PESA: Phrase Pair Extraction as Sentence Splitting

Stephan Vogel
InterACT, Language Technologies Institute, Carnegie Mellon University
5000 Forbes Av., Pittsburgh, PA 15213
vogel+@cs.cmu.edu

Abstract

Most statistical machine translation systems use phrase-to-phrase translations to capture local context information, leading to better lexical choice and more reliable local reordering. The quality of the phrase alignment is crucial to the quality of the resulting translations. Here, we propose a new phrase alignment method, not based on the Viterbi path of word alignment models. Phrase alignment is viewed as a sentence splitting task. For a given spitting of the source sentence (source phrase, left segment, right segment) find a splitting for the target sentence, which optimizes the overall sentence alignment probability. Experiments on different translation tasks show that this phrase alignment method leads to highly competitive translation results.

1 Introduction

Statistical machine translation (SMT) is currently the most promising approach to large vocabulary text translation. In the spirit of the Candide system developed in the early 90s at IBM (Brown et al., 1993), a number of statistical machine translation systems have been presented in the last few years (Wang and Waibel, 1998), (Och and Ney, 2000), (Yamada and Knight, 2000). These systems share the basic underlying principles of applying a translation model to capture the lexical and word reordering relationships between two languages, complemented by a target language model to drive the search process through translation model hypotheses. The primary differences among systems lie in the structure of their translation models. Whereas the original IBM system was based on purely word-based translation models, modern systems try to incorporate more complex structure.

Most state of the art data-driven translation systems use phrase translations as the primary building blocks to capture local context information, leading to better lexical choice and more reliable local reordering. The quality of the translations is largely dependent on the quality of phrase pairs extracted from bilingual corpora. Different phrase alignment methods have been developed. Most of them rely on word-to-word alignment. A short introduction will be given in Section 2. In (Vogel 2004) a new phrase alignment algorithm was introduced, which is not based on calculating a Viterbi alignment. In this paper a detailed study of this approach, including several extensions, will be presented.

2 Extracting Phrase Translations from Bilingual Corpora

A simple approach to extract phrase translations from a bilingual corpus is to harvest the Viterbi path generated by a word alignment model. A number of probabilistic word alignment models have been proposed (Brown et al., 1993) (Vogel et al., 1996) (Och and Ney, 2000) and shown to be effective for statistical machine translation.

Phrase alignment is essentially a post-processing step to word alignment. For any word sequence in the source sentence the Viterbi alignment can be used to read of the indices or the corresponding target words. The smallest and the largest index are then the boundaries for the entire target phrase aligned to the source phrase.

Many word alignment models are not symmetric with respect to source and target language. The IBM type and the HMM alignment models view alignment as a function, which aligns each source word to exactly one target word. On the other hand, target word can be aligned to many (even non-consecutive) source words. The concept of word fertility was introduced to incorporate this aspect into the alignment model. To make up for this asymmetry in

word, models training can be done in both directions: source to target and target to source. This results in two Viterbi paths for each sentence pair. Different ways have been explored to combine the information from those alignments. (Och and Ney, 2000) described experiments using the intersection, the union and a combination using heuristics. (Koehn, 2003) studied different combination schemes and concluded that using the right one has a bigger impact on the resulting performance of the translation system than the underlying word alignment model itself.

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases and align those phrases in an integrated way.

3 Phrase Alignment as Sentence Splitting

Instead of searching for all possible phrase alignments in a sentence pair we can pose a simpler problem. We want to find a translation for singe a source phrase $\tilde{f} = f_1...f_l$. Assume that we have a sentence pair in our bilingual corpus which contains this phrase in the source sentence. We are looking now for a sequence of words $\tilde{e} = e_1...e_k$ in the target sentence, which is a translation of the source phrase. The traditional approach is to calculate the Viterbi alignment, using for example IBM4 alignment, and read off the source phrase. Here, we describe an alternative approach to phrase alignment, based on calculating a constrained word alignment.

3.1 Constrained Word Alignment

We modify the IBM1 alignment model in the following way:

- for words inside the source phrase we sum only over the probabilities for words inside the target phrase candidate, and for words outside of the source phrase we sum only over the probabilities for the words outside the target phrase candidates;
- the position alignment probability, which for the standard IBM1 alignment is $1/I$, where I is the number of words in the target sentence, is modified to $1/(k)$ inside the source phrase and to $1/(I-k)$ outside the source phrase.

