

Two-level Parallel Text Extraction from Comparable Corpora

Xiaojun Zhang

Supervisor: Prof. Dr. Hans Uszkoreit
Dr. Maxime Amblard



26th October, 2011

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 26 Oct, 2011

Signature _____

Acknowledgement

Honestly, the past two years are the best time of my life till now, which includes my pursuit of the master's degree. I am lucky to have the people who gave me help, love and happiness. Now, I am glad to take this chance to mention their names and express my gratitude.

First, I would like to thank the committee of The European Masters Program in Language and Communication Technologies (LCT), who provided me generous financial support. Especially, I appreciate Prof. Hans Uszkoreit and PD Dr. Valia Kordoni, who offered me such a fantastic opportunity to join this program, introduced me into the world of language technology, and made an impact on my life with computational linguistic science.

I would like to express my gratitude to all those who help me to complete this thesis. I want to thank the Prof. Hans Uszkoreit and Dr. Maxime Amblard for giving me permission to commence this thesis, to do the necessary research work and to use departmental data. I am deeply indebted to my direct advisor Dr. Jia Xu from the Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) who helped, encouraged and guided me in all the period of research for and writing of this thesis.

I have furthermore to thank the coordinator of LCT Program, Mrs. Bobbye Pernice who Mr. Patrick Blackburn who have assisted with my living situations and residency in Europe. Many thanks to my former supervisor Prof. Yao Tianfang, who recommended me to this program, is always kind and generous to student. I sincerely thank Xu Feiyu and Li Hong, who afforded me a short-term internship in DFKI during 2010's summer, gave me sufficient freedom to practice as software engineer.

Special thanks to Wang Qiaoyan, who took care of my life and support me during the hard time.

To my parents, Jiang Jian and Zhang Dan.

Abstract

Providing sufficient parallel corpora essentially boosts the quality of machine translation system. Parallel texts – as the most important resource in statistical machine translation (SMT) – appear to be limited in quantity, genre and language coverage. Recent research work has focused on exploring comparable corpora, which contain bilingual information. This information compensates the existing parallel texts with additional vocabularies and phrase translation candidates. Therefore, our goal is to find a new method that exploits comparable corpora for collecting parallel data.

Munteanu and Marcu (2005, 2006) have developed two systems for mining parallel fragments and sentences from comparable corpora. However, they left several issues unsolved: 1) in the work of Munteanu and Marcu (2006), they cannot measure the correlation of extracted fragments due to a lack of metrics that could determine whether the pair is equivalent translation; 2) each of their presented solutions is restricted to just one of the two relevant levels of extraction: sentential and sub-sentential fragments. To address these problems, we propose a modified IBM Model 1 for fragment detection, and use two-level classifiers for further verifying both sentences and sub-sentential fragments; in this two-level classification step, more features are investigated and utilized for improving the accuracy of the results.

The evaluation is conducted in similar-domain and out-domain translation test corpora of the German-English language pair. We compare the proposed method with the re-implemented system of Munteanu and Marcu (2006). The results show that our framework achieves BLEU score improvements of up to 0.98 %. Moreover, our experiments on different domains and training corpus sizes show the potential of future enhancement.

Contents

Declaration	1
Acknowledgement	2
Abstract	4
1 Introduction	9
1.1 Machine Translation and Parallel Corpus	9
1.2 Comparable Corpus and its Usage	10
1.3 Motivation and Goal	10
1.4 Proposed Method	12
2 Related work	14
2.1 Explore Comparable Corpora on Different Granularities	14
2.1.1 Sentence Level	14
2.1.2 Sub-sentential Level	15
2.1.3 Other Levels	15
2.2 Explore Comparable Corpus with Statistical Models	16
2.2.1 Generative Model	16
2.2.2 Discriminative Approach	17
3 Framework of Two-level Parallel Text Extraction	18
3.1 Overview	18
3.2 Generate Candidate Sentences	19
3.3 Build Lexicon Table by Using Log-Likelihood-Ratios	20
3.4 Fragment Extraction	21
3.4.1 Signal Filter	22

3.4.2	IBM1 Extractor	23
3.5	Two-level classifier	27
3.5.1	Maximum Entropy Classification	28
3.5.2	Feature Setting	29
4	Experiment	31
4.1	Data Set	31
4.2	Lexicon Building and Sentence Candidate Selection	32
4.3	Fragment Extraction	33
4.4	Two-level Classifier	34
4.4.1	Building Dev Corpus	34
4.4.2	Training and Test	36
4.4.3	Apply into Framework	37
4.5	SMT Evaluation	38
4.5.1	Baseline System	38
4.5.2	Result and Comparison	38
5	Discussion	40
5.1	Domain Adaptation	40
5.2	Influence of the Initial Training Corpus	41
5.3	Comparison with Munteanu’s Framework	42
6	Conclusion and Future Work	44
A	Appendix	46
	Bibliography	51

List of Tables

3.1	Failure Samples from IBM1 Extractor	27
4.1	Description of Initial Corpus	31
4.2	Description of Wikipedia Raw Corpus	32
4.3	Build LLR lexicon from Giza Alignment	33
4.4	Fragment Extraction by Two Methods	34
4.5	List of fragment samples from two methods	34
4.6	Constitute of Dev Corpus	36
4.7	ME Performance on Dev Corpus: Fragment	36
4.8	ME Performance on Dev Corpus: Sentence	37
4.9	Numbers of Extractions	37
4.10	BLEU Score Evaluation with Initial Corpus News-Commentary	39
4.11	BLEU Score Evaluation with Initial Corpus Europarl	39
5.1	Domain Constitution in Balanced Evaluation Data	41
A.1	Samples of Sentence Pair Extraction	48
A.2	Samples of Fragment Pair Extraction	50

List of Figures

1.1	A Sample Result of Extracted Sentence Pair	11
1.2	A Sample Result of Fragment Extraction	11
1.3	A Parallel Fragment Extraction System in Munteanu 2006	12
3.1	An Overview of Two-level Parallel Corpus Extraction System	19
3.2	A Sample Diagram of Signal Filter	23
4.1	An Example of Phrase Table in Moses	35
5.1	Influence of the Training Corpus Size	42

Chapter 1

Introduction

1.1 Machine Translation and Parallel Corpus

Parallel texts are important resource in natural language processing (NLP). They provide indispensable training data for statistical machine translation (Brown et al., 1990; Och and Ney, 2002) and have been found helpful for other topics, such as question answering (Staff, 2003), cross-language information retrieval (Oard, 1997), and annotation projection (Diab and Resnik, 2002; Yarowsky et al., 2001a,b).

However, parallel corpora are limited in quantity, genre and language coverage. They are available only between few languages, such as English, French, German, Chinese, and Arabic. Furthermore, there are only a few parallel corpora publicly available for under-resourced languages, e.g., Romanian, Greek and Latvian. This problem always exists despite the most up-to-date effort of parallel texts mining from the web, e.g., Smith et al. (2010) designed an information extraction system to collect parallel texts from the web by considering URLs, hypertext formation, contents and monolingual texts.

Hence, other methods and approaches are necessary to solve the above problems. One potential solution to the scarcity of parallel corpora is to exploit non-parallel texts with the same topic, so-called comparable corpora. Such corpora can be obtained by taking advantage of existing methods for exploring bilingual documents via the web, for example, the entries of Wikipedia.

1.2 Comparable Corpus and its Usage

In contrast to parallel corpus, a comparable corpus is one which selects similar texts in more than one language or variety. There is no agreement on the nature of the similarity yet. However, the texts are collected according to a set of criteria, e.g., the texts are from the same domain and temporal duration; they contain overlapping information (Munteanu and Marcu, 2005, 2006).

Comparable corpus is a relatively new concept in machine translation, NLP and corpus linguistics in general. Research in comparable corpora in NLP started about 15 years ago with the first work on general lexica (Rapp, 1995) and named entity translation derivation from noisy parallel corpora (Fung, 1995). Those investigations were motivated by the scarcity of linguistic resources (namely, parallel training corpora) in statistical machine translation (SMT). The authors supposed (which has been proven by recent experiments in (Goutte et al., 2009; Munteanu and Marcu, 2005, 2006)) that the quantity of training data has an impact on the performance of SMT, and a comparable corpus can compensate for the shortage of parallel corpora.

1.3 Motivation and Goal

The latest research has also shown that adding extracted parallel phrases and sentences from non-parallel corpora to the training data improves the MT performance in view of un-translated word coverage (Hewavitharana and Vogel, 2008). Furthermore, it has been also proven experimentally that the under-resourced language pairs would gain benefit from the exploitation of comparable corpora. These experiments showed performance improvements of more than 50% by using BBC news comparable corpora for English, Arabic and Chinese over a raw baseline MT system. The authors stated that the impact of comparable corpora on SMT performance is almost equivalent to that of human-translated data with the similar size and domain (Munteanu, 2006).

Based on the above background and SMT research tendency, we are motivated to investigate comparable corpus in order to, on one hand, mining more aligned texts to improve the current best MT system; on another hand, providing possibility to detect parallel texts automatically for under-resourced language pairs.

In our task, we firstly assume that the texts are already document aligned. A

good example is the bilingual corpus from Wikipedia, two articles are written by different languages and authors. However, the contents are somehow similar. In case they describe a person, a temporal sequence of his/her growth is always applied; if a historical event is depicted, authors almost try to give explanation with basic narrative elements such as cause, involvement, development, result and impact. Naturally, the similar phrases and sentences are picked and used in these bilingual articles. Therefore, these similarity could be viewed as *parallel* and could be applied into SMT system. In our work, we'll focus on exploring these document-aligned corpus, sentences and sub-sentential fragments are aimed to be extracted.

Die 16th Street Baptist Church ist eine Kirche der Baptisten-gemeinde in Birmingham (Alabama)
 Birmingham in Alabama , die Überwiegend von Afroamerikanern frequentiert wird .

Sixteenth street baptist church is a large , predominantly African American baptist church in
 Birmingham , Alabamabirmingham in the U.S. state of Alabama .

Figure 1.1: A Sample Result of Extracted Sentence Pair


Der name des Kulturmagazins **leitet sich von** der Bezeichnung **eines hellen kräftigen rottons** aus **der Pantone-farbsystemfarbskala** ab .

 The publication 's name **refers to** a pantone color code ; in **the pantone matching system** , 032c refers to **a bold red** .

Figure 1.2: A Sample Result of Fragment Extraction

Example as in Figure 1.1 shows an aligned sentence pair from Wikipedia bilingual articles. It is obvious that this pair could provide MT system with more word-alignment, vocabulary, language model and phrase table. Therefore, the first task is to find the **sentence translations** among document-aligned corpus.

