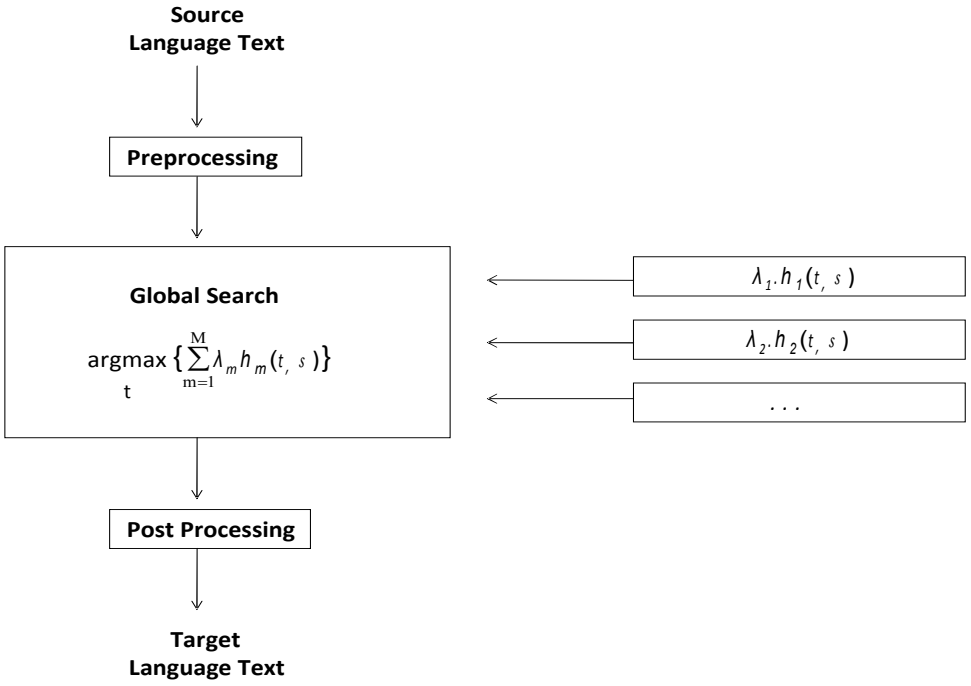

Figure 2.6: Architecture of the translation approach based on the Maximum Entropy framework.

The *maximum entropy framework* proposed in [K. Papineni and Ward, 1998] for the problem of natural language understanding has been successfully applied to the SMT task as shown in Och and Ney [2002]. Using this approach, the posterior probability $\Pr(t|s)$ is modeled as the log linear combination of the set of features and is considered

a generalization of the noisy channel paradigm. In this framework, we have a set of M feature functions h_m, and for each feature function there exists a model parameter (scaling factor) λ_m, with $m = 1...M$, Mathematically:

$$\begin{aligned} \hat{t} &= \arg\max_t Pr(t|s) && (2.11) \\ &= \arg\max_t \frac{exp(\sum_{m=1}^{M} \lambda_m h_m(t,s))}{\sum_{t'} exp(\sum_{m=1}^{M} \lambda_m h_m(t',s))} && (2.12) \\ &= \arg\max_t \sum_{m=1}^{M} \lambda_m h_m(t,s)\} && (2.13) \end{aligned}$$

The model coefficients are trained using the maximum class posterior criterion or with respect to the translation quality measured as error criterion as in Och [2003]. Figure 2.6 shows the architecture of a SMT system in the maximum entropy framework.

Note that the noisy channel approach is a special case of this framework. Namely, it is the case for which:

$$\begin{aligned} h_1(s,t) &= log\,\Pr(t) && (2.14) \\ h_2(s,t) &= log\,\Pr(s|t) && (2.15) \\ \lambda_1 &= \lambda_2 = 1 && (2.16) \end{aligned}$$

Och and Ney [2002] showed that using these feature functions (i.e., the same models as in source channel approach), but optimizing the feature weights λ_1 and λ_2, translation quality is significantly improved.

2.5 Language Modeling

Statistical language modeling has been used for a variety of natural language processing (NLP) applications including speech recognition [Chen and Goodman, 1996; Roark et al., 2007], part-of-speech tagging, syntactic parsing and Information retrieval [Song and Croft, 1999] etc. Most of the research in language modeling was performed by the large vocabulary speech recognition community, and the SMT community has mostly borrowed the models that were popularized for speech recognition. The main focus of SMT research has been on translation modeling [Lopez, 2008b], but it is well known that improvements in the language model generally lead to improved translation per-

formance [Brants et al., 2007].

Typically, so-called back-off n-gram models are used in SMT, however alternative language models exist and have shown to improve SMT quality, some of these include for instance continuos space LM based on neural networks [Schwenk, 2007; Schwenk et al., 2006], syntax based models based on context free grammars [Charniak et al., 2003; Marcu et al., 2006; Wu et al., 1998].

The language model tries to estimate the likelihood of the target language sentence, the more common the sentence is, the more probable that it is a good translation in terms of fluency since the source language is not taken into account. N-gram language modeling consists of dividing the sentences into fragments that are small enough to be frequent but large enough to contain some language information. Probability estimates of each fragment are calculated based on occurrence counts.

Formally the language model $\Pr(t)$ for a sentence with n words is, defined as the joint probability of the sequence of all the words in that sentence:

$$\Pr(t) = \Pr(w_1, w_2, ..., w_n) \tag{2.17}$$

Chain rule is then applied to decompose the joint probability into a set of conditional probabilities:

$$\Pr(t) = \Pr(w_1)\Pr(w_2|w_1)\Pr(w_3|w_1w_2)\Pr(w_4|w_1w_2w_3)...\Pr(w_n|w_1...w_{n-1}) \tag{2.18}$$

This is simplified by applying the Markov assumption that one can approximate the probability of a given word given its entire history by computing the probability of a word given the last $n-1$ words. Thus having $n = 2$ we get a bigram model, whereas $n = 3$ gives a trigram model and so on. For example, a trigram model would consider two previous words:

$$\Pr(t) = \Pr(w_1)\Pr(w_2|w_1)\prod_{i=3}^{n}\Pr(w_i|w_{i-1}w_{i-2}) \tag{2.19}$$

Generally, the larger n, the more information we get about the context of a specific sequence (larger discrimination). The smaller the n, the more cases will be seen in the training data, therefore better statistical estimates will be obtained (more reliability). In practice n varies between 3 and 5. The $\Pr(t)$ computation, that is, the probability of a word w given $n-1$ previous words is estimated using *Maximum Likelihood Estimation*

(MLE), for example for a bigram as:

$$\Pr(w_3|w_1w_2) = \frac{count(w_1w_2w_3)}{count(w_1w_2)} \tag{2.20}$$

Even with very large training corpora, it will always happen that one request a probability for an unseen sequence of n words. N-grams that do not occur would be assigned zero probability and will void an entire sentence probability. To avoid zero counts (and $\Pr(t) = 0$) probability estimates are usually smoothed using a smoothing algorithm. The simplest one is to add one to all the counts of n-grams, this is known as *add-one smoothing*, for example for bigrams:

$$\Pr(w_j|w_i) = \frac{count(w_iw_j) + 1}{count(w_i) + V} \tag{2.21}$$

where V is the vocabulary or number of types (all different words seen in the corpus)

The smoothing strategies include *interpolation* and *back-off* models. The common idea among all smoothing techniques is to take some of the probability "mass" from the known n-grams and redistribute it to the unseen ones. A good overview of $n-$gram smoothing techniques is presented in Chen and Goodman [1996].

2.5.1 Perplexity

The quality of a language model can be judged by its impact on an application, but it is very hard to measure and impossible to use for direct optimization. Thus, a common alternative is to estimate the quality of the LM independent of the application. The perplexity measure is related to the probability that the model assigns to the test data, and is defined as:

$$PP_p(T) = 2^{H_p(T)} \tag{2.22}$$

Here $H_p(T)$ is the cross-entropy of the language model on data T. Perplexity is a function of both the model and the text [Rosenfeld, 1997]. It can be interpreted as the branching factor of a language model, models with lower values of perplexity are better models. Perplexity can be roughly interpreted as the geometric mean of the branch out factor of the language: a language with perplexity X has roughly the same difficulty as another language in which every word can be followed by X different words with equal probabilities.

2.5.2 Training and decoding tools

Manual annotation of word alignments is an expensive and frustrating task. Recent popularity of statistical MT can be attributed to the availability of automatic training and decoding tools. In 1999, a freely available tool GIZA was released as part of the EGYPT toolkit. GIZA implemented IBM models to generate Viterbi algorithm [Al-Onaizan et al., 1999]. An improved version of the toolkit appeared in 2001 and 2003 by the name of GIZA++ [Och and Ney, 2003].

Wang and Waibel [1997] presents a stack decoder from the IBM model 2 based on the $A*$ search algorithm. Tillmann and Ney [2000] and Tillmann and Ney [2003] present Dynamic Programming based decoders for the IBM model 2 and model 4. Germann et al. [2001] compares the speed and quality of a stack based, a greedy and an integer programming decoder based on IBM model 4.

Dechelotte [2007] presents a translation system based on the IBM-4 decoder. Hoang and Koehn [2008] present a description of the open source Moses decoder which is often used in the SMT research community. The experiments reported in thesis have also been conducted using the Moses decoder.

2.6 MT Evaluation

'More has been written about MT evaluation over the past 50 years than about MT itself', a remark attributed to Yorick Wilks in [Hovy et al., 2002] summarizes the importance of the subject. The goal of MT evaluation is to judge the correctness of an SMT output. This judgement is made by ranking the *adequacy* of the translation in conveying the source language meaning and the *fluency* of expression in the target language [White, 1994]. The cost of human evaluation makes it infeasible for use in iterative system development, where regular evaluation is required to determine system performance. Also, human judgments by adequacy and fluency are not much reliable due to the high inter-judge variations and are almost not used any more. Nowadays, the focus is on doing system comparisons. This need has resulted in emergence of various evaluation metrics, however, so far the MT community has not accepted a unified evaluation criteria.

The automatic metrics make use of a set of test sentences for which we already have human translations, called *reference translations*. The intuition behind these metrics is that MT must be good if it resembles human translation of the same sentence [Papineni et al., 2002]. The metrics perform partial string match between the SMT output and

the reference translations. However, having a single reference translation may bias the evaluation towards a particular translation style, so use of multiple reference translations is generally preferred to take into account the diversity of translation styles.

One of the earliest evaluation metric, borrowed from ASR evaluation is the *word error rate (WER)* [Och et al., 1999a], also known as Levenshtein or edit distance. WER scores the sentences based on the number of insertions, deletions and substitutions required to transform the output sentence to the reference sentence. But since WER does not recognize word orderings, this makes it less appropriate for MT evaluation because a word that is translated correctly but is in the wrong location will be penalized as a deletion (in the output location) and an insertion (in the correct location). This problem motivated the use of *position independent word error rate (PER)*, which is similar to WER but does not penalize re orderings as it regards the output and the reference sentence as unordered bags of words rather than totally ordered strings [Och et al., 1999a].

Translation edit rate (TER) is also an often used evaluation metric which allows block movements of words and thus takes into account the reordering of words and phrases in translation [Snover et al., 2006]. It also measures the amount of editing that would have to be performed to change a hypothesis so that it exactly matches the reference. Specifically,

$$TER = \frac{\text{number of edits}}{\text{average number of reference words}} \tag{2.23}$$

The edit operations employed by TER include the classical insertion, deletion and substitution of single words as in WER, with an additional shift operation which permits shifts of word sequences. All edits, including shifts of any number of words, by any distance, have equal cost of 1.

Translation edit rate plus (TERp) is an extension of TER. It uses all the edit operations of TER, matches, insertions, deletions, substitutions and shifts as well as three new edit operations: stem matches, synonym matches and phrase substitutions. Unlike TER, the TERp implementation assigns a varying cost to substitution so that a lower cost is used if the two words are synonyms, share the same stem, or are paraphrases of each other. TERp identifies words in the hypothesis and reference that share the same stem using the Porter stemming algorithm. Two words are determined to be synonyms if they share the same synonym set according to Word Net. Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis

if that phrase pair occurs in the TERp paraphrase phrase table. With the exception of phrase substitutions, the edit operations have fixed cost regardless of the word in question [Snover et al., 2009].

The most widely used MT evaluation metric is BLEU, short for *bilingual evaluation under study* [Papineni et al., 2002]. The metric works by measuring the n-gram co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is a precision oriented metric as it considers the number of n-gram matches as a fraction of the total number of n-grams in the output sentence.

The NIST evaluation metric Doddington [2002], is based on the BLEU metric but with some alterations. BLEU simply calculates n-gram precision considering each n-gram of equal importance, NIST however calculates how informative a particular n-gram is, the rarer a correct n-gram, the more weight it is given. NIST score also differs in its calculation of the brevity penalty. Small variations in translation length do not impact the overall NIST score as much.

2.7 Information Retrieval

We used an IR toolkit for our research as described in the later chapters. We give here just an overview of the IR process, it is not intended to explain the whole field.

The meaning of the term *information retrieval* can be very broad. Getting the credit card out of the wallet to get the credit card number is also a form of information retrieval. Formally, as defined by Manning et al. [2009]:

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from with in large collections (usually stored on computers).

Information retrieval is fast becoming the dominant form of information access, rather in modern expression the word "search" has tended to replace the term information retrieval [Manning et al., 2009]. IR is normally termed the science of searching for documents, for information within documents, and for metadata about documents, as well as searching within the relational databases and the world wide web. Thus the IR systems operate at various scales ranging from providing search capabilities over billion of documents stored on millions of computers to personal information retrieval systems.

The IR systems give users the access to relevant information based on the information need that the user expresses as a query to the system. Generally speaking, the IR process begins when the system is asked some information, known as the *query*. Queries are formal statements of information needs and often follow a language template to define the information need. A query does not usually identify a single object/information set in the collection to search in, but rather several objects may match the query with varying degrees of relevance. The information/objects to search in are usually represented in the form of database information systems. The information in these databases, often called *indexes*, is stored in such a fashion such as to make query search fast and efficient. Often documents are themselves not stored in the indexes but are instead stored in the system by document surrogates or metadata. Most IR systems compute a numeric score to define how well each object matches the query and rank the objects according to this score, showing the top ranking results to the user.

Information about the terms in a collection of documents is indexed in such a way that they can be quickly accessed later using a term or a document as a reference. The documents are usually parsed into terms and added to the index (database) after considerations like, whether the word is important enough to add (stop words), whether to add the word as is or its stem form (so "stem", "stemming", and "stems" would all become the same term) and whether to recognize certain words as acronyms. English stop words, i.e. frequently used words, such as "a" or "the", are normally not indexed because they are so common that they are not useful to query on. Retrieval algorithms then use the collected information in the index for their scoring calculations and decide which documents to return for a given query.[1]

2.8 Steps in the IR Process

An IR system prepares for retrieval by indexing the documents and formulating queries resulting in document and query representations respectively. These representations are then matched and scored by the system and the top scoring results are returned (matching algorithm). The search process often goes through several iterations using the phenomenon of relevance feedback [Salton, 1971], which is a technique where new query keys are automatically extracted from relevant documents.

Retrieval models are roughly divided as exact match modes and best match models [Belkin and Croft, 1987]. Exact models return only the documents that match exactly some well-defined query, for e.g.. models based on booleans logic. Best match models

[1] http://sourceforge.net/apps/trac/lemur/wiki/TitleIndex

on the other hand can return documents based on partial match to the query too, such models include the vector space models and the probabilistic models.

2.8.1 Indexing: Creating Document Representations

Indexing (also called cataloging or metadata assignment) is the process of making statements about a document in accordance with the conceptual schema. This means preparing the raw document collection into an easily searchable representation of documents. This transformation of documents to indexed form involves the use of regular expressions, parsers, stop word list and miscellaneous filters. This is normally done in the following steps :

1. *Document Linearization:* is the process of reducing the document to a stream of terms. This is done by removal of markup and format information followed by tokenization.

2. *Filtration :* Involves deciding the most important terms for document representation. Frequently used terms or stop words are removed from text streams. A cost-effective approach involves removing all terms that appear commonly in the document collection and which will not improve relevant retrieval. This is often accomplished using a generic stop-word list.

3. *Stemming :* Stemming refers to the process of reducing terms to their stems or root variant. Thus, "computer", "computing", "compute" is reduced to "comput" and "walks", "walking" and "walker" is reduced to "walk". As can be seen, the root forms are not necessarily 'real words'. This fact is usually hidden from the user, because stemming only affects the internal representation of documents and queries [Talvensaari, 2008]. Not all systems use the same type of stemmer.

2.8.2 Query Formulation: Creating Query Representations

Retrieval is to predict the degree of relevance of a document with respect to a query description. The query description is normally transformed into a formal query representation that expresses the information need in terms of the system's conceptual schema. A query can be in the form of a "bag of words" by simply giving the features in an unstructured list or it can be a "structured query" combining various features using boolean operators.

For the research presented in this dissertation, we used the query language provided by the toolkit, which is modeled on the InQuery IR system. The InQuery IR system

is based on the network inference model of IR [Turtle and Croft, 1991]. In this model relevance is seen as the probability that a document satisfies an information need. It allows the use of various document and query representations to determine the existence of a belief. Consequently, this framework allows the use of a variety of IR models, e.g. the vector space model or boolean query system. The InQuery query language is very flexible and can be used for a wide range of query forms, from free-text querying to strictly structured queries with boolean and word proximity operators and anything in between.

2.8.3 IR Evaluation

For the purpose of evaluation, the chosen IR algorithm is applied to either document or query preprocessing, document-query matching or all of these, depending on the algorithm. It then returns an ordered list of responses called the retrieved set or ranked list. Various performance metrics which are based on *recall* and *precision* are used in evaluation.

Let R define the set of relevant documents for a test topic, and A the set of the documents retrieved for the topic by the selected IR algorithm, then *recall* is the set of relevant documents that have been retrieved, i.e.

$$Recall = \frac{\mid R \cap A \mid}{\mid R \mid} \tag{2.24}$$

While *precision* is the set of retrieved documents that are relevant, that is,

$$Precision = \frac{\mid R \cap A \mid}{\mid A \mid} \tag{2.25}$$

Retrieval performance varies widely from search to search along with the requirements for precision and recall which vary from query to query; thus making meaningful evaluation difficult. Standard practice is to compute a single measure of goodness that combines precision and recall. Some of these include van Rijsbergen's F-measure, averaging, mean average precision, reciprocal rank and error, etc.

Note that IR as used in this work is not equivalent to IR like google. Some available tools for IR include DryadLINQ toolkit, Lemur toolkit, PolyIRTK toolkit, Ivory toolkit and JAFER toolkit amongst many others. We have used the Lemur toolkit [Ogilvie and Callan, 2001] in our research.

2.9 Comparable corpora

Given the shortage of parallel corpora, research turned to explore various corpora which led to emergence of several terms to describe these corpora. The research first turned to noisy parallel corpus [Fung, 1995a; Fung and Mckeown, 1994], i.e., corpus composed of documents which fail to meet all the constraints of parallel corpora, that is documents having missing or misplaced translations. Subsequently, the research focused on non-parallel corpus before defining and adopting *comparable corpora* as a case study.

The concept of a comparable corpus and its use depend largely on the point of view of the experimenter and the subject of his research, it would be presumptuous to propose a universal definition. Comparable corpora are of various natures, covering a continuum between truly parallel and completely unrelated texts. Fung and Cheung [2004], for example, while giving a classification of various corpora (parallel, noisy parallel, comparable) classify their corpus as very-non-parallel since the documents in their corpus are composed of very disparate topics.

Bowker and Pearson [2002] define comparable corpora as consisting of set of texts in different languages that are not translations of each other. [McEnery and Xiao., 2007] however gives a more concise definition by defining comparable corpus as a corpus containing components that are collected using the same sampling frame and similar balance and representatives, e.g., the same proportions of the texts of the same genres in the same domains in a range of different languages. The sub corpora of a comparable corpus are not translations of each other. Rather, their comparability lies in their same sampling frame and similar balance.

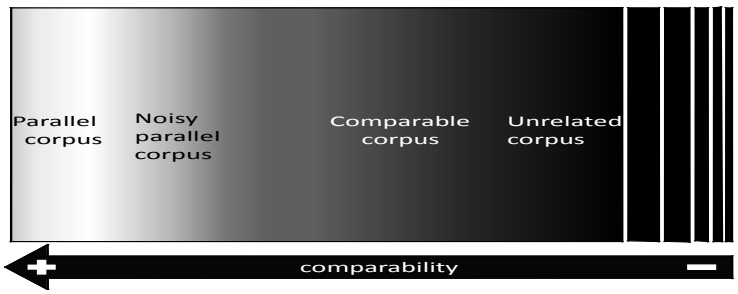

Figure 2.7: Comparability among multilingual corpora (taken from Prochasson [2009]).

This definition leads to the notion of the degree of comparability of document subparts of a comparable corpus. This degree of comparability is relative to the amount

of features (qualitative and quantitative) shared among the documents in the corpus. Figure 2.7 schematizes the notion of comparability with respect to the definition of corpus. This figure shows that parallel documents are fully comparable, they have, by definition, everything in common except the writing language. In contrast, the unrelated corpora are poorly comparable.

Thus, a comparable corpus is a collection of texts composed independently in the respective languages and combined on the basis of similarity of content. These are documents in one to many languages, that are comparable in content and form in various degrees and dimensions. Whereas a parallel corpus, also called bitext, consists in bilingual/multilingual texts aligned at the sentence level. An example of comparable documents is shown in figure 2.8, these are documents from the AFP English and French news reports. We see that the content of the two documents is slightly different but overlapping. The lines show the sentences which could be potential parallel sentences, the sentences which our approach aims to find.

Typically, comparable corpora don't have any information regarding document pair similarity. Generally, there exist many documents in one language which don't have any corresponding document in the other language. Also, when the correspondence information among the documents is available, the documents in question are not literal translations of each other. Thus, extracting parallel data from such corpora requires special algorithms designed for the corpora in question.

Potential sources of comparable corpora are multilingual news reporting agencies like AFP, Xinhua, Al-Jazeera, BBC etc, or multilingual encyclopedias like Wikipedia [Bharadwaj and Varma, 2011; Otero and Lopez, 2010; Smith et al., 2010], Encarta , etc. Some of these comparable corpora are widely available from LDC,[1] in particular the Gigaword corpora: or over the web for many languages and domains, e.g., Wikipedia. Comparable texts can also be found in large quantities over the world wide web, which is a continually growing resource [Fung et al., 2010; Hong et al., 2010; Ishisaka et al., 2009; Kilgarriff and Grefenstette, 2003; Nie et al., 1999; Resnik and Smith, 2003; Zhang et al., 2006].

This dissertation reports the results using the comparable news corpora. The approach presented in this thesis is independent of the type of comparable corpus and is applicable to all types of comparable corpus, with modifications (in indexing and querying scheme), with respect to the corpus at hand.

