Wider Context by Using Bilingual Language Models in Machine Translation

Jan Niehues[1], **Teresa Herrmann**[1], **Stephan Vogel**[2] **and Alex Waibel**[1,2]
[1]Institute for Anthropomatics, KIT - Karlsruhe Institute of Technology, Germany
[2] Language Techonolgies Institute, Carnegie Mellon University, USA
[1]firstname.lastname@kit.edu [2]lastname@cs.cmu.edu

Abstract

In past Evaluations for Machine Translation of European Languages, it could be shown that the translation performance of SMT systems can be increased by integrating a bilingual language model into a phrase-based SMT system. In the bilingual language model, target words with their aligned source words build the tokens of an n-gram based language model. We analyzed the effect of bilingual language models and show where they could help to better model the translation process. We could show improvements of translation quality on German-to-English and Arabic-to-English. In addition, for the Arabic-to-English task, training an extra bilingual language model on the POS tags instead of the surface word forms led to further improvements.

1 Introduction

In many state-of-the art SMT systems, the phrase-based (Koehn et al., 2003) approach is used. In this approach, instead of building the translation by translating word by word, sequences of source and target words, so-called phrase pairs, are used as the basic translation unit. A table of correspondences between source and target phrases forms the translation model in this approach. Target language fluency is modeled by a language model storing monolingual n-gram occurrences. A log-linear combination of these main models as well as additional features is used to score the different translation hypotheses. Then the decoder searches for the translation with the highest score.

A different approach to SMT is to use a stochastic finite state transducer based on bilingual n-grams (Casacuberta and Vidal, 2004). This approach was for example successfully applied by Allauzen et al. (2010) on the French-English translation task. In this so-called n-gram approach the translation model is trained by using an n-gram language model of pairs of source and target words, called tuples. While the phrase-based approach captures only bilingual context within the phrase pairs, in the n-gram approach the n-gram model trained on the tuples is used to capture bilingual context between the tuples. As in the phrase-based approach, the translation model can also be combined with additional models like, for example, language models using log-linear combination.

Inspired by the n-gram-based approach, we introduce a bilingual language model that extends the translation model of the phrase-based SMT approach by providing bilingual word context. In addition to the bilingual word context, this approach enables us also to integrate a bilingual context based on part of speech (POS) into the translation model. When using phrase pairs it is complicated to use different kinds of bilingual contexts, since the context of the POS-based phrase pairs should be bigger than the word-based ones to make the most use of them. But there is no straightforward way to integrate phrase pairs of different lengths into the translation model in the phrase-based approach, while it is quite easy to use n-gram models with different context lengths on the tuples. We show how we can use bilingual POS-based language models to capture longer bilingual context in phrase-based translation

systems.

This paper is structured in the following way: In the next section, we will present some related work. Afterwards, in Section 3, a motivation for using the bilingual language model will be given. In the following section the bilingual language model is described in detail. In Section 5, the results and an analysis of the translation results is given, followed by a conclusion.

2 Related Work

The n-gram approach presented in Mariño et al. (2006) has been derived from the work of Casacuberta and Vidal (2004), which used finite state transducers for statistical machine translation. In this approach, units of source and target words are used as basic translation units. Then the translation model is implemented as an n-gram model over the tuples. As it is also done in phrase-based translations, the different translations are scored by a log-linear combination of the translation model and additional models.

Crego and Yvon (2010) extended the approach to be able to handle different word factors. They used factored language models introduced by Bilmes and Kirchhoff (2003) to integrate different word factors into the translation process. In contrast, we use a log-linear combination of language models on different factors in our approach.

A first approach of integrating the idea presented in the n-gram approach into phrase-based machine translation was described in Matusov et al. (2006). In contrast to our work, they used the bilingual units as defined in the original approach and they did not use additional word factors.

Hasan et al. (2008) used lexicalized triplets to introduce bilingual context into the translation process. These triplets include source words from outside the phrase and form and additional probability $p(f|e, e')$ that modifies the conventional word probability of f given e depending on trigger words e' in the sentence enabling a context-based translation of ambiguous phrases.

