Improving Statistical Word Alignment with a Rule-Based Machine Translation System

WU Hua, WANG Haifeng
Toshiba (China) Research & Development Center
5/F., Tower W2, Oriental Plaza, No.1, East Chang An Ave., Dong Cheng District
Beijing, China, 100738
{wuhua, wanghaifeng}@rdc.toshiba.com.cn

Abstract

The main problems of statistical word alignment lie in the facts that source words can only be aligned to one target word, and that the inappropriate target word is selected because of data sparseness problem. This paper proposes an approach to improve statistical word alignment with a rule-based translation system. This approach first uses IBM statistical translation model to perform alignment in both directions (source to target and target to source), and then uses the translation information in the rule-based machine translation system to improve the statistical word alignment. The improved alignments allow the word(s) in the source language to be aligned to one or more words in the target language. Experimental results show a significant improvement in precision and recall of word alignment.

1 Introduction

Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) (Brown et al. 1993). Besides being used in SMT, it is also used in translation lexicon building (Melamed 1996), transfer rule learning (Menezes and Richardson 2001), example-based machine translation (Somers 1999), etc. In previous alignment methods, some researches modeled the alignments as hidden parameters in a statistical translation model (Brown et al. 1993; Och and Ney 2000) or directly modeled them given the sentence pairs (Cherry and Lin 2003). Some researchers used similarity and association measures to build alignment links (Ahrenberg et al. 1998; Tufis and Barbu 2002). In addition, Wu (1997) used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.

Generally speaking, there are four cases in word alignment: word to word alignment, word to multi-word alignment, multi-word to word alignment, and multi-word to multi-word alignment. One of the most difficult tasks in word alignment is to find out the alignments that include multi-word units. For example, the statistical word alignment in IBM translation models (Brown et al. 1993) can only handle word to word and multi-word to word alignments.

Some studies have been made to tackle this problem. Och and Ney (2000) performed translation in both directions (source to target and target to source) to extend word alignments. Their results showed that this method improved precision without loss of recall in English to German alignments. However, if the same unit is aligned to two different target units, this method is unlikely to make a selection. Some researchers used preprocessing steps to identity multi-word units for word alignment (Ahrenberg et al. 1998; Tiedemann 1999; Melamed 2000). The methods obtained multi-word candidates based on continuous N-gram statistics. The main limitation of these methods is that they cannot handle separated phrases and multi-word units in low frequencies.

In order to handle all of the four cases in word alignment, our approach uses both the alignment information in statistical translation models and translation information in a rule-based machine translation system. It includes three steps. (1) A statistical translation model is employed to perform word alignment in two directions[1] (English to Chinese, Chinese to English). (2) A rule-based English to Chinese translation system is employed to obtain Chinese translations for each English word or phrase in the source language. (3) The translation information in step (2) is used to improve the word alignment results in step (1).

A critical reader may pose the question "why

[1] We use English-Chinese word alignment as a case study.

not use a translation dictionary to improve statistical word alignment?" Compared with a translation dictionary, the advantages of a rule-based machine translation system lie in two aspects: (1) It can recognize the multi-word units, particularly separated phrases, in the source language. Thus, our method is able to handle the multi-word alignments with higher accuracy, which will be described in our experiments. (2) It can perform word sense disambiguation and select appropriate translations while a translation dictionary can only list all translations for each word or phrase. Experimental results show that our approach improves word alignments in both precision and recall as compared with the state-of-the-art technologies.

2 Statistical Word Alignment

Statistical translation models (Brown, et al. 1993) only allow word to word and multi-word to word alignments. Thus, some multi-word units cannot be correctly aligned. In order to tackle this problem, we perform translation in two directions (English to Chinese and Chinese to English) as described in Och and Ney (2000). The GIZA++ toolkit is used to perform statistical alignment. Thus, for each sentence pair, we can get two alignment results. We use S_1 and S_2 to represent the alignment sets with English as the source language and Chinese as the target language or vice versa. For alignment links in both sets, we use *i* for English words and *j* for Chinese words.

$$S_1 = \{(A_j, j) \mid A_j = \{a_j\},\ a_j \geq 0\}$$

$$S_2 = \{(i, A_i) \mid A_i = \{a_i\},\ a_i \geq 0\}$$

Where, $a_x (x = i, j)$ represents the index position of the source word aligned to the target word in position *x*. For example, if a Chinese word in position *j* is connected to an English word in position *i*, then $a_j = i$. If a Chinese word in position j is connected to English words in positions i_1 and i_2, then $A_j = \{i_1, i_2\}$.[2] We call an element in the alignment set an *alignment link*. If the link includes a word that has no translation, we call it a *null link*. If $k (k > 1)$ words have null links, we treat them as *k* different null links, not just one link.