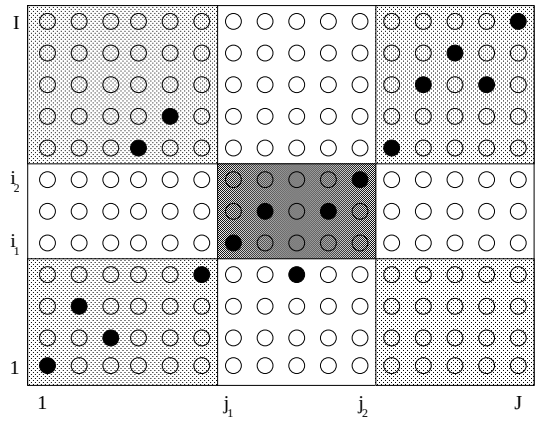

Figure 1: *Phrase alignment as sentence splitting, source and target word indices alogn the x and y axes, respectively.*

This is depicted in Figure 1. Given the source sentence (on the x-axis) and the source phrase running from position j_1 to j_2, we need to find the boundaries i_1 and i_2 in the target sentence (y-axis), which give the best alignment probability when restricting the calculation of the word alignment to the shaded areas. Filled spots indicate the Viterbi path. Notice that the center area (darker shade) does not need to include all the target words which are aligned to the source phrase according to this Viterbi path. No heuristics are applied to rule out this kind of phrase pair.

We calculate this constrained alignment probability in the following way:

$$\begin{aligned} p_{i_1,i_2}(f|e) = & \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \\ & \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{1}{k} p(f_j|e_i) \\ & \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \end{aligned} \quad (1)$$

and optimize over the target side boundaries i_1 and i_2.

$$(i_1, i_2) = \underset{i1,i2}{\operatorname{argmax}} \{p_{i_1,i_2}(f|e)\}$$

It should be mentioned that left, right, or even both segments can be empty. The alignment calculation is then modified accordingly. This means also that the entire sentence can be used as a phrase, which then can be aligned to the entire target sentence.

3.2 Looking from Both Sides

Similar to $p_{i_1,i_2}(f|e)$, we can calculate $p_{i_1,i_2}(e|f)$, now multiplying along the target words and summing over the source words.

$$p_{i_1,i_2}(e|f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1..j_2)} \frac{1}{J-l} p(e_i|f_j) \times \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{1}{l} p(e_i|f_j) \times \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1..j_2)} \frac{1}{J-l} p(e_i|f_j) \quad (2)$$

Again, the indices on the source side, i.e. j_1 and j_2 are kept fixed, whereas the indices on the target side are modified. To find the optimal target phrase we interpolate the log probabilities and take the pair (i_1, i_2) which gives the highest probability.

$$(i_1, i_2) = \underset{i_1,i_2}{\text{argmax}}\{(1-c) log(p_{(i_1,i_2)}(f|e)) + \quad (3)$$

$$c \cdot log(p_{(i_1,i_2)}(e|f))\} \quad (4)$$

Single source words are treated identically, i.e. just as phrases of length 1. The target translation can then be one or several words.

3.3 Multiple Phrase Alignment Scores

The probabilities calculated in 1 and 2 are indicators how good $(e_{i_1}, ..., e_{i_2})$ is a translation of $(f_{j_1}, ..., f_{j_2})$ within the sentence pair $(\mathbf{f}, \mathbf{e})$. Esp. in long sentences the overall alignment scores can override a poor alignment within the phrase pair.

We experimented therefore with additional scores, which look only at the phrase pair itself:

- Phrase translation scores:

$$p(\tilde{f}|\tilde{e}) = \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} p(f_j|e_i) \quad (5)$$

  and correspondingly for the inverse direction. Here the probabilities are used as given in the lexicon. Therefore, these probabilities do not depend on the sentence pair used to align the phrase pair.

- Phrase translation scores with renormalized lexical probabilities: The same calculation as before, but recalculating the lexicon probabilities to make them sum up to 1 for each column in the alignment matrix.

$$p_r(\tilde{f}|\tilde{e}) = \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{p(f_j|e_i)}{\sum_{i'=1}^{I} p(f_j|e_{i'})} \quad (6)$$

  And similarly for the direction $p_r(\tilde{e}|\tilde{f})$. The same phrase pair can then get different scores in different sentence pairs.

