# Bootstrapping Word Alignment via Word Packing

**Yanjun Ma, Nicolas Stroppa, Andy Way**
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland
{yma,nstroppa,away}@computing.dcu.ie

## Abstract

We introduce a simple method to pack words for statistical word alignment. Our goal is to simplify the task of automatic word alignment by packing several consecutive words together when we believe they correspond to a single word in the opposite language. This is done using the word aligner itself, i.e. by bootstrapping on its output. We evaluate the performance of our approach on a Chinese-to-English machine translation task, and report a 12.2% relative increase in BLEU score over a state-of-the art phrase-based SMT system.

## 1 Introduction

Automatic word alignment can be defined as the problem of determining a translational correspondence at word level given a parallel corpus of aligned sentences. Most current statistical models (Brown et al., 1993; Vogel et al., 1996; Deng and Byrne, 2005) treat the aligned sentences in the corpus as sequences of tokens that are meant to be words; the goal of the alignment process is to find links between source and target words. Before applying such aligners, we thus need to segment the sentences into words – a task which can be quite hard for languages such as Chinese for which word boundaries are not orthographically marked. More importantly, however, this segmentation is often performed in a *monolingual* context, which makes the word alignment task more difficult since different languages may realize the same concept using varying numbers of words (see e.g. (Wu, 1997)). Moreover, a segmentation considered to be "good" from a monolingual point of view may be unadapted for training alignment models.

Although some statistical alignment models allow for 1-to-$n$ word alignments for those reasons, they rarely question the monolingual tokenization and the basic unit of the alignment process remains the word. In this paper, we focus on 1-to-$n$ alignments with the goal of simplifying the task of automatic word aligners by *packing* several consecutive words together when we believe they correspond to a single word in the opposite language; by identifying enough such cases, we reduce the number of 1-to-$n$ alignments, thus making the task of word alignment both easier and more natural.

Our approach consists of using the output from an existing statistical word aligner to obtain a set of candidates for word packing. We evaluate the reliability of these candidates, using simple metrics based on co-occurence frequencies, similar to those used in associative approaches to word alignment (Kitamura and Matsumoto, 1996; Melamed, 2000; Tiedemann, 2003). We then modify the segmentation of the sentences in the parallel corpus according to this packing of words; these modified sentences are then given back to the word aligner, which produces new alignments. We evaluate the validity of our approach by measuring the influence of the alignment process on a Chinese-to-English Machine Translation (MT) task.

The remainder of this paper is organized as follows. In Section 2, we study the case of 1-to-$n$ word alignment. Section 3 introduces an automatic method to pack together groups of consecutive



*Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 304–311,
Prague, Czech Republic, June 2007. 

| | | 1: 0 | 1: 1 | 1: 2 | 1: 3 | 1: $n$ ($n > 3$) |
|---|---|---|---|---|---|---|
| IWSLT | Chinese–English | 21.64 | 63.76 | 9.49 | 3.36 | 1.75 |
| IWSLT | English–Chinese | 29.77 | 57.47 | 10.03 | 1.65 | 1.08 |
| IWSLT | Italian–English | 13.71 | 72.87 | 9.77 | 3.23 | 0.42 |
| IWSLT | English–Italian | 20.45 | 71.08 | 7.02 | 0.9 | 0.55 |
| Europarl | Dutch–English | 24.71 | 67.04 | 5.35 | 1.4 | 1.5 |
| Europarl | English–Dutch | 23.76 | 69.07 | 4.85 | 1.2 | 1.12 |

Table 1: Distribution of alignment types for different language pairs (%)

words based on the output from a word aligner. In Section 4, the experimental setting is described. In Section 5, we evaluate the influence of our method on the alignment process on a Chinese to English MT task, and experimental results are presented. Section 6 concludes the paper and gives avenues for future work.