[2] In the following of this paper, we will use the position number of a word to refer to the word.

Based on S_1 and S_2, we obtain their intersection set, union set and subtraction set.

Intersection: $S = S_1 \cap S_2$

Union: $P = S_1 \cup S_2$

Subtraction: $\mathrm{F} = \mathrm{P} - S$

Thus, the subtraction set contains two different alignment links for each English word.

3 Rule-Based Translation System

We use the translation information in a rule-based English-Chinese translation system[3] to improve the statistical word alignment result. This translation system includes three modules: source language parser, source to target language transfer module, and target language generator.

From the transfer phase, we get Chinese translation candidates for each English word. This information can be considered as another word alignment result, which is denoted as $S_3 = \{(k, C_k)\}$. C_k is the set including the translation candidates for the k-th English word or phrase. The difference between S_3 and the common alignment set is that each English word or phrase in S_3 has one or more translation candidates. A translation example for the English sentence "He is used to pipe smoking." is shown in Table 1.

English Words	Chinese Translations
He	他
is used to	习惯
pipe	烟斗，烟筒
smoking	吸，吸烟

Table 1. Translation Example

From Table 1, it can be seen that (1) the translation system can recognize English phrases (e.g. is used to); (2) the system can provide one or more translations for each source word or phrase; (3) the translation system can perform word selection or word sense disambiguation. For example, the word "pipe" has several meanings such as "tube", "tube used for smoking" and "wind instrument". The system selects "tube used for smoking" and translates it into Chinese words "烟斗" and "烟筒". The recognized translation

[3] This system is developed based on the Toshiba English-Japanese translation system (Amano et al. 1989). It achieves above-average performance as compared with the English-Chinese translation systems available in the market.

candidates will be used to improve statistical word alignment in the next section.

4 Word Alignment Improvement

As described in Section 2, we have two alignment sets for each sentence pair, from which we obtain the intersection set S and the subtraction set F. We will improve the word alignments in S and F with the translation candidates produced by the rule-based machine translation system. In the following sections, we will first describe how to calculate monolingual word similarity used in our algorithm. Then we will describe the algorithm used to improve word alignment results.

4.1 Word Similarity Calculation

This section describes the method for monolingual word similarity calculation. This method calculates word similarity by using a bilingual dictionary, which is first introduced by Wu and Zhou (2003). The basic assumptions of this method are that the translations of a word can express its meanings and that two words are similar in meanings if they have mutual translations.

Given a Chinese word, we get its translations with a Chinese-English bilingual dictionary. The translations of a word are used to construct its feature vector. The similarity of two words is estimated through their feature vectors with the cosine measure as shown in (Wu and Zhou 2003). If there are a Chinese word or phrase w and a Chinese word set Z, the word similarity between them is calculated as shown in Equation (1).

$$sim(w, Z) = \underset{w' \in Z}{Max}(sim(w, w')) \quad (1)$$

4.2 Alignment Improvement Algorithm

As the word alignment links in the intersection set are more reliable than those in the subtraction set, we adopt two different strategies for the alignments in the intersection set S and the subtraction set F. For alignments in S, we will modify them when they are inconsistent with the translation information in S_3. For alignments in F, we classify them into two cases and make selection between two different alignment links or modify them into a new link.

In the intersection set S, there are only word to word alignment links, which include no multi-word units. The main alignment error type in this set is that some words should be combined into one phrase and aligned to the same word(s) in the target sentence. For example, for the sentence pair in Figure 1, "used" is aligned to the Chinese word "习惯", and "is" and "to" have null links in S. But in the translation set S_3, "is used to" is a phrase. Thus, we combine the three alignment links into a new link. The words "is", "used" and " to" are all aligned to the Chinese word "习惯", denoted as (is used to, 习惯). Figure 2 describes the algorithm employed to improve the word alignment in the intersection set S.