However, in general, the highly paralleled sentences are not available in comparable corpora. Another example as in Figure 1.2 describes a non-parallel sentence pair. Since this sentence pair contains plenty untranslated texts, it is unlikely that any parallel sentence detection method would consider this pair as useful. If we use these sentences in MT training process, the amount of noises might do more harm than good, because they predicate incorrect word alignment. The best way to make use of these non-parallel pairs is to extract only the translated fragments. Consequently, the other task is to find the appropriate boundaries of non-parallel sentence pairs, extract the enclosed contents and align them as **parallel fragment**.

Identifying parallel sub-sentential fragments is a difficult task. It requires the ability to recognize translational equivalence in very noisy environments, namely the sentence pairs that express different (although overlapping) content. However, a good solution to this problem would have a strong impact on parallel data acquisition efforts. Enabling the exploitation of corpora that do not share parallel sentences would greatly increase the amount of training corpus.

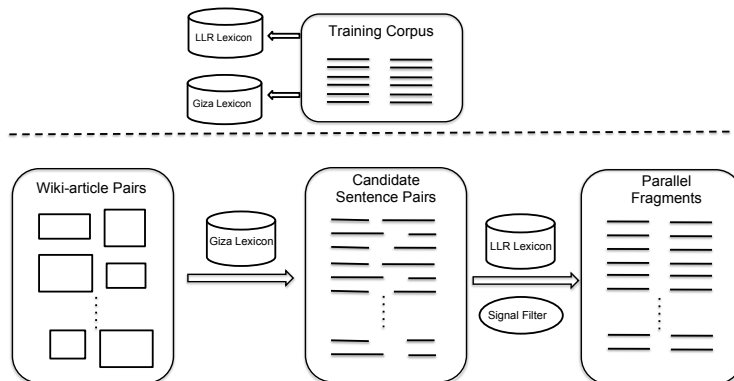


Figure 1.3: A Parallel Fragment Extraction System in Munteanu 2006

Extracting fragments among comparable corpora is the goal of Munteanu’s framework as well (Munteanu and Marcu, 2006). It could be depicted as Figure 1.3. Starting with two large monolingual corpora (a non-parallel corpus) which consist of documents, they aligned similar document pairs by using information retrieval technology. For each document pair, they generated all possible sentence pairs and passed them through a simple word-overlap-based filter and obtained parallel fragments. Additionally, they presented a method for computing a probabilistic lexicon based on the log-likelihood ratios (LLR) statistic, which produces a more reliable lexicon. They demonstrated that using this lexicon helps improve the results in the extraction step.

In the end, the extracted data together with initial parallel texts would feed into SMT system to evaluate the improvements.

1.4 Proposed Method

However, there are several unsolved issues in Munteanu’s work. We list them as follows:

- Word-based filter is too simple. To be specific, two-direction signal filter approach shows no connection guarantee between the extracted fragments. The method that source and target signal are processed separately is unreasonable; a joint analysis should be able to produce better results.
- Their system only extract sub-sentential level parallel texts. It remains the sentences as undiscovered.
- Although they demonstrate a better MT system in the end, the benchmark is consist of very small size training corpus, which is far from practical application.

To solve these problems, we propose a new framework to extract more accurate two-level data. As showed in Figure 3.1, we change two-direction signal filter as one-direction signal filter plus an IBM1 extractor, it addresses the problem of disconnection. Moreover, we modify the objective function based on IBM model 1 to eliminate length-bias, which ensure the efficiency of algorithm and constrain the time complexity in quadratic. In addition, we treat both sentences and sub-sentential fragments as candidates pairs, therefore two classifiers are built to select highly paralleled data. In this framework, the candidates are passed to a maximum entropy (ME) classifier to decide whether they are translational equivalences.

In Chapter 2, we introduce several recent advanced research works; the differences between these state-of-the-art methods are analyzed. Chapter 3 illustrates our framework in details, including the process of calculating LLR lexicon, modifying IBM Model 1 and building two-level ME classifiers. After that, we set up experiments in Chapter 4. The baseline system, Munteanu’s framework and our method are all implemented. Furthermore, we present and compare the SMT results with various sizes of training corpora and different domains. Besides, we design another set of experiment on dev-corpus to examine each feature in the classification step. Next, we discuss the phenomena and the reasons behind them in Chapter 5. In the end, we draw a conclusion and list the future work in Chapter 6.

Chapter 2

Related work

In this chapter, we investigate the current comparable corpus exploring methods with two aspects, different text levels and statistical models. Because we intend to explore comparable corpora in both sentence and sub-sentential level in our task, we list the related work sorted by different granularities. There are several methods aim to find parallel data among noisy corpus. They vary according to different metrics of "noisiness" and granularity. Furthermore, in our framework, we apply both generative process as IBM Model 1 and discriminative model as Maximum Entropy classifier. Therefore, in this section we review the research work which are related to these two statistical methods. Most of them are based on supervised or semi-supervised approaches, and utilize the extracted results into MT evaluation as benefit.

2.1 Explore Comparable Corpora on Different Granularities

2.1.1 Sentence Level

The previous most relevant research aims at mining comparable corpora for parallel sentences. The earliest efforts in this level are in Zhao and Vogel (2002) and Utiyama and Isahara (2003). Both works extend algorithms that designed to perform sentence alignment: they use dynamic programming to align sentences between documents assumed to be similar. Empirically, these approaches are only applicable to corpora that are almost "noisy-parallel", i.e., the documents which are fairly

identical, both in content and in sentence ordering. A similar extension of these method is Champollion (Ma, 2006), which uses dynamic programming as well to fetch a global optimal.

Munteanu and Marcu (2005) analyze sentence pairs without considering their context. Alternatively, they combine each source sentence with multiple possible targets, and classify all possible sentence pairs by applying lexicon features. This straightforward method enables them to find sentences from fairly dissimilar document pairs, and to handle any amount of reordering, which makes the method applicable to real comparable corpora. Nevertheless, the run time complexity turns out to be an obvious issue if they treat each possible pair as candidate.

The researches reported by Fung and Cheung (2004a,b); Fung (2004); Wu and Fung (2005) are aimed at very non-parallel corpora. They also match each source document with several target ones and examine all possible sentence pairs; but the list of document pairs is not fixed. After one pass of sentence extraction, the document pairs are enlarged with additional ones, then the system iterates. Therefore, they consider the document pairs that are not similar.

2.1.2 Sub-sentential Level

One common limitation of the above methods is that they are proposed to find only full sentences. However, in comparable corpora, fully parallel sentences are rare. Our methodology coupled with detect sub-sentential fragments. This is a difficult task, requiring the ability to recognize fragment translations even in non-parallel sentence pairs.

The work on sub-sentential fragments is introduced by Deng et al. (2006) and Xu et al. (2005). However, they obtain parallel fragments from parallel sentence pairs by chunking/splitting them and aligning the remaining parts appropriately. While we obtain them from comparable or non-parallel sentence pairs.

2.1.3 Other Levels

A lot of the work involving comparable corpora has focused on investigating word translations (Diab and Finch, 2000; Fung and Yee, 1998; Gaussier et al., 2004; Koehn and Knight, 2000; Rapp, 1999; Shao and Ng, 2004). They intend to derive a more powerful lexicon, which is crucial as well in our work.

Another related research effort is that of Resnik and Smith (2003), whose system is designed to discover parallel document pairs on the web. Document-level investigation is meaningful as well; it provides the possibility to find the appropriate document pairs among large scale corpora.

Our work falls between these two directions; we attempt to discover parallelism at the level of sub-sentential fragments and sentences.

2.2 Explore Comparable Corpus with Statistical Models

2.2.1 Generative Model

In most prior work (Brown et al., 1993; Vogel et al., 1996), generative models are used to approximate the translation process. Given a sentence in one language (the *source*, denoted as $s = s_1^m$), we can find a probability distribution over sentences in the other language (the *target*, denoted as $t = t_1^n$). While these models do allow for a certain degree of deviation between sentences, the deviations are assumed to be systematic (e.g. the German word *die* must often be inserted when generating based on an English string). In noisy comparable sentences, the situation is remarkably different: words may be inserted or deleted seemingly at random depending on what information each sentence happened to include.

Quirk et al. (2007) describes two models to handle these phenomena: a conditional model of loose translation, and a joint model of simultaneous generation. For the conditional model: First, the target sentence length is drawn according to an unspecified distribution. Next, for each target position, the position of the source word that generated this word is picked. Then the target word in that position is drawn according to the source word that generated that position. For the joint model: They generate a source-only fragment, generate a target-only fragment, or generate a bilingual fragment together. To further simplify the story, they assume that the fragments are again generated left-to-right in both the source and target sentences. Their models are able to retrieve large amount sub-sentential fragment, but those extractions unfortunately cannot achieve better result on MT evaluation.

In contrast to their work, our framework applies IBM Model 1 to generate one side fragment instead of picking two sides simultaneously. The conditional model

acts as both an objective function and a generative process to detect the proper boundary.

2.2.2 Discriminative Approach

Tillmann (2009) extends previous work on extracting parallel sentence pairs from comparable data (Munteanu and Marcu, 2005). For a given source sentence S , a maximum entropy (ME) classifier is applied to a large set of candidate target translations. Furthermore, a beam-search algorithm is used to abandon target sentences as non-parallel if they fall outside the beam. With this combination, they manage to waive pre-filtering at the document level (Fung, 2004; Munteanu and Marcu, 2005; Resnik and Smith, 2003; Snover et al., 2008; Utiyama and Isahara, 2003). The original implementation techniques are extended for the ME classifier and beam search algorithm in their work, i.e. feature function values are cached along with IBM Model 1 probabilities. Such a search-driven approach makes fewer assumptions about the data and may increase the number of extracted entities. Their experiments show the potential of digging two monolingual corpora. Our framework applies ME model as well as a refinement step. We apply almost the same features as they had, additional with our own normalized IBM 1 score.

Xu et al. (2005) develop a method to segment long parallel sentences into several short parts using a novel normalized IBM1 Model. The main idea is to use the word alignment information to find the optimal split point in a sentence pair and separate it into two pairs. This split process iterates until meets halt criteria. They demonstrate a signification BLEU score increase on German-Chinese translation system. The purpose is very different compared to our goal, but we borrow their idea of normalized IBM1 model into our framework, which helps to eliminate length bias in classification.