## 2 The Case of 1-to-$n$ Alignment

The same concept can be expressed in different languages using varying numbers of words; for example, a single Chinese word may surface as a compound or a collocation in English. This is frequent for languages as different as Chinese and English. To quickly (and approximately) evaluate this phenomenon, we trained the statistical IBM word-alignment model 4 (Brown et al., 1993),[1] using the GIZA++ software (Och and Ney, 2003) for the following language pairs: Chinese–English, Italian–English, and Dutch–English, using the IWSLT-2006 corpus (Takezawa et al., 2002; Paul, 2006) for the first two language pairs, and the Europarl corpus (Koehn, 2005) for the last one. These asymmetric models produce 1-to-$n$ alignments, with $n \geq 0$, in both directions. Here, it is important to mention that the segmentation of sentences is performed totally independently of the bilingual alignment process, i.e. it is done in a *monolingual* context. For European languages, we apply the maximum-entropy based tokenizer of OpenNLP[2]; the Chinese sentences were human segmented (Paul, 2006).

In Table 1, we report the frequencies of the different types of alignments for the various languages and directions. As expected, the number of 1: $n$ alignments with $n \neq 1$ is high for Chinese–English ($\simeq 40\%$), and significantly higher than for the European languages. The case of 1-to-$n$ alignments is, therefore, obviously an important issue when dealing with Chinese–English word alignment.[3]

### 2.1 The Treatment of 1-to-$n$ Alignments

Fertility-based models such as IBM models 3, 4, and 5 allow for alignments between one word and several words (1-to-$n$ or 1: $n$ alignments in what follows), in particular for the reasons specified above. They can be seen as extensions of the simpler IBM models 1 and 2 (Brown et al., 1993). Similarly, Deng and Byrne (2005) propose an HMM framework capable of dealing with 1-to-$n$ alignment, which is an extension of the original model of (Vogel et al., 1996).

However, these models rarely question the monolingual tokenization, i.e. the basic unit of the alignment process is the word.[4] One alternative to extending the expressivity of one model (and usually its complexity) is to focus on the *input representation*; in particular, we argue that the alignment process can benefit from a simplification of the input, which consists of trying to reduce the number of 1-to-$n$ alignments to consider. Note that the need to consider segmentation and alignment at the same time is also mentioned in (Tiedemann, 2003), and related issues are reported in (Wu, 1997).

### 2.2 Notation

While in this paper, we focus on Chinese–English, the method proposed is applicable to any language

[1]More specifically, we performed 5 iterations of Model 1, 5 iterations of HMM, 5 iterations of Model 3, and 5 iterations of Model 4.

[2]`http://opennlp.sourceforge.net/`.

[3]Note that a 1: 0 alignment may denote a failure to capture a 1: $n$ alignment with $n > 1$.

[4]Interestingly, this is actually even the case for approaches that directly model alignments between phrases (Marcu and Wong, 2002; Birch et al., 2006).

pair – even for closely related languages, we expect improvements to be seen. The notation however assume Chinese–English MT. Given a Chinese sentence $c_1^J$ consisting of $J$ words $\{c_1, \ldots, c_J\}$ and an English sentence $e_1^I$ consisting of $I$ words $\{e_1, \ldots, e_I\}$, $A_{C \to E}$ (resp. $A_{E \to C}$) will denote a Chinese-to-English (resp. an English-to-Chinese) word alignment between $c_1^J$ and $e_1^I$. Since we are primarily interested in 1-to-$n$ alignments, $A_{C \to E}$ can be represented as a set of pairs $a_j = \langle c_j, E_j \rangle$ denoting a link between one single Chinese word $c_j$ and a few English words $E_j$ (and similarly for $A_{E \to C}$). The set $E_j$ is empty if the word $c_j$ is not aligned to any word in $e_1^I$.