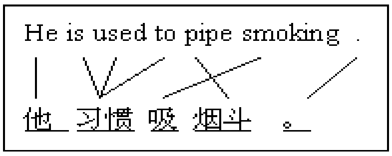

Figure 1. Multi-Word Alignment Example

Input: Intersection set S, Translation set S_3,
Final word alignment set WA

For each alignment link (i, j) in S, do:

(1) If all of the following three conditions are satisfied, add the new alignment link $(ph_k, w) \notin WA$ to WA.

a) There is an element $(ph_k, C_k) \in S_3$, and the English word i is a constituent of the phrase ph_k.

b) The other words in the phrase ph_k also have alignment links in S.

c) For each word s in ph_k, we get $T = \{t \mid (s,t) \in S\}$ and combine[4] all words in T into a phrase w, and the similarity $sim(w, C_k) > \delta_1$.

(2) Otherwise, add (i, j) to WA.

Output: Word alignment set WA

Figure 2. Algorithm for the Intersection Set

In the subtraction set, there are two different links for each English word. Thus, we need to select one link or to modify the links according to the translation information in S_3.

For each English word i in the subtraction set, there are two cases:

[4] We define an operation "*combine*" on a set consisting of position numbers of words. We first sort the position numbers in the set ascendly and then regard them as a phrase. For example, there is a set {{2,3}, 1, 4}, the result after applying the combine operation is (1, 2, 3, 4).

Case 1: In S_1, there is a word to word alignment link $(i, j) \in S_1$. In S_2, there is a word to word or word to multi-word alignment link $(i, A_i) \in S_2$ [5].

Case 2: In S_1, there is a multi-word to word alignment link $(A_j, j) \in S_1 \,\&\, i \in A_j$. In S_2, there is a word to word or word to multi-word alignment link $(i, A_i) \in S_2$.

For Case 1, we first examine the translation set S_3. If there is an element $(i, C_i) \in S_3$, we calculate the Chinese word similarity between *j* in $(i, j) \in S_1$ and C_i with Equation (1) shown in Section 4.1. We also combine the words in A_i ($(i, A_i) \in S_2$) into a phrase and get the word similarity between this new phrase and C_i. The alignment link with a higher similarity score is selected and added to WA.

Input: Alignment sets S_1 and S_2 Translation unit $(ph_k, C_k) \in S_3$
(1) For each sub-sequence[6] s of ph_k, get the sets $T_1 = \{t_1 \mid (s, t_1) \in S_1\}$ and $T_2 = \{t_2 \mid (s, t_2) \in S_2\}$ (2) Combine words in T_1 and T_2 into phrases w_1 and w_2 respectively. (3) Obtain the word similarities $ws_1 = sim(w_1, C_k)$ and $ws_2 = sim(w_2, C_k)$. (4) Add a new alignment link to WA according to the following steps. a) If $ws_1 > ws_2$ and $ws_1 > \delta_1$, add (ph_k, w_1) to WA; b) If $ws_2 > ws_1$ and $ws_2 > \delta_1$, add (ph_k, w_2) to WA; c) If $ws_1 = ws_2 > \delta_1$, add (ph_k, w_1) or (ph_k, w_2) to WA randomly.
Output: Updated alignment set WA

Figure 3. Multi-Word to Multi-Word Alignment Algorithm

If, in S_3, there is an element (ph_k, C_k) and *i* is a constituent of ph_k, the English word *i* of the alignment links in both S_1 and S_2 should be combined with other words to form phrases. In this case, we modify the alignment links into a multi-word to multi-word alignment link. The algorithm is described in Figure 3.

For example, given a sentence pair in Figure 4, in S_1, the word "whipped" is aligned to "突然" and "out" is aligned to "抽出". In S_2, the word "whipped" is aligned to both "突然" and "抽出" and "out" has a null link. In S_3, "whipped out" is a phrase and translated into "迅速抽出". And the word similarity between "突然抽出" and "迅速抽出" is larger than the threshold δ_1. Thus, we combine the aligned target words in the Chinese sentence into "突然抽出". The final alignment link should be (whipped out, 突然 抽出).