Chapter 3

Framework of Two-level Parallel Text Extraction

3.1 Overview

As we introduced in Figure 1.3, Munteanu and Marcu (2006) aims to extract fragments among candidate sentence pairs. A two-direction signal filter method is applied to extract f' by given target sentence E , and derive e' by given source sentence F . It remains risk that f' and e' are not actually parallel. Moreover, through investigation, we found comparable corpus would contain plenty parallel sentences, which tend to be more helpful for SMT. Based on the above observations, we extend Munteanu's system into a two-level parallel corpus extraction system. Figure 3.1 shows the framework and the workflows of our method. We retain previous component such as the candidate sentence selection and the method of obtaining lexicons. The bold oval fields, i.e. IBM1 Extractor, Fragment Classifier and Sentence Classifier, highlight our contributions.

In our framework, training corpus plays a very important role. Besides helping to prepare GIZA and LLR lexicon, it provides a more precise IBM1 Extractor and two-level classifier, which are used in the further steps.

As the figure shows, the main difference between our method and Munteanu's is that, after candidate sentences are selected, a one-direction signal filter together with a reverse-direction IBM1 Extractor is applied to search for appropriate fragment pairs. In addition, a refinement step is added after fragment pairs are extracted by using fragment classifier. Here we build this classifier with maximum entropy

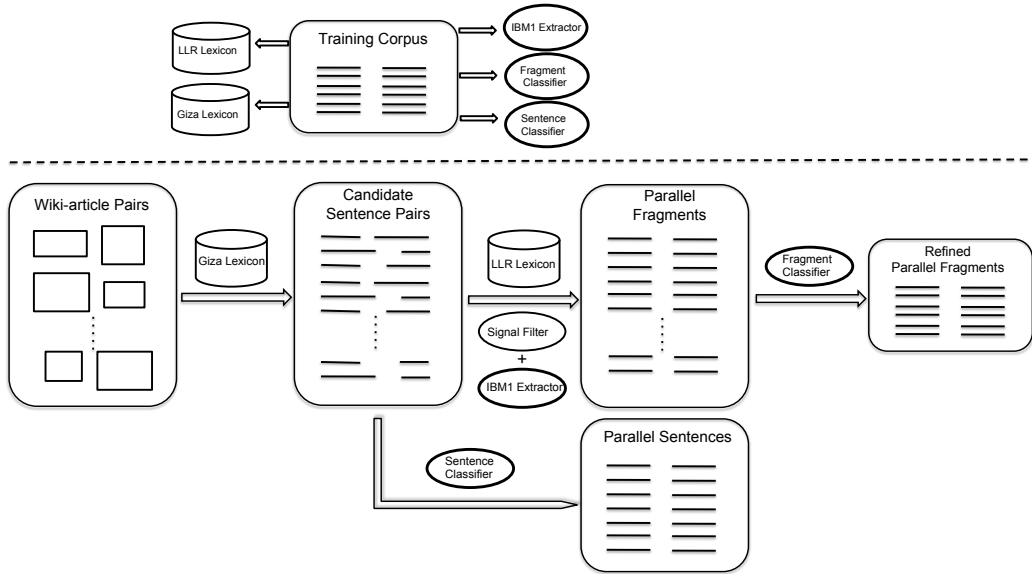


Figure 3.1: An Overview of Two-level Parallel Corpus Extraction System

model; real-value or binary features are extracted to represent vectors of fragment pair. With the maximum entropy classifier, the trained model decides whether this pair is parallel or not. Likewise, a sentence classifier is applied directly to look for parallel sentence among the candidates.

In the following sections, we will explain each module in our framework in details.

3.2 Generate Candidate Sentences

For each foreign document and its document-aligned English article, we firstly take all possible sentence combinations and pass them into a GIZA-lexicon filter.

The GIZA-lexicon is derived from initial training corpus, by running the GIZA++ implementation of the IBM word alignment models (Brown et al., 1993). In this lexicon table, one word in source language t_i may aligned with multiple words in target language e_j ; each pair is given a positive value, which indicates the conditional probability $p(e_j|t_i)$, and vice versa. The lexicon entries with high-probability value are qualified translation, and the rest entries predicate the relevance between two words. Thus, we use this acceptable lexicon as a source to select candidate sentences.

The process of selection is simple and trivial. Firstly, it verifies that the length

ratio of the sentence pair is not greater than two. Secondly, it ensures that at least half the words in each sentence have a translation in the other sentence according to the GIZA-lexicon. Pairs cannot satisfy these two criteria are passed and filtered out.

This step removes the highly noisy sentence pairs which may have no connection and turn out to be less possible to have parallel phrases or sentences. Meanwhile, it takes risk to remove good pairs that filtered by this limited lexicon. Because a lexicon trained by word alignment cannot be 100% trusted. The out-of-vocabulary (OOV) problem will affect the recall. Even worse, this step also could accept several wrong pairs, because the word-overlap condition is weak; for instance, a stop word in source language would connect to a large amount words in target language. Consequently, it contributes to word-overlap and biases the filtering result.

3.3 Build Lexicon Table by Using Log-Likelihood-Ratios

Although there are lots of drawbacks when use GIZA lexicon, it is acceptable as a sentence pair filter. However, in the further steps such as parallel fragment/sentence extraction, the precision of lexicon should be considered. In addition, besides distinguishing the translation probability between source language and target language, we expect to measure the probability that two words are not translations of each other. Based on these concern, we apply a more powerful Log-Likelihood-Ratio (LLR) lexicon (Munteanu and Marcu, 2006). The method of calculating the probabilistic translation lexicon LLR-Lex is firstly from Dunning (1993). It has also been used by Moore (2004) as a measure of word association. In general, this method gives a measure of association between two samples in documents by their co-occurrence. And we apply it in our framework as the translation probability.

The LLR statistic gives a measure of the likelihood of this hypothesis. The LLR score of a word pair is low when these two distributions are very similar (the words are irrelevant), and high otherwise (the words are strongly associated). However, the LLR scores could indicate either a positive association or a negative one; If word e and f are independent, we expect that $p(e|f) = p(e|\neg f) = p(e)$; otherwise, they are regarded as positive association.

From the empirically observation, this lexicon size reduced the suspicious con-

nection of each target word e . The number of connection in Giza-lexicon is 12 in average, but in LLR lexicon it reduces to 10; on the other hand, it contributes the negative valued entries, which improves the correctness and affects the further step in signal filtering and classification.

3.4 Fragment Extraction

Intuitively, our framework tries to distinguish between source fragments that have a translation on the target side, and fragments that do not. After we obtain candidate sentence pairs, one very naive method to extract the parallel pairs by using lexicon information is to examine all possible substrings from both target and source language sentences. A measure criteria is applied, threshold is fixed empirically. Thus we could ‘generate’ all possible sub-level fragments.

However, the drawback of this method is obvious: The search space is too large and redundant. Suppose the length of source and target sentence is m and n respectively, the time complexity of generating pairs is $O(m^2n^2)$ (efficient IBM Model 1 algorithm was introduced by Xu et al. (2005). It achieved run-time of $O(mn)$, however, it is not applicable in our problem). Therefore, we need more efficient method to prune the search space and maintain the precision and recall as much as possible.

In the example in Figure 1.2, it shows a sentence pair from wikipedia article, they are written in German and English. Apparently, it cannot be treated as parallel sentence, since a lot of words cannot find their translation from the other side. However, the bold and connected parts are good fragments to be considered as parallel, i.e. **eines hellen kräftigen rottens, a bold red**. An intuitive thought is that, mining from two language sides simultaneously is inefficient, therefore we fix one fragment from $F \rightarrow E$ as e' , and in return we use e' to find the most possible fragment in the reverse direction $E \rightarrow F$ as f' . So that we obtain a pair e' and f' , in which it could be both accurate and correct.

For the first step extraction from $F \rightarrow E$, we apply a signal filter invented by Munteanu and Marcu (2006); for the second step from $E \rightarrow F$, we invent a novel IBM1 model to find the global optimal. They will be explained with details in next two subsection.

3.4.1 Signal Filter

The first step aims to find the appropriate fragment e' in E according to source language sentence F . One approach is to consider the target sentence as a numeric signal, where translated words correspond to positive values (come from LLR-lex entries), and the others to negative ones. We want to obtain the part of the sentence where the signal is positive. This can be achieved by applying a smoothing filter after the signal, and selecting those fragments of the sentence for which the corresponding filtered values are positive.

The details of the procedure are presented in Figure 3.2. Here let the German sentence be the source sentence, and the English one be the target. Firstly, we derived a word alignment, by greedily linking each English word with its best translation candidate from the German sentence. The link of initial signal for each $e \in E$ is the $f \in F$, which maximize the $p^+(e|f)$ or $p^-(e|f)$; $p^+(e|f)$ means there is one translation according to LLR-lex, while the $p^-(e|f)$ shows the probability is negative, which indicates the best situation should be a non-translation; In worst case, if there is no either positive or negative entry f, e in LLR-lex, we set the probability by -1 , which means no combination from F with e is available in LLR-lex. Thus the initial signals are generated, each word get a $[-1, 1]$ real value (the red data point in the figure).

We obtain the filtered signal (black data point in the figure) by applying an averaging filter, which sets the value at each point to be the average of several values surrounding it. The idea of filter is simple. Because one or two word in target sentence may have no entry in LLE lex, but the surrounding words indicate strong connection. In this case, we apply filter and fill the gap for these target words, as a result, they still get chance to be extracted. This method is efficient in both run-time complexity and implementation, and could solve the OOV problem. In our experiments, we set the window size be 5 after experiments on dev corpus.

