Automatic Filtering of Bilingual Corpora for Statistical Machine Translation

Shahram Khadivi and Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55
52056 Aachen, Germany
{khadivi,ney}@cs.RWTH-Aachen.de

Abstract. For many applications such as machine translation and bilingual information retrieval, the bilingual corpora play an important role in training the system. Because they are obtained through automatic or semi automatic methods, they usually include noise, sentence pairs which are worthless or even harmful for training the system. We study the effect of different levels of corpus noise on an end-to-end statistical machine translation system. We also propose an efficient method for corpus filtering. This method filters out the noisy part of a corpus based on the state-of-the-art word alignment models. We show the efficiency of this method on the basis of the sentence misalignment rate of the filtered corpus and its positive effect on the translation quality.

1 Introduction

Bilingual corpora play an important role in developing statistical machine translation systems. But, since the manual compilation of bilingual corpora is a very expensive process, most of available bilingual corpora are generated in an automatic way. The parallel documents[1] are becoming more and more available, mainly on the Web, so that there is a need for automatic methods for bilingual corpus compilation. The automatically generated corpora usually include noise, sentence pairs which are worthless or even harmful for training the system. The noise might be due to any difference between the contents of source and target documents, non-literal translation, or errors in aligning documents, paragraphs, and sentences.

Related Work

Automatically generated bilingual corpora usually contain a considerable number of noisy sentence pairs. These noisy sentence pairs may have a negative impact on the training of the statistical machine translation or bilingual information retrieval systems. Due to this problem, various researchers have investigated different methods for corpus filtering. Here, we give a brief overview of the important works done in this field.

[1] documents available in more than one language but with the same content

In [1], the authors remove parallel documents for which their respective file size differ largely or for which, after applying sentence alignment, a relatively large number of empty alignments appear. They also make use of the length similarity between sentence pairs as well as the existence of bilingual dictionary entries in a sentence pair.

In [2] and [3], the authors make use of a literalness criterion for each sentence pair to filter a noisy Japanese-English corpus. They measure the literalness between source and target sentences by referring to a translation dictionary and counting the number of times that the translation dictionary entries occurred only in the source sentence, only in the target sentence, or in both source and target sentences.

In [4], the author studies the use of a noisy corpus in addition to a large clean training corpus in order to improve the translation quality. He identifies the noisy sentence pairs by accumulating five alignment scores for each sentence pair based on the following features: three different sentence length features and two lexical features based on IBM model 1 score. The sentence pairs which have a score less than a threshold are considered as noise.

In [5], the effect of parallel sentence extraction from in-domain comparable corpora on the machine translation performance has been studied. They align each sentence in a source document to all possible target sentences in several associated target documents. The associated target documents are the most similar documents to the source document among a relatively large number of documents. Then, they filter out noisy sentences by using two classifiers: a simple rule-based classifier and a maximum entropy based classifier. They show the significant improvement of the end-to-end translation quality, when the extracted corpus is added to the baseline out-of-domain corpus.

All above authors have investigated different methods for corpus filtering and showed the positive effect of corpus filtering on their statistical machine translation systems. But, the impact of the level of corpus noise on training the statistical machine translation models has not been specifically investigated. In this paper, we study the effect of different levels of corpus noise on an end-to-end statistical machine translation system. We also introduce an efficient approach for corpus filtering, and we show that which specific model among different statistical machine translation models must be trained on the filtered corpus.

The remaining part of this paper is organized as follows. Section 2 deals with our statistical translation engine. In section 3, we describe the corpus compilation procedure. In section 4, we introduce the length-based and translation likelihood-based filtering. In section 5, we deal with the experiment results and the related discussion.

2 Statistical Machine Translation

In statistical machine translation , we are given a source language (‘French’) sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language (‘English’) sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we will

choose the sentence with the highest probability:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \ \left\{ Pr(e_1^I | f_1^J) \right\} \tag{1}$$

$$= \underset{e_1^I}{\operatorname{argmax}} \ \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \tag{2}$$

The decomposition into two knowledge sources in Equation 2 is known as the source-channel approach to statistical machine translation [6]. It allows an independent modeling of target language model $Pr(e_1^I)$ and translation model $Pr(f_1^J|e_1^I)$[2]. The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence. It can be further decomposed into alignment and lexicon model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. We have to maximize over all possible target language sentences.