## 3 Automatic Word Repacking

Our approach consists of packing consecutive words together when we believe they correspond to a single word in the other language. This bilingually motivated packing of words changes the basic unit of the alignment process, and simplifies the task of automatic word alignment. We thus minimize the number of 1-to-$n$ alignments in order to obtain more comparable segmentations in the two languages. In this section, we present an automatic method that builds upon the output from an existing automatic word aligner. More specifically, we (i) use a word aligner to obtain 1-to-$n$ alignments, (ii) extract candidates for word packing, (iii) estimate the reliability of these candidates, (iv) replace the groups of words to pack by a single token in the parallel corpus, and (v) re-iterate the alignment process using the updated corpus. The first three steps are performed in both directions, and produce two *bilingual dictionaries* (source-target and target-source) of groups of words to pack.

### 3.1 Candidate Extraction

In the following, we assume the availability of an automatic word aligner that can output alignments $A_{C \to E}$ and $A_{E \to C}$ for any sentence pair $(c_1^J, e_1^I)$ in a parallel corpus. We also assume that $A_{C \to E}$ and $A_{E \to C}$ contain 1: $n$ alignments. Our method for repacking words is very simple: whenever a single word is aligned with several consecutive words, they are considered candidates for repacking. Formally, given an alignment $A_{C \to E}$ between $c_1^J$ and $e_1^I$, if $a_j = \langle c_j, E_j \rangle \in A_{C \to E}$, with $E_j = \{e_{j_1}, \ldots, e_{j_m}\}$ and $\forall k \in [\![1, m-1]\!], j_{k+1} - j_k = 1$, then the alignment $a_j$ between $c_j$ and the sequence of words $E_j$ is considered a candidate for word repacking. The same goes for $A_{E \to C}$. Some examples of such 1-to-$n$ alignments between Chinese and English (in both directions) we can derive automatically are displayed in Figure 1.

| | |
|---|---|
| 白葡萄酒: white wine | closest: 最 近 |
| 百货公司: department store | fifteen: 十 五 |
| 抱歉: excuse me | fine: 很 好 |
| 报警: call the police | flight: 次 航班 |
| 杯: cup of | get: 拿 到 |
| 必须: have to | here: 在 这里 |

Figure 1: Example of 1-to-$n$ word alignments between Chinese and English

### 3.2 Candidate Reliability Estimation

Of course, the process described above is error-prone and if we want to change the input to give to the word aligner, we need to make sure that we are not making harmful modifications.[5] We thus additionally evaluate the reliability of the candidates we extract and filter them before inclusion in our bilingual dictionary. To perform this filtering, we use two simple statistical measures. In the following, $a_j = \langle c_j, E_j \rangle$ denotes a candidate.

The first measure we consider is co-occurrence frequency ($COOC(c_j, E_j)$), i.e. the number of times $c_j$ and $E_j$ co-occur in the bilingual corpus. This very simple measure is frequently used in associative approaches (Melamed, 1997; Tiedemann, 2003). The second measure is the alignment confidence, defined as

$$AC(a_j) = \frac{C(a_j)}{COOC(c_j, E_j)},$$

where $C(a_j)$ denotes the number of alignments proposed by the word aligner that are identical to $a_j$. In other words, $AC(a_j)$ measures how often the

[5] Consequently, if we compare our approach to the problem of collocation identification, we may say that we are more interested in precision than recall (Smadja et al., 1996). However, note that our goal is not recognizing specific sequences of words such as compounds or collocations; it is making (bilingually motivated) changes that simplify the alignment process.

aligner aligns $c_j$ and $E_j$ when they co-occur. We also impose that $|E_j| \leq k$, where $k$ is a fixed integer that may depend on the language pair (between 3 and 5 in practice). The rationale behind this is that it is very rare to get reliable alignment between one word and $k$ consecutive words when $k$ is high.

The candidates are included in our bilingual dictionary if and only if their measures are above some fixed thresholds $t_{cooc}$ and $t_{ac}$, which allow for the control of the size of the dictionary and the quality of its contents. Some other measures (including the Dice coefficient) could be considered; however, it has to be noted that we are more interested here in the filtering than in the discovery of alignment, since our method builds upon an existing aligner. Moreover, we will see that even these simple measures can lead to an improvement of the alignment process in a MT context (cf. Section 5).