Now we can extract the positive span from filtered signal as fragments. However, this approach attempts to produce short 'positive fragments, which may refer to a translation but still risky. For example, in Figure 3.2 **code** obtain a positive value but cannot find correct translation in German sentence. Another fault is more serious, as in this figure, the oval and the rectangle field display the two direction result from signal filter. Apparently, they contain too much noises as parallel text. To avoid such mistakes, we disregard fragments which are less than λ words, λ is

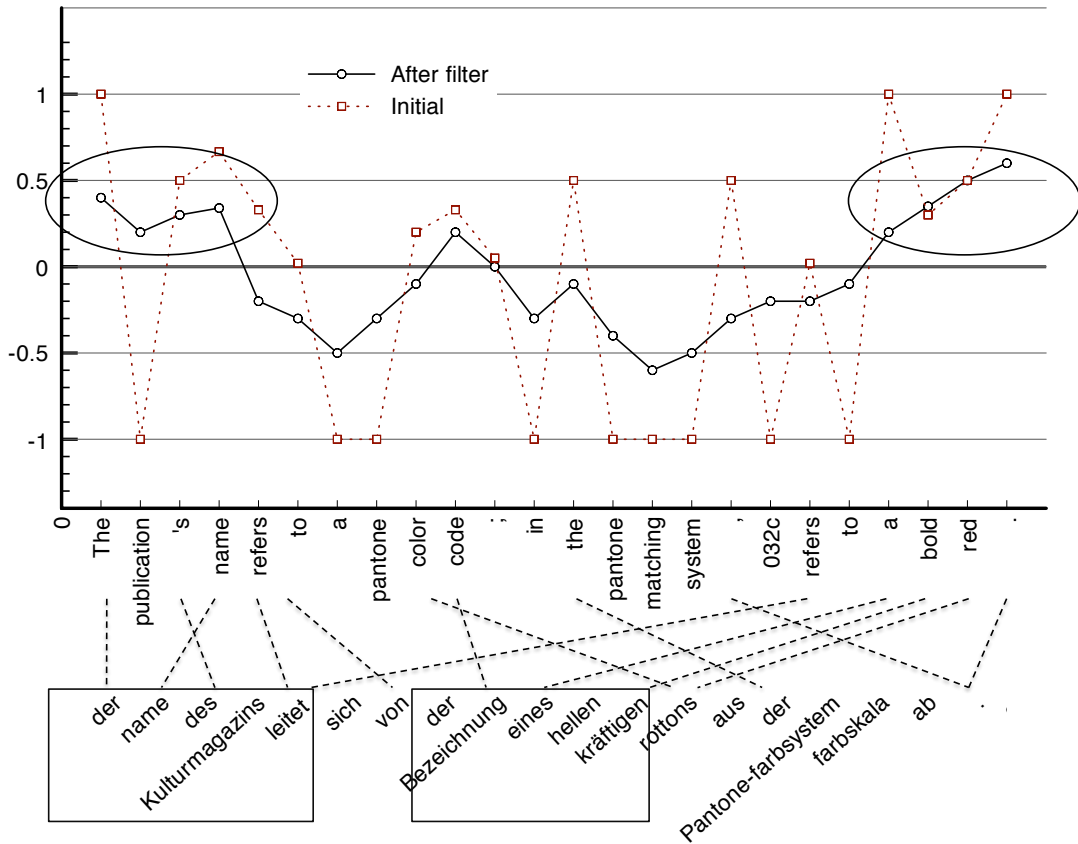


Figure 3.2: A Sample Diagram of Signal Filter

set to 3 in our experiment.

3.4.2 IBM1 Extractor

So far we get e' as a fragment from $F \rightarrow E$, the fragment extraction is partially finished. Now the method to get fragment from F as f' is obvious, we could select arbitrary substring from F , and find one f' which maximize a objective function: $P(f'|e')$. Here comes two issues:

- Which objective function should be chose?
- Arbitrary substring is inefficient, is there any heuristic method to prune search space?

The first problem looks similar when we estimate the translation model $Pr(\mathbf{f}|\mathbf{e})$ in statistical machine translation(SMT). In SMT, this probability is predicated by guessing possible word alignments between \mathbf{f} and \mathbf{e} , this part remains the same. The only difference is translation probability: SMT still don't know the lexicon table, but in our problem, we could use LLR-lex as prior data.

IBM1 Model in SMT

The word alignment is an essential part in translation model, as we will explain in next section. Here we give out the common definition and basic notions in SMT. As the name implies, it aligns the English word to null/one/multiple foreign word between one sentence pair. i.e. (*Le programme a ete mis en application* \longleftrightarrow *And(null) the(1) program(2) has(3) been(4) implemented(5,6,7)*), the numbers in brackets indicate an *acceptable* word alignment.

More formally, $\mathbf{e} = e_1^l = e_1, e_2, \dots, e_l$ denotes a english sentence with l words; likewise, $\mathbf{f} = f_1^m = f_1, f_2, \dots, f_m$ denotes the foreign sentence with m words. $\mathbf{a} = a_1^m = a_1, a_2, \dots, a_m$ presents the alignment from \mathbf{f} to \mathbf{e} , each $a_i, 1 \leq i \leq m$ indicates e_{a_i} is a translation of f_i ; If no f word is connected to e , then let $a_i = 0$.

According to Brown et al. (1990), the translation probability $Pr(\mathbf{f}|\mathbf{e})$ is able to work out as:

$$Pr(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f_j|e_i) \quad (3.1)$$

Modify IBM Model 1 as Fragment Extractor

Now we can see the difference mentioned before: Instead of applying iterative parameterization, $t(f|e)$ is already prepared. It's almost a decode problem in SMT: by given string e' and translation table $t(f|e)$, generate f' , such that

$$\arg \max_{f' \in \mathcal{H}(F)} P(f'|e') \quad (3.2)$$

Here the hypothesis $\mathcal{H}(F)$ is constraint to the set of sub-string of F. Furthermore, we assume the length of f' and e' are m and $l+1$. According to Eq (3.1), we have:

$$P(f'|e') = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f'_j|e'_i) \quad (3.3)$$

It seems reasonable to apply Eq(3.2) and Eq(3.3) directly in searching the global optimal. However, one obvious failure is, it always returns the maximal value with fragment length equal to 1. Because each j , $\sum_{i=0}^l t(f'_j|e'_i)$ is less than $l + 1$, so that the multiply of them would always derive smaller value, therefore f' with length 1 is optimal. Here we give a brief proof. Assume the fragment e' is given, and we attempt to extract $f_p^q = f_p, f_{p+1}, \dots, f_q$ from source sentence F , then we have :

$$P(f_p^q|e') = \frac{\epsilon}{(l+1)^{q-p+1}} \prod_{j=p}^q \sum_{i=0}^l t(f_j|e'_i) \quad (3.4)$$

$$= \epsilon \cdot \prod_{j=p}^q \frac{\sum_{i=0}^l t(f_j|e'_i)}{(l+1)} \quad (3.5)$$

$$\equiv \epsilon \cdot \prod_{j=p}^q \frac{d(f_j)}{l+1} \quad (3.6)$$

In Eq (3.5), $t(f_j|e'_i)$ is posterior translation probability by given e'_j , thus $d(f_j)$ is a fixed value and less than $l + 1$. Thus, it's obvious that $\frac{d(f_j)}{l+1} \leq 1$. As a result, $P(f_p^q|e') \leq P(f_p^{q-1}|e')$. So far, it is obvious that the maximization is equivalent to find the sub-string from f with length 1. Which is definitely NOT what we want.

To address the length-bias problem, we propose a new method to adjust the extraction length:

$$G(f'|e') = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f'_j|e'_i) \frac{1}{1 + |m - l - 1|/(l+1)} \quad (3.7)$$

The benefit is straightforward, the term $|m - l - 1|$ could be viewed as offset between target fragment and source fragment. The larger the offset is, the smaller the factor comes. Thus it could avoid the whole length move towards either too long or too short. (The variations of *long* and *short* are not symmetric, nevertheless, it is able to compensate the length bias.) Secondly, the division $l + 1$ allows the offset vary according to the length of e' . If the e' is a fragment with larger length, the scale of looking for f' with length around $|e'|$ is extended. For instance, one English fragment **10 Downing street** . with length 4, which is sufficient short; When tree German fragment are extracted with almost the same IBM1 score, the only difference is their length. In this situation, the length-bias factor plays an important role to

keep seeking the one with the length close to its given fragment. However, if the e' is large as *to the office of first lord of the treasury rather than to*, the extraction would accept more various length fragment around 12. This provide a dynamic strategy to automatic eliminate the length-bias.

The second problem we mentioned at the beginning is about the search space. Although we propose a modified IBM Model 1 as objective function, we still need to protest brute-force search in fragment generation. In our framework, we initialize the length of f' to be $|e'|$, and adjust the offset-length from 0 to $\lambda_2 * |e'|$ on both side. The term $\lambda_2 * |e'|$ indicates a window to extend f' 's size, and its size could vary dynamically according to $|e'|$ as well.

After the above two problems are solved, we now describe the complete modified IBM1 model as reverse extraction method in Algorithm 1. The procedure is simple and easy to understand, $G(f|e')$ indicates the modified calculation in Eq (3.7). The time complexity is still quadratic, but it reduces the inner loop from l to $l \cdot \lambda$. Also we could add estimation function to prune efficiently, such as assuming all the further extension will obtain a maximal IBM1 score $P(f'|e')$ up to 1, if the objective function still cannot exceed the current optimal because of the length-bias factor, it would quit and return the current best result.

Algorithm 1 IBM1 Extractor

Input: Source sentence $F = f_1, f_2, \dots, f_l$;

Fragment $e' = e_1, e_2, \dots, e_m$;

$t(e|f)$ as lexicon table;

window ratio λ ;

Output: f' as a sub-string from F

$n = \lambda \cdot l$

for $i = 1$ **to** l **do**

for $j = i - n/2$ **to** $i + n/2$ **do**

$f' = f_i, f_{i+1}, \dots, f_j$

$f_{max} = \arg \max_{f \in \{f', f_{max}\}} G(f|e')$

end for

end for

return f_{max}

We should notice that, the complexity to calculate each $G(f|e')$ is $O(lm)$. But if

we play a trick and store them in advance, it won't take extra cost each time. Thus the whole run-time in Algorithm 1 is $O(lm + l^2)$, which is acceptable and somehow efficient after prune.

3.5 Two-level classifier

Now we partially accomplished our task, a modified IBM Model 1 is proposed to extract parallel fragments by given a sentence pair. The main achievement so far is to guarantee the connection between target fragment and source fragment. The probability is maximized and the time complexity is efficient.

However, we still cannot ensure the quality of these extraction pairs. How parallel could they be? Let's step back to the fragment derivation approach: A signal filter together with LLR-lex firstly filter out a substring e' from E , which is considered to be very likely to link to a potential fragment f' from F ; than a generative IBM1 model is utilized and come up with f' under a reverse direction $F \rightarrow E$. Although we admit the connection between them are strongest among this sentence pair. But there is no assurance that e' is a definite good candidate. Signal filter could generate 'false' candidates, which shouldn't be regarded in further steps. Even in this case, the IBM1 extractor still could pick up a global optimal. Table 3.1 shows a list of examples, the left side are results from English sentence, the right side is it's corresponding IBM1 extraction.

English fragment	German fragment
<i>the magazine the berlin magazine</i>	<i>das Magazin zum best magazine</i>
<i>to the hall ,</i>	<i>, die zu den</i>
<i>in the world ,</i>	<i>, die in dem</i>
<i>in the beginning of the movie</i>	<i>die Wahrheit über den Streich des</i>

Table 3.1: Failure Samples from IBM1 Extractor

To solve this problem, we need a more powerful step to distinguish and evaluate how parallel each fragment pair are. In addition, as an initial goal, we need to select parallel sentences directly upon candidate sentences. With these two concerns, we add a classifier on both sentence and fragment candidates.