An extension to the classical source-channel approach is the direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$. Using a log-linear model [7], we obtain:

$$Pr(e_1^I|f_1^J) = \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right) \cdot Z(f_1^J)$$

Here, $Z(f_1^J)$ denotes the appropriate normalization constant. The term $h_m(e_1^I, f_1^J)$ denotes various models which are involved in the translation process. And each model is weighted by its model scaling factor λ_1^M. As a decision rule, we obtain:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \ \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

This approach is a generalization of the source-channel approach. It has the advantage that additional models or feature functions can be easily integrated into the overall system. The model scaling factors λ_1^M are trained according to the maximum entropy principle, e.g. using the GIS algorithm. Alternatively, one can train them with respect to the final translation quality measured by some error criterion [8].

3 Corpus Compilation

Manual collection of bilingual text corpora might be very expensive, e.g. in the EUTRANS project [9]. An efficient and cheap approach for collecting parallel text is mining the Internet for parallel documents [10] and [11]. For the purpose of bilingual corpus generation, we download the parallel documents in HTML format from *the European Union Website* [3] which exists in all official languages of the European Union.

[2] The notational convention will be as follows: we use the symbol $Pr(\cdot)$ to denote general probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(\cdot)$.

[3] `http://europa.eu.int`

The documents are already aligned at the document level. After extracting the plain text from HTML documents, we apply a hierarchical rule-based method for text tokenization. Then, by using an automatic sentence aligner program, we form a bilingual corpus. The last step is the corpus filtering to obtain a high quality bilingual corpus for training the translation models. The paragraph/sentence alignment, and the corpus filtering will be described in the next sub-section and Section 4, respectively.

Sentence Alignment

We employ an improved version of the Gale and Church algorithm [12] to align the paragraphs and sentences in the parallel documents. The Gale and Church algorithm is a dynamic programming method for aligning corresponding sentences in two parallel documents. This method is based on the length of the two sentences to be aligned. The length of the sentence is measured in term of its characters.

We extend the algorithm in the following ways. In order to improve the sentence alignment quality, we make use of a small dictionary. Each entry in the dictionary is used as an anchor point. In other words, each entry in the dictionary tells the sentence alignment algorithm if the source word (phrase) of a given entry exists in source sentence then the target word (phrase) might be in the target sentence. To each entry of the dictionary, two values have been assigned: presence bonus and absence (mismatch) penalty. By introducing these two values, the dictionary entries can be taken into account in alignment algorithm more easily and efficiently.

We also integrate several heuristics to the Gale and Church algorithm. Some of them are as follows:

- If the source sentence starts with a digit or an item symbol, it is very likely that the target sentence starts with the same digit or symbol.
- The number of numerical sequences in two aligned sentences should be very close to each other.
- The number of parenthesis pairs in two aligned sentences should be very close to each other.

There are a few other sentence alignment algorithms available like [13], [14], [15], [16], and *Champollion*, the sentence aligner of the Linguistic Data Consortium[17]. In [18], there is an evaluation on different sentence alignment methods on the task of Portuguese and English parallel texts. They mentioned that due to the very similar performance of the methods, choosing the best sentence aligner is not an easy task. In addition, they found that the performance of the sentence aligners vary for different tasks. Due to this result and the aim of our research which is studying the effect of noise on statistical machine translation system, finding and employing the best sentence aligner is not a crucial matter in this task.

To obtain a bilingual corpus from the documents available in *the European Union Website*, we apply the developed sentence aligner in two steps: aligning the paragraphs of documents, and aligning the sentences within two aligned paragraphs. The following alignment mappings between source and target sentences are permitted: 1:1, 1:0, 0:1, 2:2, 2:1, 1:2. In this paper, we refer to the bilingual corpus obtained from *the European Union Website* as the EU corpus.