### 3.3 Bootstrapped Word Repacking

Once the candidates are extracted, we repack the words in the bilingual dictionaries constructed using the method described above; this provides us with an updated training corpus, in which some word sequences have been replaced by a single token. This update is totally naive: if an entry $a_j = \langle c_j, E_j \rangle$ is present in the dictionary and matches one sentence pair $(c_1^J, e_1^I)$ (i.e. $c_j$ and $E_j$ are respectively contained in $c_1^J$ and $e_1^I$), then we replace the sequence of words $E_j$ with a single token which becomes a new lexical unit.[6] Note that this replacement occurs even if no alignment was found between $c_j$ and $E_j$ for the pair $(c_1^J, e_1^I)$. This is motivated by the fact that the filtering described above is quite conservative; we trust the entry $a_i$ to be correct. This update is performed in both directions. It is then possible to run the word aligner using the updated (simplified) parallel corpus, in order to get new alignments. By performing a deterministic word packing, we avoid the computation of the fertility parameters associated with fertility-based models.

Word packing can be applied several times: once we have grouped some words together, they become the new basic unit to consider, and we can re-run the same method to get additional groupings. However, we have not seen in practice much benefit from running it more than twice (few new candidates are extracted after two iterations).

It is also important to note that this process is bilingually motivated and strongly depends on the language pair. For example, *white wine*, *excuse me*, *call the police*, and *cup of* (cf. Figure 1) translate respectively as *vin blanc*, *excusez-moi*, *appellez la police*, and *tasse de* in French. Those groupings would not be found for a language pair such as French–English, which is consistent with the fact that they are less useful for French–English than for Chinese–English in a MT perspective.

### 3.4 Using Manually Developed Dictionaries

We wanted to compare this automatic approach to manually developed resources. For this purpose, we used a dictionary built by the MT group of Harbin Institute of Technology, as a preprocessing step to Chinese–English word alignment, and motivated by several years of Chinese–English MT practice. Some examples extracted from this resource are displayed in Figure 2.

有: there is
想要: want to
不必: need not
前面: in front of
一: as soon as
看: look at

Figure 2: Examples of entries from the manually developed dictionary

## 4 Experimental Setting

### 4.1 Evaluation

The intrinsic quality of word alignment can be assessed using the Alignment Error Rate (AER) metric (Och and Ney, 2003), that compares a system's alignment output to a set of gold-standard alignment. While this method gives a direct evaluation of the quality of word alignment, it is faced with several limitations. First, it is really difficult to build a reliable and objective gold-standard set, especially for languages as different as Chinese and English. Second, an increase in AER does not necessarily imply an improvement in translation quality (Liang et al., 2006) and vice-versa (Vilar et al., 2006). The

[6] In case of overlap between several groups of words to replace, we select the one with highest confidence (according to $t_{ac}$).

relationship between word alignments and their impact on MT is also investigated in (Ayan and Dorr, 2006; Lopez and Resnik, 2006; Fraser and Marcu, 2006). Consequently, we chose to extrinsically evaluate the performance of our approach via the translation task, i.e. we measure the influence of the alignment process on the final translation output. The quality of the translation output is evaluated using BLEU (Papineni et al., 2002).

### 4.2 Data

The experiments were carried out using the Chinese–English datasets provided within the IWSLT 2006 evaluation campaign (Paul, 2006), extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). This multilingual speech corpus contains sentences similar to those that are usually found in phrase-books for tourists going abroad. Training was performed using the default training set, to which we added the sets devset1, devset2, and devset3.[7] The English side of the test set was not available at the time we conducted our experiments, so we split the development set (devset 4) into two parts: one was kept for testing (200 aligned sentences) with the rest (289 aligned sentences) used for development purposes.