Other approaches address this problem by integrating word sense disambiguation engines into a phrase-based SMT system. In Chan and Ng (2007) a classifier exploits information such as local collocations, parts-of-speech or surrounding words to determine the lexical choice of target words, while Carpuat and Wu (2007) use rich context features based on position, syntax and local collocations to dynamically adapt the lexicons for each sentence and facilitate the choice of longer phrases.

In this work we present a method to extend the locally limited context of phrase pairs and n-grams by using bilingual language models. We keep the phrase-based approach as the main SMT framework and introduce an n-gram language model trained in a similar way as the one used in the finite state transducer approach as an additional feature in the log-linear model.

3 Motivation

To motivate the introduction of the bilingual language model, we will analyze the bilingual context that is used when selecting the target words. In a phrase-based system, this context is limited by the phrase boundaries. No bilingual information outside the phrase pair is used for selecting the target word. The effect can be shown in the following example sentence:

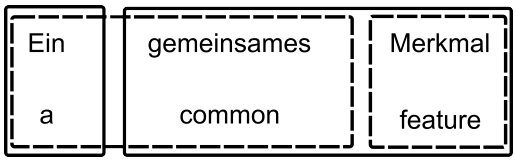
> *Ein gemeinsames Merkmal aller extremen Rechten in Europa ist ihr Rassismus und die Tatsache, dass sie das Einwanderungsproblem als politischen Hebel benutzen.*

Using our phrase-based SMT system, we get the following segmentation into phrases on the source side: *ein gemeinsames*, *Merkmal*, *aller*, *extremen Rechten*. That means, that the translation of *Merkmal* is not influenced by the source words *gemeinsames* or *aller*.

However, apart from this segmentation, other phrases could have been conceivable for building a translation:
ein, *ein gemeinsames*, *ein gemeinsames Merkmal*, *gemeinsames*, *gemeinsames Merkmal*, *Merkmal aller*, *aller*, *extremen*, *extremen Rechten* and *Rechten*.

As shown in Figure 1 the translation for the first three words *ein gemeinsames Merkmal* into *a common feature* can be created by segmenting it into *ein gemeinsames* and *Merkmal* as done by the

Figure 1: Alternative Segmentations



phrase-based system or by segmenting it into *ein* and *gemeinsames Merkmal*. In the phrase-based system, the decoder cannot make use of the fact that both segmentation variants lead to the same translation, but has to select one and use only this information for scoring the hypothesis.

Consequently, if the first segmentation is chosen, the fact that *gemeinsames* is translated to *common* does effect the translation of *Merkmal* only by means of the language model, but no bilingual context can be carried over the segmentation boundaries.

To overcome this drawback of the phrase-based approach, we introduce a bilingual language model into the phrase-based SMT system. Table 1 shows the source and target words and demonstrates how the bilingual phrases are constructed and how the source context stays available over segment boundaries in the calculation of the language model score for the sentence. For example, when calculating the language model score for the word *feature* $P($ *feature_Merkmal* | *common_gemeinsames*$)$ we can see that through the bilingual tokens not only the previous target word but also the previous source word is known and can influence the translation even though it is in a different segment.

4 Bilingual Language Model

The bilingual language model is a standard n-gram-based language model trained on bilingual tokens instead of simple words. These bilingual tokens are motivated by the tuples used in n-gram approaches to machine translation. We use different basic units for the n-gram model compared to the n-gram approach, in order to be able to integrate them into a phrase-based translation system.

In this context, a bilingual token consists of a target word and all source words that it is aligned to. More formally, given a sentence pair $e_1^I = e_1...e_I$ and $f_1^J = f_1...f_J$ and the corresponding word alignment $A = \{(i, j)\}$ the following tokens are created:

$$t_j = \{f_j\} \cup \{e_i | (i, j) \in A\} \quad (1)$$

Therefore, the number of bilingual tokens in a sentence equals the number of target words. If a source word is aligned to two target words like the word *aller* in the example sentence, two bilingual tokens are created: *all_aller* and *the_aller*. If, in contrast, a target word is aligned to two source words, only one bilingual token is created consisting of the target word and both source words.