[1] http://www.ldc.upenn.edu/

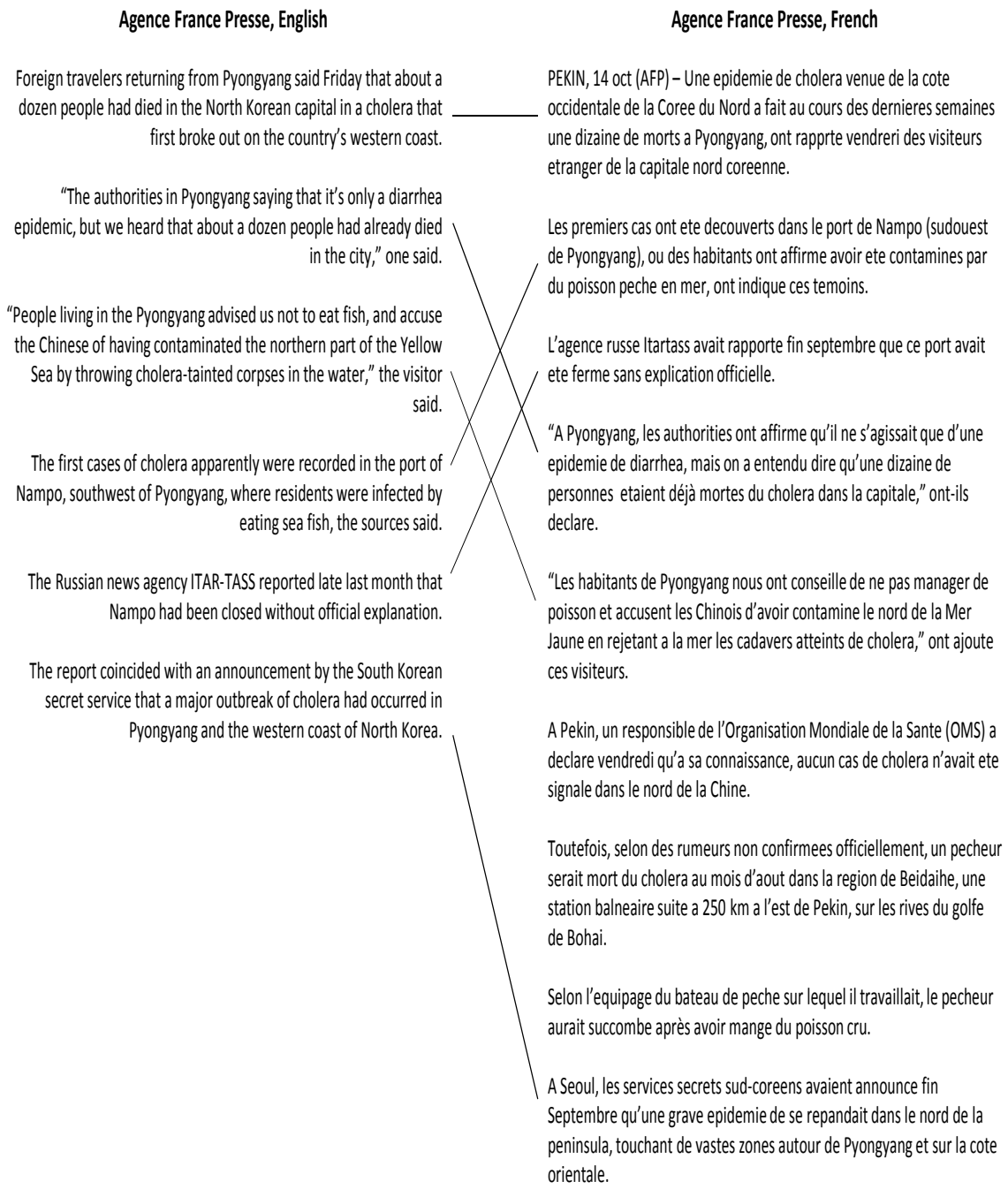

Figure 2.8: Example of two comparable documents from the AFP English and French corpora having many matching sentences. Note that this is not always the case. (Figure taken from Munteanu [2006]).

2.9.1 Classification of Comparable News Corpora

The work presented in this thesis uses comparable corpora in the news domain, so we present a brief classification of such corpora. Different comparable news corpora exhibit different levels of parallelism. At the parallel end of the scale are the corpora for which documents in one language are fully translated in the other language. An example of such corpora are the news reports produced by the "Le Monde Diplomatique", which reports in 17 languages, and some of these articles are translated in several languages, while others are region specific and only exist in the language of that region [Munteanu, 2006]. Getting parallel data from such corpora is easy. All that needs to be done is identification of parallel document pairs and then sentence alignment algorithms would suffice to give the parallel sentences.

Then there are corpora which have some documents translated, some not fully translated but still related and thus sharing many parallel sentences, and others not translated at all. Example of such corpora are the news feeds produced by agencies like Xinhua News and Agence France Presse. Such corpora contain parallel sentences at sentence level in the documents and need special treatment for their extraction: these are the sentences that our approach extracts.

Then there are news corpora which exhibit little parallelism, both at document and sentence level. Munteanu [2006] report news articles produced by the BBC to be an example of such corpora. Normally this is the case for the news agencies where rather than translating an article from one language to the other, independent reporting is done in each language, thus the chances of finding full sentences translations of each other decreases. However, since there are often news articles reporting the same event in different languages, such corpora are the ideal candidates for extracting parallel parts of sentences or sub-sentential segments as has been done by [Cettolo et al., 2010; Munteanu and Marcu, 2006; Quirk et al., 2007; Shinyama and Sekine, 2003; Wang and Callison-Burch, 2011].

2.9.2 Research using Comparable Corpora

The ease of availability of these comparable corpora and the potential for parallel corpus as well as dictionary creation has sparked an interest in trying to make maximum use of these comparable resources. Categorizing broadly, these include learning lexicons and new word translations, finding parallel sentence fragments and full parallel sentences.

The approaches to learn translations of unknown words are generally based on the assumption that translationally equivalent words appear in similar contexts. Thus the

general approach mainly followed is: compute the source word context (from monolingual source corpus) and transfer it in the space of the target language. Then compare with the context of all the words in target language (computed from target language monolingual corpus) and choose the closest matching target word based on context proximity. The various research efforts differ mainly on the methods for computing contexts and their similarities. Some of these research efforts in dictionary learning and identifying word translations include [Andrade et al., 2011; Chiao and Zweigenbaum, 2002; Diab and Finch, 2000; Fung, 1995b; Fung and Yee, 1998; Gaussier et al., 2004; Koehn and Knight, 2002; Morin and Prochasson, 2011; Pekar et al., 2006; Peter and Lynne, 2000; Rapp, 1995; Sadat et al., 2003; Shao and Ng, 2004; Sharoff et al., 2006; Xabier et al., 2008].

The temporal structure of comparable corpora has been used to find translation /transliteration of named entities (i.e., names of people and locations). The intuition being that the named entities that co-occur often in documents from same time periods are more likely to be mutual translations. Some these works include [Alegria et al., 2006; Huang et al., 2005; Ji, 2009; Klementiev and Roth, 2006; Sproat Richard and Zhai, 2006; Udupa et al., 2009].

Some of the other works include extracting phrasal alignments [Kumano et al., 2007], word sense disambiguation [Kaji, 2003], acquiring synonyms [Shimohata and Sumita, 2005], parallel fragment extraction [Cettolo et al., 2010; Munteanu and Marcu, 2006; Quirk et al., 2007; Shinyama and Sekine, 2003; Wang and Callison-Burch, 2011], extracting lay paraphrases of specialized expressions [Deléger and Zweigenbaum, 2009], language and translation model adaptation [Snover et al., 2008], improving SMT performance using extracted parallel sentences [Abdul-Rauf and Schwenk, 2009a,b; B.Lu et al., 2010; Do et al., 2010; Gahbiche-Braham et al., 2011; Munteanu and Marcu, 2005].

2.9.3 Finding parallel sentences

The research most relevant to the work presented in this thesis is that focused on finding parallel sentences from comparable corpora. This problem has been approached in two ways: (1) by extending the sentence alignment algorithms (finding parallel documents and then sentence align them) [Utiyama and Isahara, 2003; Zhao and Vogel, 2002] and (2) by designing approaches specific to comparable corpora which identify similar document pairs and find all the possible sentence pairs [Abdul-Rauf and Schwenk, 2009b; Do et al., 2009; Fung and Cheung, 2004; Gahbiche-Braham et al., 2011; Munteanu and Marcu, 2005; Wu and Fung, 2005]. Our approach falls in the later category, i.e., we

devise a method to find parallel sentence pairs based on the peculiarities of comparable data.

We report here the previous works which use the web as a source of semi parallel sentences or using other sources of comparable corpora. We present the two sources of data separately since web data has features like URL addresses and HTML structures which help in analyzing the document pairs, whereas our approach focuses on comparable data without the need for additional structure information.

2.9.3.1 Finding parallel sentences from the web corpora

Nie et al. [1999] report *PTMiner* to mine parallel corpora from the web using URL pattern matching and several other criteria like HTML structure, file length etc. They report promising results from a collection of English-Chinese texts.

Resnik and Smith [2003] use their STRAND based structural filtering system which filters candidate parallel pairs by determining a set of pair-specific structural values from the underlying HTML page. They report a precision of 98% and a recall of 61% on their developed English-Chinese parallel corpus.

Zhang et al. [2006] use a multiple feature parallel text identifier via a k-nearest neighbor classifier to identify Chinese-English parallel pairs from the internet. They report 95% precision and 97% recall rate.

Ishisaka et al. [2009] report developing a Japanese-English parallel corpus by collecting open source software manuals from the web. They report usefulness of the corpus by conducting SMT experiments.

Fung et al. [2010] focus on the web as a resource for extracting potential parallel sentences by crawling comparable and parallel web sites. They propose a sentence extraction architecture inspired by various earlier works. Interesting results are reported using French and English Wikipedia articles.

Hong et al. [2010] present a method to directly search sentence pairs from the web. They propose a method which discovers document pairs from the web by using ranked query formulation using cue words from the source document. Document filtering is then done using the IBM model 1 alignment.

2.9.3.2 Finding parallel sentences from comparable corpora

Some of the works aimed at discovering parallel sentences from comparable corpora include [Masuichi et al., 2000] in which they report a method to extract bilingual text pairs from pseudo comparable corpora that they create. They apply a bootstrapping

approach to an existing cross-language information retrieval method based on the Information Mapping approach.

Zhao and Vogel [2002] use a generative method to find parallel sentences from the Xinhua comparable news corpus. They use sentence and lexicon-based methods combined under a maximum likelihood criterion. They report improvements in word alignments using their found parallel sentences.

Utiyama and Isahara [2003] use cross-language information retrieval techniques and dynamic programming to extract sentences from an English-Japanese comparable news corpus. They identify similar article pairs, and then, treating these pairs as parallel texts, align their sentences on a sentence pair similarity score and use dynamic programming (DP) to find the least-cost alignment over the document pair. Yang and Li [2003] use an approach based on DP to identify potential parallel sentences in title pairs in an English-Chinese comparable corpus. Longest common sub sequence, edit operations and match-based score functions are subsequently used to determine confidence scores.

Fung and Cheung [2004] approach the problem by using a cosine similarity measure to match Chinese and English documents. They work on "very non-parallel corpora". They generate all possible sentence pairs and select the best ones based on a threshold on cosine similarity scores. Using the extracted sentences they learn a dictionary and iterate over with more sentence pairs, improving performance with bootstrapping. In a different approach to this problem, Wu and Fung [2005] use inversion transduction grammars (ITG) along with cross language information retrieval techniques to find parallel sentences from 'very nonparallel quasi-comparable' corpora. They work on Chinese and English Topic Detection and Tracking (TDT) data reporting an average precision of 65%.

Munteanu and Marcu [2005] use a bilingual lexicon to translate each of the words of the source sentence. These translations are then used to query the database to find matching translations using information retrieval (IR) techniques. Candidate sentences are determined based on word overlap and the decision whether a sentence pair is parallel or not is performed by a maximum entropy (ME) classifier trained on parallel sentences. Bootstrapping is used and the size of the learned bilingual dictionary is increased over iterations to get better results. Following the same ideology, an unsupervised learning approach to the problem is proposed by Do et al. [2010]. They use a comparable/noisy parallel corpus to train an initial SMT system and then use this system to translate another comparable corpus to get parallel sentence pairs. The process is iterated by adding the extracted pairs to the training data and the quality

of the SMT system is improved. They experiment with TER, NIST, BLEU and PER for scoring their sentences, the PER filter was found best suited for their approach. They term their approach as unsupervised learning. They applied this unsupervised approach successfully to an under resourced language pair, French-Vietnamese.

Uszkoreit et al. [2010] mine parallel sentences from web pages and digitized books. They first translate the documents into one single language and then approach the problem as cross language near duplicate detection using only textual content of the documents (using n-gram information). Their parallel corpora improved the quality of SMT systems.

Gahbiche-Braham et al. [2011] following similar methods as our work report SMT improvements with their parallel sentences extracted from noisy parallel corpus. They use the hunalign sentence alignment tool to align the sentences of the matching document pairs. They use the extracted and translated sentences for translation model adaptation achieving significant improvements.

Some ongoing works in progress focusing on parallel corpora production to improve SMT systems include [Eisele and Xu, 2010] who work in the framework of ACCURAT project, for which the objective is to analyze and evaluate novel methods of exploiting comparable corpora, including evaluation of some already proposed methods. The intention is to provide researchers with re-implemented versions of various baseline methods. They focus their research on eighteen under-resourced European language pairs.

Chapter 3

Scheme for Parallel Sentence Generation from Comparable Corpora

3.1 Overview

This chapter presents our research work aimed at developing an efficient solution to parallel sentence mining from comparable corpora. Preparing human translated parallel corpora is not an easy task, thus research turned towards exploring the amply available comparable corpora (section 2.9.3). The work reported in this chapter is a contribution in this regard. The research question under observation was:

Improving SMT performance by efficient selection of parallel sentences in comparable corpora.

Our approach is inspired by the work of Munteanu and Marcu [2005]. They devised an algorithm to find parallel sentences from comparable corpora. They use a probabilistic bilingual lexicon to translate the words of the foreign document and use them as query to search similar documents from the English corpus. IR is used for selecting similar documents pairs. All possible sentence pairs are then generated from these document pairs and the candidate sentence pairs are selected using word overlap filter. They devise a maximum entropy classifier which judges the candidate sentences to be parallel or not and assigns a similarity score. They were able to extract good quality parallel sentences, which when used as additional bitexts by the SMT systems

resulted in nice improvements in SMT scores.

Our approach is similar to [Munteanu and Marcu, 2005] in that we also use IR, but our IR framework returns the set of potential parallel sentences which we simply score for similarity using mostly TER or TERp. We shift some effort to the query building stage, where we use proper SMT translations to emphasize both precision and recall in query creation.

In this chapter we present a detailed description of our proposed approach. We have worked on Arabic to English and French to English systems. However, please note that our approach is independent of the source-target language. Since for this research, the target language was English for both the systems, so we present the approach by referring to *source* language corpus as *foreign* language corpus and *target* language corpus as *English* corpus.

We start by describing the evaluation methodology (section 3.2) and the experimental resources used during our study (section 3.3). In section 3.4, we describe our parallel sentence extraction approach in detail. Section 3.5 presents the results for decision of the best filter suited to the task. Sections 3.6 and 3.8 give the theory and experimental evidence for our *sentence tail removal* and *dictionary creation* methods. In section 3.11, we present an alternative sentence selection experiment and the impact of SMT quality on our method in section 3.12. We conclude the chapter by presenting a theoretical and empirical comparison of our method with [Munteanu and Marcu, 2005].

3.2 Evaluation Methodology

We evaluate the data extracted by our approach by measuring its impact on the performance of SMT systems. Essentially, we use the extracted data as additional training material, and verify whether this leads to better performance.

The baseline systems for our evaluation are built using the already available human translated parallel corpus. After extracting additional data from the comparable corpus, we then train a comparative MT system on both the initial (baseline) and extracted corpora. The quality of the extracted data is measured by its impact on MT performance, i.e. by the difference in the performances of the baseline and comparative MT systems.

3.3 Experimental Resources

3.3.1 Comparable corpora

The Linguistic Data Consortium (LDC) provides large collections of monolingual data, namely the LDC Gigaword corpora. Those collections include texts from multilingual news reporting agencies. We identified agencies that provided news feeds for the languages of our interest. There are two text sources that do exist in Arabic[1] and English: the AFP and XIN collection. We chose Agence France Press (AFP) for our study on French-English.[2] Table 3.1 summarizes the characteristics of these corpora. The number of words are given after tokenization.

Source	**Arabic**	**French**	**English**
AFP	212M	333M	527M
XIN	80M	-	140M

Table 3.1: Characteristics of the Gigaword corpora used for the task (number of words).

Note that the English parts are much larger than the Arabic and French parts. Since these are also news resources, the likelihood of finding sentences that are translations of each other is high, and we aim to find those. From these collections, for each language, we create a comparable corpus by putting together articles coming from the same agency and the same time period.

3.3.2 Resources used for SMT Systems

3.3.2.1 Arabic to English

For this research we consider the translation from Arabic into English, under the same conditions as the official NIST 2008 evaluation. The used bitexts include various news wire translations[3] as well as some texts from the GALE project.[4] We also added the 2002 to 2005 test data to the parallel training data (using all reference translations). This corresponds to a total of about 5.8M Arabic words. Our baseline system is trained on these bitexts only.

[1] LDC corpus LDC2007T40 (Arabic)
[2] LDC corpora LDC2007T07 (English) and LDC2006T17 (French).
[3] LDC2003T07, 2004E72, T17, T18, 2005E46 and 2006E25.
[4] LDC2005E83, 2006E24, E34, E85 and E92.

We use the 2006 NIST eval data as development data and the official NIST 2008 eval data as internal test set. All case sensitive BLEU scores are calculated with the NIST scoring tool with respect to four reference translations. Both data sets include texts from news wires as well as newsgroups.

3.3.2.2 French to English

The translation model was trained using the news-commentary corpus (1.56M words),[1] and a bilingual dictionary of about 500k entries.[2] This system uses only a limited amount of human-translated parallel texts, in comparison to the bitexts that are available in NIST evaluations. In a different version of this system (section 3.12), the Europarl (40M words) and the Canadian Hansard corpus (72M words) were added.

In the framework of the EuroMatrix project, a test set of general news data was provided for the shared translation task of the third workshop on SMT [Callison-Burch et al., 2008], called *newstest2008* in the following. The size of this corpus amounts to 2051 lines and about 44 thousand words. This data was randomly split into two parts for development and testing. Note that only one reference translation is available. We also noticed several spelling errors in the French source texts, mainly missing accents. Those were mostly automatically corrected using the Linux spell checker. This increased the BLEU score by about 1 BLEU point in comparison to the results reported in the official evaluation [Callison-Burch et al., 2008]. The system tuned on this development data is used to translate large amounts of text of the French Gigaword corpus.

3.4 Proposed Approach

3.4.1 Introduction

The comparable corpora that we work on are loosely independent monoligual collections of documents (section 3.3.1). The parallel sentences that we seek to find could be located anywhere in these collections, in any document and in any two sentences from them. The framework that is taken up in this research is designed to restrict this potentially huge space and seek reasonably sized search space which is likely to contain good data [Munteanu and Marcu, 2005].

The high level architecture of our proposed approach is shown in figure 3.1. We start by translating the source language part of the comparable corpus to the target

[1] Available at http://www.statmt.org/wmt08/shared-task.html

[2] The different conjugations of a verb and the singular and plural form of adjectives and nouns are counted as multiple entries.

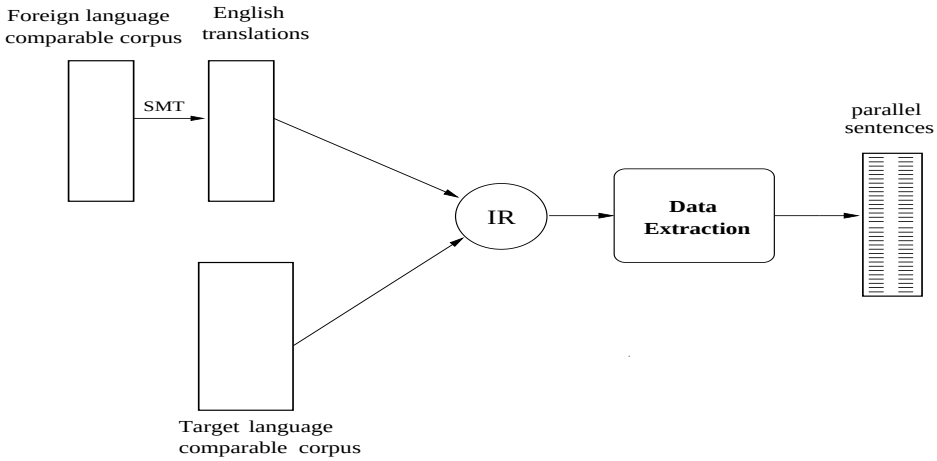

Figure 3.1: High level Architecture of the parallel sentence extraction system.

language using a SMT system (section 3.4.2). These translated texts are then used to perform information retrieval from the target language side of the comparable corpus (section 3.4.3), thus obtaining candidate sentence pairs. Parallel sentences are then filtered out by comparison with the automatic translations. This is done using simple filters like WER, TER and TERp (section 3.4.4).

We have evidence that the quality of the SMT system does not seem to affect the information retrieval process (section 3.12). Thus, a fairly simple SMT system built on small amounts of bitexts can be used to extract good parallel sentences from a comparable corpus. The resources required by our system are minimal : bitexts and monolingual data in the target language to train a standard SMT system. In the following sections we will explain each component of our system as depicted in figure 3.1.

3.4.2 Translating the foreign language corpus

This is the first step of our approach, i.e. devising precise queries of the foreign (source) language corpus. These are then used to find the potential matching sentences in the target language (English). The SMT systems used in our experiments are based on the Moses SMT toolkit [Koehn et al., 2007]. The corpora used for Arabic-English and French-English systems are detailed in section 3.3.2. To build these systems, first Giza++ is used to perform word alignments in both directions. Second, phrases and lexical re-orderings are extracted using the default settings of the Moses SMT toolkit. The 4-gram back-off target LM is trained on the English part of the bitexts and the

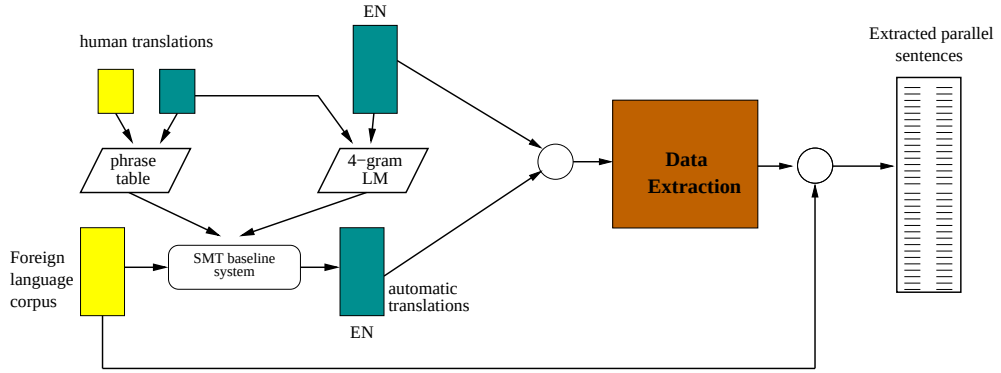

Figure 3.2: Detailed framework of the SMT system used for translations.

Gigaword corpus of about 3.2 billion words. Therefore, it is likely that the target language model includes at least some of the translations of the foreign language corpus. This can help to obtain good quality translations. Figure 3.2 shows the graphical representation of this process and its integration in the complete system.

A strong feature of our approach is use of proper SMT translations as queries to perform IR in the English side of the comparable corpus. Doing so we are able to emphasize both precision and recall while formulating queries for IR. Though the IR process treats the queries as bag of words, use of proper sentences has the same advantage as the phrase based systems have over word based systems (i.e. translation for each query word is with respect to the context of the sentence).

3.4.3 Finding the best matching sentence

Once we have the English translations of the foreign language corpus, we have the two text collections in the same language. Now, if for most of these translations, we are able to find the matching sentences from the English corpus, we have a potential parallel corpus. However, given the huge size of the corpora at hand, we need to design this search framework in a fashion that it is accurate as well as efficient. We use information retrieval for this and model our search space using the temporal information of these corpora to achieve best results.

We presented a rather detailed general overview of information retrieval in section 2.7. Here, we give a brief overview of the process for the sake of clarity. Information retrieval as defined by Manning et al. [2009] is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from with in large collections (usually stored on computers). IR is normally termed the science of

searching for documents, for information within documents, and for metadata about documents, as well as searching within the relational databases and the world wide web. Likewise, for us the aim of the IR process is to be able to find the matching sentence from the English side of the comparable corpus, if it exists, or in the other case, the nearest matching sentence. The subsequent filters of our system are robust enough to score the sentences based on similarity and filter out good sentences.