As a pre-processing step, the English sentences were tokenized using the maximum-entropy based tokenizer of the OpenNLP toolkit, and case information was removed. For Chinese, the data provided were tokenized according to the output format of ASR systems, and human-corrected (Paul, 2006). Since segmentations are human-corrected, we are sure that they are good from a monolingual point of view. Table 2 contains the various corpus statistics.

| | | Chinese | English |
|---|---|---|---|
| Train | Sentences | 41,465 | |
| | Running words | 361,780 | 375,938 |
| | Vocabulary size | 11,427 | 9,851 |
| Dev. | Sentences | 289 (7 refs.) | |
| | Running words | 3,350 | 26,223 |
| | Vocabulary size | 897 | 1,331 |
| Eval. | Sentences | 200 (7 refs.) | |
| | Running words | 1,864 | 14,437 |
| | Vocabulary size | 569 | 1,081 |

Table 2: Chinese–English corpus statistics

### 4.3 Baseline

We use a standard log-linear phrase-based statistical machine translation system as a baseline: GIZA++ implementation of IBM word alignment model 4 (Brown et al., 1993; Och and Ney, 2003),[8] the refinement and phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) using Phramer (Olteanu et al., 2006), a 3-gram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002) on the English side of the training data and Pharaoh (Koehn, 2004) with default settings to decode. The log-linear model is also based on standard features: conditional probabilities and lexical smoothing of phrases in both directions, and phrase penalty (Zens and Ney, 2004).

## 5 Experimental Results

### 5.1 Results

The initial word alignments are obtained using the baseline configuration described above. From these, we build two bilingual 1-to-$n$ dictionaries (one for each direction), and the training corpus is updated by repacking the words in the dictionaries, using the method presented in Section 2. As previously mentioned, this process can be repeated several times; at each step, we can also choose to exploit only one of the two available dictionaries, if so desired. We then extract aligned phrases using the same procedure as for the baseline system; the only difference is the basic unit we are considering. Once the phrases are extracted, we perform the estimation of the features of the log-linear model and unpack the grouped words to recover the initial words. Finally, minimum-error-rate training and decoding are performed.

The various parameters of the method ($k$, $t_{cooc}$, $t_{ac}$, cf. Section 2) have been optimized on the development set. We found out that it was enough to perform two iterations of repacking: the optimal set of values was found to be $k = 3$, $t_{ac} = 0.5$, $t_{cooc} = 20$ for the first iteration, and $t_{cooc} = 10$ for the second

[7]More specifically, we choose the first English reference from the 7 references and the Chinese sentence to construct new sentence pairs.

[8]Training is performed using the same number of iterations as in Section 2.

| | BLEU[%] |
|---|---|
| Baseline | 15.14 |
| n=1. with C-E dict. | 15.92 |
| n=1. with E-C dict. | 15.77 |
| n=1. with both | **16.59** |
| n=2. with C-E dict. | **16.99** |
| n=2. with E-C dict. | 16.59 |
| n=2. with both | 16.88 |

Table 3: Influence of word repacking on Chinese-to-English MT

iteration, for both directions.[9] In Table 3, we report the results obtained on the test set, where $n$ denotes the iteration. We first considered the inclusion of only the Chinese–English dictionary, then only the English–Chinese dictionary, and then both.

After the first step, we can already see an improvement over the baseline when considering one of the two dictionaries. When using both, we observe an increase of $1.45$ BLEU points, which corresponds to a $9.6\%$ relative increase. Moreover, we can gain from performing another step. However, the inclusion of the English–Chinese dictionary is harmful in this case, probably because 1-to-$n$ alignments are less frequent for this direction, and have been captured during the first step. By including the Chinese–English dictionary only, we can achieve an increase of $1.85$ absolute BLEU points ($12.2\%$ relative) over the initial baseline.[10]

**Quality of the Dictionaries** To assess the quality of the extraction procedure, we simply manually evaluated the ratio of incorrect entries in the dictionaries. After one step of word packing, the Chinese–English and the English–Chinese dictionaries respectively contain 7.4% and 13.5% incorrect entries. After two steps of packing, they only contain 5.9% and 10.3% incorrect entries.