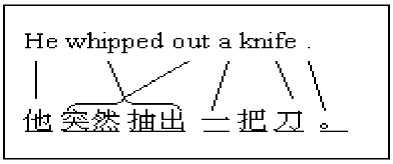

Figure 4. Multi-Word to Multi-Word Alignment Example

For Case 2, we first examine S_3 to see whether there is an element $(i, C_i) \in S_3$. If true, we combine the words in A_i ($(i, A_i) \in S_2$) into a word or phrase and calculate the similarity between this new word or phrase and C_i in the same way as in Case 1. If the similarity is higher than a threshold δ_1, we add the alignment link (i, A_i) into WA.

If there is an element $(ph_k, C_k) \in S_3$ and *i* is a constituent of ph_k, we combine the English words in A_j ($(A_j, j) \in S_1$) into a phrase. If it is the same as the phrase ph_k and $sim(j, C_k) > \delta_1$, we add (A_j, j) into WA. Otherwise, we use the multi-word to multi-word alignment algorithm in Figure 3 to modify the links.

After applying the above two strategies, there are still some words not aligned. For each sentence pair, we use *E* and *C* to denote the sets of the source words and the target words that are not aligned, respectively. For each source word in *E*, we construct a link with each target word in *C*. We use $L = \{(i, j) \mid i \in E, j \in C\}$ to denote the alignment candidates. For each candidate in *L*, we look it up in the translation set S_3. If there is an element $(i, C_i) \in S_3$ and $sim(j, C_i) > \delta_2$, we

[5] (i, A_i) represents both the word to word and word to multi-word alignment links.

[6] If a phrase consists of three words $w_1 w_2 w_3$, the sub-sequences of this phrase are $w_1, w_2, w_3, w_1 w_2, w_2 w_3$.

add the link into the set WA.

5 Experiments

5.1 Training and Testing Set

We did experiments on a sentence aligned English-Chinese bilingual corpus in general domains. There are about 320,000 bilingual sentence pairs in the corpus, from which, we randomly select 1,000 sentence pairs as testing data. The remainder is used as training data.

The Chinese sentences in both the training set and the testing set are automatically segmented into words. The segmentation errors in the testing set are post-corrected. The testing set is manually annotated. It has totally 8,651 alignment links including 2,149 null links. Among them, 866 alignment links include multi-word units, which accounts for about 10% of the total links.

5.2 Experimental Results

There are several different evaluation methods for word alignment (Ahrenberg et al. 2000). In our evaluation, we use evaluation metrics similar to those in Och and Ney (2000). However, we do not classify alignment links into sure links and possible links. We consider each alignment as a sure link.

If we use S_G to indicate the alignments identified by the proposed methods and S_C to denote the reference alignments, the precision, recall and f-measure are calculated as described in Equation (2), (3) and (4). According to the definition of the alignment error rate (AER) in Och and Ney (2000), AER can be calculated with Equation (5).

$$precision = \frac{|S_G \cap S_C|}{|S_G|} \quad (2)$$

$$recall = \frac{|S_G \cap S_C|}{|S_C|} \quad (3)$$

$$fmeasure = \frac{2*|S_G \cap S_C|}{|S_G| + |S_C|} \quad (4)$$

$$AER = 1 - \frac{2*|S_G \cap S_C|}{|S_G| + |S_C|} = 1 - fmeasure \quad (5)$$

In this paper, we give two different alignment results in Table 2 and Table 3. Table 2 presents alignment results that include null links. Table 3 presents alignment results that exclude null links. The precision and recall in the tables are obtained to ensure the smallest AER for each method.

	Precision	Recall	AER
Ours	0.8531	0.7057	0.2276
Dic	0.8265	0.6873	0.2495
IBM E-C	0.7121	0.6812	0.3064
IBM C-E	0.6759	0.7209	0.3023
IBM Inter	0.8756	0.5516	0.3233
IBM Refined	0.7046	0.6532	0.3235

Table 2. Alignment Results Including Null Links

	Precision	Recall	AER
Ours	0.8827	0.7583	0.1842
Dic	0.8558	0.7317	0.2111
IBM E-C	0.7304	0.7136	0.2781
IBM C-E	0.6998	0.6725	0.3141
IBM Inter	0.9392	0.5513	0.3052
IBM refined	0.8152	0.6926	0.2505

Table 3. Alignment Results Excluding Null Links

In the above tables, the row “Ours” presents the result of our approach. The results are obtained by setting the word similarity thresholds to $\delta_1 = 0.1$ and $\delta_2 = 0.5$. The Chinese-English dictionary used to calculate the word similarity has 66,696 entries. Each entry has two English translations on average. The row “Dic” shows the result of the approach that uses a bilingual dictionary instead of the rule-based machine translation system to improve statistical word alignment. The dictionary used in this method is the same translation dictionary used in the rule-based machine translation system. It includes 57,684 English words and each English word has about two Chinese translations on average. The rows “IBM E-C” and “IBM C-E” show the results obtained by IBM Model-4 when treating English as the source and Chinese as the target or vice versa. The row “IBM Inter” shows results obtained by taking the intersection of the alignments produced by “IBM E-C” and “IBM C-E”. The row “IBM Refined” shows the results by refining the results of “IBM Inter” as described in Och and Ney (2000).

