A Holistic Approach to Bilingual Sentence Fragment Extraction from Comparable Corpora

Mahdi Khademian[1], Kaveh Taghipour[1], Saab Mansour[2], Shahram Khadivi[1]

[1]Human Language Technology Laboratory, Department of Computer Engineering and IT
Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran
[2]Human Language Technology and Pattern Recognition Group
Computer Science Department, RWTH Aachen University, Aachen, Germany
khademian@aut.ac.ir, k.taghipour@aut.ac.ir, mansour@cs.rwth-aachen.de, khadivi@aut.ac.ir

Abstract

Achieving accurate translation, especially in multiple domain documents with statistical machine translation systems, requires more and more bilingual texts and this need becomes more critical when training such systems for language pairs with scarce training data. In the recent years, there have been some researches on new sources of parallel texts that are documents which are not necessarily parallel but are comparable. Since these methods search for possible translation equivalences in a greedy manner, they are unable to consider all possible parallel texts in comparable documents. This paper investigates a different approach for this need by considering relationships between all words of two comparable documents, which works fairly well even in the worst case of comparability. We represent each document pair in a matrix and then transform it to a new space to find parallel fragments. Evaluations show that the system is successful in extraction of useful fragment pairs.

Keywords: Parallel Fragment Extraction, Hough Transform, Statistical Machine Translation

1. Introduction

In statistical machine translation, parallel corpora are essential for training translation models. We require large datasets for training statistical machine translation (SMT) models and for estimating parameters. Nowadays, there are developed frameworks for creating SMT systems.

For many language pairs, currently we have acceptable translation systems not superior than human translations but very useful for getting enough information from different sources in different languages, especially the Web. Training SMT systems for new language pairs requires considerable amount of parallel sentence pairs for the languages. Therefore, finding new sources of parallel corpora becomes as an important activity in the field of SMT. Recently, there has been a great interest in automatically creating parallel texts from bilingual documents which are called comparable corpora (Resnik and Smith, 2003; Adafre and de Rijke, 2006; Munteanu and Marcu, 2006; Tillmann, 2009; Tillmann and Xu, 2009; Smith et al., 2010). The exact definitions and corresponding attributes are provided in (Fung and Cheung, 2004).

(Resnik and Smith, 2003) considers the Web as a massive data repository, which is useful to find information in multiple languages because of redundancy and repetition of data available on the Web. The research mines Internet archive data to extract parallel texts. There are similar studies such as (Munteanu and Marcu, 2006) and (Tillmann and Xu, 2009) in which parallel texts are extracted from multilingual news feeds even reported on different news agencies. In addition, recently, human gathered information of Wikipedia has attracted attention of some researchers to find more parallel sentences (Adafre and de Rijke, 2006; Smith et al., 2010).

Mining comparable corpora to collect parallel sentence pairs (or fragment pairs) has two major problems. The first problem is the document alignment and the second is the alignment of parallel sentences. In this paper, we do not concern the former and assume that all documents are aligned. After the document alignment procedure, sentence alignment or fragment extraction is performed.

Presented idea of this paper, addresses the second problem. We try to extract parallel fragments from comparable corpora by representing each document pair as a matrix of word level correspondence and then try to find fragments in a holistic manner by transforming the matrix to another space. Since there is no initial assumption on fragment boundaries in documents and using no heuristic methods for boundary detection, the method is in fact a holistic approach to fragment extraction. The rest of this paper is organized as follows: Section 2 lists the recent studies in the literature. Section 3 focuses on the the idea behind this approach. In section 4, details of the proposed method are explained. Section 5 describes data, experimental setup and evaluations and finally, section 6 concludes the paper.

2. Backgrounds

Early successful researches in SMT were based on parallel documents. In parallel documents most of the source sentences are translated to the target language and there are 1-1 (one-to-one), 1-2, 2-1 and 2-2 relationships between sentences. An effective statistical sentence aligner is developed in (Gale and Church, 1991) and successive researchers have proposed new alignment techniques for parallel corpora in (Melamed, 1996) and (Moore, 2002). (Gale and Church, 1991) finds best sentence alignments based on a dynamic programming fashion method which tries to maximize overall translation probability. These methods use sentence position, length and word counts of sentence to find the alignment. Most of these efforts worked well on parallel corpora. The authors of (Fung and Cheung, 2004) have stated that as bilingual corpora become less parallel, it is better to rely on information about word translations rather than sentence length and position. Therefore, new methods for handling comparable corpora situations are needed.