The existence of unaligned words is handled in the following way. If a target word is not aligned to any source word, the corresponding bilingual token consists only of the target word. In contrast, if a source word is not aligned to any word in the target language sentence, this word is ignored in the bilingual language model.

Using this definition of bilingual tokens the translation probability of source and target sentence and the word alignment is then defined by:

$$p(e_1^I, f_1^J, A) = \prod_{j=1}^{J} P(t_j | t_{j-1}...t_{j-n}) \quad (2)$$

This probability is then used in the log-linear combination of a phrase-based translation system as an additional feature. It is worth mentioning that although it is modeled like a conventional language model, the bilingual language model is an extension to the translation model, since the translation for the source words is modeled and not the fluency of the target text.

To train the model a corpus of bilingual tokens can be created in a straightforward way. In the generation of this corpus the order of the target words defines the order of the bilingual tokens. Then we can use the common language modeling tools to train the bilingual language model. As it was done for the normal language model, we used Kneser-Ney smoothing.

4.1 Comparison to Tuples

While the bilingual tokens are motivated by the tuples in the n-gram approach, there are quite some differences. They are mainly due to the fact that the

Source	Target	Bi-word	LM Prob
ein	a	a_ein	P(a_ein | <s>)
gemeinsames	common	common_gemeinsames	P(common_gemeinsames | a_ein, <s>)
Merkmal	feature	feature_Merkmal	P(feature_Merkmal | common_gemeinsames)
	of	of_	P(of_ | feature_Merkmal)
aller	all	all_aller	P(all_aller | of_)
aller	the	the_aller	P(the_aller | all_aller, of_)
extremen	extreme	extreme_extremen	P(extreme_extremen)
Rechten	right	right_Rechten	P(right_Rechten | extreme_extremen)

Table 1: Example Sentence: Segmentation and Bilingual Tokens

tuples are also used to guide the search in the n-gram approach, while the search in the phrase-based approach is guided by the phrase pairs and the bilingual tokens are only used as an additional feature in scoring.

While no word inside a tuple can be aligned to a word outside the tuple, the bilingual tokens are created based on the target words. Consequently, source words of one bilingual token can also be aligned to target words inside another bilingual token. Therefore, we do not have the problems of embedded words, where there is no independent translation probability.

Since we do not create a a monotonic segmentation of the bilingual sentence, but only use the segmentation according to the target word order, it is not clear where to put source words, which have no correspondence on the target side. As mentioned before, they are ignored in the model.

But an advantage of this approach is that we have no problem handling unaligned target words. We just create bilingual tokens with an empty source side. Here, the placing order of the unaligned target words is guided by the segmentation into phrase pairs.

Furthermore, we need no additional pruning of the vocabulary due to computation cost, since this is already done by the pruning of the phrase pairs. In our phrase-based system, we allow only for twenty translations of one source phrase.

4.2 Comparison to Phrase Pairs

Using the definition of the bilingual language model, we can again have a look at the introductory example sentence. We saw that when translating the phrase *ein gemeinsames Merkmal* using a phrase-based system, the translation of *gemeinsames* into *common* can only be influenced by either the preceeding *ein # a* or by the succeeding *Merkmal # feature*, but not by both of them at the same time, since either the phrase *ein gemeinsames* or the phrase *gemeinsames Merkmal* has to be chosen when segmenting the source sentence for translation. If we now look at the context that can be used when translating this segment applying the bilingual language model, we see that the translation of *gemeinsames* into *common* is on the one hand influenced by the translation of the token *ein # a* within the bilingual language model probability $P(\textit{common\_gemeinsames} \mid \textit{a\_ein}, <s>)$.