3.4.3.1 Proposed IR scheme

We used the Lemur IR toolkit [Ogilvie and Callan, 2001] for our sentence extraction procedure. The toolkit has special defined formats for the data to be indexed (it accepts html, trecweb, trectext, trecalt, doc, ppt, pdf and txt formats) and also a well defined query language. The usual format is to use parameter files to define the various parameters used while indexing and querying. It allows two types of indexing formats, the *Key File Index* and the *Indri Index*. These two types differ in the form they store and retrieve data from the disk. We chose the Indri indexing scheme as it provides some extra capabilities like storing field and annotation data that can be searched on. This feature enabled us to index our documents in such a way that using the specialized *Indri Query Language* we can ask the index to tell us the best matching document as well as to return the best matching sentence from that document,[1] thus getting the results as a ranked list of sentences. By these means we can retrieve the best matching sentences from the English side of the comparable corpus.

With our scheme (figure 3.3), we are able to emphasize both precision and recall in the sentence selection process by IR contrary to Munteanu and Marcu [2005] where they emphasize only recall in their article selection step. According to them their document matching procedure is imprecise due to noise in dictionary (the query will contain many wrong words). Whereas our use of proper SMT translations lets us create precise queries. Using an isolated translation of each word as query compromises precision, whereas full sentence queries provide better precision as obviously phrase-based is better than word-based translation. To emphasize recall, for each sentence, we focus not just to retrieve the best matching sentence but rather a set of the best matching sentences from each closest matching document. We limit the IR results to 5 results per sentence, based on our experience of getting the best sentences in the top of the list i.e. if good matches exist. The information retrieval framework is on its own recall oriented as it always returns a sentence. Thus for each of our query sentence we

[1] http://www.lemurproject.org/lemur/indexing.php

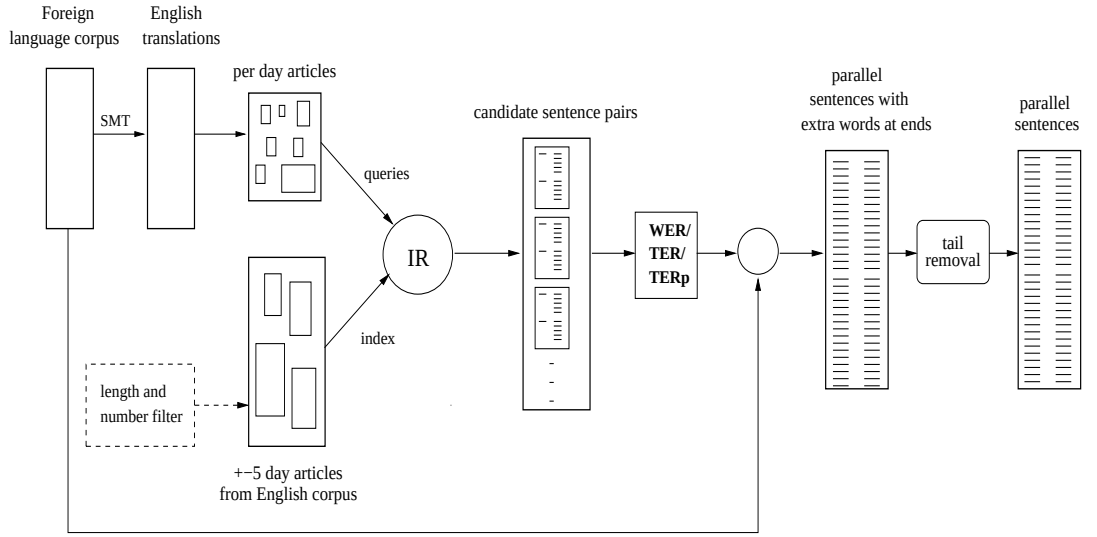

Figure 3.3: Detailed architecture of the parallel sentence extraction system. The source language side of the comparable corpus is translated into the target language (English in our case).

always get 5 sentences, as judged best matches by the IR framework. Our subsequent filters do the job of filtering out the good parallel sentences from the bad ones.

Our sentence extraction process is based on a simple heuristic that, considering the corpus at hand i.e. news corpus, we can safely assume that a news item reported on day X in the foreign language corpus will be most probably found in the day X-5 and X+5 time period in the English corpus. We experimented with various time periods for the extraction process and found a window size of ± 5 days to be the most efficient in terms of both time and the quality of the retrieved sentences. This is also inline with the windowing scheme used by Munteanu and Marcu [2005]. We cleaned the corpora by removing the tables, stock results and sports results etc. These were mostly very long sentences comprising too many numbers. These are of no use as SMT training data.

We first collect all sentences from the SMT translations corresponding to the same day (we call them query sentences) and then the corresponding articles from the English Gigaword corpus in a ± 5 day window (search space for IR). These are made using the date and ID information for each sentence of both corpora. These day-specific files are then used for information retrieval using an IR toolkit. The top 5 scoring sentences are returned by this IR process. At the end of this step, we have 5 potentially matching sentences for each of our query sentence.

SMT Query: **20060623:** Even in Moscow where the situation is incomparably better the children of infected mothers are pariahs.
IR Results: **1) 20060628:** Even in Moscow, where the situation is incomparably better, children of infected mothers are often treated as pariahs **2) 20060626:** For the 2002 climax in Yokohama, it was an easy decision to make, but with the incomparable Pierluigi Collina now retired, FIFA have a problem **3) 20060619:** Ties between the two powerhouses have expanded after apartheid South Africa shunned diplomatic relations with China, allying instead with Taiwan, another international pariah during that epoch **4) 20060623:** He may be viewed in the West as a pariah figure, but Iran's hardline President Mahmoud Ahmadinejad continues to enjoy a strong stature at home, analysts say **5) 20060621:** Syed Hamid said the action demonstrated that Myanmar – which has become an international pariah for its reluctance to abandon military rule and improve its human rights record – does not want ASEAN to play a bridging role

Table 3.2: IR results for a query sentence dated 20060623 from French AFP. The first match found by the IR process is the best match, even though it is located at a distance of 5 days. Note that the $4rth$ result found by IR is from the same day as the query sentence, but it is not relevant.

The information retrieval step is the most time consuming task in the whole system. The time taken depends upon various factors like size of the index to search in (search space), length of the query sentence etc. To have a time estimate, using a ± 5 day window typically required 9 seconds per query vs 15 seconds per query when a ± 7 day window was used. The number of results retrieved per sentence also had an impact on retrieval time with 20 results taking 19 seconds per query, whereas 5 results taking 9 seconds per query. Query length also affected the speed of the sentence extraction process. In the scenario at hand, each word (nouns, verbs and even dates and numbers) other than the stop words, could be a potential keyword, so no explicit efforts were made to reduce query length.

We show some samples of the results obtained in this step in tables 3.2, 3.3 and 3.4. Table 3.2 shows the case where the IR returns a correct match as the first sentence.

SMT Query: **20071115:** The cable channel also plans to expand its coverage and its production units in the United Arab Emirates.
IR Results: **1) 20071116:** While most Americans would not have watched the debate on cable channel CNN , its impact could be important in dictating wider coverage , in local media markets in key states , of the accelarating Democratic race **2) 20071115:** President Pervez Musharraf 's government took the channels off cable as part of strict curbs on the media , leaving millions of Pakistanis starved of news about the political crisis gripping their country **3) 20071120:** The final award of the night went to Gore , honoring both his more recent work as an environmental campaigner and his role in launching cable and satellite channel Current TV , which relies mainly on user-generated content **4) 20071117:** She said it was necessary to explain to the two channels the policy of the United Arab Emirates , of which Dubai is part , and guidelines applying to media operating out of the free zone **5) 20071117:** She said it was necessary to explain to the two channels the policy of the United Arab Emirates , of which Dubai is part , and guidelines applying to media operating out of the free zone , but she declined to go into details of the criteria they would have to abide byo in order to resume broadcasting

Table 3.3: An example of non-relevant IR results obtained for a query sentence. Query sentence dated 20071115 taken from Arabic AFP.

Note that this was the only matching sentence available in the search space as the other four sentences do not match. Had their existed other matching sentences, the IR process would have returned them in the ranked list as in the case for the sentence shown in table 3.4. Here, since the exact same sentence has been reported on multiple days, the IR process returns the top 5 occurrences. Table 3.3 shows a typical sentence for which the best match found has some matching words only and is clearly not a matching sentence. Sentences like this, which occur pretty often, would be assigned a very low score in the following step and will eventually be discarded by the process.

We experimented with two schemes to use these sentences returned by the IR framework. One was to simply take the first sentence as the best sentence (it is already the best sentence according to IR score) and use it as a potential matching sentence. The

SMT Query: **20060701:** According to Palestinian statistics 126 women and 300 minors aged under 18 years are held in Israeli prisons among some 9 400 Palestinian prisoners.
IR Results: **1) 20060626:** According to Palestinian statistics, 126 women and 300 minors under the age of 18 are being detained in Israel out of a total of some 9,400 Palestinians in Israeli prisons **2) 20060626:** According to Palestinian statistics, 126 women and 300 minors under the age of 18 are being detained in Israel out of a total of some 9,400 Palestinians in Israeli prisons **3) 20060628:** According to Palestinian statistics, 126 women and 300 minors under the age of 18 are being detained in Israel out of a total of some 9,400 Palestinians in Israeli prisons **4) 20060630:** According to Palestinian statistics, 126 women and 300 minors under the age of 18 are being detained in Israel out of a total of some 9,400 Palestinians in Israeli prisons **5) 20060627:** According to Palestinian statistics, 126 women and 300 minors under the age of 18 are being detained in Israel out of a total of some 9,400 Palestinians

Table 3.4: IR results for a query sentence dated 20060701 from French AFP. The best match is the first sentence (5 days apart from the query sentence), however, in this case this sentence was reported exactly as it is, on multiple dates, so the toolkit found the best match in all top 5 results.

other scheme was to compute TER between the query sentence and all 5 result sentences and then use the sentence having the lowest TER score as a potential matching sentence. A comparison of the two schemes is presented in section 3.11. Since there was no apparent advantage of one scheme over the other, for the experiments presented in this dissertation, we used the first sentence as per IR score to be the best sentence. Using more than one sentence per query could be a potential option too (though this amounts to duplicating the source sentence, but parallel corpora often contain multiple translations). We did not do much experimentation with this option so it can be considered a potential future work.

SMT Query : "Democracy cannot be imposed from above. That is a contradiction in terms," she said. **IR Result :** "Democracy cannot be imposed from above. That is a contradiction in terms," she said.		
WER	TER	TERp
0	0	0
SMT Query : " They are 14 over seven hospitals in the region , " said Christian Lahccen , head of Air France Canada , at a press conference . **IR Result :** " There are 14 spread over seven hospitals in the region , " Christian Lahccen , head of Air France Canada , said in a news conference .		
WER	TER	TERp
22.22	17.86	10.80
SMT Query : The Organization of Petroleum Exporting Countries (Opep) held on April 24 in Vienna an extraordinary ministerial meeting during which it will consider a possible reduction of its production of crude , said Tuesday , a source close to the Opep . **IR Result :** The Organisation of Petroleum Exporting Countries (OPEC) is to hold an extraordinary ministerial meeting here on April 24 to consider a possible reduction of its oil production , a source close to OPEC said here on Tuesday .		
WER	TER	TERp
61.54	47.50	43.80

Table 3.5: Some example sentences with their respective WER, TER and TERp scores.

3.4.4 Parallel Sentence Generation (Filters)

Once information retrieval is done, the sentence pairs are passed through simple filters to measure the degree of similarity between the SMT translation and the retrieved sentences. Based on the similarity scores, the pairs are classified as parallel or non parallel.

Gale and Church [1993] based their alignment program on the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. We initially used the same logic in our selection of the candidate sentence pairs. However our observation was that the filters that we use implicitly place a penalty when the length difference between two sentences is too large. Thus, using this inherent property, we did not apply any explicit sentence length filtering.

We chose three filters for our study, WER (Levenshtein distance), Translation Edit Rate (TER)[Snover et al., 2006] and the relatively new Translation Edit Rate plus (TERp) [Snover et al., 2009]. A brief explanation of these filters is given in MT evaluation metrics section (chapter2, section 3.2). The choice of the filters was done in accordance to the task in consideration. WER measures the number of operations required to transform one sentence into the other (insertions, deletions and substitutions). A zero WER would mean the two sentences are identical, subsequently lower WER sentence pairs would be sharing most of the common words. However there are many correct translations for any given foreign sentence. These correct translations could differ not only in word choice but also in the order in which these words appear, something that WER is incapable of taking into account. This shortcoming is addressed by TER which allows block movements of words and thus takes into account the reordering of words and phrases in translation. TERp is an extension of Translation Edit Rate and was one of the top performing metrics at the NIST Metric MATR workshop.[1]

The TER filter allows shifts if the two strings (the word sequence in the translated and the IR retrieved sentence) match exactly, however TERp allows shifts if the words being shifted are exactly the same, are synonyms, stems or paraphrases of each other, or any such combination. This allows better sentence comparison by incorporation of a kind of linguistic information about words. Table 3.5 shows some example sentences and their corresponding WER, TER and TERp filter scores. The first sentence pair, being identical, gets a score of zero from all the filters. For the second and third sentence TERp score is less than TER and WER score, as expected based on their calculation methods. In the case of the second sentence, TERp employed one phrase substitution: (in a news conference → at a press conference). In the third sentence 2 phrase substitutions were done : (organisation → organization) and (its oil production → its production). We report these phrase substitutions as per output of the TERp algorithm. Theoretically, we can expect TERp to select the better matching sentences than the other two filters. We used these 3 filters in our sentence selection algorithm and compared their performance, in the following section we present our results.

3.5 Experimental results: Choice of filters

As detailed in the previous section, we chose WER, TER and TERp filters to decide whether the sentences are parallel or not. For this matter we compute the WER, TER

[1] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/

and TERp scores between the automatic translation and the extracted IR sentences. Since these filters have different scoring schemes we get different sets of sentences for the same filter threshold for the three filters, except for the identical sentences, where all three filters assign a zero score (see table 3.5 on page 64). We aim to determine the filter best suited for this task. In the following sections we show the results of our experiment on the two pairs of languages that we worked on.

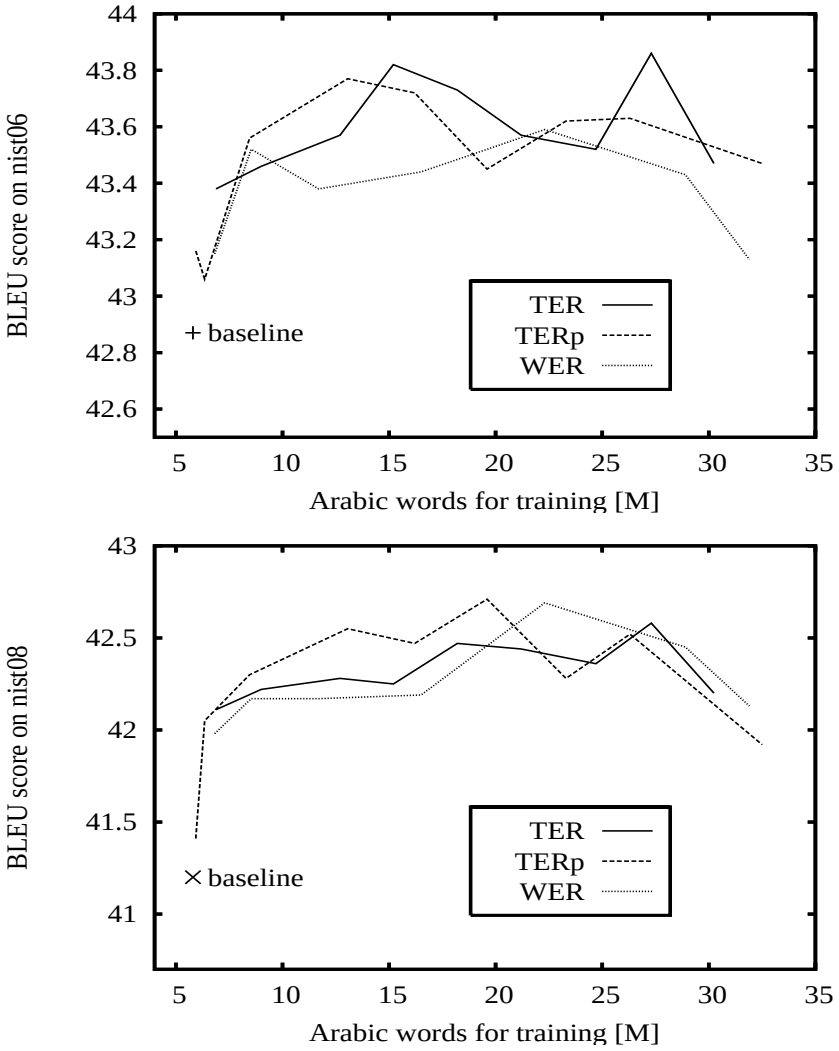

Figure 3.4: BLEU scores on the NIST06 (development) and NIST08 (test) data respectively using WER, TER and TERp filters as a function of the total **Arabic** words. Sentences were extracted from the XIN comparable corpus.

3.5.1 Arabic to English

In this section we report the results when translating from Arabic into English. For the Arabic-English task we had two comparable corpora, AFP and XIN. For our experiments for filter comparison we report the results using the XIN corpus. Figure 3.4 shows the results obtained in function of the total number of words added from the XIN comparable corpus. These experiments were performed by adding our extracted sentences to only 5.8M words of human-provided translations. We find that on the development set (NIST06) TER and TERp filters are almost close by. WER filter performs bad on development data but is close to TER on test data. Whereas, on the test set (NIST08) sentences selected by the TERp filter outperform the other filters.

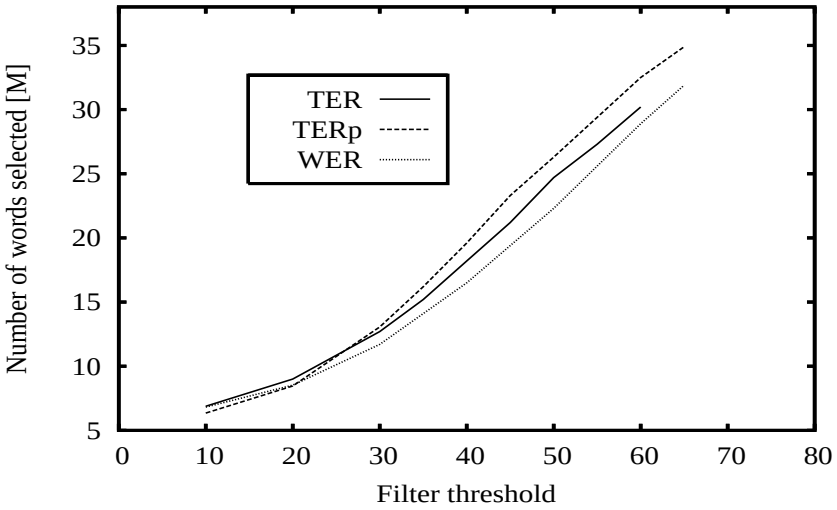

Figure 3.5: Comparison of TER, TERp and WER in terms of number of words selected for the same filter threshold.

Interestingly, for the same filter threshold TERp selected more sentences, followed by TER and then WER, for example for the filter threshold of 60, WER and TER select 28.9M and 30.2M words respectively, whereas TERp selects 32.5M words. This is also evident from figure 3.5 which presents a comparison of the three filters in terms of number of words selected for the same filter threshold. This holds particularly for thresholds greater than 30.

Based on these results, we consider TER and TERp filters to be best suited for experiments on Arabic-English.

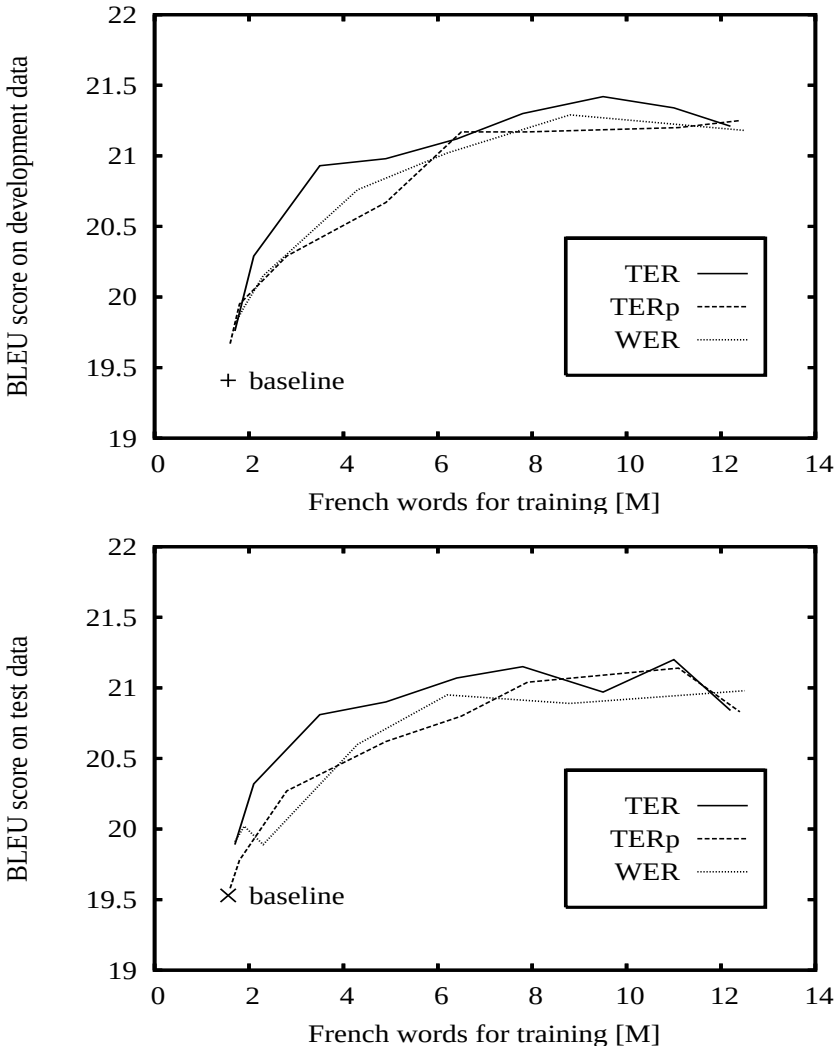

Figure 3.6: BLEU scores on the development and test data using an WER,TER or TERp filter as a function of total **French** words.

3.5.2 French to English

For our experiments with the French-English language pair, like in the previous case, we built various SMT systems by adding sentences with different filter thresholds to the already available 1.56M of human translated news commentary corpus (baseline corpus). Figure 3.6 shows the results of our experiments on the development and the test data respectively.

As evident from the figures, TER selected sentences score slightly better than those selected by WER and TERp. Also, as can be seen in the two figures, irrespective of the

Arabic: بدا الاف الموظفين فى فرز الاصوات التى تم تسجيلها فى عشرات الاف الماكينات الالك ترونية فى 855 بلدة ومدينة عبر البلاد فى الساعة الثامنة صباحا . **Query:** *Thousands of officials began counting the votes registered in tens of thousands of electronic machines in 855 towns and cities across the country at 8 a.m.* **Result:** *Thousands of officials began counting the votes registered in tens of thousands of electronic machines in 855 towns and cities across the country at 8 a.m.* ***thursday.***
Arabic: كان ويكرمسنغ يشير بذلك الى الجمود الحالى بين حكومته ومتمردى جبهة نمور تحرير ايلام التاميلية . **Query:** *ويكرمسنغwas referring to the current stalemate between his government and the Liberation Tigers of Tamil Eelam .* **Result:** *Wickremesinghe was referring to the current stalemate between his government and the Liberation Tigers of Tamil Eelam* ***(LTTE) REBELS .***
Arabic: اتخذ بونو هذا الموقف بعد ان طالب بعض المشرعين الحكومة باعادة التفكير فى التواجد العسكرى الاسبانى فى افغانستان . **Query:** *Bono adopted this position after some legislators asked the government to rethink the Spanish military presence in Afghanistan .* **Result:** *Bono adopted this attitude after some legislators asked the government to reconsider the Spanish military presence in Afghanistan .* ***(SPAIN-AFGHANISTAN) .***

Table 3.6: Some examples of an Arabic source sentence, the SMT translation used as query and the potential parallel sentence as determined by information retrieval. Bold parts are the extra tails at the end of the sentences which we automatically removed.

filter, the addition of our extracted sentences results in considerable improvements in BLEU score as compared to the baseline score (the crosses on the two figures represent the score with 1.56M baseline corpus). Based on the performance, we chose TER filter as standard for our experiments with the French-English language pair.