Most of the researches done in previous years are considering comparable corpora rather than parallel documents. Webpages hierarchical structures and some criteria such as document length is used in (Resnik and Smith, 2003) to construct candidate documents and then extract texts based on document structures, mainly HTML tags. (Fung and Cheung, 2004) present a method which works similarly in two phases of document alignment and then sentence alignment. This method does not take sentence positions into account and mainly uses the number of translated words in sentence pairs to extract parallel sentences. Similarly, there are two steps in the work presented in (Munteanu and Marcu, 2006). Publication dates and some vector based features are used for matching news articles. Then, the method uses IBM word translation probabilities and alignment models to calculate log-likelihood ratios for sub-sentential fragment extraction based on an inspired signal processing method. Another research, published in (Adafre and de Rijke, 2006), uses two approaches to find similar sentences in the Wikipedia pages across multiple languages. Primarily, it uses inter language and cross language hyperlinks of documents to populate and match documents of Wikipedia. This approach uses translation systems to translate one sentence into the target language and then compare machine translated result with the target document. By using similarity measures it detects parallel sentences across documents. The second approach of this research incorporates bilingual lexicons which are created from multiple language titles of the same documents. By applying the created lexicon and similarity measure, parallel sentences are extracted. (Smith et al., 2010) trained a ranking model for sentence extraction. It uses some features such as alignment probability, length of aligned sentences and word fertilities in addition to the markup features of Wikipedia.

Many of these researches focused on parallel sentence extraction while some others could extract parallel sentence fragments. This paper presents s different approach for parallel fragment extraction from comparable corpora. Next section presents the idea and some related examples.

3. BiText matrix creation from comparable documents

The problem of fragment extraction deals with some sort of search and sentence (or fragment) pair selection. In other words, it is an assignment problem for assigning sentences in source document to target ones. This assignment is carried based on an association measure among words or sentences in document pairs.

The first step in our approach, is to represent a document pair, in a two dimensional matrix of association scores between each source and target word. For example, we can use IBM 1 word alignment model (Brown et al., 1993) for scoring degrees of association between each two words. A sample document pair is illustrated in figure 1.

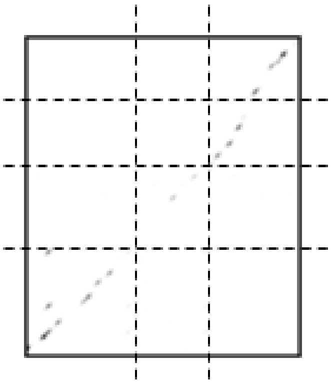

Figure 1: The matrix of a sample document pair scored with IBM model 1. Higher scores are illustrated with darker points and dashed lines are sentence boundaries.

If we assume the matrix in a Cartesian coordination system with the origin at the bottom-left, the horizontal axes covers source text words from left to right and the vertical one relates to the words of the target text upward. Each entry in this matrix shows the score between a source and target word. We call the region around the entries with higher scores, translation regions. Roughly speaking, two translation regions appear in figure 1. Therefore, the matrix suggests that three of four sentences in the target text are possibly translations of two sentences in the source text, or at least the two texts have some parallel sentence fragments. In this example, the first sentences of source and target languages are as follows (sentences are selected from Europarl corpus test set (Koehn, 2005)):

- **English:** my group and I therefore suggest that, tomorrow if possible, or another day if necessary, we look into finding a way to devote an hour to this extremely important subject, and to adopt a resolution.
- **French:** je suggre donc avec mon groupe, de chercher, demain si possible, un autre jour si ncessaire, comment il serait possible de consacrer une heure ce sujet, combien important et d'adopter une rsolution.