On the other hand, it is also influenced by the translation of the word *Merkmal* into *feature* encoded into the probability $P(\textit{feature\_Merkmal} \mid \textit{common\_gemeinsames})$. In contrast to the phrase-based translation model, this additional model is capable of using context information from both sides to score the translation hypothesis. In this way, when building the target sentence, the information of aligned source words can be considered even beyond phrase boundaries.

4.3 POS-based Bilingual Language Models

When translating with the phrase-based approach, the decoder evaluates different hypotheses with different segmentations of the source sentence into phrases. The segmentation depends on available phrase pair combinations but for one hypothesis translation the segmentation into phrases is fixed. This leads to problems, when integrating parallel POS-based information. Since the amount of differ-

ent POS tags in a language is very small compared to the number of words in a language, we could manage much longer phrase pairs based on POS tags compared to the possible length of phrase pairs on the word level.

In a phrase-based translation system the average phrase length is often around two words. For POS sequences, in contrast, sequences of 4 tokens can often be matched. Consequently, this information can only help, if a different segmentation could be chosen for POS-based phrases and for word-based phrases. Unfortunately, there is no straightforward way to integrate this into the decoder.

If we now look at how the bilingual language model is applied, it is much easier to integrate the POS-based information. In addition to the bilingual token for every target word we can generate a bilingual token based on the POS information of the source and target words. Using this bilingual POS token, we can train an additional bilingual POS-based language model and apply it during translation. In this case it is no longer problematic if the context of the POS-based bilingual language model is longer than the one based on the word information, because word and POS sequences are scored separately by two different language models which cover different n-gram lengths.

The training of the bilingual POS language model is straightforward. We can build the corpus of bilingual POS tokens based on the parallel corpus of POS tags generated by running a POS tagger over both source and target side of the initial parallel corpus and the alignment information for the respective words in the text corpora.

During decoding, we then also need to know the POS tag for every source and target word. Since we build the sentence incrementally, we cannot use the tagger directly. Instead, we store also the POS source and target sequences during the phrase extraction. When creating the bilingual phrase pair with POS information, there might be different possibilities of POS sequences for the source and target phrases. But we keep only the most probable one for each phrase pair. For the Arabic-to-English translation task, we compared the generated target tags with the tags created by the tagger on the automatic translations. They are different on less than 5% of the words.

Using the alignment information as well as the source and target POS sequences we can then create the POS-based bilingual tokens for every phrase pair and store it in addition to the normal phrase pairs. At decoding time, the most frequent POS tags in the bilingual phrases are used as tags for the input sentence and the translation is done based on the bilingual POS tokens built from these tags together with their alignment information.

5 Results

We evaluated and analyzed the influence of the bilingual language model on different languages. On the one hand, we measured the performance of the bilingual language model on German-to-English on the News translation task. On the other hand, we evaluated the approach on the Arabic-to-English direction on News and Web data. Additionally, we present the impact of the bilingual language model on the English-to-German, German-to-English and French-to-English systems with which we participated in the WMT 2011.

5.1 System Description

The German-to-English translation system was trained on the European Parliament corpus, News Commentary corpus and small amounts of additional Web data. The data was preprocessed and compound splitting was applied. Afterwards the discriminative word alignment approach as described in (Niehues and Vogel, 2008) was applied to generate the alignments between source and target words. The phrase table was built using the scripts from the Moses package (Koehn et al., 2007). The language model was trained on the target side of the parallel data as well as on additional monolingual News data. The translation model as well as the language model was adapted towards the target domain in a log-linear way.

The Arabic-to-English system was trained using GALE Arabic data, which contains 6.1M sentences. The word alignment is generated using EMDC, which is a combination of a discriminative approach and the IBM Models as described in Gao et al. (2010). The phrase table is generated using Chaski as described in Gao and Vogel (2010). The language model data we trained on the GIGAWord

V3 data plus BBN English data. After splitting the corpus according to sources, individual models were trained. Then the individual models were interpolated to minimize the perplexity on the MT03/MT04 data.