The concept "two-level" in our framework is scenario-oriented. Currently we are dealing with two scenarios: candidate sentences from naive word-overlap algorithm,

and extracted fragments from generative method. Even we characterize them under the same feature set, they cannot be treated under the same model due to the differential hypotheses. To be concrete, sentence pair may be easier to classify by their length ratio or translation overlap, while fragment pair intend to share the approximate length (as we applied IBM1 extractor). In this case, they rely more on fertility or longest translation span.

Thus the hypotheses cannot be uniform, and one hyperplane is not sufficient to divide two genre data sets. To handle this diversity, we used two separate ME model for each classification respectively. The feature sets remain the same, while we parametrize on different training set and derive two models. More details are explained in section 4.4.

3.5.1 Maximum Entropy Classification

Maximum entropy method is selected to categorize the candidate pairs into 'parallel' or 'non-parallel'. Maximum Entropy (ME) principle comes from the thermodynamics, in information theory, it states that subject to known constraints, the probability distribution which best represents the current state of knowledge is the one with largest entropy. Berger et al. (1996) introduce this method into SMT.

There are dependent variable categories $0, 1, \dots, J$. One regression is run for each category $0, 1, \dots, J$ to predict the probability of x (the dependent variable for any observation) being in that category. The regressions are, for $k = 1, 2, \dots, J$:

$$P(y = k|x) = \frac{\exp[\lambda_i \cdot f_i(y, x)]}{1 + \sum_{j=1}^J \exp[\lambda_j f_j(y, x)]} \quad (3.8)$$

meanwhile, an adding-up constraint should be satisfied on reference category:

$$P(y = 0|x) = \frac{1}{1 + \sum_{j=1}^J \exp[\lambda_j f_j(y, x)]} \quad (3.9)$$

In which y is the observed outcome by the observation x , $f_i(y, x)$ is the i th feature vector of the observation, and λ_i is the i th vector of the regression coefficients. The unknown parameters in each vector λ_i are typically jointly estimated by maximum a posteriori (MAP) estimation, which is an extension of maximum likelihood using regularization of the weights. The solution is typically found using an iterative

procedure such as iteratively re-weighted least squares [IRLS] or, more popular, a quasi-Newton method such as the L-BFGS method (Byrd et al., 1994). The optimal values of these parameters are obtained by maximizing the likelihood of the training data. Therefore two different principles – maximum entropy and maximum likelihood – end up to the same result.

3.5.2 Feature Setting

The most important issue by apply this model is feature setting. Features should express the properties of the instances. Good feature set will highly distinguish which category the instance belongs to. Here we fully utilize the previous lexicon table, follow the system in Marcu and Meananu (2005) to investigate the lexicon-based features.

Suppose in a correct alignment between two non-parallel sentences, most words would have no translation equivalents; however, in an alignment between parallel sentences, words should be aligned. Although there would be exceptions due to the robustness of lexicon we generated before, this principle is still basic and applicable.

Also, we observed that a word in an alignment which tends to connect lots of other words, imply a non-parallelism. We introduce the term **fertility** as how many foreign words each native word produces (Brown et al, 1993). In most case, these connections were produced due to a lack of correct alternatives.

Another phenomenon to indicate the parallelism is long contiguous connected spans, which denotes the number of long substring in a sentence that fully connected in the aligned sentence. For example, Figure 3.2 shows a $F \rightarrow E$ candidate, the English strings *The publication 's name* are continuously connected by its German counterparts. Long contiguous connected spans could indicate parallel fragment/sentence pairs. In contrast, long contiguous un-connected spans predicate non-parallelism.

Here we also add IBM1 model as real-value feature, IBM1 score of both $p(f|e)$ and $p(e|f)$ would be strong indicative about parallelism. If these two values are equally high, it implies a parallel pair. But we still need to balance the lengths of the two pairs, we normalize the alignment probability by the source sentence length

and adjust its weight with a parameter β as described in Xu et al. (2005),

$$P'(f'|e') = \left[\frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=0}^l t(f'_j|e'_i) \right]^{\beta \frac{1}{m+1} + 1 - \beta} \quad (3.10)$$

To summarize, our classifier uses the following features to characterize candidate sentences/fragments pairs:

LENGTH: lengths of the sentences, as well as the length difference and length ratio;

TRANS: number and percentage of connected words for both $F \rightarrow E$ and $E \rightarrow F$;

FERT: Top three fertilities and there percentage in both F and E ;

UNCONNECT: Length and percentage of longest substring which are not connected;

CONTIG: Length and percentage of longest contiguous span which are connected;

DIGITAL: Unmatched digital number;

IBM1: Normalized IBM1 score as described in Eq (3.10).

Once we get the real-value of feature vectors, ME principle is applied as Eq (3.8), a log-linear combination function is parametrized with positive and negative samples. The same equation is used to classify the instances after all parameters are fixed. After training, two separate models are generated for different scenarios. Sentence classification would apply the model trained on positive and negative sentence pairs; likewise, fragment classification would use the model retrieved from fragment pairs.

So far we finished the introduction of the whole framework. To conclude, this system is a semi-supervised approach, the sentence aligned corpus as initial data would provide word-alignment lexicon, this lexicon in return explore fully in document for two-level parallel pairs. In next section, we set up experiments to analyze the whole workflow, and evaluate the contribution from each component.

Chapter 4

Experiment

To demonstrate that our novel framework outperforms the current state-of-the-arts method as in Munteanu and Marcu (2006), we set up experiments and compare with baseline system. Two systems are built with different size training data set to study how important the initial corpus is, and how the LLR-lexicon will influence the whole extraction process. We practice the extraction workflow based on Wikipedia document-aligned corpus ¹, which may raise the problem of domain adaptation. In the end, we test our newly extractions on several SMT evaluation corpus in section 4.5.

4.1 Data Set

We use German-English parallel sentences from Europarl-v6 (Koehn) and news-commentary data ² as initial training data, they are sentence aligned, Table 4.1 shows the information about this corpus. News-commentary (NC) and Europarl-v6 (EP6) are with 10 times size different.

Corpus	Sentence Pair No.	Token No.
NC	100K	16M
EP6	1.5M	290M

Table 4.1: Description of Initial Corpus

¹Document-aligned Wikipedia corpus is provided by DFKI from Accurat project.

²<http://www.statmt.org/wmt10/training-parallel.tgz>

With these initial corpus, we explore the parallel target on wikipedia corpus, this corpus is consists of various topics, i.e. politic, music, sport, events, architecture, etc. Table 4.2 describes this corpus from text aspect. All wikipedia document are split into sentence by NLTK tool (Loper and Bird, 2002), and tokenized by additional script from WMT share task ³.

Lanuage	Document Pair No.	Sentence No.	Token No.
DE	362,342	4.1M	1.6G
EN		5.0M	2.7G

Table 4.2: Description of Wikipedia Raw Corpus

4.2 Lexicon Building and Sentence Candidate Selection

Initial training corpus are sent to GIZA++ ⁴ (Och and Ney, 2003) for building lexicon tables. GIZA++ is a statistical machine translation toolkit that is used to train the most popular word alignment model. This toolkit also contains the source for the mkcls tool which generates the word classes necessary for training some of the alignment models.

GIZA++ runs with standard process: 5 iterations in IBM1 model, 3 iterations in both IBM3 and IBM4 model, 5 iterations in HMM model. The alignment process runs in both directions, and then symmetrize the alignments using the refined heuristic. After that, a Giza lexicon is trained and could be used as a resource for both LLR-lex generation and sentence filter. In this lexicon table, one word in source language t_i may aligned with multiple words in target language e_j ; each pair is given a positive real value, which indicates the conditional probability $p(e_j|t_i)$.

We list the comparison of Giza and LLR lexicon in Table 4.3. From this table, it proves the advantages of LLR: firstly, the average connection of each entry decline at least 15% on both initial corpus; secondly, with positive and negative value, LLR gives a more truthful lexicon to tell the correlation.

³<http://homepages.inf.ed.ac.uk/jschroe1/how-to/scripts.tgz>

⁴<http://code.google.com/p/giza-pp/>

Initial Corpus	Vocab	Giza Lexicon	LLR-Lexicon		Sentence Pair
			pos	neg	
NC	99K	0.5M	0.4M	98K	9.3M
EP6	186K	2.7M	2.3M	126K	80.6M

Table 4.3: Build LLR lexicon from Giza alignment ,under column LLR-Lexicon,pos represent $p^+(e|f)$, neg represent $p^-(e|f)$; the last column is candidate sentence generation based on Giza lexicon.

Once we derive GIZA lexicon, it is used to select sentence candidates from documents pairs by the heuristic rules in section 3.2. As expected, the size of sentence pair got a quadratic growth, as showed in Table 4.3. We observe that, a richer lexicon would produce more sentence pair, which will increases the quantity of ultimate extraction. However, we notice that the average size of large initial corpus is also larger, which indicate a noisy word alignment. Thus, the candidate pair size is enlarged, but also raises risk to filter in sentence pairs which have not actual alignment fragment.

4.3 Fragment Extraction

We apply two fragment extraction method, one is two-direction signal filter introduced by Munteanu, another is singal filter + reverse IBM1 extractor as our novel method. The former will be used in final evaluation as baseline. Table 4.4 gives about the quantity of pairs it extracted by applying different initial corpus. The average length ratio in Munteanu’s method is much larger then our result. Because the signal filter simply extracts fragments based on lexicon, it couldn’t guarantee any alignment between the extractions.

We give two sets of samples as intermediate result of the extractions, showed in Table 4.5 . Obviously, the Munteanu’s method suffers from the problem we just mentioned. These sets of examples convince the IBM1 extractor together with signal filter outperform the previous approach.

Another observation is, neither method could extract fragment with high recall. Compared with initial candidate sentence pairs, only 0.3% of them is able to produce parallel fragment, which still remains much room for improvement.

	Initial corpus	Pairs No.	Token No.	Ave Length Ratio
Muteanu2006	NC	119K	4.0M	1.34
	EP6	625K	18.4M	
IBM1 Extractor	NC	30K	1.2M	1.10
	EP6	242K	7.9M	

Table 4.4: Fragment extraction with munteanu’s and our methods. The last column is the average length ratio for parallel extraction.