4 Corpus Filtering

The automatic methods for corpus compilation are noise-prone. The main reasons are sentence alignment errors and the existence of the noise in the original source and target documents. The latter is very obvious in comparable corpora which make them more complicated for obtaining bilingual sentence pairs. But also in parallel documents, two corresponding documents may not be fully suitable for a bilingual corpus extraction due to free translation or summarization of the original text, selective translation of important parts of the original document, or existence of useless data like tables in the parallel documents. An additional type of noise is caused by the use of a third language in the original documents. For example, in the EU Spanish corpus, we may have a sequence of words in French.

Thus, there is high probability that an automatically aligned corpus contains many noisy sentence pairs which are worthless or even harmful for statistical machine translation . Therefore, we need a filtering scheme to remove the noisy sentence pairs from the corpus. We will describe two different methods for corpus filtering:

- a *length-based filtering* which makes use of length constraints of sentence pairs,
- a *translation likelihood-based* which makes use of the translation likelihood measure for the given sentence pairs.

4.1 Length-Based Filtering

We develop a length-based filtering algorithm to remove presumably harmful or worthless sentence pairs. The rules are based upon the length of the source and target sentences and work as follows:

1. The lengths of source sentence and target sentence must not differ largely. When I and J denote the lengths of target and source sentences, respectively, the above rule can be expressed by the following detailed rules:
   - $(6 \cdot I > J \wedge I < 6 \cdot J)$
   - $(I < 3 \vee J < 3 \vee (I < 2.2 \cdot J \wedge J < 2.2 \cdot I))$
   - $(I < 10 \vee J < 10 \vee (I < 2 \cdot J \wedge J < 2 \cdot I))$
2. At least one alphabetical character must occur in each sentence of a sentence pair.
3. The sentence end symbols in source sentence and target sentence must be similar.
4. The source sentence and the target sentence must not be empty.

In addition, we also identify the language of the source and target sentences to be in the language which is supposed. The developed language identification system is a maximum entropy based language classifier for identifying the language of each text line using the YASMET toolkit [19]. The maximum entropy features are the most frequent trigrams of each language.

The main problem of this filtering scheme is that it also removes many correct or useful sentence pairs from the corpus. In other words, the length-based method can clean the corpus with a high precision but with a low recall.

4.2 Translation Likelihood-Based Filtering

In order to filter out worthless or harmful sentence pairs from the compiled bilingual corpus in a more systematic scheme, we make use of the translation probability of each sentence pair which is produced by word alignment models [20] and [21]. For this purpose, we train IBM model 1, Hidden Markov model, and IBM model 4 in a successive manner using the maximum likelihood algorithm on the whole corpus (unclean corpus). The final parameter values of a simpler model serve as starting point for a more complex model. We train these models in both directions, from source to target and from target to source. Hence, for each sentence pair we have two probabilities: $\sum_{a_1^J} p(f_1^J, a_1^J|e_1^I)$ and $\sum_{b_1^I} p(e_1^I, b_1^I|f_1^J)$ where a_1^J / b_1^I is an alignment which describes a mapping from the source / target position j / i to the target/source position a_j / b_i . By scaling these probabilities with the source and target sentence lengths, we arrive at the following score for each sentence pair (f_1^J, e_1^I) in the corpus:

$$\begin{aligned}\text{Score}(f_1^J, e_1^I) =& \frac{1}{J} \log \sum_{a_1^J} p(f_1^J, a_1^J|e_1^I) + \\ & \frac{1}{I} \log \sum_{b_1^I} p(e_1^I, b_1^I|f_1^J)\end{aligned} \tag{3}$$

A very good approximation and computationally efficient variation of Equation 3 is achieved by calculating the Viterbi alignment instead of the summation over all alignments, i.e. by replacing the the $\sum$ operator with $\max$ operator in the equation:

$$\begin{aligned}\text{Score}(f_1^J, e_1^I) =& \frac{1}{J} \log \max_{a_1^J} p(f_1^J, a_1^J|e_1^I) + \\ & \frac{1}{I} \log \max_{b_1^I} p(e_1^I, b_1^I|f_1^J)\end{aligned} \tag{4}$$

Now, we have a score for each sentence pair. We empirically determine the threshold value for discriminating correct sentence pairs from incorrect sentence pairs.