Generally, the results excluding null links are better than those including null links. This indicates that it is difficult to judge whether a word has counterparts in another language. It is because the translations of some source words can be omitted. Both the rule-based translation system and the bilingual dictionary provide no such information.

It can be also seen that our approach performs the best among others in both cases. Our approach achieves a relative error rate reduction of 26% and 25% when compared with "IBM E-C" and "IBM C-E" respectively[7]. Although the precision of our method is lower than that of the "IBM Inter" method, it achieves much higher recall, resulting in a 30% relative error rate reduction. Compared with the "IBM refined" method, our method also achieves a relative error rate reduction of 30%. In addition, our method is better than the "Dic" method, achieving a relative error rate reduction of 8.8%.

In order to provide the detailed word alignment information, we classify word alignment results in Table 3 into two classes. The first class includes the alignment links that have no multi-word units. The second class includes at least one multi-word unit in each alignment link. The detailed information is shown in Table 4 and Table 5. In Table 5, we do not include the method "Inter" because it has no multi-word alignment links.

	Precision	Recall	AER
Ours	0.9213	0.8269	0.1284
Dic	0.8898	0.8215	0.1457
IBM E-C	0.8202	0.7972	0.1916
IBM C-E	0.8200	0.7406	0.2217
IBM Inter	0.9392	0.6360	0.2416
IBM Refined	0.8920	0.7196	0.2034

Table 4. Single Word Alignment Results

	Precision	Recall	AER
Ours	0.5123	0.3118	0.6124
Dic	0.3585	0.1478	0.7907
IBM E-C	0.1682	0.1697	0.8311
IBM C-E	0.1718	0.2298	0.8034
IBM Refined	0.2105	0.2910	0.7557

Table 5. Multi-Word Alignment Results

All of the methods perform better on single word alignment than on multi-word alignment. In Table 4, the precision of our method is close to the "IBM Inter" approach, and the recall of our method is much higher, achieving a 47% relative error rate reduction. Our method also achieves a 37% relative error rate reduction over the "IBM Refined" method. Compared with the "Dic" method, our approach achieves much higher precision without loss of recall, resulting in a 12% relative error rate reduction.

Our method also achieves much better results on multi-word alignment than other methods. However, our method only obtains one third of the correct alignment links. It indicates that it is the hardest to align the multi-word units.

6 Discussion

Readers may pose the question "why the rule-based translation system performs better on word alignment than the translation dictionary?" For single word alignment, the rule-based translation system can perform word sense disambiguation, and select the appropriate Chinese words as translation. On the contrary, the dictionary can only list all translations. Thus, the alignment precision of our method is higher than that of the dictionary method. Figure 5 shows alignment precision and recall values under different similarity values for single word alignment including null links. From the figure, it can be seen that our method consistently achieves higher precisions as compared with the dictionary method. The t-score value (t=10.37, p=0.05) shows the improvement is statistically significant.

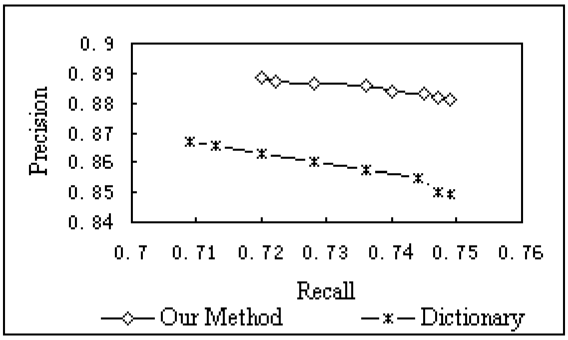

Figure 5. Recall-Precision Curves

For multi-word alignment links, the translation system also outperforms the translation dictionary. The result is shown in Table 5 in Section 5.2. This is because (1) the translation system can automatically recognize English phrases with higher accuracy than the translation dictionary; (2) The translation system can detect separated phrases while the dictionary cannot. For example, for the sentence pairs in Figure 6, the solid link lines describe the alignment result of the rule-base translation system while dashed lines indicate the alignment result of the translation dictionary. In example (1), the phrase "be going to"