3.6 Sentence tail removal

In this section we report a feature that we implemented in order to correct a category of error that we came across in our extracted sentence pairs. Two main classes of errors are known when extracting parallel sentences from comparable corpora: firstly, cases where the two sentences share many similar words and same word order but actually convey different meanings, and secondly, cases where the two sentences are (exactly) parallel except at sentence ends where one sentence has more information than the other. The first case is not easy to detect or correct, linguistically it comes much under the semantic error category, when the two sentences share most of the content words

1 **French:** *Au total, 1,634 million d'électeurs doivent désigner les 90 députés de la prochaine législature parmi 1.390 candidats présentés par 17 partis, dont huit sont représenté s au parliament.* **Query:** In total, 1,634 million voters will designate the 90 members of the next parli ament among 1.390 candidates presented by 17 parties, eight of which are represented in parliament. **Result:** Some 1.6 million voters were registered to elect the 90 members of the legisl ature from 1,390 candidates from 17 parties, eight of which are represented in parliament***, several civilian organisations and independent lists.***
2 **French:** *De son coté, Mme Nicola Duckworth, directrice d'Amnesty International pour l'Europe et l'Asie centrale, a déclaré que les ONG demanderaient à M.Poutine de mettre fin aux violations des droits de l'Homme dans le Caucase du nord.* **Query:** For its part, Mrs Nicole Duckworth, director of Amnesty International for Eur ope and Central Asia, said that NGOs were asking Mr Putin to put an end to human rights violations i n the northern Caucasus. **Result:** Nicola Duckworth, head of Amnesty International's Europe and Central Asia de partment, said the non-governmental organisations (NGOs) would call on Putin to put an end to human rights abuses in the North Caucasus ***, including the war-torn province of Chechnya.***
3 **French:** *"Ill a été capturé à Tikrit dans une zone résidentielle", a dit ce responsable.* **Query:** "He was captured in Tikrit in a residential area," said the official. **Result:** "He was captured in Tikrit in a residential area," the official ***said.***
4 **French:** *Je comprends leur préoccupation, mais je me sens blessée", a-t- elle dit au Straits Times.* **Query:** I understand their concern, but I feel hurt," she told the straits times. **Result:** I understand their worries, but I feel hurt," she told the straits times ***newspaper.***
5 **French:** *Plus de 40 pays ont adopté des programmes Vision 2020.* **Query:** More than 40 countries have adopted the vision 2020. **Result:** More than 40 countries have adopted the Vision 2020***, programmes.***

Table 3.7: Some examples of a French source sentence, the SMT translation used as query and the potential parallel sentence as determined by information retrieval. Bold parts are the extra tails at the end of the sentences which we automatically removed.

yet express different meanings. However, for the second case of errors, our use of proper SMT translations gave us the advantage of trying to detect and correct such sentences.

For two sentences differing at sentence end, it is possible to detect the"extra tail"

by identifying exclusive insertions at the end of one sentence. Using WER, we detected the extra insertions at the end of the IR result sentence and removed them. Some examples of such sentences along with tails detected and removed are shown in tables 3.6 and 3.7 for Arabic-English and French-English respectively. The bold parts are the extra insertions at the end of the IR returned sentence that we automatically detected and removed. This scheme is easily applied to the IR returned sentence. In the opposite case, this could be done by using the word alignments between the foreign sentence and the SMT translation (which the SMT translation framework gives us). However, we applied this scheme to the IR sentence only.

By detecting insertions at the end, it is not always the correct portions that get chopped off, table 3.7 sentence 3 is an example of one of such cases. In sentence 4, the sentence was a better translation with the word "newspaper", but it got removed being the extra tail. In sentence 5 in the same table, the SMT translation (query) erroneously does not contain the word "programmes", whereas the retrieved IR sentence does (and is an exact translation of the French sentence) but it gets removed because it does not match the query. These examples also show that the problem of finding only true translation pairs is a hard one. However, our task of getting useful MT training data does not require a perfect solution; as we will see and have seen in section 3.5, even such noisy training pairs can help improve translation systems' performance.

3.7 Experimental results: Sentence tail removal

3.7.1 Arabic to English

To investigate the effect of sentence tail removal on the Arabic-English systems, we proceed as before by building SMT systems using our extracted sentences with and without tail removal and comparing their performance. Figure 3.7 shows these results on the development and test data for the three filters. From the figure, despite the fact that tail removal works better for WER filter, WER filter is actually never better than TER and TERp, this was also the case for WER in the previous results (section 3.5). Table 3.8 gives an idea of the percentage of sentences on which tail removal was applicable (based on the TER score).

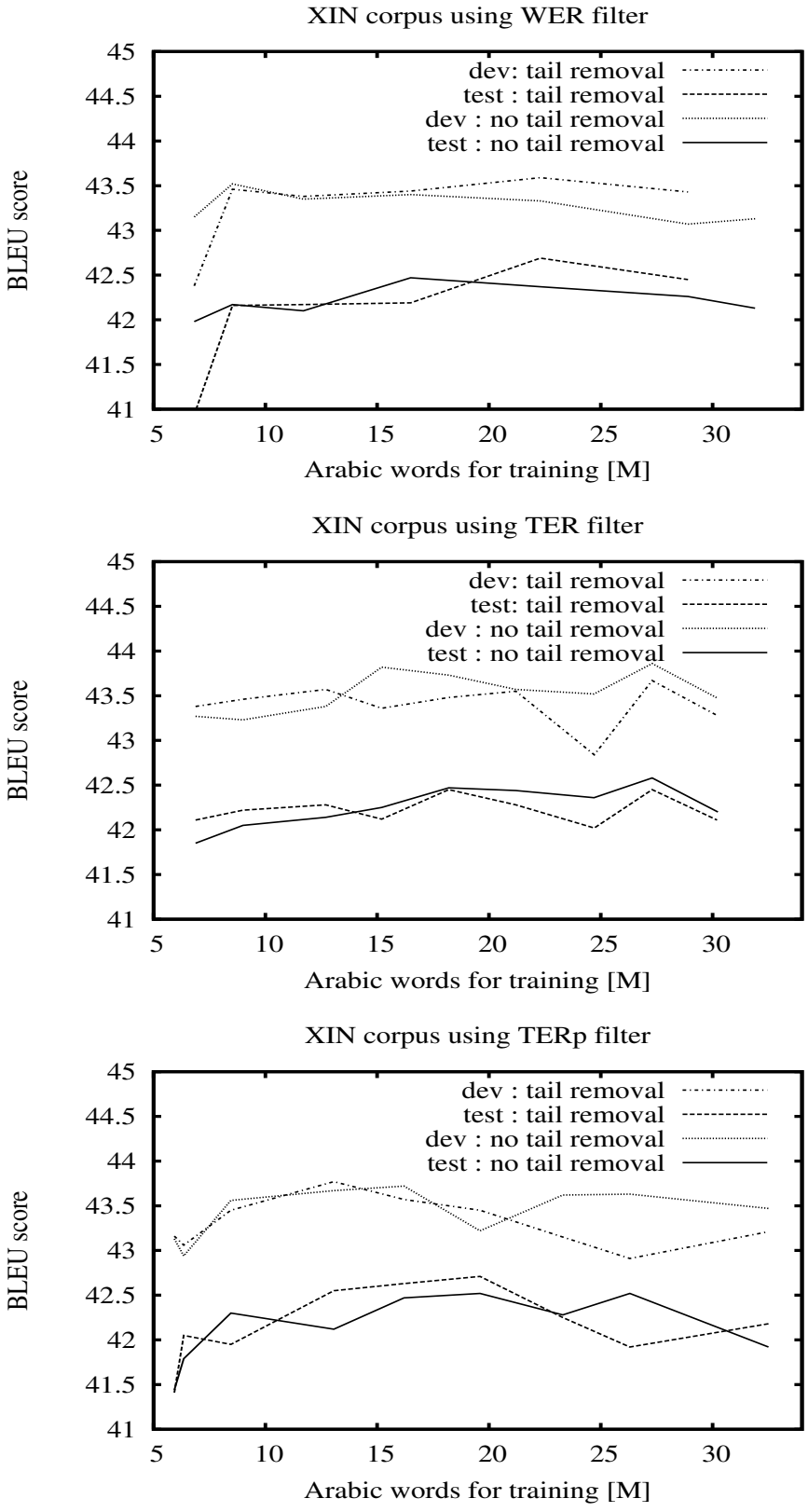

Figure 3.7: Effect of sentence tail removal using sentences selected by WER, TER and TERp filters (Arabic-English).

Tail removal if applied to smaller data sets (better scoring sentences), improves the results with some exceptions. With larger data sets, there are forcibly too many errors, so it doesn't help much. This is evident from figure 3.7, for sentences filtered using the TER filter. Our data sets are sorted by the filter score, thus the smaller data sets mostly contain short and almost parallel sentences (having very low filter scores), sentences having tails of the sort as shown in lines 4-6 in table 3.6 and sentences 3, 4 and 5 in table 3.7. Removing such tails resulted in an improvement in BLEU score. The bigger data sets had sentences with lower filter scores, i.e. not parallel sentences rather sentences which had some matching words, extra ends detected in such cases by WER were not true tails, but just different texts at the end (since the sentences were not parallel but just sharing some common words).

TER threshold	10	20	30	40	50	60
Tail removal(%)	1.45	5.52	8.76	11.1	12.5	13.1

Table 3.8: Percentage of sentences based on the TER filtered XIN sentences showing how often tail removal is applied.

3.7.2 French to English

To determine the effect of sentence tail removal on the French-English systems, we built SMT systems using the automatically extracted sentences with extra parts removed. We were able to get slight improvements in development and test data scores for some data sets but not for all as shown in figure 3.8. These improvements were spread over all filter thresholds i.e. we got slight improvements but also slight deteriorations in the BLEU scores irrespective of the filter threshold. Thus, for the French-English language pair there is no clear tendency whether tail removal is beneficial or not.

Our observation is that no method, either tail removal or choice of filter is best over all sizes of extracted data.

3.8 Dictionary Creation

Using proper SMT translations as queries gives us another benefit of being able to build a dictionary, often with proper nouns not usually found in the dictionaries. In our translations for Arabic-English, we keep the unknown words as they are, i.e. in Arabic. This enables us to extract word translations from the comparable corpora.

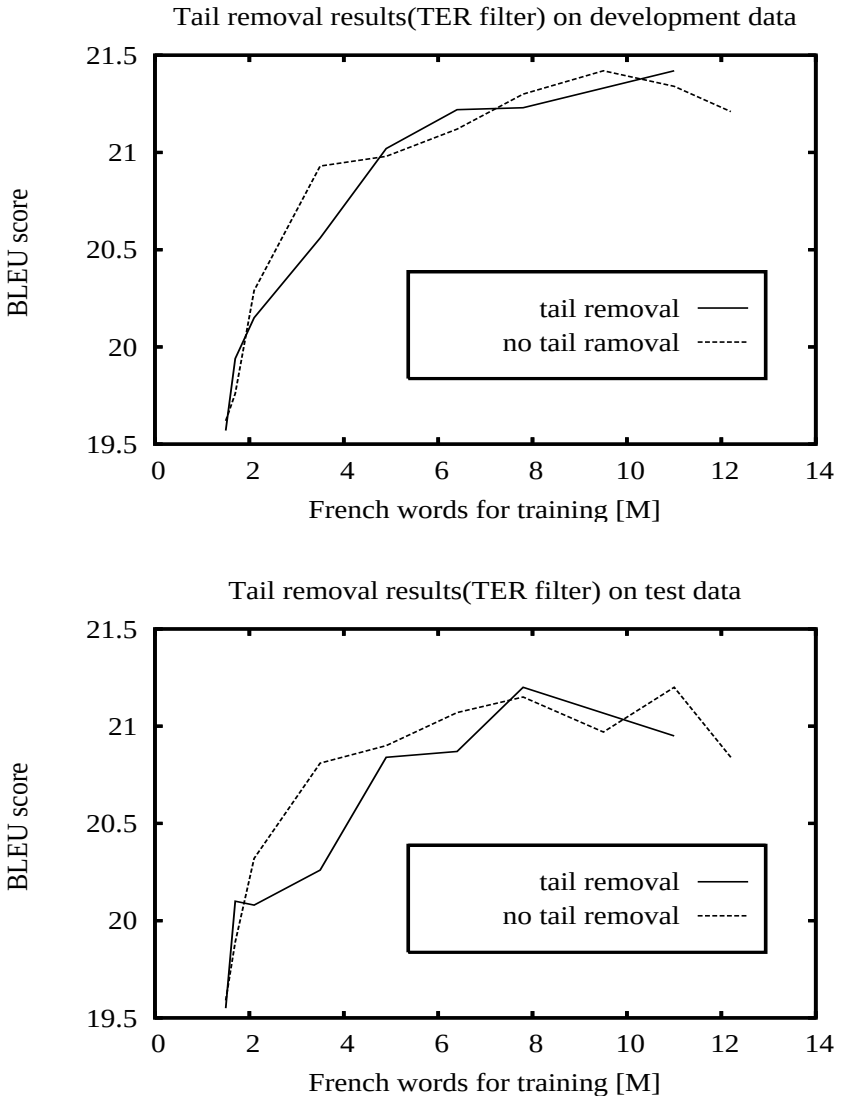

Figure 3.8: Effects of tail removal on sentences selected using TER filter (French-English).

Consider the case of a translation with one unknown word in Arabic, if all the other words around align well with the English sentence that we found with IR, we could conclude the translation of the unknown Arabic words.

Let Q and S denote the query and the sentence retrieved by IR respectively:

$$Q = (q_1, q_2, ...q_{i-1}, \mathbf{q_i}, q_{i+1}, ..., q_m) \tag{3.1}$$

وصل جون **ابيزيد** الى هنا امس الثلاثاء . **Query:** John **ابيزيد** *arrived here on* Tuesday . **IR Result:** John **Abizaid** *arrived here on* Tuesday .
وفى **نارانثيواث** اصيب شرطيان *فى* الهجوم بالقنابل . **Query:** At **نارانثيواث** *Two policemen were* injured in the bomb attack . **IR Result:** In **Narathiwat** , *two policemen were* injured in the bomb attack .
وقد وصل **راجاباكسه** الى هنا يوم السبت *فى* زيارة دولة للهند استغرقت ثلاثة ايام . **Query:** The **راجاباكسه** *arrived here Saturday* on a three-day visit to India . **IR Result:** **Rajapakse** *arrived here Saturday* on a three-day visit to India .

Table 3.9: Sample sentence pairs for the dictionary building technique. The bold parts are the word/translation pair detected by our technique and the italic parts are the preceding and following matching words of the sentences.

$$S = (s_1, s_2, ...s_{j-1}, \mathbf{s_j}, s_{j+1}, ..., s_n) \tag{3.2}$$

The query and the IR result are both in English, with an Arabic word in the query (q_i in equation 3.1). Now, if q_i in Q aligns with s_j in S and the two or three preceding and following English words of Q and S match (q_{i-1}, $q_{i-2} = s_{j-1}$, s_{j-2} and q_{i+1}, q_{i+2}, $q_{i+3} = s_{j+1}$, s_{j+2}, s_{j+3}), we can safely conclude that s_j is the translation/transliteration of the Arabic word q_i. Table 3.9 shows some sample sentences. The word and its expected translation is shown in bold, the preceding and following matching parts of the two sentences are shown in italic. Table 3.6 on page 69, lines 5 and 6 is another example of such a sentence, from where the proper noun "Wickrelesinghe" could be extracted along with its translation, which in this case is a transliteration of the word.

With this idea of dictionary creation, we were able to make an Arabic-English dictionary. Our observation is that most of the words found this way are the names of persons, places and new terms which are normally not found in the traditional Arabic-English dictionaries. We were able to find about 64K and 0.2M such words using XIN and AFP corpora respectively. Table 3.10 shows a sample of the extracted word pairs. Some of these words are transliterations of the Arabic words.

Table 3.11 shows the results obtained when we added our dictionaries to 5.8M of baseline corpus. The dictionary extracted from AFP provided a gain of 0.27 and

English word from SMT	Arabic unknown word
PetroChina	بتروتشاينا
Bolotine	بولوتين
Amrozi	امروزى
Bulldozers	البولدوزورات
Schulte	شولتى
Jiuxuan	جيوشيوان
Dijmarescu	ديجماريسكو
Aliasghar Soltanieh	اليسجار سلطانيه

Table 3.10: Examples of some words found by our dictionary building technique.

Bitexts	**Total words (Arabic)**	**BLEU score**	
		Dev	**Test**
Baseline	5.8M	42.87	41.20
+dicXIN	5.9M	42.68	41.14
+dicAFP	6.02M	**43.14**	**41.53**
+dicXIN+dicAFP	6.08M	**42.88**	**41.31**

Table 3.11: Results of adding our extracted dictionary words to the baseline corpus.

0.33 BLEU points on development and test data respectively. Strangely the dictionary extracted from XIN did not prove to be much helpful. A combination of the two dictionaries also helped improve score but not as much as dicAFP alone.

3.9 Machine Translation Improvements

Our main goal was to be able to create an additional parallel corpus to improve machine translation quality. Thus, we evaluate our extracted corpora by showing that adding them to the training data of a baseline MT system improves its performance.

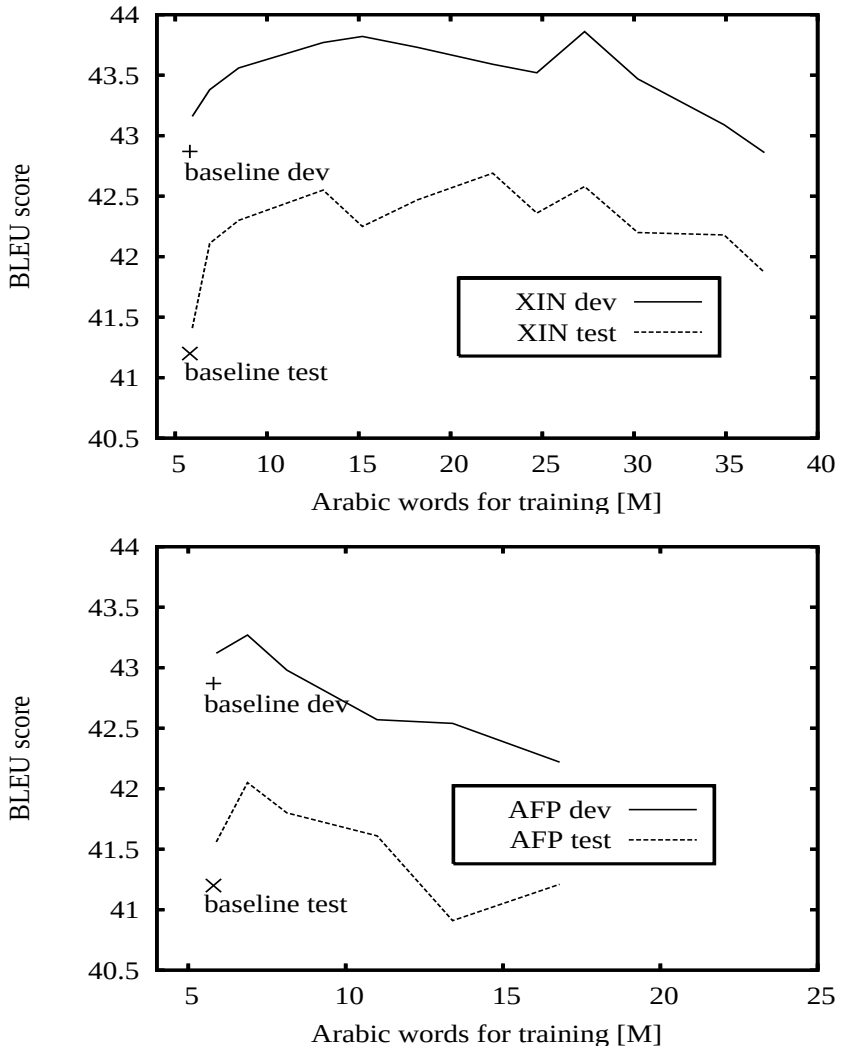

Figure 3.9: BLEU scores when using 5.8M human translated bitexts and our extracted bitexts from AFP and XIN comparable corpora.

3.9.1 Arabic to English

In this section we report the results of our experiments conducted by using sentences extracted from the AFP and XIN comparable corpora. The results are summarized in figure 3.9, which shows the trend obtained in function of the total number of Arabic words. These experiments were performed by adding our extracted sentences to only 5.8M words of human-provided translations. This figure shows results of the best operating points, selected on the Dev data, using the TER and TERp filters (with and without sentence tail removal) for our extracted sentences.

Bitexts	Total words (Arabic)	BLEU score Eval06	Eval08
Baseline	5.8M	42.87	41.20
+AFP-T-40-tr	6.88M	43.27	42.05
+XIN-T-55-wt	27.3M	43.86	42.58
+AFP-T-40-tr+XIN-T-55-tr	28.3M	43.79	42.60

Table 3.12: Summary of BLEU scores for the best systems built on sentences extracted from the XIN and AFP corpora. The name of the bitext indicates the filter threshold used, XIN-T-55-wt means sentences selected from the XIN corpus based on TER filter threshold of 55, -wt indicates with tails and -tr indicates tail removal.

It is evident from figure 3.9 that sentences extracted from XIN were much better than those from AFP. Even though the AFP corpus is almost 3 times bigger than the XIN corpus (see tables 3.1 (page 55) and 3.15 (page 81)), the number of good sentences found are much smaller than for XIN. Also, in the parallel data from Munteanu and Marcu [2005], provided by the LDC, 60% of the sentences are from XIN and the rest from AFP (see section 3.13 on SMT results using their data).

The best operating point on the development data for XIN was achieved by adding 21.5M sentences selected by the TER filter without sentence tail removal to the baseline corpus (27.3M in total). This gives a gain of 0.99 BLEU points on NIST06 (dev data) and 1.38 BLEU points on NIST08 (test data). These results are presented tabularly in table 3.12. The name of the bitext indicates the filter threshold used, for example, XIN-T-55-wt means sentences selected from the XIN corpus based on TER filter threshold of 55, -wt indicates with tails and -tr indicates tail removal. Using 1.08M (6.88M total) of AFP corpus (the best score on dev data), improvements of 0.40 and 0.85 BLEU points were achieved on the development and test data respectively. Using a combination AFP and XIN, we were able to achieve an improvement of 0.92 and 1.40 BLEU points on dev and test data respectively, which is slightly more on the test data than with XIN alone. Our observation was that generally sentences extracted from the XIN comparable corpus helped to improve SMT translation better than the sentences from AFP alone or the combination of AFP and XIN.

Bitexts	Total words	BLEU score Dev	BLEU score Test
News	1.56M	19.36	19.44
News+AFP-T-60-wt	9.5M	**21.42**	**20.97**
News+dict	2.4M	20.62	20.31
News+dict+AFP-T-55-wt	8.6M	**21.63**	**21.51**
News+Eparl	41.7M	22.17	22.23
News+Eparl+AFP-T-70-wt	52.4M	**22.19**	22.13

Table 3.13: Summary of BLEU scores for the best systems on the development data with the news-commentary corpus, the bilingual dictionary and the Europarl corpus. The naming convention used is CorpusName-FilterName-FilterThreshold. All corpora are without tail removal, thus having the suffix -wt (with tails).

3.9.2 French to English

As our goal was to improve SMT performance by creating' parallel texts for domains which do not have enough parallel corpora. Therefore, only the news-commentary bitexts and the bilingual dictionary were used to train an SMT system that produced the queries for information retrieval. Experiments by adding our extracted bitexts to the baseline corpus showed significant improvements in BLEU score.