Figure 1 is an example of matrices created from comparable documents. This matrix is called a BiText of two documents. Appearance of a sequential linear segment in a BiText is an indicator for potential translation equivalence between segments in the source and target documents. We can expect diagonal sequences of higher scores in the BiText matrices of parallel documents.

A BiText of two parallel documents would be similar to figure 2. This example is extracted from the Universal Declaration of Human Rights in English and French languages from the UN website. As expected, a diagonal sequence which indicates translational equivalence between two texts is appeared in the BiText. Another BiText of the same document pair but with permuted source sentences is shown in figure 3.

Mining parallel texts from comparable documents or even parallel corpora has some difficulties. One issue of mining

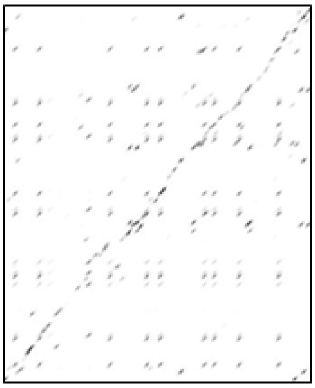

Figure 2: A BiText matrix from two texts

in parallel corpora is the type of relation between sentences. As an example, in some methods there is only assumption of one-to-one relations between sentences in the source and target documents while the others assume one-to-two, two-to-one or higher order relationships between sentences in the source and target documents.

The proposed method of extraction does not consider sentence boundaries and obviously has no assumption of any kind of relationships between sentences. For example, in figure 1, we can see that the third and fourth sentences in target document are translations of the third sentence in the source text. This observation shows that, the proposed method also works well for sentence fragment extraction. For example the following fragment pairs are extracted from preamble of the Universal Declaration of Human Rights based on its BiText matrix (see figure 2).

- **English:** is essential to promote
- **French:** est essentiel d'encourager
- **English:** of friendly relations between nations
- **French:** de relations amicales entre nations considrant
- **English:** all peoples and all nations, to
- **French:** tous les peuples et toutes les nations afin

Another problem in mining parallel texts from comparable corpora is the displacement of translations between two documents. In other word, we can expect a translation of a sentence in the source document everywhere in the target document if exists. Figure 3 shows that a BiText matrix reveals every possibility for parallel fragments.

Therefore the method can recognize fragments anywhere in the documents. The experiments show that the proposed method works also on the worst case of comparability; it is capable of extracting tiny parallel fragment pairs in such documents.

By considering linear sequences of word pairs with higher scores as parallel, the next main step for fragment extraction is the detection of such sequences in the BiText matrices. This problem can be handled by transforming the matrix to a new space. Next section describes the technical aspects of the transformation.

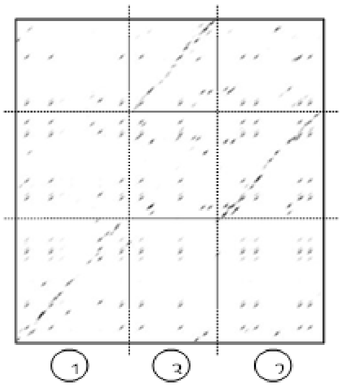

Figure 3: A BiText created from a permuted text of preamble of the Declaration of Human Rights

4. The proposed method

The three major steps of the proposed method are as follows:

- Scoring word pairs to form BiTexts
- Required preprocessing
- Fragment Extraction

The BiText matrix is created based on a translation model which eventually evaluates two words association in the source and target documents. This process is described in section 4.1. We can employ some techniques in the scoring step to reduce noise and achieve better linear sequence detection. These techniques are described in section 4.3. Linear sequence detection and parallel fragment extraction is done by applying Hough transformation on the matrix. This issue is covered in section 4.2. Finally, there are some computational remarks and parameter tuning issues which are discussed in the section 4.4.

4.1. Scoring Word Pairs to Form BiTexts

The idea of using BiText matrices is roughly described in section 3. The words in source and target documents form a matrix of scores. The scores are calculated based on a translation model which can also incorporate word alignments and even reordering models in addition to translation model probabilities.