The translation likelihood scores can also be utilized for corpus weighting instead of corpus filtering. It means that the sentence pairs with a better score will get a higher weight. It reduces the impact of noisy sentence pairs on training the statistical machine translation models.

5 Results

Here we will present the results of the corpus filtering for two corpora, Xerox and EU. The language pair for both corpora is Spanish-English. The Xerox corpus (Table 1) is a noise free corpus which has been manually aligned, it is composed of technical manuals describing various aspects of Xerox hardware and software installation, administration,

Table 1. Statistics of the Xerox corpus

	English	Spanish
Train: Sentences	56K	
Words	665K	753K
Vocabulary Size	8K	11K
Vocabulary Singletons	2K	3K
Test: Sentences	1125	
Words	8K	10K

Table 2. Statistics of the EU corpus

	English	Spanish
Train: Sentences	975K	
Tokens	19M	22M
Vocabulary Size	73K	94K
Vocabulary Singletons	25K	32K
Test: Sentences	2000	
Words	48K	54K

usage, etc. The EU corpus (Table 2) has been automatically aligned and is a noisy corpus (details in Section 3).

To study the effects of noise on an end-to-end statistical machine translation , we use the Xerox corpus. Then, by introducing artificial noise on this corpus, we study the effect of noise on statistical machine translation . We make use of the EU corpus as a real case study for analyzing the effect of translation likelihood-based filtering.

We evaluate the proposed filtering scheme by two criteria, sentence alignment evaluation and end-to-end translation quality.

5.1 Sentence Alignment Evaluation

To evaluate the sentence alignment quality, we select the EU corpus as a noisy corpus. We generate two clean corpora by applying each of the two corpus filtering methods to the noisy corpus. In each corpus, we keep just one instance per sentence pair, then the sentence alignment evaluation will be more accurate. We randomly selected 400 sentence pairs from each corpus. Then, we asked an expert to judge sentence alignment accuracy in all sentence pairs by assigning correct or incorrect to each pair. The details of sentence alignment evaluation for the length-based filtered corpus and translation likelihood-based filtered corpus are shown in in Table 3.

The results show that the translation likelihood-based filtering is better than the length-based filtering in removing incorrectly aligned sentence pairs. At the same time, we observed that the number of filtered sentence pairs in the translation likelihood-based filtering is less than the length-based filtering. This observation along with the significance test confirm the superiority of translation likelihood-based filtering over length-based filtering.

Table 3. Sentence Alignment Evaluation

Filtering method	Misalignment error rate [%]
Length-based	5.0
Translation likelihood-based	3.2

In order to measure the efficiency of translation likelihood-based filtering , we perform another experiment on the Xerox corpus. We introduce different levels of noise to the clean Xerox corpus, by randomly scrambling a given amount of the sentence pairs, i.e. we randomly select two sentence pairs with about the same length and then we make them noisy by exchanging their target parts. After making the corpus noisy, we apply the translation likelihood-based filtering to the corpus and measure its accuracy in identifying the noisy sentence pairs. This type of noise can not be identified by any length-based filtering approaches. Table 4 shows the results of this experiment, the first column is the level of introduced artificial noise to the clean corpus and the second column is the accuracy of identifying the noisy sentence pairs. The results show even with 80% noise in the corpus the translation likelihood-based filtering is able to identify the noisy sentence pairs with the accuracy about 90%.

Table 4. Sentence Alignment Evaluation of the Xerox Artificial Noisy Corpus

Fraction of incorrect sentence pairs [%]	Filtering error rate [%]
20	10.4
40	11.9
60	13.0
80	11.6

5.2 Translation Results

In this section, we study the effect of corpus noise on the translation quality of an end-to-end statistical machine translation . We make use of a phrase-based translation engine [7]. In all translation experiments, we will report the baseline translation results in BLEU score [22].

In the first experiment, we study the effect of different levels of corpus noise on the translation quality. Again, we use the Xerox corpus which is a clean corpus, and introduce different levels of artificial noise to the corpus with the same method as described in the last section. Then, we train the statistical machine translation models on each of the artificially noisy corpora. The translation results are shown in Table 5. The first column shows the percentage of noisy sentence pairs in the corpus. The second column shows the translation scores in BLEU when the full corpus is utilized for training the

system. The last column shows the translation results when only the clean part of the corpus is used for training.