To make the sentence level scores 1 and 2 comparable with the phrase level scores 5 and 6 the former all normalized to the source phrase length. A weighted combination of the different scores can then be used to find the best phrase translation candidates.

The different scores capture slightly different aspects of alignment. The sentence alignment score tells how well the phrase can be aligned within a given sentence pair; the phrase translation score captures how good of a translation a target phrase is, given a word lexicon; the score in Equation 6 lies somewhat in-between, as the phrase translation probabilities depend - via renormalization of the lexicon probabilities - on the sentence pair. All these scores can be combined and used to find the optimal boundaries (i_1, i_2) for the target phrases:

$$(i_1, i_2) = \underset{i_1,i_2}{\text{argmax}}\{\sum_{k=1}^{6} c_k log(P_k)\} \quad (7)$$

where the P_k are those 6 alignment scores and the scaling factors sum up to 1.

3.4 N-best Translation Candidates

Word and phrase alignment models never work perfectly. Often the models give highest score to an alignment which is not correct or not the best one. In the phrase alignment stage, not all information used in the translation system is available. Therefore, we delay the decision about which one is the best translation for a given source phrase. We take not only the best translation candidate, but all candidates whose scores lie are within a given margin of the best one.

If a phrase occurs in several sentences in the bilingual corpus then translations from all these sentences are collected into one list. Pruning is then applied to this combined list. Additional pruning can be performed using the individual scores. For example, if any one of the scores is much worse than the corresponding score of the best candidate, then this target phrase is not used.

4 Implementation Details

Finding the translation of a phrase with the above described alignment approach is computationally rather expensive. The optimization process requires that alignment scores for I^2 target phrases are calculated. However, most boundary pairs i_1, i_2 will not make much sense, leading to translations which are too short or too long, or just being in the wrong part of the target sentence. Therefore, we restrict the search in the optimization process.

4.1 Restricting the Target Phrase Length

Unnecessary calculations can be excluded by restricting the search with the following conditions:

$$1/c_1 \cdot l - c_2 <= k <= c_1 \cdot l + c_2$$

c_1 and c_2 are constants, typically 1.5 and 1. For example, a 1-word phrase can be translated into a phrase consisting of 1 to 2.5 words, which is rounded to 3 words. The translation of a 5 word phrase can range from 2 to 9 words.

Sometimes source and target sentences in a parallel corpus differ significantly in the number of words. For example languages with agglutinative morphology typically require fewer words. To take this into account an appropriate skew factor can be applied in the calculation of the boundaries in Equation 8. Using a fertility based phrase length model we expect that tighter bounds will be possible without loos of translation quality.

4.2 Estimating the Center of the Target Phrase

An additional speed-up is possible by first estimating the approximate position of the target phrase within the target sentence. For each word f_j in the source phrase, the expected center position i_c of the translation for this word is calculated as:

$$i_c(f_j) = \sum_{i=1}^{I} \frac{1}{i} p(f_j|e_i) \quad .$$

The center for a multi-word phrase is then the average of the values calculated for the individual words.

$$i_c(f_{j_1}...f_{j_2}) = \frac{1}{l} \sum_{j=j_1}^{j_2} i_c(f_j) \quad .$$

Given this approximate position of the target phrase the overall optimization is then restricted to a range around this center position. With minimum target phrase length $k_{min} = 1/c_1 \cdot k - c_2$ and maximum target phrase length $k_{max} = c_1 \cdot k + c_2$, we get as interesting range for the target phrase boundaries:

$$\begin{aligned} i_1 &\in \{i_c - k_{max}, ..., i_c\} \\ i_2 &\in \{i_1 + k_{min}, ..., i_1 + k_{max}\} \quad . \end{aligned}$$

Of course, i_1 and i_2 are subject to the restrictions that they stay within the sentence boundaries.

4.3 Incremental Calculation

With expected center, minimal length and maximal length of the target phrase, fewer calculations of the constrained alignment are required. Still, for a 3 word phrase starting for example at position five we still have 36 valid i_1, i_2) pairs, for which the constrain alignment need to be calculated. Fortunately, we do not need to redo all the summation. If we store for each column

$$S(j, i_1, i_2) = \sum_{i=i_1}^{i_2} p(f_j|e_i)$$

then we have for the lower boundary

$$S(j, i_1 + 1, i_2) = S(j, i_1, i_2) \pm p(f_j|e_{i_1})$$

with $+$ if $j \in (j_1, ...j_2)$, i.e. inside the source phrase, and $-$ for $j \notin (j_1, ...j_2)$, i.e. the outside range. Similarly

$$S(j, i_1, i_2 + 1) = S(j, i_1, i_2) \pm p(f_j|e_{i_1+1})$$

for changing the upper boundary of the target phrase. This reduces computation significantly, esp. for longer sentences.