	English	German
Munteanu2006	virus , thought to	des Virus zu finden
	perception of what the world	Realität ist und was nicht ,
	national historic landmark in 2006	einer National historic Landmark
	intro twelve extremely short tracks , and four	zwölf extrem kurze Tracks , und vier
IBM1 Extractor	with the sun	mit der Sonne
	Frankfurt is a	Frankfurt ist eine
	the second team plays in the 2nd	die zweite Mannschaft spielt in der 2
	the album consists of an instrumental	das Album besteht aus einem Instrumentalen in
	in the same year , the first ever german	in den nächsten Jahren den Sprung in die erste

Table 4.5: List of fragment samples from two methods

4.4 Two-level Classifier

In this section, training corpus is selected based on initial corpus and intermediate phrase table from Moses (Koehn et al., 2007). Features are selected as described in 3.5.2. Here we empirically evaluate the influence of hypothesis and the effect of each feature. Finally, we build two separate classifiers for sentence and fragment pairs.

4.4.1 Building Dev Corpus

The challenging part in collecting training and test corpus is: How to build the hypothesis that depicts almost the same distribution as in practice. For sentences, it is easier because candidate sentence are very noisy, it won’t be hard to classifier

correct pairs; however, in fragments, we already get almost parallel pairs by this novel method, thus our target is changed to looking for high quality instance from good bases. Here we use only Europarl-v6 as develop corpus.

Sentence Level

We randomly select 11,000 $F \leftrightarrow E$ pairs from training corpus, label them as positive. To generate negative samples, we hold F side from the positive instances, and seek the English sentence in Europarl-v6 which is not its alignment, but satisfied the heuristic rules in candidate filtering. Furthermore, we split 10,000 sentence pairs from both positive and negative instances as training set, and left the remaining set as test.

Fragment Level

Fragment source is not easy to retrieve. One method is using the intermediate output from Moses. Moses stores the extracted phrases on disk. For each pair $\langle e, f \rangle$, the phrase translation probability $\varphi(f|e)$ and $\varphi(e|f)$ are estimated. In addition, phrase translation scoring functions are computed. Figure 4.1 is a example of phrase table, the five real numbers after the second separator are:

1. inverse phrase translation probability $\varphi(f|e)$;
2. inverse lexical weighting $lex(f|e)$;
3. direct phrase translation probability $\varphi(e|f)$;
4. direct lexical weighting $lex(e|f)$;
5. phrase penalty , always $exp(1) = 2.718$.

in europa ,	in europe	0.61371	0.20755	0.20743	0.492402	2.718
europaeischen	in europe	0.06848	0.07538	0.00081	0.461284	2.718
im europaeischen	in europe	0.579275	0.00901	0.24722	0.162482	2.718

Figure 4.1: An Example of Phrase Table in Moses

Here we only take translation probability φ as reference. The phrase pair $\langle f, e \rangle$ obtains probabilities $\varphi(f|e) = 1$ and $\varphi(e|f) = 1$ are considered as fully paralleled phrases. The one obtains probabilities $\varphi(f|e) < 0.3$ and $\varphi(e|f) < 0.3$ are selected as non-parallel phrases. To investigate how the constitute of positive and negative

samples would affect the performance of classification, we build two different ratio dev corpus— **Set A** and **Set B**, as listed in Table 4.6. The only difference between them is the size of negative instance number.

		Training Pairs		Test Pairs	
		pos	neg	pos	neg
Sentence		10K		1K	
Fragment	Set A	10K	10K	1K	
	Set B	10K	20K	1K	

Table 4.6: Constitute of Dev Corpus

4.4.2 Training and Test

We empirically investigate how the feature setting impact the result of classification with gradually adding features. Table 4.7 describes the procedure of Fragment classification on both Set A and Set B. Likewise, Table 4.8 lists all the result on Sentence classification.

Features	Set A			Set B		
	Precision	Recall	F-measure	Precision	Recall	F-measure
LENGTH& TRANS	0.794	0.786	0.790	0.790	0.784	0.787
+ FERT	0.801	0.793	0.797	0.796	0.793	0.794
+ UNCONNECT	0.817	0.804	0.810	0.820	0.801	0.811
+ CONTIG	0.831	0.821	0.826	0.839	0.825	0.831
+ DIGITAL	0.853	0.831	0.842	0.857	0.833	0.845
+ IBM1	0.855	0.830	0.846	0.858	0.830	0.849

Table 4.7: ME Performance on Dev Corpus: Fragment

As we can see, basic features such as translation number and sentence length ratio are useful in practice. With comparison of other additional features, longest continuous span is important and significantly boost the performance, while the top 3 fertilities are not so helpful as we expect.

Sentence classification already achieves a high performance due to its hardness. In contrast, it leaves much room to improve in fragment task. Another observation is that, on different ratios of negative and positive training instance, fragment

Features	Sentence Pairs		
	Precision	Recall	F-measure
LENGTH & TRANS	0.890	0.901	0.895
+ FERT	0.902	0.905	0.903
+ UNCONNECT	0.932	0.920	0.926
+ CONTIG	0.935	0.922	0.928
+ DIGITAL	0.947	0.931	0.939
+ IBM1	0.950	0.931	0.940

Table 4.8: ME Performance on Dev Corpus: Sentence

classifiers don't perform with significant difference. Thus in the framework, we use model trained under Set A.

4.4.3 Apply into Framework

Once the two-level classifiers are prepared, we use them into the whole framework. Table 4.9 shows the quantity of pairs which are classified as parallel. We make a brief conclusion from this table as follows:

1. A large amount of candidate sentences are abandoned. In contrast, it remains lots of fragment pairs, which indicates a high quality of IBM1 extractor;
2. The final fragment extraction based on news-commentary corpus is far less than Europarl corpus. It demonstrates the importance of lexicon. The more vocabulary a lexicon covers, the more capable to explore large amount data pairs.

		before		after	
		Pairs	Token	Pairs	Token
NC	Sentence	9.3M	1.2G	25K	1.9M
	Fragment	30K	1.2M	12K	0.4M
EP6	Sentence	80.6M	8.7G	71K	7.0M
	Fragment	242K	7.9M	48K	1.4M

Table 4.9: Numbers of Extractions

Table A.2 gives sample outputs of extracted fragments. Without surprise, although we make great effort to improve the fragment extraction, there are noisy pairs included.

4.5 SMT Evaluation

In this section, we add the extracted fragments and sentences into existing training corpus as new parallel data. Phrase based translation process is applied by Moses-ems. Moses-ems is an efficient automated tool of SMT translation, the process include preprocess, word-alignment and symmetrization, phrase extraction and scoring, decoding and reranking. Finally, it provide standard eval script to evaluate the output with reference, and give a score (it could be BLEU, multi-bleu, nist, etc).

In our experiment, we test the new result on two test corpus: Balanced from Accurat project; and Test2010 from WMT shared task 2010 ⁵. Because the new extractions are mainly from wikipedia, the domain is diverse and the vocabulary maybe not overlaps on news-test. Therefore, two test sets with different domain constitute would help us to investigate how significant the domain adapted.

4.5.1 Baseline System

We take News-commentary and Europarl as two raw-baseline, which compared with our training corpus by adding new extractions onto them.

Moreover,because our work is an extension based on Munteanu 2006, we build their system as baseline as well. As showed in Fig 1.3, two direction signal filter process is applied without further classification. With this simple system, we extracted 625K fragment pairs. Although LLR lexicon is helpful in lexicon level filtering, a lot of them are considered as non-parallel under manual evaluation.

Despite we implement their method on DE-EN language translation instead of RO-EN, the parameters and threshold don't changed a lot, i.e. in signal filter algorithm, the minimal fragment length λ is set to be 4, and the surrounding window size n is 5.

4.5.2 Result and Comparison

We evaluate the performance on BLEU score. Table 4.10 shows the result and comparison of our system with raw baseline, and Munteanu's system. From this table, we observe that by adding Munteanu's extractions onto initial training corpus, the BLEU score increase slightly. Because the size of initial corpus is large, Munteanu's

⁵<http://www.statmt.org/wmt10/test.tgz>

method got 1.14 BLEU gain on 10M training corpus, but this increment tends to be less or even harmful when applied on larger initial corpus. In contrast, our new proposed method outperforms theirs, and increase the baseline up to 0.98 BLEU score on *Balanced* test corpus.

However, Munteanu’s result is harmful on Test2010, also our framework doesn’t improve the baseline much. This could be explained by the domain adaptation, since we only extract data from multi-domain corpus, it could contribute to the same domain test corpus, instead of news domain. Although we cannot achieve significant improvement, our framework is still robust than Munteanu’s.

To test our framework on the current best MT system, we set the initial corpus as a merge of Europarl v6 and News-Commentary, because from experiment, this training corpus could build SMT system with highest BLEU score. To investigate whether our method would help to improve the current best system, we integrate the extracted texts into the training data. Table 4.11 shows the result. As we expected, due to the few extractions, the result cannot vary a lot. In multi-domain test, it only enhances 0.05 – 0.10 BLEU score, which is reasonable but not agreeable. However, the cross-domain test seems unfortunately negative.

	Balanced		Test2010	
	MaxLen=80	MaxLen=100 +splitter	MaxLen=80	MaxLen=100 +splitter
Baseline	19.45	20.25	15.23	15.69
Munteanu06	19.54	20.32	14.94	15.05
Improved	20.21	21.23	15.54	15.98

Table 4.10: BLEU Score Evaluation with Initial Corpus News-Commentary

	Balanced		Test2010	
	MaxLen=80	MaxLen=100 +splitter	MaxLen=80	MaxLen=100 +splitter
Baseline	27.39	28.57	18.31	18.52
Munteanu06	26.60	27.45	16.43	16.95
Improved	27.58	28.69	18.20	18.47

Table 4.11: BLEU Score Evaluation with Initial Corpus Europarl

Chapter 5

Discussion

From the experiments we observe that, as Munteanu’s work has predicted, the proportional extracted pairs from comparable corpus do provide better translation results. In this chapter, we attribute the improvement to two main facts: (1) the domain similarity between the training corpus and the comparable data, (2) the portion of extractions as additional parallel texts. Furthermore, we discuss the reason why our MT results outperform other’s system as in Munteanu and Marcu (2006).

5.1 Domain Adaptation

We notice that domain plays a relative important role. As we mentioned, Test2010 are provided by WMT shared task. 1,000 test sentence pairs are selected from the news domain. In contrast, we list the consist of domain of test set Balanced in Table 5.1.

By comparing Table 4.10 and Table 4.11, Balance test set always achieves more improvements than Test2010. In the News-Commentary’s set, it gains 0.98 versus 0.29; In the Europarl’s set, it improves +0.12 instead of -0.05. Although in the latter experiment, the inclination and the declination of two test sets are not significant, it clearly shows how domain have an influence on the test results. Because comparable corpus are from wikipedia, which consists with articles about music, film, laws, IT, etc. It could be considered to have similar domain constitution with Balanced set.