For both tasks the reordering was performed as a preprocessing step using POS information from the TreeTagger (Schmid, 1994) for German and using the Amira Tagger (Diab, 2009) for Arabic. For Arabic the approach described in Rottmann and Vogel (2007) was used covering short-range reorderings. For the German-to-English translation task the extended approach described in Niehues et al. (2009) was used to cover also the long-range reorderings typical when translating between German and English.

For both directions an in-house phrase-based decoder (Vogel, 2003) was used to generate the translation hypotheses and the optimization was performed using MER training. The performance on the test-sets were measured in case-insensitive BLEU and TER scores.

5.2 German to English

We evaluated the approach on two different test sets from the News Commentary domain. The first consists of 2000 sentences with one reference. It will be referred to as Test 1. The second test set consists of 1000 sentences with two references and will be called Test 2.

5.2.1 Translation Quality

In Tables 2 and 3 the results for translation performance on the German-to-English translation task are summarized.

As it can been seen, the improvements of translation quality vary considerably between the two different test sets. While using the bilingual language model improves the translation by only 0.15 BLEU and 0.21 TER points on Test 1, the improvement on Test 2 is nearly 1 BLEU point and 0.5 TER points.

5.2.2 Context Length

One intention of using the bilingual language model is its capability to capture the bilingual contexts in a different way. To see, whether additional bilingual context is used during decoding, we analyzed the context used by the phrase pairs and by the n-gram bilingual language model.

However, a comparison of the different context lengths is not straightforward. The context of an n-gram language model is normally described by the average length of applied n-grams. For phrase pairs, normally the average target phrase pair length (avg. Target PL) is used as an indicator for the size of the context. And these two numbers cannot be compared directly.

To be able to compare the context used by the phrase pairs to the context used in the n-gram language model, we calculated the average left context that is used for every target word where the word itself is included, i.e. the context of a single word is 1. In case of the bilingual language model the score for the average left context is exactly the average length of applied n-grams in a given translation. For phrase pairs the average left context can be calculated in the following way: A phrase pair of length 1 gets a left context score of 1. In a phrase pair of length 2, the first word has a left context score of 1, since it is not influenced by any target word to the left. The second word in that phrase pair gets a left context count of 2, because it is influenced by the first word in the phrase. Correspondingly, the left context score of a phrase pair of length 3 is 6 (composed of the score 1 for the first word, score 2 for the second word and score 3 for the third word). To get the average left context for the whole translation, the context scores of all phrases are summed up and divided by the number of words in the translation.

The scores for the average left contexts for the two test sets are shown in Tables 2 and 3. They are called avg. PP Left Context. As it can be seen, the context used by the bilingual n-gram language model is longer than the one by the phrase pairs. The average n-gram length increases from 1.58 and 1.57, respectively to 2.21 and 2.18 for the two given test sets.

If we compare the average n-gram length of the bilingual language model to the one of the target language model, the n-gram length of the first is of course smaller, since the number of possible bilingual tokens is higher than the number of possible monolingual words. This can also be seen when looking at the perplexities of the two language models on the generated translations. While the perplexity of the target language model is 99 and 101 on Test 1 and 2, respectively, the perplexity of the bilin-

gual language model is 512 and 538.

Metric	No BiLM	BiLM
BLEU	30.37	30.52
TER	50.27	50.06
avg. Target PL	1.66	1.66
avg. PP Left Context	1.57	1.58
avg. Target LM N-Gram	3.28	3.27
avg. BiLM N-Gram		2.21

Table 2: German-to-English results (Test 1)

Metric	No BiLM	BiLM
BLEU	44.16	45.09
TER	41.02	40.52
avg. Target PL	1.65	1.65
avg. PP Left Context	1.56	1.57
avg. Target LM N-Gram	3.25	3.23
avg. BiLM N-Gram		2.18

Table 3: German-to-English results (Test 2)

5.2.3 Overlapping Context

An additional advantage of the n-gram-based approach is the possibility to have overlapping context. If we would always use phrase pairs of length 2 only half of the adjacent words would influence each other in the translation. The others are only influenced by the other target words through the language model. If we in contrast would have a bilingual language model which uses an n-gram length of 2, this means that every choice of word influences the previous and the following word.