From now on, we simply use word translation probabilities calculated based on IBM model 1 (Brown et al., 1993). Thus, there is a need for an initial parallel corpus for training the IBM model 1 and then scoring matrices for the selected comparable documents. In addition, it is appropriate to normalize association levels to have a homogeneous set of scores.

4.2. Linear Sequence Detection

The method used in this paper is called the Hough Transformation (Duda and Hart, 1972). Here the process for linear sequence recognition and then fragment extraction based on the Hough Transformation is described briefly; we will not delve into the details of this procedure and just provide some overview of the overall process. The overall procedure is carried out in four main steps:

- matrix transformation
- finding points with higher values in the transformation space
- linear sequence detection based on the scores found in the previuos step
- fragment extraction

4.2.1. Transformation

The Standard Hough Transform (SHT) uses the parametric representation of a line with the following equation:

$$\rho = x.cos(\theta) + y.sin(\theta) \quad (1)$$

In this equation theta is the angle of perpendicular line from origin to the line to be projected into the transformed space and rho is the mathematical distance of the line from origin. This suggests that every linear sequence in Cartesian space is transformed into a point in the target space (variables in the target space are theta and rho).

4.2.2. Finding Points with Higher Values

As it is mentioned in the previous sub-section, points with higher values in transformed space are related to linear sequences in the BiText matrix. One method for finding these points is to grid the transformed space and then score and sort the grid boxes according to the values of their inside points to make a sorted list. The second method is repeatedly finding the greatest scores and then removing the neighbors. After this step, a list of points related to their corresponding theta and rho is populated.

4.2.3. Linear Sequence Detection

Each point in the Hough transformed space is related to a linear sequence in the BiText matrix. For each point found in the previous step we can do the following procedure to detect its related sequence in the BiText matrix:

- Extract and sort entries in the initial BiText matrix related to their corresponding value in the Hough transformed space to form a sequence.
- Two constraints are considered when extracting fragments. The first is responsible for eliminating short sequences and the other helps us fill gaps between fragmented sequences to form longer ones.

4.2.4. Fragment extraction

The detected linear sequences are in translation regions and are related to a series of source and target word pairs. Thus, for each sequence there is a corresponding source fragment and another in the target document. By applying some smoothing mechanism to the sentence boundaries, parallel complete sentences can be extracted from comparable corpora based on linear sequences.

4.3. Required Preprocessing

Figure 4 shows the initial BiText matrix of preamble of the Universal Declaration of Human Rights which was depicted before in the figure 2. These horizontal and vertical noisy entries are related to popular words in the word translation probability table of IBM model 1. Examples for these words are stop words and punctuations. It is apparent that noisy points in the BiText matrix could contribute to form many lines in variety of degrees. Considering these words would result in creation of semi-smoothed transformed matrix in which distinguishing useful points with high values is difficult. The negative effect of popular words in the dictionary is eliminated by setting their association levels in the BiText matrix to zero. But which words should be selected for this purpose? By counting number of occurrences of words in dictionary and applying a threshold, for example the top 1% of the most popular words, these words are selected. This simple strategy removes most of the noise in the BiText matrix (figure 2). Even after omitting such words, there is still a small degree of noise in the matrix. Remaining noise has a negative effect by forming inappropriate horizontal and vertical lines. In addition, the noise constructs long sequences and which are not useful.

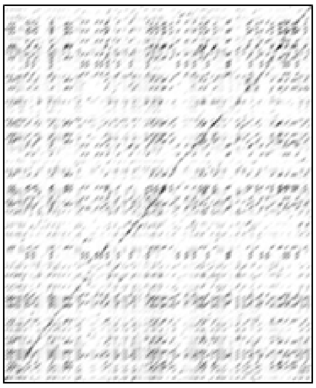

Figure 4: Initial BiText matrix of figure 2

A long vertical sequence means that one source word is translated into a series of target words (for example 300 words), which is meaningless. This observation suggests that we should limit ranges in which we try to find points with higher value in the transformed matrix.