Domain	Percentage (%)
General information about European Union	12
Scientific and educational journal	12
Official and legal documents	12
News and magazine articles	24
Information technology	18
Letters	5
Fictions	5

Table 5.1: Domain Constitution in Balanced Evaluation Data

5.2 Influence of the Initial Training Corpus

We observe that the improvement vary dramatically according to the size of initial parallel corpora. This could be explained by the common sense in SMT: the translation accuracy (e.g., BLEU score) increase monotonically by providing more training data, but the speed would be slow down. That means, if we choose large parallel texts as baseline, it is difficult to obtain significant improvement when adding small portion parallel training data.

This rule also applies when we add the extracted pairs. And it would be even more difficult to achieve better translation results due to the unexpected noises. Although we make efforts to select and filter the most reliable pairs among candidates, and consider them as parallel, it still unfortunately fails to guarantee the quality. Therefore, the noises are indispensable and interfere with the ideal improvement.

In general, the poor quality pairs would enlarge the lexicon table with false translation entries, extend the phrase table with incorrect alignment. However, they could contribute to language model and compensate the decoder, because they are all grammatical correct sentences or fragments. While it is too hard to investigate how these poor quality jeopardize the MT system, we could only design experiments to show the approximate limitation by adding the extracted texts.

We therefore design a set of experiments to study how the initial training corpus effect the ultimate translation quality. Since the extracted texts are few, we only vary the size of training corpus. We randomly select sentence from Europarl corpus with token size of 20M, 40M, 60M and 160M (full), and we compare the MT results on these four corpora by adding extracted texts with the raw baseline system without additional data. Consequently, we derive a curve diagram as in Figure 5.1. It depicts

the BLEU score changes after adding the same set of extracted parallel texts. As we can see, our method obtains significant improvement when the initial training corpus size is small (20M), and the score difference declines when the corpus size grows. After approximate 50M, the improvement almost doesn't exist, and even become harmful. We believe this phenomenon due to the noises generated by our framework.

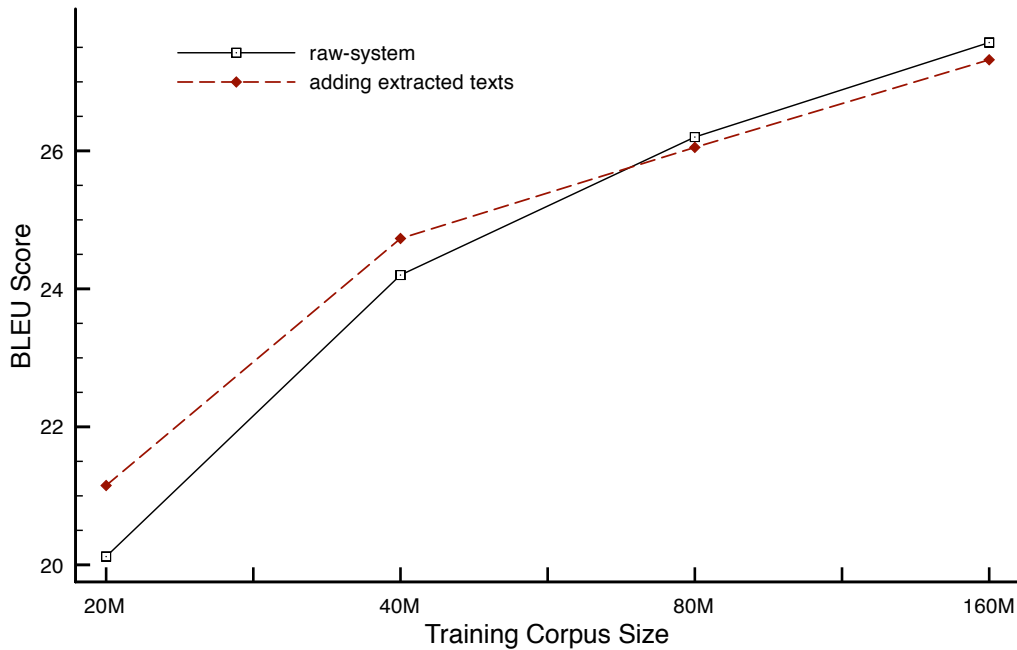


Figure 5.1: BLEU Score Changes by Adding Extracted Texts into Different Size Training Corpus (Random Selection from Europarl)

From the above figure, we prove the mentioned rule in the comparable corpora scenario. The most important is, it shows our framework's limitation, which means it couldn't always afford help on all SMT system. Although the improvement based on system with large training corpora cannot be satisfied, it is still applicable on the under-resourced language pairs.

5.3 Comparison with Munteanu's Framework

At first glance, we may be surprised about the poor performance of Munteanu's output. Their framework is actually harmful in most case, especially on the training

corpus of Europarl and test set of Test2010. The BLEU score decreases 1.57, which is significant. The only improvement is based on News-commentary training corpus and Balanced test set. Even though, the improvement is as little as ignorable.

However, in Munteanu and Marcu (2006), their experiments were conducted only on very small size training corpus (1M-10M tokens), but we design our experiments based on larger training corpus (20M-400M tokens). Those poor extracted text pairs cannot compensate in our training data. Moreover, the comparable corpus, the training data and the test set in Munteanu’s work are all from news domain, which is easier to obtain an improvement according to our previous discussion. In contrast, these three corpora are from different domains in our experiment, which appears to be more challenging and difficult. Now we understand why Munteanu’s framework cannot achieve the expected results as they used to be. As a result, we show the robustness and the adaptivity of our framework.

Chapter 6

Conclusion and Future Work

We have presented a novel extension of Munteanu’s framework to extract parallel data, described how to leverage such a model to extract parallel sub-sentential fragments and sentences from comparable corpora, and demonstrated the impact of these two-level extractions on a machine translation system. Reimplementation of the model should not be challenging; each component is realized as an independent package for further adaptation and improvement.

We retained the approach of building LLR lexicon in Munteanu and Marcu (2006). Our experiment was able to demonstrate that using this lexicon helps improve the lexicon translation quality for the German-English language pair. In addition, we managed to modified the IBM Model 1 for the extraction step, which proved to be an efficient approach to find proper parallel fragments. By experiments we demonstrated that this IBM1 extractor combined with signal filter outperforms the original two-direction signal filter. Moreover, the normalized IBM1 feature has been integrated into a discriminative model to boost the accuracy of classification.

In German-English SMT experiments, we have shown that our framework leads a better result. The significant improvement has been achieved on a relatively small corpus. That is, after adding extracted parallel texts into the original news-commentary training corpus, the BLEU score could be enhanced by 0.98 compared to the raw system without any additional data. Sufficient evidence showed our framework outperforms Munteanu’s in two aspects: obtaining better domain adaptation and generating more reliable translation results.

However there are many potential improvements to explore. If we limit ourselves to the problem of extracting fragments from comparable articles, there are several

points in this pipeline that could benefit from optimization. Firstly, the signal filtering function is somehow simple; more advanced filters could function better and eliminate the weakness of applying heuristics (such as demanding that the extracted fragments have at least 3 words). Secondly, despite the better lexicon, the greatest source of errors are still false translation probabilities, i.e. with punctuations or same-class words. Giving attention to such faults should help get rid of these errors and achieve improvements in the extraction step.

Another improvement would be: Bootstrap our models instead of applying only single pass extraction; retrain the noisy models, and re-extract new fragments. This method could further enlarge the vocabulary of the MT system. Although only small numbers might be collected by each iteration, the gradual extension of vocabulary is able to lead a significant difference. However, the drawback of this method is obvious – it would be too slow to train Giza lexicon every time. Other walk-arounds avoiding to re-generate the whole lexicon should be investigated.

So far, our methods didn't show reliable effects when large parallel training corpora are provided. The reasons could be attributed to two sides: 1) Domain adaptability: Wikipedia corpora with various domain articles cannot afford an significant improvement on news test domain; 2) Low recall: only 0.3% of parallel texts are extracted, which cannot essentially help a strong MT system trained by large training corpus. In the future, we might investigate large scale comparable corpora on the news domain, or eliminate the domain dissimilarity between training and test data set. Besides, improving the framework to achieve higher recall would be another challenging topic as the future work.

Appendix A

Appendix

The following two tables list the sample result of our framework.

No.	English Sentence	German Sentence
1	10 years was initially formed in knoxville, tennessee in 1999 with singer mike underdown, drummer brian vodinh, bassist lewis cosby, and guitarists ryan "tater" johnson and matt wantland.	10 years wurde 2000 von sänger mike underdown, schlagzeuger brian vodinh, bassist lewis cosby und den gitarristen ryan tater johnson und matt wantland gegründet.
2	1993 sb is a trans-neptunian object of the plutino class.	1993 sb ist ein transneptunisches objekt der plutino-klasse.
3	it was designated as a national historic landmark in 2006.	im jahr 2006 wurde die kirche zu einer national historic landmark erklärt.
4	their first studio album ifni was released in 2004.	2004 veröffentlichte die gruppe ihr erstes studioalbum ifni.
5	in 2009 they released their third studio album el dorado , the first of their albums to be released worldwide.	ende des jahres nahmen sie ihr drittes studioalbum el dorado auf, das im januar weltweit 2009 erschien.
6	in september 2008 they invited the percussionist johnny kalsi of the dhol foundation in london to play with them.	september 2008 luden sie den percussionisten johnny kalsi von der dhol foundation in london nach berlin ein.
7	18 scorpii was identified in september 2003 by astrobiologist margaret turnbull from the university of arizona in tucson as one of the most promising nearby candidates for hosting life based on her analysis of the habcat list of stars.	18 scorpii wurde im september 2003 durch die astrobiologin margaret turnbull von der university of arizona in tucson als einer der aussichtsreichsten nahen kandidaten, um leben zu beherbergen, festgelegt.