For the reverse direction, i.e. when calculating $p_{i_1,i_2}(e|f)$ the situation is even simpler. For each row in the alignment matrix as shown in Figure 1 we need to calculate and store the inside and outside sum only once, as the boundaries j_1 and j_2 remain constant. When calculating the product in Equation 6 the appropriate precalculated sum can be taken.

5 Calculating Phrase Translation Probabilities

One general problem with using phrase translations in a statistical machine translation system is that most phrase pairs are seen only a

few times, even in very large corpora. This is especially true for longer phrases. Therefore, probabilities based on occurrence counts have little discriminative power. Selecting one translation over the others is left to the language model within the decoder. Although the language model does occasionally select the appropriate translation, it is beneficial to provide more meaningful scores from the translation model.

There is a second problem with phrase translation probabilities based on phrase pair frequency: the resulting probabilities are not compatible with the word translation probabilities.

To get more discriminative probabilities, which are also compatible with word-for-word translations, we calculate phrase translation probabilities based on a statistical lexicon for the constituent words in the phrase according to

$$p(\tilde{f}|\tilde{e}) = \prod_j \sum_i p(f_j|e_i) \quad (8)$$

where the word probabilities $p(f_j|e_i)$ are estimated using any of the standard word alignment models.

6 Just-In-Time Phrase Alignment

We would like to use long phrases whenever they are available. This makes the phrase translation table rather large, especially for large corpora. We therefore use a different approach. The phrase alignment is calculated only for the phrases required to translate a test sentence. In other words, phrase alignment is done on-the-fly during decoding. This requires that the bilingual corpus is loaded into the decoder and an index generated for fast search. A very efficient implementation for finding all occurrences of all the n-grams seen in a test sentence, without any restriction to length and in very large corpora is described in (Zhang, Vogel, 2005).

Given the list of occurrences of phrases available for translating a test sentence, the phrase alignment is calculated and all translation candidates are inserted into a translation lattice. The search algorithm searches locates the best path in this lattice using all available information (see Section 7.2).

When run-time is critical a mixed approach can be used: the translations for the short phrases with high frequency can be extracted from the bilingual corpus during training. Only the alignments for longer and less frequent phrases need then be calculated at decoding time. This way the calculation of online alignments can be cut down to 10-20% (Zhang, Vogel, 2005).

7 Experiments

7.1 The Corpora

We report a number of experiments carried out on different translation tasks:

- TIDES: Translation of Chinese and Arabic news into English. The Chinese system was trained on a 130 million word corpus, the Arabic system on 80 million words. The corpus was primarily used to compare the new phrase alignment to our previous system.
- BTEC: This is the Basic Traveler Expression Corpus (Takezawa et al. 2002). This corpus is a multilingual corpus, existing in We report results for Chinese-English for a small corpus translation task, which was one of the tasks in the IWSLT 2004 spoken language evaluation campaign. This 20,000 sentence corpus allows for fast experimentation and was used to study different aspects of the proposed phrase alignment approach.

7.2 Decoder

The decoder used in the translation experiments is a beam search decoder, which allows for restricted word reordering (Vogel 2003). The different models used in the decoder are:

- The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus according the the new alignment method described in this paper.
- A trigram language model. The SRI language modeling toolkit was used to train the models (SRI-LM Toolkit). Modified Kneser-Ney smoothing was used throughout.
- A word reordering model, which assigns higher costs to longer distance reordering. We approximate the jump probabilities of the HMM word alignment mode (Vogel et al., 1996) by a simple Gaussian distribution:
$$p(j|j') = e^{-|j-j'|}$$
where j is the current position in the source sentence and j' is the previous position.

- A very simple sentence length model, which gives a constant bonus for each word generated. This is essentially used to compensate for the tendency of the language model to prefer shorter translations.

Each model score is multiplied by a scaling factor, which can be modified to tune the overall system.

7.3 The Evaluation Method

We report results using the well-known automatic evaluation metrics Bleu (Papineni, 2001) and NIST (MTeval, 2004). They compare the system output with several human translations and use n-gram matches to calculate a translation quality score. This score functions as a weighted precision: how many of the generated n-grams are correct. To balance high precision, a length penalty is applied to translations which are too short compared to the reference translations.