The baseline BLEU score is 19.36 (news-commentary texts only). The best BLEU score of 21.42 on the development data is obtained when adding 7.9M words of our automatically aligned bitexts to the baseline corpus (9.5M in total). This corresponds to an increase of 2.06 points BLEU on the development set (19.36 $\rightarrow$ 21.42) and an increase of 1.53 BLEU points on the test set (19.44 $\rightarrow$ 20.97), as shown in table 3.13, first two lines. Our experiments showed that adding the dictionary improves the baseline system (third line in Table 3.13), but we get much better improvement with the extracted data only. We then added our extracted corpus to the collection of News-commentary (1.56M) and Europarl(40.1M) bitexts but we did not get any gain here.

Using our extracted corpus with larger baselines, we got promising improvements for WMT 2010 [Lambert et al., 2010] and WMT 2011 [Schwenk et al., 2011] evaluations. Some of these results from [Lambert et al., 2010] are shown in table 3.14. We present them here to demonstrate the significance of our extracted sentences even for

Bitexts	Total words (Fr)	BLEU score	
		Dev	Test
Eparl+NC	52M	22.80 (0.03)	25.31 (0.2)
Eparl+NC+Extracted	68M	**22.97 (0.03)**	**26.20 (0.1)**
Eparl+NC+UN	275M	23.38 (0.1)	26.30 (0.2)
Eparl+NC+News	111M	23.46 (0.1)	26.95 (0.2)
Eparl+NC+News+Extracted	127M	**23.62 (0.01)**	**27.04 (0.06)**
Eparl+NC+10^9_1+News	242M	23.77 (0.04)	27.11 (0.04)
Eparl+NC+10^9_1+News+Extracted	258M	23.75 (0.05)	**27.24 (0.05)**

Table 3.14: Summary of BLEU scores when adding our extracted sentences to larger SMT systems. The values are the average over 3 MERT runs performed with different seeds. The numbers in parentheses are the standard deviation of these three values [Lambert et al., 2010].

improving bigger systems. The resources used for the results reported in table 3.14 are the latest versions of News Commentary (NC) and Europarl (Eparl) corpora (version 5). *News* denotes the corpus produced by using automatic translations of French News corpus [Schwenk, 2008]. 10^9_1 denotes a subset of French-English Gigaword (10^9) corpus.[1] *Extracted* denotes the bitexts produced using our scheme. This *Extracted* corpus contains the parallel sentences extracted from the French AFP and APW news texts from the French and English LDC Gigaword corpora. For development and test sets *news-test2008* and *newstest2009* were used respectively.

When 16M of our extracted bitexts are added to Eparl+NC, we obtain a system of similar performance as the system trained on Eparl+NC+UN (2nd and 3rd lines in table 3.14), while our extracted bitext are 10 times smaller than the UN corpus. The extracted bitexts prove beneficial in improving SMT performance, even when added to larger corpora, (111M of Eparl+NC+News) where an improvement 0.16 and 0.09 BLEU points is gained on development and test sets respectively. Finally when added

[1]Two filters were applied to select this subset. One is an IBM model 1 [Brown et al., 1993] based lexical filter trained on a corpus composed of Eparl, NC, and UN data. The other filter is an n-gram language model (LM) cost of the target sentence normalised with respect to its length. This filter was trained with all monolingual resources available except the 10^9 data. Subset, 10^9_1 are the selected sentence pairs with a lexical cost inferior to 4 and an LM cost inferior to 2.3.

to Eparl+NC+10^9_1+News, the extracted bitexts provide a gain of 0.13 points.

3.10 Characteristics of the comparable corpus

In approaches based on comparable corpora, it is very typical to retrieve only small fraction of parallel sentences of the overall corpus since many sentences actually don't have a translation. Nevertheless, these small amounts have proved to be very beneficial when used as additional bitexts [Abdul-Rauf and Schwenk, 2009b; Do et al., 2009; Fung and Cheung, 2004; Gahbiche-Braham et al., 2011; Munteanu and Marcu, 2005; Wu and Fung, 2005]. However, the amount of good sentences found depends upon the comparable corpus at hand. In the study of Munteanu and Marcu [2006] two different comparable corpora were used, the BBC comparable corpora which was truly non-parallel and the EZZ corpus which was comparatively more parallel in degree. They report better SMT improvements using the parallel sentences from the EZZ corpus, whereas the BBC corpus produced improvements by using the extracted sentence fragments rather than full sentences. Clearly, the EZZ corpus yielded more parallel sentences. From our own experiments with two different Arabic comparable corpora, XIN and AFP, we found the XIN comparable corpus to be far more productive than the AFP comparable corpus. Table 3.15 shows the amount of sentences extracted from each corpus:

Source	**Comparable Corpus**	**Extracted Words**
	Arabic	
AFP	212M	16.8M
XIN	80M	35.1M
	French	
AFP	333M	12.3M

Table 3.15: Amount of words (Arabic/French) extracted from the XIN and AFP comparable corpora (number of words). We are considering the extracted sentences to be "good" till the TER threshold of 70.

The XIN corpus seems to be more comparable than AFP. Independent reporting in the two language versus human translation could be one of the factors. Also, as reported in section 3.9.1, the sentences extracted from the XIN corpus improve SMT performance to a much greater extent than those extracted from the AFP corpus (figure 3.9).

An interesting observation is that the French AFP corpus, despite being more in quantity than Arabic AFP corpus, yields less words. Information retrieval using the AFP corpus was a much lengthy task. Though the English corpus (140M) for XIN was also almost double the Arabic corpus (80M), information retrieval generally took much less time as compared to AFP. There were some corpus specific factors that we think contributed to this. For example, for each day AFP had twice the number of articles than XIN, also the sentences were much longer in AFP. According to our general observation AFP covered a general range of news, having many stock details and sports results along with the everyday news, whereas XIN comprised of specific regional news data.

3.11 An alternative sentence selection experiment

In our scheme we compute the TER, WER and TERp between the SMT translation and the IR returned sentence. By default, we choose the first sentence from the 5 IR returned sentences, as it is the one with the highest score according to the IR toolkit. We experimented with an alternative way of selecting the IR returned sentences. We computed the TER between the SMT query sentence and the 5 IR returned sentences, and chose the pair with the lowest TER score.

Using the XIN comparable corpus, we found that the first sentence (the one with highest IR score) got selected 24% of the time, followed by the 2nd, 3rd, 4rth and 5th sentence with very close percentages of 19.3%, 18.7%, 18.6% and 19.2% respectively. We conducted SMT experiments using the corpus made from these parallel sentences. In figure 3.10 we show the BLEU scores obtained on the dev and test data using our default settings (*1-best IR:* choosing the first sentence returned by IR) and (*5-best IR:* choosing the IR sentence based on lowest TER between SMT output and the 5 IR returned sentences). As the figure shows, there is no clear advantage for one approach.

An other alternative approach to selecting sentences could have been the use of K-best SMT translations. However, the expected benefit of this method is rather small since the sentences in the k-best list typically vary only by some words. Note also that the differences in word order, choice of articles, singular vs plural and tense etc. have no affect on the IR process. Therefore, we anticipate that using the k-best output of the SMT system would not change significantly the sentences retrieved by IR (though, the post IR selection of the sentences would change for various filter thresholds). Also, note that computing the TER between the retrieved sentence and k-best translations to choose the best sentence pair is not useful as even though this will give us the best

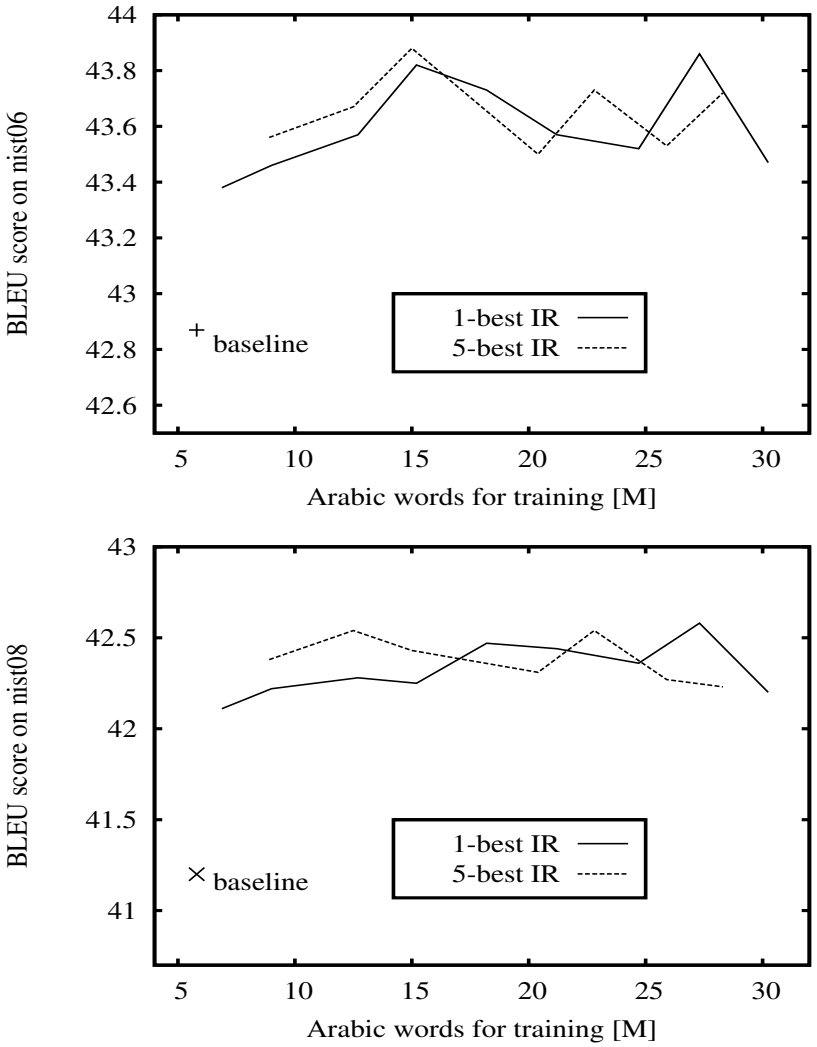

Figure 3.10: BLEU scores on the NIST06 (development) and NIST08 (test) data as a function of total **Arabic** words. **1-best IR:** choosing the first sentence returned by IR and **5-best IR:** choosing the IR sentence based on lowest TER between SMT output and the 5 IR returned sentences.

translation with respect to the retrieved sentence, but in the end, in the bilingual corpus we would be using the Foreign language side of the sentence, which would always be the same.

3.12 Effect of SMT Quality

It is a well known fact that the amount of parallel corpus used to train an SMT system directly determines the quality of the translations. In general, the more the parallel corpora used to train the system, the better the translation quality. Our scheme is easy to implement using available open source tools. Thus for our scheme, the question to be answered is whether we need a state-of-the-art SMT system built from large amounts of parallel corpora or an SMT system built from minimal amounts of parallel corpora would suffice to obtain good queries?

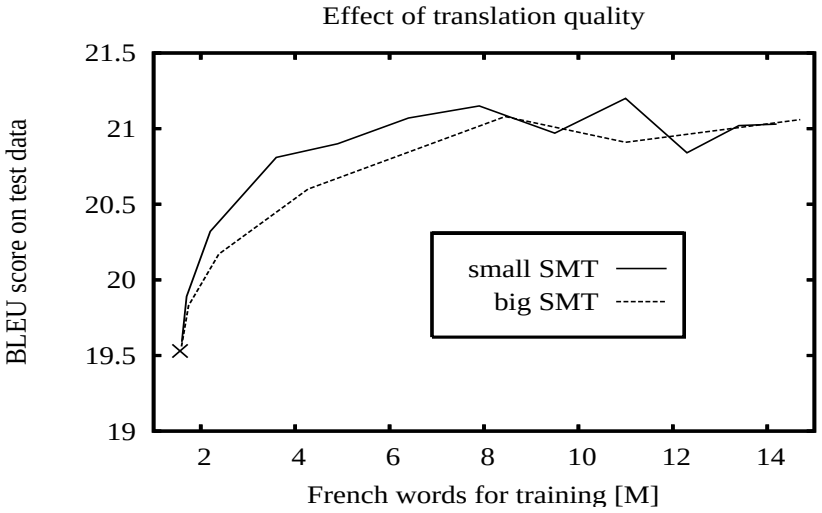

Figure 3.11: BLEU scores on test data when using queries produced by the small and big SMT systems.

To investigate the effect of translation quality on the overall result of the scheme, we built two SMT systems using the French-English language pair trained on already available human translated corpora: One big system trained on 116M words and a small system trained on only 2.4M words, as detailed in section 3.3.2.2 (page 56). Parallel sentence extraction was done using the translations performed by these two SMT systems as IR queries. Our experiments showed no apparent gain when using the best SMT system, in fact our approach works well with an SMT system trained on small amount of bitexts as depicted in figure 3.11.

Figure 3.12 depicts the number of words selected for each TER threshold by the two systems. Interestingly, using the big SMT system the number of words selected for each threshold (TER $\geq$ 30) is always greater than those selected by the small SMT system. A better SMT system makes less errors and produces better sentences, thus

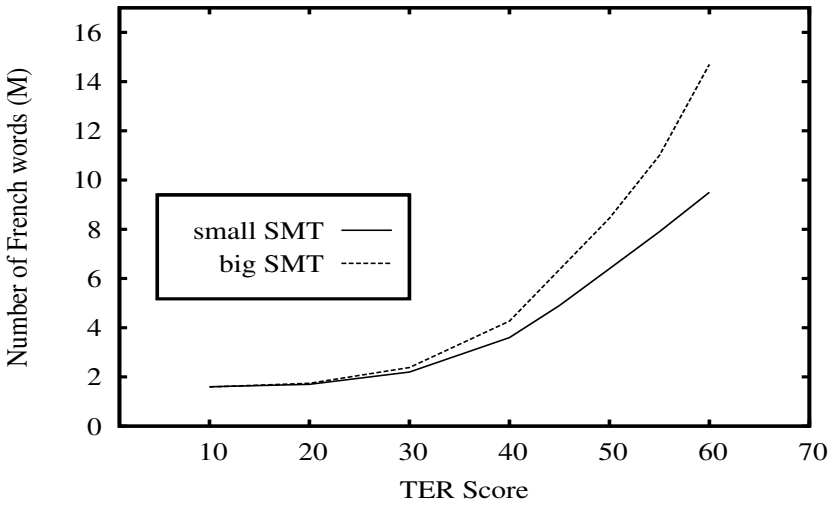

Figure 3.12: Number of words selected for each TER threshold for both the big and small SMT systems.

more sentences are selected for same threshold. Nonetheless, the point to be noted is that the difference in performance of the SMT system by the addition of the parallel corpora created by the two systems (big and small) is very little.

We found no experimental evidence that the improved automatic translations yielded better alignments of the comparable corpus. Thus strengthening our claim of usability of our approach for language pairs with limited amount of parallel corpora to start with. Our scheme still works good with a worse SMT system.

3.13 Comparison with previous work

We conducted a comparison of our approach with the technique proposed in [Munteanu and Marcu, 2005] using the same data as they use for their experiments. In this section we report our comparison on a theoretical as well as empirical basis.

3.13.1 Theoretical Comparison

Munteanu and Marcu [2005] use a bilingual probabilistic lexicon to get the translations of the source documents. For each word of the document they choose the word with the highest probability from their dictionary. Using these translations, they do information retrieval against the English comparable corpora using the Lemur IR toolkit [Ogilvie and Callan, 2001] (we use the same IR toolkit). For each document and the set of

	Their approach	**Our approach**
Query creation	Probabilistic dictionary ***-precision***	SMT ***+precision***
Information retrieval	Matching document retrieval	Matching sentence retrieval
Post processing	1. Sentence pair **creation** (word overlap filter) 2. Maximum Entropy classifier	1. TER/TERp/WER filter 2. Sentence tail removal
Needed resources	1. Parallel sentences 2. ME classifier 3. bilingual dictionary	Average SMT system

Table 3.16: Theoretical comparison of our scheme with that of Munteanu and Marcu [2005]

associated documents, they take all possible sentence pairs and pass them through a word overlap filter. The word overlap filter selects the candidate sentence pairs based on the ratio of lengths of the two sentences, and the fact that half the words in each sentence have the translation in the other sentence according to the dictionary. These candidate sentence pairs are then classified as parallel or not by a maximum entropy (ME) classifier, based on task specific feature functions and trained on already available human translated parallel sentences. They use the technique of bootstrapping and learn a dictionary over several iterations, the size of which increases iteratively to get better results.

Our approach is easier to implement than that of Munteanu and Marcu [2005] as we only use open source tools. The theoretical comparison of the two approaches is shown in table 3.16 . We shift our effort to the pre-IR stage, i.e. query creation, where we use a proper SMT system built from small amounts of parallel texts.

We are able to emphasize both precision and recall in this step, contrary to Munteanu and Marcu [2005] who only emphasize recall in their article selection step. For them, using isolated word translations as queries compromises precision, whereas for us, full sentence queries provide better precision as obviously phrase-based translation is better than word-based translation.

Our post-IR processing is much simpler than their ME classifier. Our IR framework retrieves best matching sentences for each query sentence (section 3.4.3.1, page 59),

rather than best matching document, as in their case, so we don't need to do any explicit candidate sentence pair creation which they do as they retrieve matching documents. Our retrieved sentences are already in the form of sentence pairs. The filters that we use inherently penalize when the sentence length ratio between the two sentences is too high. Thus, after the IR step, all we need to do is pass our sentence pairs through the TER or TERp filter and select the best scoring sentences. As detailed in section 3.6 having the full SMT translation in English along with the IR retrieved sentence, we get a chance to remove additional sentence tails, a common case of errors in such tasks.

3.13.2 Experimental Comparison

The Linguistic Data Consortium (LDC) provides the parallel texts extracted with the algorithm published by Munteanu and Marcu [2005] by the name LDC2007T08.[1] We will call this data the ISI corpus. This corpus contains 1.1M sentence pairs (about 35M words) which were automatically extracted and aligned from the monolingual Arabic and English Gigaword corpora (we used the same), a confidence score being provided for each sentence pair. Since we cannot reproduce their ME classifier, we conduct a comparison based on the SMT scores produced by using the two corpora in identical experimental conditions. To be able to compare our approach with theirs, we filtered our data according to the time interval of their data (date information was provided for each sentence pair). Since their corpus comprised of a combination of AFP and XIN, we conducted separate experiments using the extracted sentences found from AFP and XIN.

We conducted SMT experiments by adding our extracted sentences (using the same time frame as Munteanu and Marcu [2005]) to the already available 5.8M of human-translated sentences (as done in previous experiments). Similarly, SMT experiments were performed by adding the ISI parallel data to the 5.8M baseline parallel corpus. These results are shown graphically in figure 3.13. This figure shows results of the best operating points, selected on the Dev data, using the TER and TERp filters (with and without sentence tail removal) for our extracted sentences. As the figure shows, we perform better on both XIN and AFP with few exceptions on larger data sets with ISI performing better. The best scores obtained are shown in table 3.17, these are selected based on development data score. The name of the bitext indicates the filter threshold used: ISIXIN-0.99 means sentences selected from ISI XIN corpus using a threshold of 0.99 according to their provided score, similarly WeXIN-T20-wt denotes

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T08

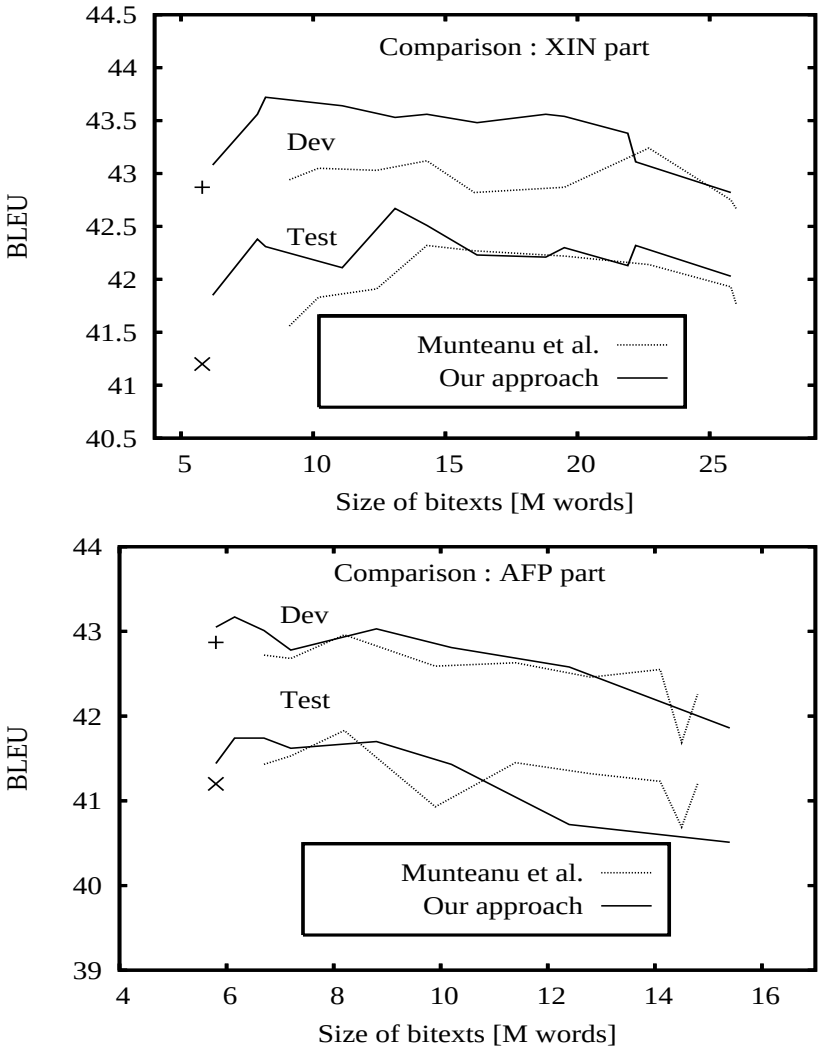

Figure 3.13: BLEU scores on the NIST06 and NIST08 data using the ISI parallel corpus and our comparative extracted bitexts in function of total number of Arabic words. Crosses at 5.8M words represent baseline dev and test scores.

our sentences extracted from XIN corpus using TER filter threshold of 20 and without any tail removal (-wt : with tails).

Using the sentences extracted from XIN corpus, we were able to achieve an improvement of 0.85 and 1.11 BLEU points, whereas using ISI's XIN part resulted in an improvement of 0.37 and 0.94 BLEU points on development and test data respectively. Generally, adding our sentences extracted from XIN helped to improve SMT performance better than those from the ISI corpus and this using much less words (8.2M

Bitexts	Total words (Arabic)	BLEU score Dev	Test
Baseline	5.8M	42.87	41.20
XIN			
+ISIXIN-0.96	22.7M	43.24	42.14
+WeXIN-T20-wt	8.2M	43.72	42.31
AFP			
+ISIAFP-0.99	8.2M	42.96	41.83
+WeAFP-P30-wt	6.15M	43.17	41.74

Table 3.17: Summary of BLEU scores for the best systems built on sentences selected from ISI's and our XIN and AFP corpora (-wt indicates sentences used with tails, i.e. without tail removal).

vs 22.7M total words). For the AFP part, their sentences performed a little better than ours for larger data sets on the test set. Using our data being able to achieve an improvement of 0.30 and 0.54 BLEU points and theirs achieving 0.09 and 0.63 BLEU point improvement on development and test sets respectively (selected on the best score achieved on the dev set).

The trend in the BLEU score in figure 3.13 shows that our sentence selection scheme is comparable (and often better) in terms of results with that of ISI. In ISI's framework, the confidence scores provided are based on the IR and maximum entropy classifier scoring scheme, whereas our filters score the sentences based on linguistic sentence similarity. Moreover our scheme does not require any complex operations, just simple filters.

3.14 Future Perspectives

The IR framework has a number of possible formulations for indexing and retrieval. We think that a comparison of some of them would be a potential future work.

An interesting extension would be to use more than one sentence per query as returned by IR. Though this amounts to duplicating the source sentence, but parallel corpora often contain multiple translations. Several methods can be used to choose the potential sentences, e.g. computing a similarity score (TER, TERp, PER, WER

or even cross-entropy, BLEU and NIST etc.) between the query and each retrieved sentence, and using all the sentences above a certain threshold. Another way of doing this implicitly using the IR framework would be to implement the index such that one sentence constitutes a document and then retrieve n-best results. We have used this scheme in the framework of our work on monolingual corpora (chapter 4), however applying this design for comparable corpora needs to be investigated.