8	in 1955, the betriebssportgemeinschaft turbine potsdam was founded.	1955 wurde die bsg turbine potsdam gegründet.
9	1-heptanol has a pleasant smell and is used in cosmetics for its fragrance.	1-heptanol dient in der parfümindustrie als zusatzstoff für nelken- und jasmin-düfte.
10	the film consists entirely of alfred hitchcock 's 1960 psycho (1960 film) psycho slowed down to approximately two frames a second, rather than the usual 24.	24 hour psycho ist eine videoinstallation des schottischen künstler douglas gordon aus dem jahr 1993, bei dem er den filmklassiker psycho (1960) psycho von alfred hitchcock auf eine abspiellänge von 24 stunden dehnt.
11	1981 saw the introduction of a single 2nd bundesliga of 20 teams.	1981 wurde die eingleisige 2. bundesliga mit 20 vereinen eingeführt.
12	it consisted of cousins kk (born kai mc-donald) and gangstad (born darius barnett).	sie besteht aus kk (geboren als k. mcdonald) und thad (geboren als d.barnett)
13	their second album, classic 220 on arista records, came eight full years after their debut.	acht jahre später veröffentlichten sie dann ihr zweites album classic 220, das auf arista records erschien.
14	the investigations ceased in 1955 and were closed in 1975.	die untersuchung wurde 1955 deaktiviert und 1975 endgültig eingestellt.
15	gray played high school basketball at emmaus high school in pennsylvania 's highly competitive lehigh valley conference, where he was a standout center (basketball) center.	gray war ein herausragender center auf der emmaus high school in emmaus, pennsylvania, bevor er von der university of pittsburgh verpflichtet wurde.
16	ogden moved to jersey city, new jerseyjersey city in 1829 and resumed the practice of law.	er zog daraufhin 1829 nach jersey city, wo er die letzten 10 jahre seines lebens verbrachte.
17	in 1987 he made a short comeback, suffering his lone loss by a knockout in seven to bobby joe young , a fringe contender of the era.	1987 machte er jedoch ein comeback, verlor aber durch kogegen bobby joe young, die erste und einzige niederlage seiner profilaufbahn.
18	in 1996, pryor was elected to the international boxing hall of fame.	im jahr 1996 wurde er in die international boxing hall of fame aufgenommen.
19	during his career in the entertainment industry, russo was manager for the manhattan transfer and bette midler from 1972 to 1979 whose breakout film, the rose (film) the rose, he produced.	während seiner karriere in der unterhaltungsindustrie war russo von 1972 bis 1979 manager von bette midler und the manhattan transfer und produzent der filme the rose und trading places (die glücksritter).

20	later, callahan persuaded siskind to join him as part of the faculty of the iit institute of design in chicago (founded by lazlo moholy-nagy as the new bauhaus).	callahan überredete ihn später, sich der fakultät des iit institute of design (new bauhaus ; illinois institute of technology) in chicago anzuschließen.
21	on march 1, 2007, the minnesota wild traded a seventh-round draft pick in 2008 nhl entry draft2008 to the new jersey devils for voros.	nachdem sein vertrag in minnesota zum ende der saison 2007 ausgelaufen war, unterschrieb er am 1. juli 2008 als free agent (nhl) free agent bei den new york rangers.
22	he graduated from the university of pennsylvania in 2001, where he was a member of sigma nu fraternity.	nach beendung seines studiums an der university of pennsylvania im jahr 2001 begann aaron yoo mit der schauspielerei.
23	the abasto de buenos aires was the central wholesale fruit and vegetable market in buenos aires, argentina, from 1893 to 1984.	der abasto de buenos aires war von 1893 bis 1984 der zentrale obst- und gemüsemarkt in der argentinischen hauptstadt buenos aires.
24	in 1999, he also published a collection of his poems.	eine erste sammlung mit gedichten von abbas kiarostami erschien 1999 in teheran.
25	datterode and röhrda, which had merged to form the greater community of netratal in 1972, were integrated into ringgau at the beginning of 1974.	datterode und röhrda, die sich 1972 in der gemeinde netratal zusammengeschlossen hatten, wurden zum beginn des jahres 1974 in die grogemeinde integriert.
26	the coat of arms was approved on 17 may 1977 by the hessian interior minister.	das wappen der wurde am 17. mai 1977 vom hessischen innenminister genehmigt.
27	final data collection is expected in december 2009.	die erfassung der daten soll bis ende 2009 abgeschlossen sein.
28	the first european to reach the region may have been the spainspaniard alonso de ojeda in 1499.	als erster europäer hat wahrscheinlich der spanier alonso de ojeda 1499 die region gesehen.
29	castle rätia ampla in riom, built 1227	burg rätia ampla in riom, 1227
30	the name ripperda is probably derived from the man s name rippert (rupert).	der name ripperda ist vermutlich eine ableitung des männlichen vornamens rippert (rupert).

Table A.1: Samples of Sentence Pair Extraction

No.	English Sentence	German Sentence
1	Frankfurt is a	Frankfurt ist eine
2	an der universität barcelona .	at the university of barcelona
3	location = berkeley	location = berkeley
4	in los angeles ist im	is in los angeles
5	first in the	in die erste
6	with the white stripes,	mit den white stripes , whites
7	world trade center	world trade center
8	new york,	new york.
9	the packers.	der packers.
10	anbar in iran.	anbar im iran.
11	and andersson and	sowie andersson und
12	beatles is the	beatles fällt die
13	in indonesien.	in indonesia.
14	to the film	die dem film
15	at granada (in brüssel .
16	one of daily mail	der daily mail .
17	the sculpture (der skulptur .
18	my mother got off the train	die mama aus dem zug (
19	the anti-christ .	der antichrist .
20	there were two more turns in the bundesliga, in	gab es zwei weitere drehungen in der bundesliga
21	in the same year , the first ever german	in den nächsten Jahren den Sprung in die erste
22	the album consists of an instrumental	das Album besteht aus einem Instrumentalen in
23	the second team plays in the 2nd	die zweite Mannschaft spielt in der 2
24	christopher and marco are his sons	zwei söhne , christopher und marco .
25	in of new jersey devils jersey	im trikot der new jersey devils
26	the total profits of the deutschen bank in the year 2005	des gesamten gewinns der deutschen bank im jahr 2005 der
27	who performed in casino	welche im film casino
28	the collections in Berlin	die antikensammlung berlin
29	sale of lasalle bank to bank of america	verkauf der lasalle bank an die bank of america
30	Amherst College will compete with the college about the williams	das amherst college konkurriert mit dem williams college um den

31	the Legislative Council of the colony vancouver Iceland, the	des legislativrats der kolonie vancouver island , dem
32	the government moved 1965 the road building in the suwannason	, verlegte die regierung 1965 das gebäude in die suwannason road
33	was in its time the holland house in kensington one	in dessen zeit wurde das holland house in kensington ein
34	of orlando predators , arena football	der orlando predators (arena football leagueaffl
35	division into by arthur andersen and andersen consulting	die aufteilung in arthur andersen und andersen consulting
36	berlin of june, he appeared at the opera in German	im juni erschien er in der deutschen oper in berlin .
37	a revolution in mainz	revolutionär in mainz
38	in zürich and new york , the citizen	in zürich und new york citynew york .
39	great acclaim concertante at the concertgebouw in amsterdam	ein unjubeltes konzert im concertgebouw in amsterdam .
40	owen sound attack of the ontario hockey League , and	den owen sound attack aus der ontario hockey league .
41	on science fiction and fantasy	auf science fiction und fantasy
42	mainly in los angeles, new york and london.	hauptsächlich in los angeles , new york und london .
43	french open in paris and the s.s. open	den french open in paris und den us open in
44	, The military traffic management command (; das military traffic management command (
45	for example, you might want a caption	, zB. könnten sie eine beschriftung , ein

Table A.2: Samples of Fragment Pair Extraction

Bibliography

- Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71, March 1996.
- 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. *Comput. Linguist.*, 16:79–85, June 1990.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June 1993.
- Richard H. Byrd, Richard H. Byrd, Peihuang Lu, Peihuang Lu, Jorge Nocedal, Jorge Nocedal, Ciyou Zhu, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1994.
- Yonggang Deng, Shankar Kumar, and William Byrne. Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(03):235, 2006.
- Corpora Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel. In *Proceedings of ACL*, pages 255–262, 2002.
- Mona Diab and Steve Finch. A statistical word-level translation model for comparable corpora. In *In Proceedings of the Conference on Content-based Multimedia Information Access (RIAO)*, 2000.

- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- 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, 1995.
- Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004a.
- Pascale Fung and Percy Cheung. *Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM*, pages 57–63. Association for Computational Linguistics, 2004b.
- Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 414–420, Stroudsburg, PA, USA, 1998.
- Percy Cheung Pascale Fung. *Sentence alignment in parallel, comparable, and quasi-comparable corpora*. 2004.
- E. Gaussier, J. m. Renders, I. Matveeva, C. Goutte, and H. Djean. A geometric view on bilingual lexicon extraction from comparable corpora. In *In Proceedings of ACL-04*, pages 527–534, 2004.
- Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors. NIPS Workshop Series. 2009.
- Sanjika Hewavitharana and Stephan Vogel. Enhancing a statistical machine translation system by using an automatically extracted parallel corpus from comparable sources. In *In Proceedings of the Workshop on Comparable Corpora, LREC08*, pages 7–10, 2008.
- Philipp Koehn. Europarl: A parallel corpus for statistical machine translation.

- Philipp Koehn and Kevin Knight. *Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm*. 2000.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007.
- Edward Loper and Steven Bird. Natural language processing toolkit, 2002.
<http://nltk.sourceforge.net/>.
- Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*, 2006.
- Robert C. Moore. Improving ibm word-alignment model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31:477–504, December 2005.
- 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*, ACL-44, pages 81–88, Stroudsburg, PA, USA, 2006.
- Douglas W. Oard. Cross-language text retrieval research in the usa. presented at 3rd ercim delos workshop. In *TREC-8 Experiments at Maryland: CLIR, QA and Routing*. In *Proceedings of the Eighth Text Retrieval Conference (TREC8)*, 1997.
- 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, Stroudsburg, PA, USA, 2002.

- Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March 2003.
- Chris Quirk, Raghavendra Udapa U, 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*, 2007.
- 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, Stroudsburg, PA, USA, 1995.
- Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA, 1999. ISBN 1-55860-609-3.
- Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29:349–380, September 2003.
- 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.
- Jason R. Smith, Chris Quirk, Toutanova, and Kristina. 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.
- 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, Stroudsburg, PA, USA, 2008.
- Staff. Book reviews. *Comput. Linguist.*, 29:655–664, December 2003.
- Christoph Tillmann. A beam-search extraction algorithm for comparable data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 225–228, Stroudsburg, PA, USA, 2009.

- 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 - Volume 1*, ACL '03, pages 72–79, Stroudsburg, PA, USA, 2003.
- 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*, COLING '96, pages 836–841, Stroudsburg, PA, USA, 1996.
- Dekai Wu and Pascale Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. *Natural Language Processing-IJCNLP 2005*, pages 257–268, 2005.
- Jia Xu, Richard Zens, and Hermann Ney. Sentence segmentation using ibm word alignment model 1. In *In Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation)*, pages 280–287, 2005.
- Yarowsky, Ngai David, and Grace. Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA, 2001a.
- David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–8, Stroudsburg, PA, USA, 2001b.
- 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–, Washington, DC, USA, 2002.