To analyze this influence, we counted how many borders of phrase pairs are covered by a bilingual n-gram. For Test 1, 16783 of the 27785 borders between phrase pairs are covered by a bilingual n-gram. For Test 2, 9995 of 16735 borders are covered. Consequently, in both cases at around 60 percent of the borders additional information can be used by the bilingual n-gram language model.

5.2.4 Bilingual N-Gram Length

For the German-to-English translation task we performed an additional experiment comparing different n-gram lengths of the bilingual language model. To ensure comparability between the experiments and avoid additional noise due to different optimization results, we did not perform separate optimization runs for for each of the system variants with different n-gram length, but used the same scaling factors for all of them. Of course, the system using no bilingual language model was trained independently. In Tables 4 and 5 we can see that the length of the actually applied n-grams as well as the BLEU score increased until the bilingual language model reaches an order of 4. For higher order bilingual language models, nearly no additional n-grams can be found in the language models. Also the translation quality does not increase further when using longer n-grams.

BiLM Length	aNGL	BLEU	TER
No		30.37	50.27
1	1	29.67	49.73
2	1.78	30.36	50.05
3	2.11	30.47	50.08
4	2.21	30.52	50.06
5	2.23	30.52	50.07
6	2.24	30.52	50.07

Table 4: Different N-Gram Lengths (Test 1)

BiLM Length	aNGL	BLEU	TER
No		44.16	41.02
1	1	44.22	40.53
2	1.78	45.11	40.38
3	2.09	45.18	40.51
4	2.18	45.09	40.52
5	2.21	45.10	40.52
6	2.21	45.10	40.52

Table 5: Different N-Gram Lengths (Test 2)

5.3 Arabic to English

The Arabic-to-English system was optimized on the MT06 data. As test set the Rosetta in-house test set DEV07-nw (News) and wb (Web Data) was used.

The results for the Arabic-to-English translation task are summarized in Tables 6 and 7. The performance was tested on two different domains, translation of News and Web documents. On both tasks, the translation could be improved by more than 1

BLEU point. Measuring the performance in TER also shows an improvement by 0.7 and 0.5 points.

By adding a POS-based bilingual language model, the performance could be improved further. An additional gain of 0.2 BLEU points and decrease of 0.3 points in TER could be reached. Consequently, an overall improvement of up to 1.7 BLEU points could be achieved by integrating two bilingual language models, one based on surface word forms and one based on parts-of-speech.

System	Dev	Test	
	BLEU	TER	BLEU
NoBiLM	48.42	40.77	52.05
+ BiLM	49.29	40.04	53.51
+ POS BiLM	49.56	39.85	53.71

Table 6: Results on Arabic to English: Translation of News

System	Dev	Test	
	BLEU	TER	BLEU
NoBiLM	48.42	47.14	41.90
+ BiLM	49.29	46.66	43.12
+ POS BiLM	49.56	46.40	43.28

Table 7: Results on Arabic to English: Translation of Web documents

As it was done for the German-to-English system, we also compared the context used by the different models for this translation direction. The results are summarized in Table 8 for the News test set and in Table 9 for the translation of Web data. It can be seen like it was for the other language pair that the context used in the bilingual language model is bigger than the one used by the phrase-based translation model.

Furthermore, it is worth mentioning that shorter phrase pairs are used, when using the POS-based bilingual language model. Both bilingual language models seem to model the context quite good, so that less long phrase pairs are needed to build the translation. Instead, the more frequent short phrases can be used to generate the translation.

5.4 Shared Translation Task @ WMT2011

The bilingual language model was included in 3 systems built for the WMT2011 Shared Translation Task evaluation. A phrase-based system similar to the one described before for the German-to-English results was used. A detailed system description can be found in Herrmann et al. (2011). The results are summarized in Table 10. The performance of competitive systems could be improved in all three languages by up to 0.4 BLEU points.