Chapter 4

Exploiting the Monolingual corpora

4.1 Overview

In this chapter we apply our approach to make use of available target language texts to improve SMT performance. We use information retrieval techniques to find the sentences most related to the task. Our work is inspired by the ideas of lightly-supervised training [Schwenk, 2008] and self-enhancing [Ueffing, 2006]. In the next section we present a brief introduction to our work followed by a brief presentation of previous works (section 4.3). We then describe our research in detail in section 4.4 followed by the presentation of experimental results in section 4.5 and show that we are able to improve competitive SMT systems. We conclude the chapter by a discussion on future perspectives.

4.2 Introduction

The theme of this dissertation is to devise methods to exploit the available corpora (other than parallel corpora), to be able to create training material for SMT systems and eventually improve SMT quality. In the previous chapter we showed improvements by exploiting comparable corpora. In this chapter we focus on monolingual corpora to achieve the same goal, i.e. increase the pool of SMT training data to get improvements in quality.

For our research with comparable corpora as presented in the previous chapter, we used the source language translations as queries for IR. Interestingly, the source corpus

and its translations are themselves bitexts, though not of good quality as they have been produced by an SMT system, but still the best translations can be used as additional parallel texts. Schwenk [2008] adopt this approach by translating large amounts of monolingual texts and using these translations as additional training material. They filter the translations using the SMT confidence score and use the most reliable ones. They call this approach *lightly-supervised training*. Recently, [Lambert et al., 2011] while reporting improvements using such automatic translations, showed that it is more beneficial to use translated texts which were translated from the target to the source language as this avoids the use of bad translations and translation errors are not propagated. But these approaches do no selection of appropriate data which results in a high complexity as large amounts of sentences are translated.

We propose an extension to these approaches by using IR to find the sentences most related to the in-domain development and test data. Translations of these found sentences are then used as additional bitexts. This considerably reduces the computational cost since only small amounts of data must be translated automatically. We use the available in-domain monolingual data in the target language. This data is usually available in large amounts since it is needed to train the target language model. We use IR techniques to select a small subset of relevant sentences in this LM data collection. The queries for IR are either the reference translations of the development data or the automatic translations of the test data as produced by a baseline system. This gives us the sentences which are related to the development and test data. We then translate these sentences and add them as additional bitexts. By these means we perform unsupervised training similar to Schwenk [2008], but we make an active selection of the sentences most related to the task, instead of translating blindly millions of words. Following this scheme, we were able to get appreciable improvements in SMT scores for English-French state-of-the-art SMT systems.

In figure 4.2 we show a high level architecture our two methods for comparison. Use of SMT and IR are the two main features for both the methods. When using comparable corpora to find parallel sentences, we use the translations of the source language comparable corpus and find matching sentences from the target language comparable corpus using IR. We then use the best sentence pairs based on the filter scores as additional bitexts. While, for monolingual corpus, instead, we translate the test corpus and do IR on the target language LM data using the test data translations and the target language development data as queries. Doing so, we find the sentences most related to the task. We then translate these sentences and use them as bitexts.

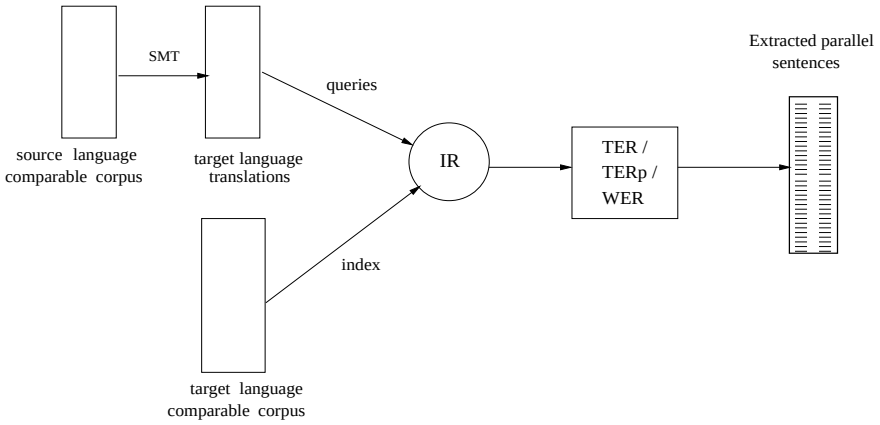

Figure 4.1: High level architecture of our approach when applied to comparable corpora.

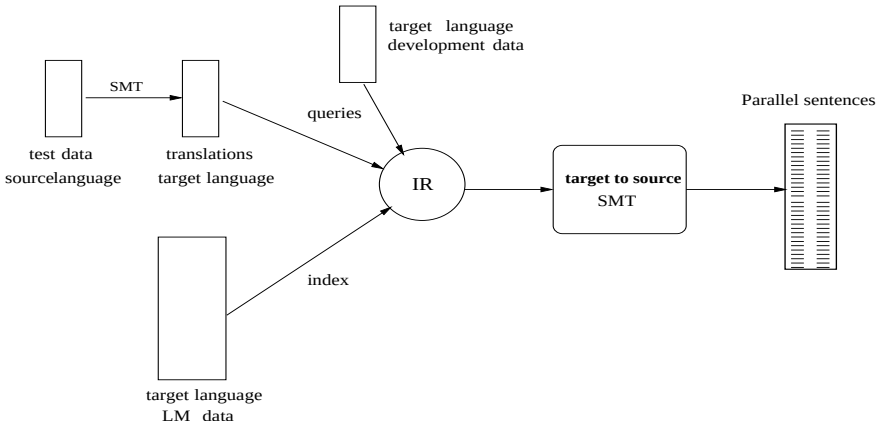

Figure 4.2: High level architecture of our approach when applied to monolingual corpora.

A direction of research very much related to our work is by [Ueffing, 2006] who translate the source side of test data and use the most reliable translations (using the SMT confidence scores) to train an additional phrase table that is used jointly with the generic phrase table. They term this approach *self-training*, and it adapts the translation model close to the test data. This could be also seen as a mixture model with the in-domain component being built on-the-fly for each test set. In follow up work, this approach was refined [Ueffing, 2007]. In contrast, while we use the development and test data to find the related sentences from target language texts, we do not strictly adapt the system to testing data, rather we add new sentences to the training data that

are most related to testing domain and a completely new model is built.

A variation of terms have been used in the literature for the use of automatic translations as bitexts, the choice of the term relates to the task and the point of view of the experimenter. [Ueffing et al., 2007] use the term *semi-supervised learning*, while [Schwenk, 2008; Schwenk and Senellart, 2009] call their method *lightly-supervised training*, [Bojar and Tamchyna, 2011] say *reverse self-training*, whereas [Lambert et al., 2011] call it *unsupervised training*. We shall be using the term unsupervised training to refer to this phenomenon.

In phrase-based machine translations systems, the translation model is represented by a large list of all known source phrases and their corresponding target language phrases. These entries are weighted using several probabilities, e.g. phrase translation probabilities in the forward and backward direction, as well as lexical probabilities in both directions. The phrases of the phrase-table are automatically extracted from sentence aligned parallel data. Adaptation of the translation model could be performed by modifying the probability distribution of the existing phrases without necessarily modifying the entries. The idea is to increase the probabilities of translations that are appropriate to the task and to decrease the probabilities of the other ones. By adding more domain specific data this is achieved by increasing the probability of the domain related phrases as compared to unrelated ones.

Unsupervised training as proposed previously [Bojar and Tamchyna, 2011; Lambert et al., 2011; Nicola Bertoldi, 2009; Schwenk, 2008] consists in translating all the monolingual data in the source or target language, to filter it according to some confidence measure, and to add this data to the existing human translations. We propose an efficient extension to unsupervised training by presenting an active data selection mechanism which helps to considerably reduce the computational cost, since only small amounts of data will need to be translated automatically. Also, note that this approach can also be used for domain adaptation by finding sentences most related to the training or testing domains.

4.3 Related Works

In this section we give a brief overview of research that has been done in the fields of unsupervised training and translation model adaptation. Ueffing [2007] presents extensive research based on the previous work in [Ueffing, 2006]. They show improvements in SMT using unsupervised training. They use an iterative approach by translating the source language test data in each iteration and re training the translation model. They

investigated several scoring criteria. Extending the work in [Schwenk, 2008], Schwenk and Senellart [2009] report using automatic translations of in-domain monolingual texts to adapt the translation model of an SMT system trained on out-of-domain corpus. By learning new task-specific phrase-pairs, they report an improvement of 3.5 points BLEU on the test set for the Arabic-French language pair.

Nicola Bertoldi [2009] also use automatic translations to adapt the translation model of an SMT system to work properly on another domain. They further use these translations to adapt the language model and reordering model. They trained a Spanish-English system using the UN corpus and used the Europarl data for adaptation, doing so they report a 5.5% improvement on the BLEU score on test set. Domain adaptation was also performed simultaneously for the translation, language and reordering model in [Chen et al., 2008]. [Bojar and Tamchyna, 2011] work on several European languages and use the automatic translations of target-side monolingual data to extend the vocabulary of the translation model. They use the term *reverse self-training* for their method.

Domain adaptation has more frequently been used using parallel corpora in approaches which try to make maximum use of the available training material. This is commonly done by modifying the statistical model and using a mixture model to optimize the coefficients to the adaptation domain. This was investigated in the framework of SMT by several authors, for instance for word alignment [Civera and Juan, 2007], for language modeling [Koehn and Schroeder, 2007; Zhao et al., 2004] and to a lesser extent for the translation model [Chen et al., 2008; Foster and Kuhn, 2007]. This mixture approach has the advantage that only few parameters need to be modified, the mixture coefficients.

Information retrieval has been previously used in the context of translation model adaptation by [Hildebrand et al., 2005], who use IR to find sentences similar to the test set from the parallel training data. They use the source side of the test data to find related sentences from the parallel corpora. [Lu et al., 2008] use a similar technique of using IR to select and weigh portions of existing training data to improve translation performance. While adaptation using existing parallel texts has shown this to be beneficial for translation quality, it does not create any new translation rules but only distributes the probability mass associated with the existing translation rules. Whereas, with our approach new translation rules can be learned since we ‘create‘ new adapted training data.

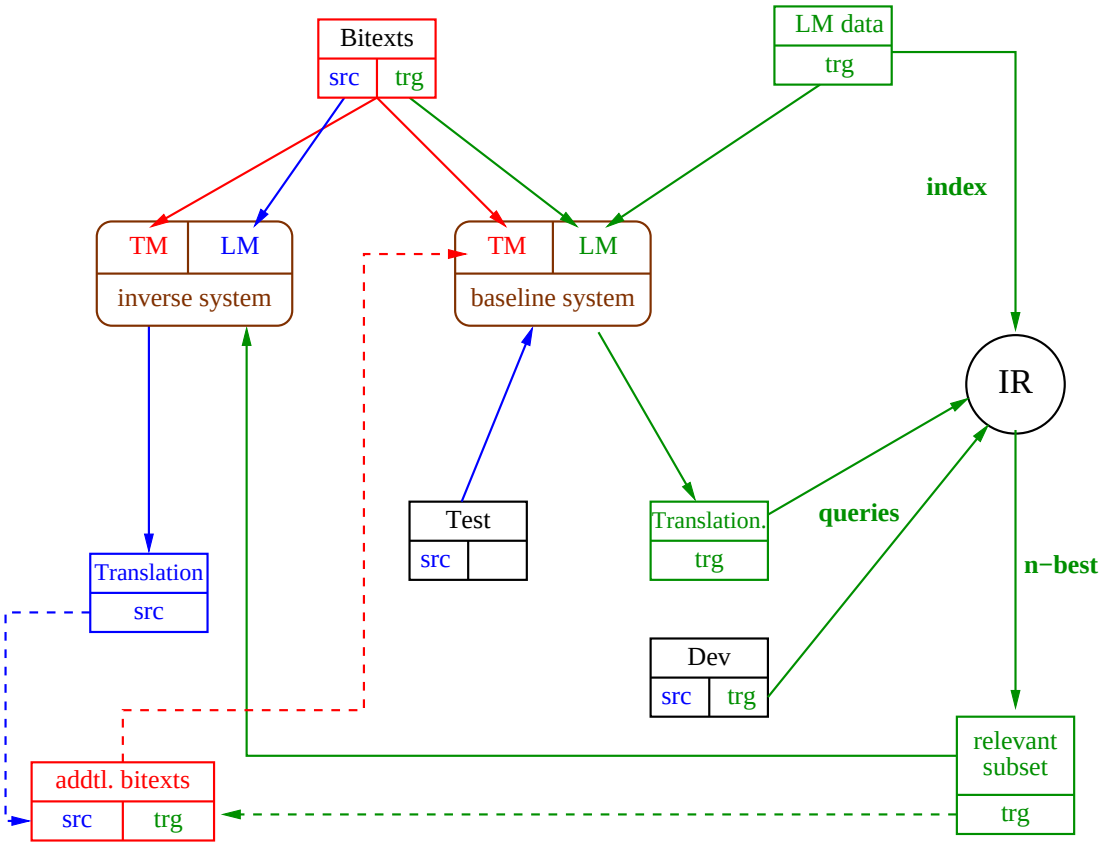

Figure 4.3: Architecture of the proposed approach.

4.4 Architecture of the approach

Figure 4.3 provides a general representation of our approach. We want to build an SMT system to translate from source to target language. For this purpose we have parallel corpora to train the translation model and monolingual data in the target language to train the language model. This data is used to build a baseline system and an additional inverse system to translate from the target to the source language. Additional source language monolingual data is used to train the LM for the inverse system. The baseline system is used to translate the test data into the target language. These translations and the reference translations of the development data are then used to retrieve relevant sentences in the large collection of language model training data. These n-best sentences are then translated back to the source language by the inverse system. Finally the obtained automatic translations together with the corresponding retrieved sentences are used as additional parallel training data and a new improved

system is built (to translate from source to target). In the research reported in thesis, we have applied this approach to translate from English to French. We present the improvements that we achieve in section 4.5.

As in the previous chapter, we used the Lemur IR toolkit for finding the matching sentences. The language model data, which was in French for our experiments was used as index. The queries for IR are the reference translations of the development data and the automatic translations of the test data as produced by the baseline system. For this task we indexed each sentence as a document, thus for each query sentence we retrieved a set of matching documents, where each document was in fact one sentence. Our indexing scheme here is different from our indexing scheme presented in the previous chapter (chapter 3, section 3.4.3.1). In that approach we indexed documents as they were in the original corpus i.e. one news article as a document composed of sentences and used the passage retrieval feature of the toolkit to retrieve sentences rather than documents. Here, our LM data has no explicit document specification, so we change our indexing scheme to fully conform to our need. We then retrieved n sentences per query. We experimented with various values of n. Note, that we can increase the size of the additional bitexts by increasing the size of n, we found the ideal value to be different for several baseline data as is presented later in section 4.5.

4.4.1 Choice of translation direction

In approaches using automatic translations as additional bitexts [Bojar and Tamchyna, 2011; Lambert et al., 2011; Nicola Bertoldi, 2009; Schwenk, 2008], the question of whether to use source language or target language adaptation data is an important one. Nicola Bertoldi [2009] raised this question and reported a better gain when adaptation data is available in the target language as compared to the source language. However for them when data is in source language they adapt only the translation model, while they adapt both the TM and LM when target language data is available. The presence of an adapted LM, doesn't solely conclude target language data to be better than source language data.

Recently, Lambert et al. [2011] showed that it is better to translate from the target to the source language instead of the inverse direction. Automatic translations can of course contain wrong translations. If we translate from source to target, these wrong translations are added to the phrase table and may be reused in future translations performed by the adapted system. However, if the automatic translations done in the target-source direction are added, the translation errors will appear in the source side

baseline	translated bitexts	Dev	Test	
		BLEU	BLEU	TER
Eparl + nc	-	26.20 (0.06)	28.06 (0.22)	56.85 (0.09)
	news fe 45M	**27.18 (0.09)**	**29.03 (0.07)**	**55.97 (0.07)**
	news ef 45M	26.15 (0.04)	28.44 (0.09)	56.56 (0.11)
Eparl + nc + subset10^9	-	26.95 (0.04)	29.29 (0.03)	55.77 (0.19)
	news fe 45M	**27.42 (0.02)**	**29.77 (0.06)**	**55.27 (0.03)**
	news ef 45M	26.75 (0.04)	28.88 (0.10)	56.06 (0.05)

Table 4.1: Translation results of the English–French systems augmented with a bitext obtained by translating news data from English to French (ef) and French to English (fe). 45M refers to the number of English running words. The values are the average over 3 MERT runs performed with different seeds. The numbers in parentheses are the standard deviation of these three values.

of the adapted phrase table. They argued that this will have less impact since it is less likely that wrong phrases will be matched when translating grammatically correct (source) sentences.

In table 4.1 we re-present results from WMT11 evaluation [Lambert et al., 2011]. The subset10^9 is a subset of the so called 10^9 French-English parallel corpus containing data crawled from Canadian and European Internet pages. This data was filtered using IBM-1 probabilities and language model scores to keep only the most reliable translations [Lambert et al., 2010]. Two subsets were built with 115M and 232M English words respectively (using two different settings of the filter thresholds).

These are the results for English-French systems adapted with news data translated from English to French (ef) and French to English (fe) direction. The experiment was repeated with two baseline corpora, the smaller baseline comprising the europarl and the news commentary bitexts and a larger baseline including the subset10^9. The results show clearly that target to source translated data are more useful than source to target translated data. The improvement in terms of BLEU score due to the use of target-to-source translated data instead of source-to-target translated data ranges from 0.5 to 0.9 points. For instance, the baseline system "eparl+nc" achieves a BLEU score of 28.06 on the test set. This could be improved to 29.03 using automatic translations in the reverse direction (French to English), while we only achieve a BLEU score of 28.44 when using automatic translation performed in the same direction as the system to be adapted.

The effect is even clearer when we try to adapt the large system "eparl+nc+subset10^9". Adding automatic translations translated from English-to-French did actually lead to a lower BLEU score (29.29 $\rightarrow$ 28.88) while we observe an improvement of nearly 0.5 BLEU in the other case. The values are the average over 3 MERT runs performed with different seeds. The numbers in parentheses are the standard deviation of these three values.

Working on English-French translation, following these lines, we chose to do IR on the French (target) monolingual data. We then translated these sentences to English using the inverse SMT system that we built.

4.4.2 Reuse of word alignments

In the previous works, for example, [Lambert et al., 2011; Schwenk, 2008; Schwenk and Senellart, 2009; Ueffing, 2006], the filtered automatic translations were added to the parallel training data and the full pipeline to build an SMT system was performed again, including word alignment with GIZA++. Word alignment of bitexts having several hundred millions of words is a very time consuming task. Table 4.2 shows the word alignment times for several baseline corpora that we worked on during this research. These have been computed using the multi-threaded version of the GIZA++ tool [Gao and Vogel, 2008]. We can see that even the smallest of these corpora, i:e *eparl+nc* takes 9 hours and 40 minutes in the alignment process. Building a new system by adding additional corpora would traditionally require the whole GIZA process to be repeated for each such experiment.

Corpus	M words	GIZA++ time
eparl+nc	58.1M	9h 40m
eparl+nc+abs+dico	60.7M	10h
eparl+nc+abs+dico+subset10^9	333.6M	43h

Table 4.2: Word alignment times taken by various corpora using mGiza with two jobs of 4 threads.

The Moses decoder provides an option to output word-to-word alignment for each sentence-translation pair, provided these word-to-word alignments are available in the phrase table. Having this additional word alignment information in the phrase table

introduces a disk usage overhead but no overhead at decoding time. So for our experiments, the word alignments from the automatic translations are added to the previously calculated alignments of the baselines texts (along with the bitexts) and a new phrase table is built. This resulted in an appreciable speed-up of the procedure.

Reusing these alignment does not just speed up the overall processing, but there have also been investigations that the alignments obtained by decoding are more suitable to extract phrases than the symmetrized word alignments produced by GIZA++. Wuebker et al. [2010] presents an approach using forced alignments and a leave-one-out technique, and to use the induced alignments to extract phrases. They report improvements with respect to word alignments produced by GIZA++. While, on the other hand, Nicola Bertoldi [2009] adapted an SMT system with automatic translations and trained the translation and reordering models on the alignments produced by Moses. They report a small drop in performance with respect to training word alignments with GIZA++. They also report a gain of almost 50% in training time. Similar ideas were also used in pivot translation [Bertoldi et al., 2008].

	Dev	**Test**	
alignment	**BLEU**	**BLEU**	**TER**
giza	27.34 (0.01)	29.80 (0.06)	55.34 (0.06)
reused giza	27.40 (0.05)	29.82 (0.10)	55.30 (0.02)
reused moses	27.42 (0.02)	29.77 (0.06)	55.27 (0.03)

Table 4.3: Results for systems trained via different word alignment configurations. The values are the average over 3 MERT runs performed with different seeds. The numbers in parentheses are the standard deviation of these three values. Translation was performed from English to French, adding 45M words of automatic translations (translated from French to English) to the baseline system "eparl+nc+subset10^9".

In WMT11 evaluation [Lambert et al., 2011], an analysis of systems built by running GIZA++ on all the data and the system trained using Moses alignments was presented. When word alignments of the baseline corpus (not adapted) are trained together with the translated data, they could be affected by the phrase pairs coming from incorrect translations. To measure this effect, we trained an additional system, for which the alignments of the baseline corpus are those trained without the translated data. For the translated data (reused moses), we re-use the GIZA++ alignments trained on all the

data. Whereas, for reused giza we do a complete training of the baseline + automatic translations, and then substitute the baseline alignments by those trained without the automatic translations, in order to check if Moses alignments did not degrade the baseline alignments.

Table 4.3 reports the results of these three alignment configurations. For these experiments 45M words of French sources and English translations were added to the *eparl+nc+subset*10^9 baseline corpus (286M words). According to the BLEU and TER scores, reusing the Moses alignments to build the adapted phrase table has no significant impact on the system performance. Given the large size of the baseline corpus (especially subset10^9), we can attribute this to the fact that the unsupervised data was small compared to the baseline. We then repeated the experiment replacing the *subset*10^9 with a smaller selection of 10^9 and arrived at the same conclusion. However, the re-use of Moses alignments saves time and resources. On the larger baseline corpus, the mGiza process lasted 46 hours with two jobs of 4 threads running and a machine with two Intel $X5650$ quad-core processors.

In the results reported in this chapter, we have made use of the word alignments obtained implicitly during the translation of the monolingual data with the Moses toolkit. In doing so for each experiment we gain considerable time.

4.4.3 Multi-pass approach

There are cases where we want to adapt our system to domains for which we only have data in the source language. A typical example is an international evaluation where we would like to adapt our system to the evaluation data. This could be done by performing cross-lingual IR which is usually implemented with a dictionary to translate some of the key words. Alternatively, we can use the unadapted system to translate the evaluation data and then use the hypothesis to retrieve relevant sentences. These sentences would be translated back to the source language and then added to the parallel training data and an adapted system would be built. This system is finally used to translate again the evaluation data. In summary we have a two pass system.

We don't expect a big difference in the sentences retrieved by IR using automatic translations as queries or the reference translations since IR techniques are usually invariant to typical errors of SMT systems like morphology or word order. Also, in the previous chapter (chapter 3, section 3.12) we show that the quality of the retrieved sentences does not depend upon the quality of translations used for IR.

This two pass approach is an interesting extension of self-learning [Ueffing, 2006].

Instead of using directly the automatic translations, we use them to perform IR and retrieve similar sentences in larger amounts of data, which are then themselves translated back to the source language. By these means we get more adaptation data and our adaptation technique suffers less from translation errors since their appear on the source side. We also create a completely new systems instead of building a small additional phrase-table as in self-learning of [Ueffing, 2006].

Finally, there are nice similarities with language model adaptation techniques used in speech recognition. Chen et al. [2001] for instance, use the hypothesis of the speech recognizer to retrieve related sentences by IR which are then used to build an additional LM. This LM is used in the second pass of the speech recognizer.