For the BTEC corpus 16 reference translations are used, while for the TIDES translation task 4 references are available. Results reported in this paper are based on case-sensitive evaluation for the TIDES data and on case-insensitive evaluation for the BTEC data.

7.4 Alignment Example

To see how the proposed phrase alignment works an example from a German-English translation task will be shown. The first step in the translation process is to find phrase translations for all the 1-word, 2-word, 3-word, etc phrases in this sentence. This is handled efficiently by the phrase search module. The result of this search is a list of phrase occurrences in the corpus.

Let us look at the following German (GE) source sentence with its gloss (GL) and an English reference translation (EN).

GE:	mit dem Zug ist es bequemer .
GL:	with the train is it more_convenient .
EN:	it is more convenient by train .

The number of phrases, which also occur in the training corpus, is 18 out of all 28 possible n-grams.

The next step is to calculate the constrained sentence alignment for each phrase occurrence, thereby obtaining translation candidates for the source phrases. For example, the phrase *mit dem Zug* occurs in 187 lines within the 58,332 line long training corpus . One occurrence is in the sentence pair:

Source:	fahren wir wieder mit dem Zug ,
Target:	let us go by train again ,

The estimation of the center of the target phrase indicates that we should search around position 4 in the target sentence. The scores (we use the negative logarithms of the probabilities) for the translation candidates are given in the following table:

Translation	Score
by train	1.75
train	3.58
go by train	5.20
go by	10.2
by train again	11.1
us go by train	11.2

We see that the best translation also has the best score. The score differences allow to prune bad translations already at this stage, using only the top 2 or 3 candidates. More translations are found in other sentence pairs, and those with the best scores are added to the translation lattice.

7.5 One-sided and Two-sided Alignment

The first experiment was done to compare the performance the performance of the new phrase alignment algorithm with a baseline system (Vogel et al. 2003), using phrase pairs extracted from the Viterbi path of HMM alignment and from the ISA alignment model (Zhang, 2003). In addition, it addressed the question if calculating a two-sided alignment gives an advantage over one-sided alignment.

Table 1: Translation results for calculating one-sided and two-sided alignment scores.

	Baseline	1-sided	2-sided
TIDES CE	22.4	26.5	28.0
TIDES AE	34.8	40.0	41.3
BTEC CE	32.7	37.7	43.5

First, the improvement over the baseline translation results turned out to be quite significant. When going from one-sided alignment

to two-sided alignment we see a remarkable improvement the small corpus. For the large corpus situation the improvement is smaller, but still worth the additional calculation.

7.6 Effect of Phrase Translation Probabilities

In Section 5 different ways to assign probabilities to phrase translation pairs have been discussed. The following experiment studied how much this impacts translation quality. Results in terms of BLEU and NIST scores are given in Table 2. The first column (Alignment) gives the condition for calculating the constraint alignment. 2S stands for the two-sided calculation in Equation 3. The statistical lexicons resulting from training IBM1 and HMM alignment models were used. In the last line (IBM1-MS) all six scores were used, calculated from the IBM1 lexicon. The second column (Decoding) shows how the phrase translation probability used in the search for the best translation of the entire test sentence was calculated.

Table 2: Translation results for the Chinese small data track.

Alignment	Decoding	Bleu	NIST
IBM1-2S	Freq	41.6	7.58
IBM1-2S	IBM1	43.5	7.67
HMM-2S	HMM	46.0	7.94
IBM1-MS	IBM1-MS	49.7	8.07

We see a clear progression in translation quality. Using the relative frequency of the phrase pairs is clearly the worst choice. Using the lexicon trained with the HMM alignment model in both the phrase alignment calculation and then the search gives an improvement over using the IBM1 lexicon. Note that this is using the HMM lexicon, but still calculating the constrained alignment in the IBM1 alignment style. The last line shows the effect of using multiple scores in the phrase alignment and using this score also in the decoder.

The subsequent experiments used the HMM/HMM configuration, i.e. using the HMM lexicon in both phrase alignment and phrase scoring.