### 5.2 Alignment Types

Intuitively, the word alignments obtained after word packing are more likely to be 1-to-1 than before. Indeed, the word sequences in one language that usually align to one single word in the other language have been grouped together to form one single token. Table 4 shows the detail of the distribution of alignment types after one and two steps of automatic repacking. In particular, we can observe that the $1{:}1$

| | | $1{:}0$ | $1{:}1$ | $1{:}2$ | $1{:}3$ | $1{:}n$ ($n > 3$) |
|---|---|---|---|---|---|---|
| C-E | Base. | 21.64 | **63.76** | 9.49 | 3.36 | 1.75 |
| | n=1 | 19.69 | 69.43 | 6.32 | 2.79 | 1.78 |
| | n=2 | 19.67 | **71.57** | 4.87 | 2.12 | 1.76 |
| E-C | Base. | 29.77 | **57.47** | 10.03 | 1.65 | 1.08 |
| | n=1 | 26.59 | 61.95 | 8.82 | 1.55 | 1.09 |
| | n=2 | 25.10 | **62.73** | 9.38 | 1.68 | 1.12 |

Table 4: Distribution of alignment types ($\%$)

alignments are more frequent after the application of repacking: the ratio of this type of alignment has increased by $7.81\%$ for Chinese–English and $5.26\%$ for English–Chinese.

### 5.3 Influence of Word Segmentation

To test the influence of the initial word segmentation on the process of word packing, we considered an additional segmentation configuration, based on an automatic segmenter combining rule-based and statistical techniques (Zhao et al., 2001).

| | BLEU[%] |
|---|---|
| Original segmentation | 15.14 |
| Original segmentation + Word packing | **16.99** |
| Automatic segmentation | 14.91 |
| Automatic segmentation + Word packing | **17.51** |

Table 5: Influence of Chinese segmentation

The results obtained are displayed in Table 5. As expected, the automatic segmenter leads to slightly lower results than the human-corrected segmentation. However, the proposed method seems to be beneficial irrespective of the choice of segmentation. Indeed, we can also observe an improvement in the new setting: 2.6 points absolute increase in BLEU ($17.4\%$ relative).[11]

[9]The parameters $k$, $t_{ac}$, and $t_{cooc}$ are optimized for each step, and the alignment obtained using the best set of parameters for a given step are used as input for the following step.

[10]Note that this setting (using both dictionaries for the first step and only the Chinese dictionary for the second step) is also the best setting on the development set.

[11]We could actually consider an extreme case, which would consist of splitting the sentences into characters, i.e. each character would be blindly treated as one word. The segmentation

### 5.4 Exploiting Manually Developed Resources

We also compared our technique for automatic packing of words with the exploitation of manually developed resources. More specifically, we used a 1-to-$n$ Chinese–English bilingual dictionary, described in Section 3.4, and used it in place of the automatically acquired dictionary. Words are thus grouped according to this dictionary, and we then apply the same word aligner as for previous experiments. In this case, since we are not bootstrapping from the output of a word aligner, this can actually be seen as a pre-processing step prior to alignment. These resources follow more or less the same format as the output of the word segmenter mentioned in Section 5.1.2 (Zhao et al., 2001), so the experiments are carried out using this segmentation.

| | BLEU[%] |
|---|---|
| Baseline | 14.91 |
| Automatic word packing | **17.51** |
| Packing with “manual” dictionary | 16.15 |

Table 6: Exploiting manually developed resources

The results obtained are displayed in Table 6.We can observe that the use of the manually developed dictionary provides us with an improvement in translation quality: 1.24 BLEU points absolute (8.3% relative). However, there does not seem to be a clear gain when compared with the automatic method. Even if those manual resources were extended, we do not believe the improvement is sufficient enough to justify this additional effort.