Due to the nature of the proposed method and fragmentation of appeared sequences in the BiText matrix, extracted sentence fragments are very short. To achieving longer segments, BiText matrices can be filtered and smoothed by filters such as Gaussian blur filters. In addition, the kernel of Gaussian blur filter could be improved by the following consideration. The first example represents kinds of language pairs in which in average each source word translates to one target word. In other words, translation from the source to target language in average has the fertility of one. In the second example, fertility of translation of words in source to target language is greater than one and in the third example it is less than one. Based on language pairs we can choose covariance matrix for Gaussian blur which improves sequence detection. Some illustrative examples are drawn in figure 5.

4.4. Computational remarks and parameter tuning

There are two computational issues for transforming BiText matrices. The first one is that the BiTexts are sparse and the second one is about limiting theta range on the transformed

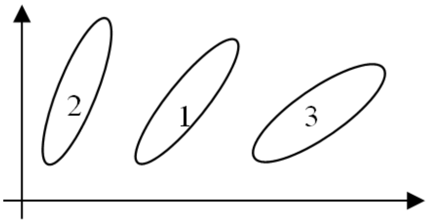

Figure 5: Some illustrative translate regions

matrix. As it is mentioned in sub-section 4.2.1 for computing Hough transform we should iterate over all entries of the BiText matrix and calculate contribution of each entry in the whole theta range (from -90 to 90 degrees in a quantized manner). On the other hand, as it is expected, the created BiText matrix is very sparse. Moreover, in extremely quasi-comparable documents we should expect only small number of points and tiny fragments. By using this assumption, there is no need to iterate over all entries of the BiText matrix and there is the need to only calculate contributions of non-zero entries.

For the second issue, figure 5 shows expectation for translation regions and therefore it shows that appeared sequences in the BiText matrix always have approximately similar constrained positive angles. Therefore, there is no need to calculate whole theta range for the transformation. It is obvious that the limitation utilized here omits the need for transforming the whole matrix and thus leads to reduced calculations significantly. In the process of fragment extraction, there are some parameters which need to be tuned. Some of the parameters are as follows: the threshold used for word elimination, the kernel for Gaussian blur filter, theta and rho quantization accuracy, theta limits, neighborhood radius for points in transformed matrix, number of points in transformed matrix, minimum sequence length and filling size for gaps. Most of these parameters can be viewed either strict or relaxed, regarding to the number of the future extracted parallel fragments. While section 4.3 provides an example for automated parameter tuning, we believe that by using relaxed parameters most possible parallel fragments could extracted and by applying some filtering mechanism we are able to extract best fragment pairs. This can be done by applying a translation model and scoring extracted parallel fragments. This method is also used in the recent researches such as (Resnik and Smith, 2003) (Munteanu and Marcu, 2006) (Tillmann, 2009) (Smith et al., 2010).

5. Experiments

During the proof of concept of this research, many possible comparable corpora were tested such as Wikipedia documents and multilingual news. For the experiments we prefer to use source in which we can align the documents easily by hyperlinks or other simple information. For example some news agencies provide a simple way to view and use their archived news in many languages.

We use news articles from the Euronews, which is an international news channel covering world events, currently presents them in 9 languages and has a public archive for its news from the year 2004. For the experiments, English-French language pair is selected and document pairs in these two languages are downloaded. The IBM model 1 translation table is trained based on Europarl parallel corpus (Koehn, 2005). For this purpose the freely available software GIZA++ is used (Och and Ney, 2003).

Two types of evaluation are carried out on the extracted fragments. The first one is done subjectively, in which around 100 fragments are selected randomly among the whole extracted fragments and judged and categorized by human experts in one of the following three categories: correct, fair and irrelevant fragment pairs. Some examples of correct, fair and irrelevant extractions are listed. The evaluation result is summarized in the table 1.

In our subjective experiment, we use a phrase based statistical machine translation system (Koehn et al., 2003). We extracted around 43000 fragment pairs from approximately 17500 news articles from the Euronews archive. If we use a randomly selected test set to evaluate the data, there will be too many out of vocabulary errors (17% of words) due to the different domains of the Europarl and extracted fragments from Euronews. Thus, we should select a test set which contains similar sentences to the extracted fragments.