Table 5. Translation Results on the Xerox Artificially Noisy Corpus

Artificial Noise [%]	the whole corpus BLEU [%]	only clean part of the corpus BLEU [%]
0	61.2	61.2
20	60.4	59.7
40	58.1	58.5
60	52.4	54.2
80	44.0	49.1

As we expected, the translation quality decreases with a growing level of noisy sentence pairs. Another important observation of this table is the difference between the translation results if we use the whole corpus or only the clean part. Even with about 40% of noise, the difference in BLEU score is about 0.4%. In summary, this table states that the corpus noise does not deteriorate the translation results.

There exist three important models in training our statistical machine translation system. They are word alignment model(s) (WA), bilingual-phrase model (BP), and language model (LM). The bilingual-phrase model must be trained on the clean part of a corpus, as there is no useful bilingual information assumed to be in the noisy sentence pairs. But, the effect of noise in training the word alignment model (WA) and its impact on the translation quality is unclear. We also study the effect of noise in training the language model (LM), as one type of corpus noise is the existence of the sentences from another language in the corpus.

In the second experiment with the Xerox corpus, we study the effect of noise in training the word alignment model (WA). We make again about 20% of the Xerox corpus noisy, then we reorder the sentence pairs in the corpus according to the translation likelihood scores in ascending order. It means that the noisy sentence pairs supposed to be at the end of the corpus, i.e. from 80% to 100%. Table 6 shows the experimental results for this noisy corpus.

The first column shows the percentage of the sentence pairs extracted from the first of the corpus. The second column depicts the translation results when only the clean part of the corpus is used for training the BP model (WA is trained with the whole corpus). The third column contains the translation results when the WA and BP models are trained both with the clean part of the corpus. In this experiment, the language model is always trained with the whole corpus, as there is no noise in the target sentences of the Xerox corpus. The contents of this table state that training the word alignment with the whole corpus (including noise) causes slightly better translation quality.

We continue the experiments with a real noisy corpus, the EU corpus (Table 2). We reorder the sentence pairs in the corpus based on the translation likelihood-based filtering score. A human expert estimates about 1.7% to 2.0% sentence misalignment rate in the corpus, depending on the accuracy of judgment. We performed a set of ex-

Table 6. Translation Results on a 20% Artificial Noisy Xerox Corpus

Fraction of Sentence Pairs[%]	filtering on BP BLEU[%]	filtering on WA+BP BLEU[%]
10.0	38.7	30.0
40.0	57.3	47.9
70.0	59.8	59.6
75.0	59.7	59.6
77.5	59.9	59.9
80.0	60.0	59.7
82.5	60.4	60.2
85.0	60.1	59.6
90.0	59.8	60.1
95.0	60.2	60.1
100.0	60.4	60.4

periments to study the effect of noise on the translation results and also its effect on different models, as shown in Table 7.

Table 7. Translation Results on the EU corpus

Fraction of Sentence Pairs[%]	filtering on BP BLEU[%]	filtering on WA+BP BLEU[%]	filtering on WA+BP+LM BLEU[%]
92.5	46.9	46.6	46.6
95.0	47.1	46.8	46.7
97.5	47.2	46.8	46.8
98.3	47.0	46.9	46.9
100.0	46.8	46.8	46.8

This table also shows the best translation results are achieved when the word alignment and language models are trained on the whole corpus. As it can be expected the filtering on language modeling training has no effect, as the level of noise in the target (monolingual) corpus is not considerable. The difference in BLEU score between the clean corpus (97.5% of the corpus) and the whole corpus is about 0.4%. However, a significance analysis [23] on the sentence level between these two systems show that the improvement is statistically significant.