## 6 Conclusion and Future Work

In this paper, we have introduced a simple yet effective method to pack words together in order to give a different and simplified input to automatic word aligners. We use a bootstrap approach in which we first extract 1-to-$n$ word alignments using an existing word aligner, and then estimate the confidence of those alignments to decide whether or not the $n$ words have to be grouped; if so, this group is considered a new basic unit to consider. We can finally re-apply the word aligner to the updated sentences.

We have evaluated the performance of our approach by measuring the influence of this process on a Chinese-to-English MT task, based on the IWSLT 2006 evaluation campaign. We report a 12.2% relative increase in BLEU score over a standard phrase-based SMT system. We have verified that this process actually reduces the number of $1\!:n$ alignments with $n \neq 1$, and that it is rather independent from the (Chinese) segmentation strategy.

As for future work, we first plan to consider different confidence measures for the filtering of the alignment candidates. We also want to bootstrap on different word aligners; in particular, one possibility is to use the flexible HMM word-to-phrase model of Deng and Byrne (2005) in place of IBM model 4. Finally, we would like to apply this method to other corpora and language pairs.

would thus be completely driven by the *bilingual* alignment process (see also (Wu, 1997; Tiedemann, 2003) for related considerations). In this case, our approach would be similar to the approach of (Xu et al., 2004), except for the estimation of candidates.

## Acknowledgment

This work is supported by Science Foundation Ireland (grant number OS/IN/1732). Prof. Tiejun Zhao and Dr. Muyun Yang from the MT group of Harbin Institute of Technology, and Yajuan Lv from the Institute of Computing Technology, Chinese Academy of Sciences, are kindly acknowledged for providing us with the Chinese segmenter and the manually developed bilingual dictionary used in our experiments.

## References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of COLING-ACL 2006*, pages 9–16, Sydney, Australia.

Alexandra Birch, Chris Callison-Burch, and Miles Osborne. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proceedings of AMTA 2006*, pages 10–18, Boston, MA.

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.

Yonggang Deng and William Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proceedings of HLT-EMNLP 2005*, pages 169–176, Vancouver, Canada.

Alexander Fraser and Daniel Marcu. 2006. Measuring word alignment quality for statistical machine translation. Technical Report ISI-TR-616, ISI/University of Southern California.

Mihoko Kitamura and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 79–87, Copenhagen, Denmark.

Philip Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54, Edmonton, Canada.

Philip Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, pages 115–124, Washington, District of Columbia.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL 2006*, pages 104–111, New York, NY.

Adam Lopez and Philip Resnik. 2006. Word-based alignment, phrase-based translation: What's the link? In *Proceedings of AMTA 2006*, pages 90–99, Cambridge, MA.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP 2002*, pages 133–139, Morristown, NJ.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of EMNLP 1997*, pages 97–108, Somerset, New Jersey.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*, pages 160–167, Sapporo, Japan.

Marian Olteanu, Chris Davis, Ionut Volosen, and Dan Moldovan. 2006. Phramer - an open source statistical phrase-based translator. In *Proceedings of the NAACL 2006 Workshop on Statistical Machine Translation*, pages 146–149, New York, NY.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of IWSLT 2006*, pages 1–15, Kyoto, Japan.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Andrea Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152, Las Palmas, Spain.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of EACL 2003*, pages 339–346, Budapest, Hungary.

David Vilar, Maja Popovic, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of IWSLT 2006*, pages 205–212, Kyoto, Japan.

Stefan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING 1996*, pages 836–841, Copenhagen, Denmark.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Jia Xu, Richard Zens, and Hermann Ney. 2004. Do we need chinese word segmentation for statistical machine translation? In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 122–128, Barcelona, Spain.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of HLT-NAACL 2004*, pages 257–264, Boston, MA.

Tiejun Zhao, Yajuan Lü, and Hao Yu. 2001. Increasing accuracy of chinese segmentation with strategy of multi-step processing. *Journal of Chinese Information Processing*, 15(1):13–18.