4.5 Experimental Evidence

In this section we provide experimental evidence by presenting the results using our approach. We first describe the experimental resources used in our experiments, followed by a brief description of baseline SMT systems (section 4.5.2). We then present in detail the results of our experiments in sections 4.5.3 and 4.5.4.

4.5.1 Experimental Resources

We consider the translation of scientific documents from English into French. This task is part of a larger project that aims in providing interactive machine translation of research papers between these two languages COSMAT.[1] We work on scientific papers in the area "Computer Science" only. We extracted in-domain monolingual and parallel data from the HAL data archive of the COSMAT project. This process is explained below, before describing the out-of-domain resources used in our research.

We first converted the pdf files of computer science data to plain text (via the TEI format) using the Grobid[2] open-source converter. These documents are nearly exclusively monolingual, but the thesis from French universities must include both an abstract in French and in English. Although in some cases the two abstracts may not be strictly parallel translations or may contain translation errors, our experiments show that these abstracts turned out to be useful parallel data.

The abstracts were first aligned at the sentence level. Then training, development and test data were selected. To avoid including incorrectly aligned sentence pairs in the development and test data, the selection was performed based on the cost of the

[1] http://www.cosmat.fr
[2] http://grobid.no-ip.org

IBM Model 1 [Brown et al., 1993] for each sentence pair. The development and test data sets were chosen at random within the subset of sentence pairs whose IBM 1 score satisfied a certain criterion. About 50k running words were selected. The rest of the data was used as training set. We also had available various bilingual dictionaries that partly contain domain specific vocabulary. Table 4.4 gives an overview of the available data.

Corpus	English	French
Bitexts:		
Out-domain:		
Europarl (eparl)	50.5M	54.4M
News Commentary (nc)	2.9M	3.3M
Crawled (subset10^9)	667M	794M
In-domain:		
Thesis abstracts (abs)	1.4M	1.6M
Various dictionaries (dico)	0.9M	1M
Development set	25.8K	28.7K
Test set	26.1K	29.2K
Monolingual data:		
LDC Gigaword	4.1G	920M
Crawled news	2.6G	612M
Scientific articles	54M	19M

Table 4.4: Data available for the English/French COSMAT task. Only the thesis abstracts and to some extent the dictionaries may be considered as in-domain data.

The parallel out-of-domain data used were the Europarl corpus (European Parliament proceedings), the News-commentary corpus (quality commentary articles about the news) and a selection[1] of the French–English 10^9 corpus (mostly crawled from bilingual Internet sites).

The monolingual data used to train our language model were the monolingual ver-

[1]We applied the same two filters as [Lambert et al., 2011] to select this subset. The first one is a lexical filter based on the IBM model 1 cost of each side of a sentence pair given the other side, normalized with respect to both sentence lengths. This filter was trained on a corpus composed of Europarl, News-commentary, and United Nations bitexts. The other filter is an n-gram language model cost of the target sentence, normalized with respect to its length. This filter was trained with all monolingual resources available except the 10^9 data.

Human bitexts	M words (French)	Dev	Test
eparl+nc	58.1M	33.52	32.80
eparl+nc+abs	59.7M	38.66	38.19
eparl+nc+abs+dico	60.7M	38.96	38.31
eparl+nc+abs+dico+subset10^9	333.6M	39.35	38.13

Table 4.5: The baseline systems used in our experiments.

sion of the bitexts, the news corpus provided at WMT 2011 (crawled from the web) and the LDC Gigaword collection.

4.5.2 Baseline SMT systems

We built standard phrase-based SMT systems using the default settings of the Moses toolkit [Koehn et al., 2007]. The multi-threaded version of the GIZA++ tool was used to compute word alignments. The parameters of Moses are tuned on the development data using the MERT tool.

A 4-gram back-off language model was trained on all the available monolingual data. This data was split into several parts, individual LMs were trained using modified Kneser-Ney smoothing as implemented in the SRILM toolkit [Stolcke, 2002] and then interpolated to get one huge LM. The corresponding interpolation coefficients were calculated to optimize the perplexity on the development data using the usual EM procedure. In addition, we have observed small improvements by keeping all observed n-grams, i.e. using a cut-off value of 1.

The BLEU scores are calculated with the tool *multi-bleu.perl* as provided in the Moses toolkit. Scoring is case sensitive and includes punctuation. Table 4.5 shows a summary of various baseline system scores. The starting point is a system trained on the Europarl and News Commentary bitexts (58.1M words) with a BLEU score of 32.80 on the test data. Adding a very small amount of in-domain data from bilingual abstracts of French PhD thesis (corpus "abs") substantially improves the BLEU score to 38.19. The use of the dictionaries has only a small impact (BLEU score of 38.31 on the test data), as a huge subset of the 10^9 bitexts (BLEU score of 38.13). Note, that though the huge subset has a relatively better score on the development data, it doesn't perform well on the test data. This clearly indicates that in-domain parallel training data can't be easily substituted by large amounts of generic parallel data.

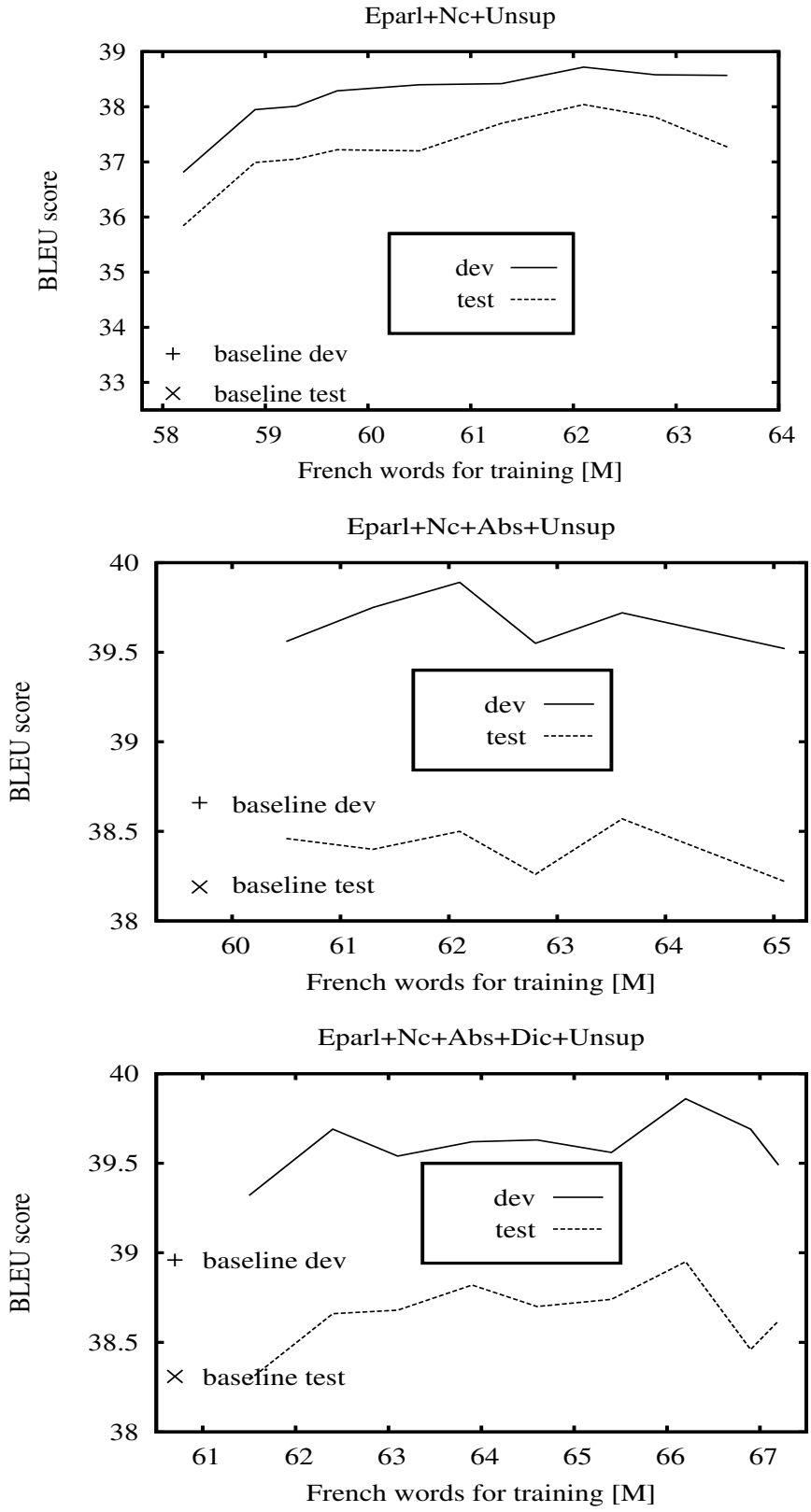

Figure 4.4: BLEU scores in function of our automatically translated French words added to the already available human translated corpora. The number of added words are increased by changing the n-best sentences selected.

4.5.3 Adding n-best automatic translations

In this section we report the results of our methods to use monolingual data in target language to adapt the translation model. We had around 19M words of scientific texts in French. We performed IR on these texts and retrieved n-best sentences matching the development and test data. We performed SMT experiments with different settings, i.e. varying the number of sentences returned by IR that will be translated back to the source language.

Bitexts	**M words (French)**	**Dev**	**Test**
eparl+nc	58.1M	33.52	32.80
eparl+nc+50-best IR	62.1M	**38.72**	**38.04**
eparl+nc+abs	59.7M	38.66	38.19
eparl+nc+abs+30-best IR	62.1M	**39.89**	**38.50**
eparl+nc+abs+dico	60.7M	38.96	38.31
eparl+nc+abs+dico+70-best IR	65.8M	**39.86**	**38.95**
eparl+nc+abs+dico+subset10^9	333.6M	39.35	38.13
eparl+nc+abs+dico+subset10^9+30-best IR	335.8M	**39.80**	**38.76**

Table 4.6: Adding the n-best unsupervised sentences to baseline corpora.

In figure 4.4 we show the general trend obtained when adding our unsupervised bitexts to the already available human translated corpora. The number of added words are increased by changing the n-best sentences selected. We see that for the initial baseline, *eparl+nc*, adding the n-best sentences is almost a continuous gain in performance until around 63.5M words when the performance starts to decline. Whereas, in the case of the other two baselines *eparl+nc+abs* and *eparl+nc+abs+dico*, which already have some in-domain data, the curves have some points where performance drops and then rises back again. Also, for these two baselines, we get a better gain on development set as compared to the test set, especially for the case of *eparl+nc+abs+dico*, where adding 0.8M of automatic translations we practically get no gain on the test set. In table 4.6 we report the results for each configuration which showed best performance

on the development data.

The first appealing result is that we seem to be able to achieve basically the same result by adding 4M words of our automatic translations than adding 1.4M words of real in-domain parallel data. This is the case for the *eparl+nc* corpus plus 4M words corresponding to the translations of the 50-best sentences retrieved for each query. Adding our automatic translations we get an improvement of 5.24 BLEU points, whereas adding real in-domain parallel data gives an improvement of 5.39 BLEU points.

Even, when this in-domain parallel data is available, we can still achieve improvements by unsupervised training, this is evident from 3rd and 4rth row in table 4.6, where we get an increase of 1.23 and 0.31 BLEU points on dev and test data respectively. Having nice results, we try to improve the system trained on the bitexts *eparl+nc+abs+dico* (60.7M words). Here we have used all the available in-domain bitexts, i.e. PhD thesis abstracts and various dictionaries. Adding about 5.5M words of selected automatic translations, we achieve an improvement in the BLEU score of about 0.5 points on the test set (38.31 $\rightarrow$ 38.95).

Finally, we add our unsupervised sentences to the big data set, *eparl+nc+abs+dico+subset*10^9. Interestingly, adding just 2.2M words corresponding to the translations of 30-best sentences, we get a gain of 0.63 BLEU points on test data. We did not do experiments with different n-best sizes for this big data set.

4.5.4 Adding automatic translations based on relative difference

We experimented with an IR score based sentence selection method. The idea is based on the observation that not all the results returned by the IR process are always related to the query sentence. Since the results are sorted by their similarity score, thus it is often at the end that the most unrelated sentences exist. We try to exclude these not so related sentences by comparing their similarity score with the first sentence (the best match). Thus, if the relative difference is greater than a certain threshold, we discard the sentence. So, instead of using the n-best hypothesis returned by IR, we select the sentences returned by each query based on a threshold on the relative difference with respect to the best answer. Therefore, the number of sentences translated back to English (source language) varies for each IR query.

Figure 4.5 shows the results of our two proposed sentence selection techniques, i.e. n-best selection and selection by relative difference. These are obtained by adding the automatic translations using the two techniques to the baseline *eparl+nc+abs+dico*. The crosses show the baseline scores. From these figures, generally the new selection

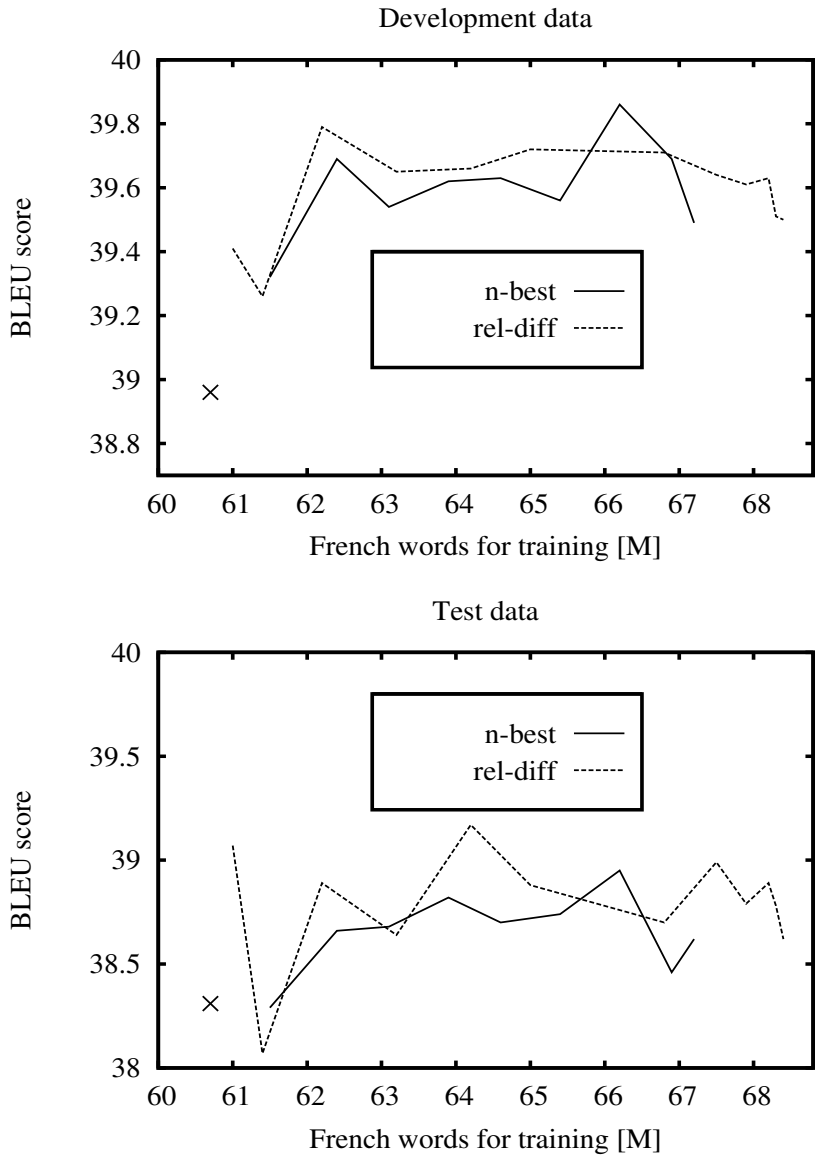

Figure 4.5: Comparison of our two adaptation techniques. Trend on BLEU score obtained by adding the automatic translations using the two techniques to the baseline *eparl+nc+abs+dico*

scheme based on relative difference seems better. Theoretically, selecting based on relative goodness of the sentences, we get the power to fine tune the process of sentence selection.

Figure 4.6 shows the number of words selected by each scheme. As can be seen, with selection by relative difference around 14% most of the sentences are selected.

Since we performed IR on in-domain French texts so the relative difference among the good and bad sentences is not very large. In case of finding related sentences from out of domain texts, the relative difference amongst the sentences would be large and a greater relative percentage might be needed to select all the sentences.

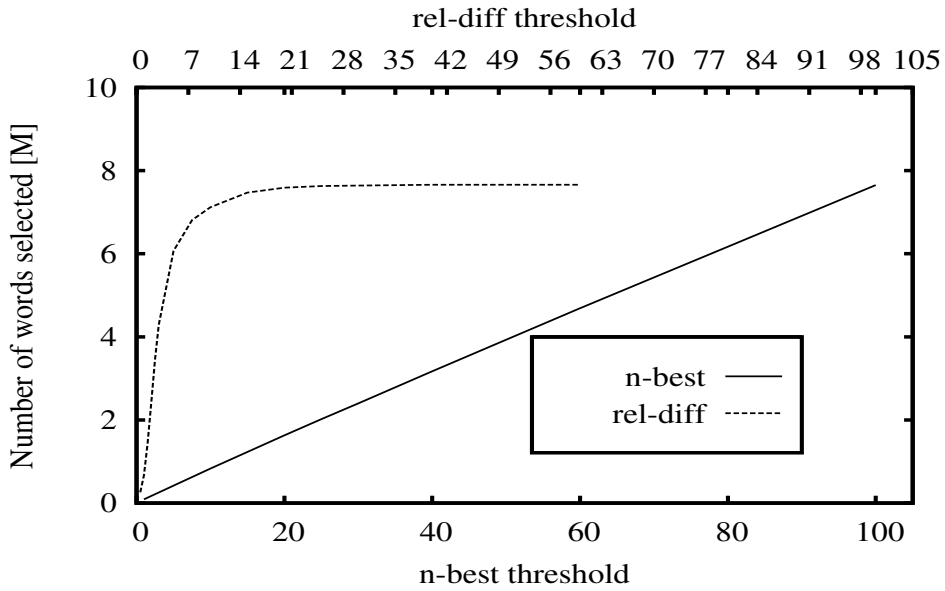

Figure 4.6: Comparison of our two adaptation techniques in terms of number of French words selected for each threshold.

Generally in the IR framework, when we retrieve n-best sentences, the relative number of reliable or related sentences vary for each query sentence. It is often the case that for one query only the top few sentence are good (e.g.. top 15%), whereas for some other query most of the sentences are related (e.g.. top 65%). Rather than taking the n-best retrieved sentences for each query, we select the number of sentences to take from each result based on its relative difference with respect to the best answer. By selecting more of the good sentences and less of the non-related sentences, we are able to do a refined selection, even from the sentences returned by IR.

4.5.5 A crude comparison of unsupervised selection schemes

In table 4.7, we present a crude comparison of the three unsupervised data selection schemes, i,e. unsupervised training as in [Schwenk, 2008], and our two schemes presented above. We call this a crude comparison because we did not perform systematic experiments to find the optimal threshold of the sentence normalized log-score of the

Bitexts	**M words (French)**	**Dev**	**Test**
eparl+nc+abs	59.7M	38.66	38.19
+ SMT threshold 0.75	64.8M	38.83	37.75
+ 30-best IR	62.1M	**39.89**	**38.50**
+ IR rel. diff 10	66.8M	39.72	38.67
eparl+nc+abs+dico	60.7M	38.96	38.31
+ SMT threshold 0.75	65.8M	39.02	38.90
+ 70-best IR	66.2M	**39.86**	**38.95**
+ IR rel. diff 1.5	62.2M	39.79	38.89
+IR rel. diff 2.5	64.2M	39.66	**39.17**
eparl+nc+abs+dico+subset10^9	333.6M	39.35	38.13
+ SMT threshold 0.75	338.4M	39.34	38.87
+ 30-best IR	335.8M	**39.80**	**38.76**
+ IR rel. diff 1.5	334.8M	39.62	39.08

Table 4.7: Comparison of different adaptation techniques when translating English scientific texts into French.

decoder for unsupervised training as proposed by [Schwenk, 2008], but simply used a value which seems to be a good compromise of quality and quantity of the translations, i.e. 0.75. Further, for our scheme of relative difference, we found 1.5% to be the best threshold for the baseline *eparl+nc+abs+dico* so we used only this data for the larger baseline, i.e. *eparl+nc+abs+dico+subset*10^9.

We see that the three methods are very close by, with the two new extensions proposed in this research to have a slight edge. The best BLEU score of our experiments (39.17) on the test data is obtained by choosing the sentences based on relative difference. The new methods proposed in this research always achieves results on the test data at least as good than the original method presented in previous works [Lambert et al., 2011; Schwenk, 2008]. However, our proposed approach has a much smaller complexity since we actively select the monolingual data to translate instead of blindly processing large amounts, followed by some filtering.

4.6 Conclusion and Perspectives

In this chapter we have presented an extension to unsupervised learning [Schwenk, 2008] by proposing a framework to actively select the sentences most relevant to the task. We attain the same performance as unsupervised training but with an evident reduction in complexity of the task.

We have proposed the use of IR to select the sentences most relevant to the testing domain. We use these sentences filtered by the IR score. Recently, techniques were proposed to select the most appropriate parallel data, e.g. Axelrod et al. [2011] who use the cross-entropy measure. It would be certainly interesting to try that approach in our setup.

When re-use the translations (target to source) as additional bitexts, we do not use any explicit filtering, partly due to the fact that we add the translations in target-to-source direction and the errors in translations are not propagated since they occur on the source side. However, it would be still be good to add only the translations which have a good sentence normalized log-score of the decoder.

Another, interesting perspective of this work would be to replace the automatic translations with human translations. By these means we try to select to most interesting data to be translated by humans in order to improve an existing system. This is usually called active learning.

Chapter 5

Conclusions

In this dissertation we have reported efficient methods to produce parallel data from amply available data resources, i.e. the comparable and monolingual corpora. We take advantage of information retrieval techniques to build a robust and efficient framework for constructing parallel data from these resources. The presented framework is capable of processing large amounts of corpora. We show that the parallel data produced with our methods helps considerably to improve the performance of statistical machine translation systems.

This research is a valuable addition to the field of comparable corpus processing. The approach described in this research makes several important contributions. Our method is to translate the source language side of the comparable corpus and use the translations to find the corresponding parallel sentences from the target language side of the comparable corpus. Thus, starting with small amounts of sentence aligned bilingual data to build an SMT system, large amounts of monolingual data are translated. These translations are then employed to find the corresponding matching sentences in the target language side of the comparable corpus, using information retrieval methods. Finally, simple filters are used to determine whether the retrieved sentences are parallel or not. By adding these retrieved parallel sentences to already available human translated parallel corpora we were able to improve Arabic-English and French-English SMT systems.

Contrary to the previous approaches as in [Munteanu and Marcu, 2005] which used small amounts of in-domain parallel corpus as an initial resource, our system exploits the target language side of the comparable corpus to attain the same goal. We add the target language comparable corpus to the LM data, thus the comparable corpus itself helps to better extract possible parallel sentences. The detailed comparison of our approach with their approach shows our approach comparable in terms of results

and also efficient because of its simple design.

We provide a scheme where for each query sentence we retrieve a probable matching sentence rather than matching documents. This feature simplifies further processing where simple sentence comparison measures are enough to determine the quality of sentence pair in terms of being potentially parallel or not. Cettolo et al. [2010] also report using the same scoring method, i.e. filtering the sentences based on TER score to select sentences for their parallel fragment extraction.

We have also presented an extension to the work of [Schwenk, 2008] by presenting a framework capable of actively selecting appropriate sentences to be used as bitexts. The main idea of our approach is to use information retrieval techniques to extract a small subset of sentences from the LM data that are mostly related to the task (by using the sentences of the development and test data as IR queries). These retrieved sentences are then translated back to the target language and the automatic translations are used as additional parallel training data to build a new improved SMT system. In contrast to previous work, we actively select the monolingual data which seems to be most appropriate for the task. By these means we considerably reduce the computational cost of unsupervised training of the translation model since only small amounts of data must be translated automatically.