[7] The error rate reductions in this paragraph are obtained from Table 2. The error rate reductions in Table 3 are omitted.

indicates the tense not the phrase "go to" as the dictionary shows. In example (2), our method detects the separated phrase "turn … on" while the dictionary does not. Thus, the dictionary method produces the wrong alignment link.

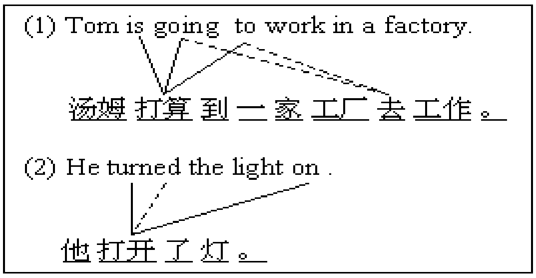

Figure 6. Alignment Comparison Examples

7 Conclusion and Future Work

This paper proposes an approach to improve statistical word alignment results by using a rule-based translation system. Our contribution is that, given a rule-based translation system that provides appropriate translation candidates for each source word or phrase, we select appropriate alignment links among statistical word alignment results or modify them into new links. Especially, with such a translation system, we can identify both the continuous and separated phrases in the source language and improve the multi-word alignment results. Experimental results indicate that our approach can achieve a precision of 85% and a recall of 71% for word alignment including null links in general domains. This result significantly outperforms those of the methods that use a bilingual dictionary to improve word alignment, and that only use statistical translation models.

Our future work mainly includes three tasks. First, we will further improve multi-word alignment results by using other technologies in natural language processing. For example, we can use named entity recognition and transliteration technologies to improve person name alignment. Second, we will extract translation rules from the improved word alignment results and apply them back to our rule-based machine translation system. Third, we will further analyze the effect of the translation system on the alignment results.

References

Lars Ahrenberg, Magnus Merkel, and Mikael Andersson 1998. *A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts.* In Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics, pp. 29-35.

Lars Ahrenberg, Magnus Merkel, Anna Sagvall Hein and Jorg Tiedemann 2000. *Evaluation of word alignment systems.* In Proc. of the Second Int. Conf. on Linguistic Resources and Evaluation, pp. 1255-1261.

ShinYa Amano, Hideki Hirakawa, Hiroyasu Nogami, and Akira Kumano 1989. *Toshiba Machine Translation System*. Future Computing Systems, 2(3):227-246.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, 19(2):263-311.

Colin Cherry and Dekang Lin 2003. *A Probability Model to Improve Word Alignment.* In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 88-95.

I. Dan Melamed 1996. *Automatic Construction of Clean Broad-Coverage Translation Lexicons.* In Proc. of the 2nd Conf. of the Association for Machine Translation in the Americas, pp. 125-134.

I. Dan Melamed 2000. *Word-to-Word Models of Translational Equivalence among Words.* Computational Linguistics, 26(2): 221-249.

Arul Menezes and Stephan D. Richardson 2001. *A Best-first Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora.* In Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation, pp. 39-46.

Franz Josef Och and Hermann Ney 2000. *Improved Statistical Alignment Models.* In Proc.of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447.

Harold Somers 1999. *Review Article: Example-Based Machine Translation.* Machine Translation 14:113-157.

Jorg Tiedemann 1999. *Word Alignment – Step by Step.* In Proc. of the 12th Nordic Conf. on Computational Linguistics, pp. 216-227.

Dan Tufis and Ana Maria Barbu. 2002. *Lexical Token Alignment: Experiments, Results and Application.* In Proc. of the Third Int. Conf. on Language Resources and Evaluation, pp. 458-465.

Dekai Wu 1997. *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora.* Computational Linguistics, 23(3):377-403.

Hua Wu and Ming Zhou 2003. *Optimizing Synonym Extraction Using Monolingual and Bilingual Resources.* In Proc. of the 2nd Int. Workshop on Paraphrasing, pp. 72-79.