7.7 Effect of Number of Alignments

Shorter phrases are often found in many sentences in the training corpus. Of all the $87,452$ 3-grams seen in the small Chinese-English corpus 20 occur 100 times or more, 74 occur 50 times or more, and 1130 3-grams occur 10 times or more. There is even a 5-gram which occurs 59 times. Different sentences may lead to different phrase translations. Restricting the number of occurrences for the phrases could therefore lead to a drop in translation quality. This can be seen in Table 3.

Table 3: Translation results for the BTEC Chinese small data track. Effect of maximum number (Max) of sentence pairs aligned for phrase pair extraction.

Max	1	2	3	5	10	50
NIST	5.47	6.17	6.87	7.44	7.66	7.94

7.8 Effect of Beam Size in Phrase Alignment

A different way to obtain multiple translations for a source phrase has been mentioned in 3.4. We use not only the best aligned target phrase, but multiple candidates. Thereby, we postpone the selection to a time when we also apply the language model and the other knowledge sources available to the translation system.

The results in Table 4 show that using multiple translations from each one occurrence of a source phrase has a big impact on translation quality.

Table 4: Translation results for the BTEC Chinese small data track. Effect of beam size in phrase alignment.

Beam	1	2	3	5
NIST	5.34	7.53	7.85	7.94

8 Summary and Future Work

Different phrase alignment methods have been proposed in the literature. The most popular ones are based on Viterbi alignment using one of the well known word alignment models. In this paper, we have presented a different approach, which requires a statistical lexicon, or some other word-occurrence statistics, but not the full Viterbi path. Given a source phrase and a sentence pair, containing this source phrase, we search for the optimal split points in the target sentence such that the probability for a constrained word alignment for the sentence pair is maximum. The constraint is simply: words inside the source phrase align only to words inside

the target phrase and words outside the source phrase align only to words outside the target phrase.

The benefits of this phrase alignment approach are manifold:

- A clear and simple optimization criterion is applied to find the optimal target phrase for a given source phrase. No heuristics need to be used as is the case when extracting the phrase alignments from Viterbi word alignments.
- No restriction in phrase length is required, especially when this phrase alignment is calculated at decoding time.
- Even using only the simple IBM1 word alignment model to train the required statistical lexicons is sufficient to achieve competitive translation results.
- The approach can easily be extended to leverage the power of higher order word alignment models.

So far the phrase alignment information is not used to update the word-to-word alignment probabilities. When using the IBM1 word alignment model, a significant amount of the probability mass is distributed over word pairs, which are clearly not correct translation pairs. By updating the lexicon using the phrase-to-phrase alignment, the probability distribution could be focused more on correct word pairs.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. *HLT/NAACL 2003*, Edmonton, Canada, May 2003.

Daniel Marcu and William Wong. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. *EMNLP 2002*, Philadelphia, PA, July 6-7, 2002.

NIST MT evaluation kit version 11a. Available at: http://www.nist.gov/speech/tests/mt/.

Franz Josef Och and Hermann Ney. Improved Statistical Alignment Models. *ACL 2000*, pp. 440-447, Hongkong, China.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. *Technical Report RC22176 (W0109-022)*, IBM Research Division, T. J. Watson Research Center.

SRILM - The SRI Language Modeling Toolkit. SRI Speech Technology and Research Laboratory. http://www.speech.sri.com/projects/srilm/

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto. Toward a broad-coverage bilingual corpus for speech translations of travel conversations in the real world. *LREC 2002*, pp. 147-152, Las Palmas, Spain, May 2002.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation. in *COLING 96*, pp. 836–841, Copenhagen, August 1996.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel. The CMU Statistical Translation System. MT Summit IX, New Orleans, LA, U.S.A., September 2003.

Stephan Vogel. SMT Decoder Dissected: Word Reordering. *Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.

Stephan Vogel, Sanjika Hewavitharana, Muntsin Kolss and Alex Waibel. The ISL Statistical Translation System for Spoken Language Translation. *International Workshop on Spoken Language Translation*, pp.65-72, Kyoto, Japan, September 2004.

Yeyi Wang and Alex Waibel. Fast Decoding for Statistical Machine Translation. *Proc. ICSLP 98*, Vol. 6, pp. 2775-2778, Sidney, Australia, 1998.

Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. *ACL 2000*, Nancy, France, 2000.

Ying Zhang, Stephan Vogel and Alex Waibel. Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. *Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.

Ying Zhang, Stephan Vogel. An Efficient Phrase-to-Phrase Statistical Machine Translation System for Arbitrarily Long Phrases and Large Corpora. *EAMT 2005*, Budapest, Hungary.