Metric	No	BiLM	POS BiLM
BLEU	52.05	53.51	53.71
avg. Target PL	2.12	2.03	1.79
avg. PP Left Context	1.92	1.85	1.69
avg. BiLM N-Gram		2.66	2.65
avg. POS BiLM			4.91

Table 8: Bilingual Context in Arabic-to-English results (News)

Metric	No	BiLM	POS BiLM
BLEU	41.90	43.12	43.28
avg. Target PL	1.82	1.80	1.57
avg. PP Left Context	1.72	1.69	1.53
avg. BiLM N-Gram		2.33	2.31
avg. POS BiLM			4.49

Table 9: Bilingual Context in Arabic-to-English results (Web data)

Language Pair	No BiLM	BiLM
German-English	24.12	24.52
English-German	16.89	17.01
French-English	28.17	28.34

Table 10: Preformance of Bilingual language model at WMT2011

6 Conclusion

In this work we showed how a feature of the n-gram-based approach can be integrated into a phrase-based statistical translation system. We performed a detailed analysis on how this influences the scoring of the translation system. We could show improvements on a variety of translation tasks covering different languages and domains. Furthermore, we could show that additional bilingual context information is used.

Furthermore, the additional feature can easily be

extended to additional word factors such as part-of-speech, which showed improvements for the Arabic-to-English translation task.

Acknowledgments

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

Alexandre Allauzen, Josep M. Crego, İlknur Durgar El-Kahlout, and François Yvon. 2010. LIMSI's Statistical Translation Systems for WMT'10. In *Fifth Workshop on Statistical Machine Translation (WMT 2010)*, Uppsala, Sweden.

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 4–6, Stroudsburg, PA, USA.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Francisco Casacuberta and Enrique Vidal. 2004. Machine Translation with Inferred Stochastic Finite-State Transducers. *Comput. Linguist.*, 30:205–225, June.

Yee Seng Chan and Hwee Tou Ng. 2007. Word Sense Disambiguation improves Statistical Machine Translation. In *In 45th Annual Meeting of the Association for Computational Linguistics (ACL-07*, pages 33–40.

Josep M. Crego and François Yvon. 2010. Factored bilingual n-gram language models for statistical machine translation. *Machine Translation*, 24, June.

Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In *Proc. of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April.

Qin Gao and Stephan Vogel. 2010. Training Phrase-Based Machine Translation Models on the Cloud: Open Source Machine Translation Toolkit Chaski. In *The Prague Bulletin of Mathematical Linguistics No. 93*.

Qin Gao, Francisco Guzman, and Stephan Vogel. 2010. EMDC: A Semi-supervised Approach for Word Alignment. In *Proc. of the 23rd International Conference on Computational Linguistics*, Beijing, China.

Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Jesús Andrés-Ferrer. 2008. Triplet Lexicon Models for Statistical Machine Translation. In *Proc. of Conference on Empirical Methods in NLP*, Honolulu, USA.

Teresa Herrmann, Mohammed Mediani, Jan Niehues, and Alex Waibel. 2011. The Karlsruhe Institute of Technology Translation Systems for the WMT 2011. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinbugh, U.K.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Demonstration Session*, Prague, Czech Republic, June 23.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-based machine translation. *Comput. Linguist.*, 32, December.

Evgeny Matusov, Richard Zens, David Vilar, Arne Mauser, Maja Popović, Saša Hasan, and Hermann Ney. 2006. The rwth machine translation system. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, Barcelona, Spain, June.

Jan Niehues and Stephan Vogel. 2008. Discriminative Word Alignment via Alignment Matrix Modeling. In *Proc. of Third ACL Workshop on Statistical Machine Translation*, Columbus, USA.

Jan Niehues, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe Translation System for the EACL-WMT 2009. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Stephan Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China.