In order to select sentences with such characteristics, first, we randomly selected 20,000 sentence pairs from the Europarl corpus. Then we trained two 3-gram language models on the source and target parts of the extracted fragments and calculated the perplexity for all 20,000 previously selected sentence-pairs.

Correct:

English: the best infrastructure in the world
French: les meilleures infrastructures du monde
English: our children are sick
French: nos enfants sont maladies
English: this conflict has to end as soon as possible, every time a child dies
French: ce conflit doit prendre fin ds que possible a chaque fois quun enfant meurt

Fair:

English: he wants, but the palestinian people
French: le veut , mais le peuple palestinien lui
English: public transport network began on monday night and
French: les transports a dbut ds lundi soir ,
English: end of november londoners and
French: 28 novembre londres doit

Irrelevant:

English: said mexico was looking increasingly
French: le mexique lquivalent du plan
English: almost 30000 londoners
French: perdu 30000 de ses

Finally, we selected the 2000 sentence pairs with the least average perplexity of source and target, out of the 20,000 pairs. We use the first 1000 sentence pairs as our tuning

Table 1: Results of the extracted parallel segments from the Euronews

number of news	300
number of extracted fragments	850
number of randomly tested fragments	100
number of correct extractions	53
number of fair extractions	29
number of irrelevant extractions	18

Table 2: Results of the objective test and measurements based on BLEU score for the tuning phase

	Tuning set	Test set
Baseline (2000 sentence pairs)	31.45	31.43
+43083 fragments	33.93	33.34

set and the second 1000, as our test set. By this selection mechanism, we could reduce the OOVs to about 3%. We also selected another 2000 sentence pairs from the Europarl corpus, randomly, as our training data for the baseline.

The second system is trained with the training data of the baseline system plus about 43000 fragment pairs. Then it is tuned and tested with the same tuning and test set of the baseline system.

A summary of the evaluations of the system on the tuning and the test set is presented in the table 2. We consider this phase as a weak objective experiment in which BLEU score for the first system with poor training set increased by around 7.5% by adding new source of training data.

As mentioned earlier, since the tuning and the test set are selected objectively rather than with a completely random selection, the BLEU scores achieved are relatively high. In addition, the amount of improvement might be misleading because of the small training set used for the baseline. Hence, concluding about the amount of improvement quantitatively is not logical, but can be seen as a support for the subjective evaluation.

Moreover, we can confidently claim that the fragments extracted with the proposed system, are useful for translation and are helpful in training statistical translation models. Thus the system can be used to collect parallel corpora of fragments, especially for the language pairs with scarce parallel data but adequate bilingual content.

6. Conclusion

This paper initially discusses the role of parallel corpora in training statistical machine translation systems and also talks about the importance of new sources for extracting parallel texts, especially for language pairs with scarce parallel corpora. It briefly reviews recent research related to comparable corpora and their achievements in gathering parallel texts from new document sources.

The main idea of this paper, forming a matrix of translation scores and finding the patterns, is explained after the backgrounds. This idea, by comparing with recent researches, is a different approach for working on comparable corpora. After presenting the proposed method, it is evaluated by a subjective and an objective test.

For the subjective evaluation, a random subset of sentence fragments which were extracted by the proposed method is evaluated by human experts. In this experiment 82% of extracted fragment pairs are evaluated as good and fair translations. The results of the objective evaluation show an improvement of translation quality and confirm the results of the subjective test.

We suggest some future researches such as extraction of weak translation pairs and filtering them by some kind of scoring mechanisms, use of sophisticated association levels between words which includes word alignment and reordering, more investigation on geometrical aspects of translation regions and applying smart kernels for handling reordering. Also, the source code and other resources will be published for future use of researchers interested in this method.

7. Acknowledgements

This material was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

S. F. Adafre and M. de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

Richard O. Duda and Peter E. Hart. 1972. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15:11–15, January.

Pascale Fung and Percy Cheung. 2004. 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. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

I. Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. *Computing Research Repository*, cmp-lg/960.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29:349–380, September.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. 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. Association for Computational Linguistics.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 93–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christoph Tillmann. 2009. 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. Association for Computational Linguistics.