Our research contributes a method by employing an SMT system itself and IR techniques to produce additional parallel corpora from easily available corpora. The idea of using proper SMT translations has also been used in [Do et al., 2010; Gahbiche-Braham et al., 2011; Uszkoreit et al., 2010].

The lack of parallel corpora is a major bottleneck in the development of SMT systems for most language pairs. The methods presented in this paper are a step towards the important goal of automatic acquisition of such corpora. We propose schemes for exploiting comparable and monolingual corpora which are generally available in huge amounts for many languages and domains. Our schemes use only open source tools, thus making them easy to implement for anyone. In this research, we have shown how they can be used to efficiently mine for parallel sentences.

5.1 Perspectives and possible extensions

The research presented in this thesis has a number of obvious extensions. We have shown that using simple filters like WER, TER and TERp, we can effectively determine good parallel sentence pairs given the SMT translation and the sentences retrieved by

IR. Employing other sentence comparison metrics could help to find a different (might be better) set of sentences, like Do et al. [2010] used TER, NIST, BLEU and PER for scoring their sentences and found PER to be best suited for their approach.

A limitation of our work is that we did not experiment with various IR evaluation schemes to find the related sentences from the corpora. It would be an interesting study to try multiple evaluation methods, like okapi, cosine measure, word confidence scores, etc. Furthermore, the IR toolkit that we used provides two indexing schemes, with different retrieval algorithms. Finding related sentences using different indexing schemes and experimenting with a variety of sentence similarity metrics is an obvious extension to this work.

A possible extension to the scheme presented in chapter 3 could be to use more than one sentence per query as returned by IR. Though this amounts to duplicating the source sentence, but parallel corpora often contain multiple translations. Several methods can be used to choose the potential sentences, e.g. computing a similarity score (TER, TERp, PER, WER or even cross-entropy, BLEU and NIST etc.) between the query and each retrieved sentence, and using all the sentences above a certain threshold. Another way of doing this implicitly using the IR framework would be to implement the index such that one sentence constitutes a document and then retrieve n-best results (as done for monolingual corpora (chapter 4)).

Our approach can also be extended to extract parallel sentences from huge amounts of corpora available on the web by identifying comparable articles using techniques such as [Fung et al., 2010; Hong et al., 2010; Nie et al., 1999; Resnik and Smith, 2003; Yang and Li, 2003].

In chapter 4, we have proposed the use of IR to select the sentences most relevant to the testing domain. We use these sentences filtered by the IR score. Recently, techniques were proposed to select the most appropriate parallel data, e.g. Axelrod et al. [2011] who use the cross-entropy measure. It would be certainly interesting to try that approach in our setup.

Another interesting perspective of the work in unsupervised learning would be to replace the automatic translations with human translations. By these means we try to select to most interesting data to be translated by humans in order to improve an existing system. This is usually called active learning.

Appendix A

Publications

- Sadaf Abdul-Rauf and Holger Schwenk, *Parallel sentence generation from comparable corpora for improved SMT*, Machine Translation, pages 1-35, 2011.

- Holger Schwenk and Sadaf Abdul-Rauf, *LIUM's Statistical Machine Translation System for the NTCIR Chinese/English Patent Translation Task*, NTCIR9 Proceedings, to appear 2011.

- Patrik Lambert, Holger Schwenk, Christophe Servan and Sadaf Abdul-Rauf, *Investigations on Translation Model Adaptation Using Monolingual Data*, Empirical Methods in Natural Language Processing, Workshop on statistical Machine Translation (EMNLP/WMT), Edinburgh, 2011.

- Holger Schwenk, Patrik Lambert, Loic Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, Kashif Shah, *LIUM SMT Machine Translation Systems for WMT 2011*, Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011.

- Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. *LIUM SMT machine translation system for WMT 2010*, In Proceedings of the Fifth Workshop on Statistical Machine Translation, pages 121-126. Association for Computational Linguistics, 2010.

- Sadaf Abdul-Rauf and Holger Schwenk, *On the use of Comparable Corpora to improve SMT performance*, EACL 2009, pages 16-23.

- Sadaf Abdul-Rauf and Holger Schwenk. 2009. *Exploiting Comparable Corpora with TER and TERp*, 2nd Workshop on Building and Using Comparable Corpora 2009, pages 46-54.
- Holger Schwenk, Sadaf Abdul-Rauf, Loic Barrault and Jean Senellart, *SPE and AMT machine Translation systems for WMT'09*, Proceedings of the Fourth Workshop on Statistical Machine Translation, March 2009, pages 130-134.
- Holger Schwenk, Yannick Esteve and Sadaf Abdul-Rauf, *The LIUM Arabic/English statistical machine translation system for IWSLT 2008*, Proceedings of IWSLT 2008, pages 63-68.

Bibliography

Sadaf Abdul-Rauf and Holger Schwenk. Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 46–54, Singapore, August 2009a. Association for Computational Linguistics. 48

Sadaf Abdul-Rauf and Holger Schwenk. 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*, EACL '09, pages 16–23, Athens, Greece, April 2009b. 22, 48, 81

Sadaf Abdul-Rauf and Holger Schwenk. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, pages 1–35, 2011. 22

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. Dan Melamed, F. J. Och, D. Purdy, N.A. Smith, and D. Yarowsky. Statistical machine translation. In *Final Report, JHU Workshop*, Baltimore, MD, December 1999. 38

I. Alegria, N. Ezeiza, and I. Fernandez. Named entities translation based on comparable corpora. In *Proceedings of Multi-Word-Expressions in a Multilingual Context Workshop in EACL*, Trento, Italy, 2006. 48

Vamshi Ambati and Stephan Vogel. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 62–65, Los Angeles, June 2010. 16

Daniel Andrade, Takuya Matsuzaki, and Junichi Tsujii. Learning the optimal use of dependency-parsing information for finding translations with comparable corpora. In

Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pages 10–18, Portland, Oregon, June 2011. Association for Computational Linguistics. 48

D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, and Louisa Sadler. *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, 1993. 20, 21

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. 111, 115

Nicholas J. Belkin and W. Bruce Croft. *Retrieval techniques*, pages 109–145. Elsevier Science Inc., New York, NY, USA, 1987. 41

Nicola Bertoldi, Mauro Cettolo, Roldano Cattoni, Boxing Chen, and Marcello Federico. Itc-irst at the 2006 tc-star slt evaluation campaign. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 19–24, 2006. 33

Nicola Bertoldi, Barbaiani Madalina, Federico Marcello, and Roldano Cattoni. Investigations on large scale lightly-supervised training for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 143–149, 2008. 100

Rohit G. Bharadwaj and Vasudeva Varma. Language independent identification of parallel sentences using wikipedia. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 11–12, New York, NY, USA, 2011. ACM. 16, 45

Michael Bloodgood and Chris Callison-Burch. Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 208–211, Los Angeles, California, June 2010. 16

B.Lu, T Jiang, K Chow, and B K. Tsou. Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 42–48, Valletta, Malta, May 2010. 48

Ondrej Bojar and Aleš Tamchyna. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. 94, 95, 97

L. Bowker and J. Pearson. *Working with Specialized Languages. A Practical Guide to Using Corpora*. Routledge, London, 2002. 44

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 858–867, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 36

P. Brown, S. Della Pietra, Vincent J. Della Pietra, and R. Mercer. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311, 1993. 19, 24, 26, 80, 103

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990. 15, 19, 24, 26

Sir E. A. Wallis Budge. *The Rosetta Stone*. British Museum Press, 1922. 19

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 70–106, Columbus, Ohio, 2008. 56

Francisco Casacuberta and Enrique Vidal. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30:205–225, June 2004. 33

Francisco Casacuberta, Enrique Vidal, and Juan Miguel Vilar. Architectures for speech-to-speech translation using finite-state models. In *Proceedings of the ACL-02 workshop on Speech-to-speech translation: algorithms and systems - Volume 7*, S2S '02, pages 39–44, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 33

Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. Mining Parallel Fragments from Comparable Texts . In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2010. 47, 48, 114

Eugene Charniak, Kevin Knight, and Kenji Yamada. Syntax-based language models for statistical machine translation. In *MT Summit IX. Intl. Assoc. for Machine Translation*, 2003. 36

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. Exploiting n-best hypotheses for smt self-enhancement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 157–160, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. 95

Langzhou Chen, Jean-Luc Gauvain, Lamel Lori, Adda Gilles, and Adda Martine. Using information retrieval methods for language model adaptation. In *Eurospeech*, pages 255–258, 2001. 102

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-1996)*, pages 310–318, 1996. 35, 37

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. 33

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33: 201–228, June 2007. 33

Yun-Chuang Chiao and Pierre Zweigenbaum. Looking for French-English translations in comparable medical corpora. In *Proceedings of American Medical Informatics Association (AMIA) Symposium*, pages 150–154, 2002. 48

Jorge Civera and Alfons Juan. Domain adaptation in statistical machine translation with mixture modelling. In *Second ACL Workshop on Statistical Machine Translation, WMT 07*, pages 177–180, June 2007. 95

Daniel Dechelotte. *Traduction automatique de la parole par methodes statistiques*. PhD thesis, Universite Paris-Sud 11, Faculte des sciences drsay, December 2007. 38

Louise Deléger and Pierre Zweigenbaum. 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, Singapore, August 2009. Association for Computational Linguistics. 48

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977. 28

Mona Diab and Steve Finch. A statistical word-level translation model for comparable corpora. In *Proc. of Conference on Content-based Multimedia Information Access*, 2000. 48

Thi N. Do, Laurent Besacier, and Eric Castelli. A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora. In *European COnference on Machine Translation (EAMT) 2010*, Saint-Raphael (France), June 2010. 18, 48, 50, 114, 115

Thi-Ngoc-Diep Do, Viet-Bac Le, Brigitte Bigi, Laurent Besacier, and Eric Castelli. Mining a comparable text corpus for a Vietnamese - French statistical machine translation system. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 165–172, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. 48, 81

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. 40

Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A survey of current paradigms in machine translation. *Advances in Computers*, 49:2–68, 1999. 19

Andreas Eisele and Jia Xu. Improving machine translation performance using comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 35–41, Valletta, Malta, May 2010. 51

G. Foster and R. Kuhn. Mixture-model adaptation for SMT. In *Empirical Methods in Natural Language Processing*, pages 128–135, 2007. 95

Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 236–243, Stroudsburg, PA, USA, 1995a. Association for Computational Linguistics. 44

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183, 1995b. 48

Pascale Fung and Percy Cheung. Mining Very-Non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP*, pages 57–63, Barcelona, Spain, July 2004. Association for Computational Linguistics. 44, 48, 50, 81

Pascale Fung and Kenneth Ward Church. K-vec: a new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, pages 1096–1102, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. 17

Pascale Fung and Kathleen Mckeown. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *In Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 81–88, 1994. 44

Pascale Fung and Lo Yuen Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 414–420, Montreal,Canada, 1998. 48

Pascale Fung, Emmanuel Prochasson, and Simon Shi. Trillions of comparable documents. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 26–34, Valletta, Malta, May 2010. 16, 45, 49, 115

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, and François Yvon. Two ways to use a noisy parallel news corpus for improving statistical machine translation. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 44–51, Portland, Oregon, June 2011. Association for Computational Linguistics. 18, 48, 51, 81, 114

William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993. ISSN 0891-2017. 17, 22, 64

Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. 99

E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Déjean. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 48

Ulrich Germann. Building a statistical machine translation system from scratch: how much bang for the buck can we expect? In *Proceedings of the workshop on data-driven methods in machine translation*, pages 1–8, Toulouse, France, 2001. Association for Computational Linguistics. 16

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 228–235, Morristown, NJ, USA, 2001. Association for Computational Linguistics. 38

Kevin Gimpel and Noah A. Smith. Generative models of monolingual and bilingual gappy patterns. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. 25, 30

Masahiko Haruno and Takefumi Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 131–138, Morristown, NJ, USA, 1996. Association for Computational Linguistics. 22

S. Hewavitharana, B. Zhao, A. S. Hildebrand, M. Eck, C. Hori, S. Vogel, and A. Waibel. The CMU Statistical Machine Translation System for IWSLT 2005. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2005. 33

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. Adaptation of the translation model for statistical machine translation based on information

retrieval. In *Proceedings of the Meeting of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, 2005. 95

Hieu Hoang and Philipp Koehn. Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 58–65, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. 38

Gumwon Hong, Chi-Ho Li, Ming Zhou, and Hae-Chang Rim. An empirical study on web mining of parallel data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 16, 45, 49, 115

Eduard Hovy, Margaret King, and Andrei Popescu-Belis. Principles of context-based machine translation evaluation. *Machine Translation*, 17:43–75, August 2002. ISSN 0922-6567. 38

Fei Huang, Ying Zhang, and Stephan Vogel. Mining key phrase translations from web corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 483–490, Vancouver, Canada, 2005. 48

John Hutchins. Example-based machine translation: a review and commentary. *Machine Translation*, 19:197–211, December 2005. 21

W. John Hutchins. Machine translation: A concise history, 2007. 19

W. John Hutchins and Harold L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992. 20, 21

Tatsuya Ishisaka, Kazuhide Yamamoto, Masao Utiyama, and Eiichiro Sumita. Development of a Japanese-English software manual parallel corpus. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 254–259, Ottawa, Ontario, Canada, August 26-30 2009. 16, 45, 49

Heng Ji. 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. Association for Computational Linguistics, August 2009. 48

S. Roukos K. Papineni and R. Ward. Maximum likelihood and discriminative training of direct translation models. In *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 189–192, 1998. 34

Hiroyuki Kaji. Word sense acquisition from bilingual comparable corpora. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, pages 32–39, Edmonton, Canada, 2003. 48

Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, 19:121–142, March 1993. 22

Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347, September 2003. 16, 45

Alexandre Klementiev and Dan Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 82–88, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 48

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, 2006. 15

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. 9, 23, 32

Philipp Koehn and Kevin Knight. Learning a translation lexicon from monolingual corpora. In *In Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002. 48

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Second ACL Workshop on Statistical Machine Translation*, pages 224–227, June 2007. 95

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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 - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 30

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Meeting of the Association for Computational Linguistics*, pages 177–180, 2007. 33, 57, 104

Tadashi Kumano, Hideki Tanaka, and Takenobu Tokunaga. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 95–103, Sweden, September 2007. 48

Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. LIUM SMT machine translation system for WMT 2010. In *Proceedings of the Fifth Workshop on Statistical Machine Translation*, July 2010. 12, 79, 80, 98

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. 92, 94, 97, 98, 99, 100, 103, 110

Adam Lopez. Statistical machine translation. *ACM Computing Surveys, Article 8*, 40 (3):1–49, August 2008a. 16, 23

Adam David Lopez. *Machine Translation by pattern matching*. PhD thesis, Institute for Advanced Computer Studies, University of Maryland, 2008b. 35

Yajuan Lu, Jin Huang, and Qun Liu. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2008. 95

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 1 edition, July 2009. 40, 58

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. Spmt: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia, July 2006. Association for Computational Linguistics. 36

Hiroshi Masuichi, Raymond Flournoy, Stefan Kaufmann, and Stanley Peters. A bootstrapping method for extracting bilingual text pairs. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 1066–1070, Saarbrücken, Germany, 2000. 49

Evgeny Matusov, Zens R, Vilar D, Mauser A, Popovic M, Hasan S, and H. Ney. The RWTH machine translation system. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, 2006. 33

Anthony McEnery and Zhonghua Xiao. Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist. Translating Europe*, chapter XX. Multilingual Matters, Clevedon, UK, 2007. 44

Emmanuel Morin and Emmanuel Prochasson. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 27–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 48

Dragos Stefan Munteanu. *Exploiting Comparable Corpora*. PhD thesis, University of southern California, California, 2006. 9, 46, 47

Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005. 12, 17, 18, 48, 50, 53, 54, 56, 59, 60, 78, 81, 85, 86, 87, 113

Dragos Stefan Munteanu and Daniel Marcu. 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*, pages 81–88, Sydney, Australia, 2006. 47, 48, 81

Marcello Federico Nicola Bertoldi. Domain adaptation for statistical machine translation. In *Forth ACL Workshop on Statistical Machine Translation*, pages 182–189, 2009. 94, 95, 97, 100

Jian Nie, Michel Simard, Pierre Isabelle, and Richard Dur. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 74–81, 1999. 16, 45, 49, 115

Douglas W. Oard. Alternative approaches for cross-language text retrieval. In *In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, 1997. 16

Franz Josef Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, Von der Fakultat fur Mathematik, Informatik, 2002. 30

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167. Association for Computational Linguistics, 2003. 35

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 295–302, Philadelphia, Pennsylvania, 2002. 15, 34, 35

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March 2003. 30, 38

Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449, December 2004. 30

Franz Josef Och, Christoph Tillmann, Hermann Ney, and Lehrstuhl Fiir Informatik. Improved alignment models for statistical machine translation. In *University of Maryland, College Park, MD*, pages 20–28, 1999a. 39

Franz Josef Och, Christoph Tillmann, Hermann Ney, and Lehrstuhl Fiir Informatik. Improved alignment models for statistical machine translation. In *University of Maryland, College Park, MD*, pages 20–28, 1999b. 32

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. Syntax for statistical machine translation. 2003. 33

Paul Ogilvie and Jamie Callan. Experiments using the Lemur toolkit. In *In Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 103–108, 2001. 43, 59, 85

Pablo Gamallo Otero and Isaac Gonzalez Lopez. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using*

Comparable Corpora: from Parallel to Non-parallel Corpora, pages 21–25, Valletta, Malta, May 2010. 16, 45

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 38, 40

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4): 247–266, 2006. 48

Bennison Peter and Bowker Lynne. Designing a tool for exploiting bilingual comparable corpora. In *2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000. 48

Katharina Probst. *Learning Transfer Rules for Machine Translation with Limited Data.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A., 2005. 21

Emmanuel Prochasson. *Comparable Corpora in Cross-Language Information Retrieval.* PhD thesis, cole Doctrale STIM, UFR Sciences & Techniques, Universit de Nantes, 2009. 9, 44

Chris Quirk, Raghavendra Udupa, and Arul Menezes. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*, pages 377–384, Copenhagen, Denmark, September 2007. 47, 48

A.G. Ramis. *Introducing Linguistic Knowledge into Statistical Machine Translation.* PhD thesis, Universitat Polit'ecnica de Catalunya, TALP Research Center, Barcelona, 2006. 21

Reinhard Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322, Cambridge, Massachusetts, 1995. 48

Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003. 16, 45, 49, 115

Brian Roark, Murat Saraclar, and Michael Collins. Discriminative n-gram language modeling. *Comput. Speech Lang.*, 21, April 2007. 35

Roni Rosenfeld. Statistical language modeling and n-grams, January 1997. URL `www.cs.cmu.edu/afs/cs/academic/class/11761-s97/WWW/tex/Ngrams.ps`. 37

Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2*, ACL '03, pages 141–144, Sapporo, Japan, 2003. 48

Gerard Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971. 41

Holger Schwenk. Continuous space language models. *Computer Speech and Language*, 21:492–518, July 2007. 36

Holger Schwenk. Investigations on large scale lightly-supervised training for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, page 18289, 2008. 18, 80, 91, 92, 94, 95, 97, 99, 109, 110, 111, 114

Holger Schwenk and Jean Senellart. Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit*, 2009. 94, 95, 99

Holger Schwenk, Daniel Déchelotte, and Jean-Luc Gauvain. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, 2006. 36

Holger Schwenk, Patrik Lambert, Loic Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, and Kashif Shah. Lium's smt machine translation systems for wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. 79

Claude Elwood Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948. 19

Claude Elwood Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656?715, October 1949. 19

Claude Elwood Shannon. Prediction and entropy of printed english. *Bell Systems Technical Journal*, pages 50–64, 1951. 19

Li Shao and Hwee Tou Ng. Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 48

Serge Sharoff, Bogdan Babych, and Anthony Hartley. Using collocations from comparable corpora to find translation equivalents. In *5th edition of International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006. 48

Mitsuo Shimohata and Eiichiro Sumita. Acquiring synonyms from monolingual comparable texts. In *IJCNLP*, pages 233–244, 2005. 48

Yusuke Shinyama and Satoshi Sekine. Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 65–71, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 47, 48

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. Translating with non-contiguous phrases. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 755–762. Association for Computational Linguistics, 2005. 25, 30

Jason R. Smith, Chris Quirk, and Kristina Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 16, 45

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, 2006. 39, 65

Matthew Snover, Bonnie Dorr, and Richard Schwartz. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing, EMNLP '08, pages 857–866, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. 48

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 259–268, Athens, Greece, March 2009. Association for Computational Linguistics. 40, 65

Harold Somers. Review article: Example-based machine translation. *Machine Translation*, 14:113–157, 1999. 21

Fei Song and Bruce W. Croft. A general language model for information retrieval. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321, New York, NY, USA, 1999. ACM Press. 35

Tao Sproat Richard and ChengXiang Zhai. Named entity transliteration with comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 73–80, Sydney, Australia, 2006. 48

Andreas Stolcke. SRILM - an extensible language modeling toolkit. In *International Conference on Speech and Language Processing*, pages II: 901–904, 2002. 104

Tuomas Talvensaari. *Comparable Corpora in Cross-Language Information Retrieval*. PhD thesis, university of Tampere, Tampere 2008, 2008. 42

Christoph Tillmann and Hermann Ney. Word re-ordering and dp-based search in statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, pages 850–856, Morristown, NJ, USA, 2000. Association for Computational Linguistics. 38

Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguist.*, 29: 97–133, March 2003. 38

Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9:187–222, July 1991. 43

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. Mint: A method for effective and scalable mining of named entity transliterations from

large comparable corpora. In *EACL*, pages 799–807. The Association for Computer Linguistics, 2009. 48

Nicola Ueffing. Using monolingual source-language data to improve MT performance. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, page 17481, 2006. 91, 93, 94, 99, 101, 102

Nicola Ueffing. Transductive learning for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 25–32, 2007. 93, 94

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21:77–94, June 2007. 94

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109, Beijing, China, 2010. 51, 114

Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 72–79, Sapporo, Japan, 2003. 48, 50

Stephan Vogel, Hermann Ney, and Christoph Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, pages 836–841, Morristown, NJ, USA, 1996. Association for Computational Linguistics. 29

Rui Wang and Chris Callison-Burch. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 52–60, Portland, Oregon, June 2011. Association for Computational Linguistics. 47, 48

Ye-Yi Wang and Alex Waibel. Decoding algorithm in statistical machine translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL-35, pages 366–372, Morristown, NJ, USA, 1997. 38

Warren Weaver. *Machine Translation of Languages: fourteen essays*. MIT Press Cambridge, 1955. 19

John S. White. The arpa mt evaluation methodologies: Evolution, lessons, and further approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205, 1994. 38

Dekai Wu. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *In Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, 80–87, Las*, pages 80–87, 1994. 17

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–403, September 1997. 33

Dekai Wu and Pascale Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Proceedings (IJCNLP 05)*, volume 3651 of *Lecture Notes in Computer Science*, pages 257–268, South Korea, 2005. Springer. 48, 50, 81

Dekai Wu, , Dekai Wu, and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL*, pages 1408–1414, 1998. 36

Joern Wuebker, Arne Mauser, and Hermann Ney. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 475–484, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 100

Saralegi Xabier, San Vicente Iflaki, and L.D.L Maddalen. Mining term translations from domain restricted comparable corpora. In *24th Conference of the Spanish Society for Natural Language Processing (SEPLN)*, pages 273–280, Madrid, Spain, 2008. 48

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530. Association for Computational Linguistics, 2001. 33

Christopher C. Yang and Kar Wing Li. Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):730–742, 2003. ISSN 1532-2882. 50, 115

Omar Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *ACL*, pages 1220–1229. The Association for Computer Linguistics, 2011. 16

Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *German Conference on Artificial Intelligence*, pages 18–32, 2002. 30, 32

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. Automatic acquisition of Chinese-English parallel corpus from the web. In *Proceedings of 28th European Conference on Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 420–431. Springer, 2006. 16, 45, 49

Bing Zhao and Stephan Vogel. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–, Maebashi, Japan, 2002. IEEE Computer Society. 48, 50

Bing Zhao, Matthias Eck, and Stephan Vogel. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, 2004. 95