More Experiments

We continue our experiments by including new scores to the Equation 4. We consider the following scores: normalized source language model ($\frac{1}{J}\log p(f_1^J)$), normalized target language model ($\frac{1}{I}\log p(e_1^I)$), and sentence length-difference penalty model

($\log(1/(1.0 + |I - J|))$). The language models seem to be useful for filtering those sentences which are from another language than the language of the corpus. The sentence-length difference penalty model explicitly penalizes the dissimilarity between source and target lengths. The translation experiments did not show any translation quality improvement when we made use of these extended models over the word alignment models. It seems that the word alignment models are robust enough against the noise of the EU corpus. We have also studied the idea of corpus weighting instead of corpus filtering by using translation likelihood filtering. The translation result when we utilized the weighted corpus for training the system was 46.9% BLEU. It means using the weighted corpus had no superiority over using the simple corpus.

6 Conclusions

In this paper, we presented an efficient approach for corpus filtering. The experiments showed that translation likelihood-based filtering is a robust method for removing noise from the bilingual corpora. It improves the sentence alignment quality of the corpus and at the same time keeps the corpus size as large as possible. It has also been shown that the translation likelihood-based filtering enhances the training corpus for the translation task. The translation quality of the filtered training corpus has statistically significant improvement over the noisy corpus. One surprising result of the experiments is that even a large percentage of incorrect sentence pairs does not seem to deteriorate the performance of a statistical machine translation system. It means that the statistical machine translation models are robust enough against the corpus noise.

Acknowledgments

This work has been funded by the European Union under the RTD project TransType2 (IST 2001 32091) and the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org).

References

1. Nie, J., Cai, J.: Filtering noisy parallel corpora of web pages. In: IEEE Symposium on NLP and Knowledge Engineering, Tucson (2001) 453–458
2. Imamura, K., Sumita, E.: Automatic construction of machine translation knowledge using translation literalness. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary (2003) 155–162
3. Imamura, K., Sumita, E.: Bilingual corpus cleaning focusing on translation literality. In: 7th International Conference on Spoken Language Processing (ICSLP-2002), Denver, Colorado (2002) 1713–1716
4. Vogel, S.: Using noisy bilingual data for statistical machine translation. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary (2003) 175–178

5. Munteanu, D.S., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In Susan Dumais, D.M., Roukos, S., eds.: HLT-NAACL 2004: Main Proceedings, Boston, Massachusetts, USA, Association for Computational Linguistics (2004) 265–272
6. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational Linguistics **16** (1990) 79–85
7. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA (2002) 295–302
8. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan (2003) 160–167
9. Vidal, E., et al.: Final report of esprit research project 30268 (EuTrans): Example-based language translation systems. Technical report (2000)
10. Resnik, P.: Mining the web for bilingual text. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), University of Maryland, College Park, MD (1999) 527–534
11. Chen, J., Nie, J.Y.: Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In: Proceedings of the sixth conference on Applied natural language processing, Seattle, Washington, Morgan Kaufmann Publishers Inc. (2000) 21–28
12. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics **19** (1993) 75–102
13. Zhao, B., et al.: Efficient optimization for bilingual sentence alignment based on linear regression. In: HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Canada (2003) 81–87
14. Melamed, I.D.: Pattern recognition for mapping bitext correspondence. In Vronis, J., ed.: Parallel Text Processing: Alignment and Use of Translation Corpora. Kluwer Academic Publishers (2000) 25–47
15. Melamed, I.D.: A geometric approach to mapping bitext correspondence. In Brill, E., Church, K., eds.: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Somerset, New Jersey, Association for Computational Linguistics (1996) 1–12
16. Simard, M., Foster, G., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In: Fourth Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI-92), Montreal, Canada (1992) 67–81
17. LDC: Champollion tool kit (2004) `http://champollion.sourceforge.net/` .
18. Caseli, H.M., Nunes, M.G.V.: Evaluation of sentence alignment methods on portuguese-english parallel texts. Scientia **14** (2003) 1–14
19. Och, F.J.: YASMET: Toolkit for conditional maximum entropy models (2001) http://www-i6.informatik.rwth-aachen.de/˜och/software/YASMET.html.
20. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: COLING '96: The 16th Int. Conf. on Computational Linguistics, Copenhagen, Denmark (1996) 836–841
21. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–311
22. Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center (2001)
23. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluationx. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada (2004) 409–412