Reverse Word Order Models

Markus Freitag, Minwei Feng, Matthias Huck, Stephan Peitz and Hermann Ney
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

In this work, we study the impact of the word order decoding direction for statistical machine translation (SMT). Both phrase-based and hierarchical phrase-based SMT systems are investigated by reversing the word order of the source and/or target language and comparing the translation results with the normal direction. Analysis are done on several components such as alignment model, language model and phrase table to see which of them accounts for the differences generated by various translation directions. Furthermore, we propose to use system combination, alignment combinations and phrase table combinations to take benefit from systems trained with different translation directions. Experimental results show improvements of up to 1.7 points in BLEU and 3.1 points in TER compared to the normal direction systems for the NTCIR-9 Japanese-English and Chinese-English tasks.

1 Introduction

In this paper, we investigate the impact of word order directions on statistical machine translation (SMT) systems. The decoding direction of a phrase based statistical machine translation engine can effect the resulting translation. This work was motivated to investigate additionally reverse alignment and language model training and to combine the benefits of both direction translations. We reverse the word order from left-to-right to right-to-left for source and target sentences to produce a reverse bilingual corpus. We have several options to use this reverse corpus. We can do all, the alignment training, the language model training and the decoding process with the reverse data or take one or two steps from this common pipeline. We compare a full reverse system with the normal system. Moreover, we compare the full reverse system with systems trained on corpora with just source or target language reversed. We analyze which methods depend on the word order and which one can give a benefit. Additionally, we train translation systems which reverse the source part only or the target part only of a training corpus. We finally do a system combination of several normal and reverse systems to show the improvement of combining the benefit of both translation directions. To make the comparison fair, the normal and reverse systems are generated with the same basic features. Only the word order is changed. Furthermore, we do alignment merging and phrase table merging of the normal and reverse systems.

This paper is structured as follows. In Section 2 we give an outline of related previous work. The reverse translation approaches and the combination algorithms are in Section 3. In Section 4, we give an overview of the translation engines. The system setups are described in Section 5. We analyze the differences of the alignment model and language model training as well as the differences of the decoding process in Section 6. The experimental results are in Section 7. Finally in Section 8, we discuss the results.

2 Related Work

Watanabe and Sumita (2002) described a right-to-left decoding method for a standard phrase-based machine translation decoder. In addition, the authors presented the bidirectional decoding method,

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 159–166.

which takes the advantages of both left-to-right and right-to-left decoding method by generating output in both ways. The experimental results showed that the right-to-left decoding is better for English-to-Japanese translation, while the left-to-right decoding is suitable for Japanese-to-English translation. It was also observed that the bidirectional method was better for English-to-Japanese translation. The authors suggested that the translation output generation should match with the underlying linguistic structure for the output language.

Finch and Sumita (2009) compared a standard phrase-based machine translation decoder using a left-to-right decoding strategy to a right-to-left decoder for many language pairs on small corpora. The authors demonstrated that for most language pairs, it was better to decode from right-to-left than from left-to-right. However, the relative performance of left-to-right and right-to-left strategies seems to be highly language dependent. The word order of the target language partially accounts for the differences in performance when decoding in different directions.

In both of the above described works, the authors only changed the decoding direction. The alignment were for both directions the same. Both works motivate to investigate the impact of a right-to-left alignment and language model training. The authors tried to combine the advantages of both left-to-right and right-to-left decoding with a bidirectional decoding method. This motivates us to use system combination, combination of alignments and combination of phrase tables to benefit both translation directions. Besides the phrase-based decoder, a hierarchical decoder is used in all experiments. Instead of many language pairs with small corpora, we focus on three language pairs on large corpora. Further, we train the systems with reverse word order on the source side only and on the target side only.

In summary, the novel contribution of this paper and differences with respect to the above two works are:

1. Retraining of the alignment model with reverse corpora.
2. Usage of both hierarchical and standard phrase-based decoders.
3. Application of system combination to take benefits from bidirectional decoding.
4. Analysis of the results, including investigations of setup with merged alignments and merged phrase tables from different directions.
5. Evaluation on large-scale tasks and data from recent public evaluation campaigns.

3 Translation Setups

3.1 Reverse Translation

For reverse translation we need to change the word order of the bilingual corpus. For example, if we reverse both source and target language, the original training example "der Hund mag die Katze . $\rightarrow$ the dog likes the cat ." is converted into a new training example ". Katze die mag Hund der $\rightarrow$. cat the likes dog the". We call this type of modification of source or target language *reversion*. A system trained of this data is called *reverse*. This modification changes the corpora and hence the language model and alignment training produces different results.

We define some reverse systems we use in our experiments. For a source sentence $f_1^J = f_1, f_2, ..., f_J$ and a target sentence $e_1^I = e_1, e_2, ..., e_I$ we define the following systems:

- **normal system**:
  - normal corpus: $f_1, f_2, ..., f_J$ and $e_1, e_2, ..., e_I$
  - alignment and language model, phrase training and decoding with normal source and target corpus
- **reverse system**:
  - reverse corpus: $f_J, f_{J-1}, ..., f_1$ and $e_I, e_{J-1}, ..., e_1$
  - alignment and language model, phrase training and decoding trained with reverse source and target corpus
- **source-reverse system**:
  - source-reverse corpus: $f_J, f_{J-1}, ..., f_1$ and $e_1, e_2, ..., e_I$
  - alignment, language model, phrase training and decoding with reverse source corpus and normal target corpus
- **target-reverse system**:
  - target-reverse corpus: $f_1, f_2, ..., f_J$ and $e_I, e_{I-1}, ..., e_1$
  - alignment and language model, phrase training and decoding with normal source and reverse target corpus
- **alignment-reverse system**:

- alignment from **reverse system** reversed back
- language model from **normal system**
- phrase training and decoding with normal data

3.2 Merging of Alignment or Phrase Table

A two-directional alignment training is done by combining the final normal and the final reverse alignment. For that we make use of the well-known merging heuristics *acl*, *iu*, *intersection* and *union* (Och and Ney, 2002) and merge the normal and reverse trained alignments.

A two-directional phrase table is a combination of two phrase tables. We use two different methods for combining phrases. *Intersection* only keeps phrases which exist in both phrase tables. The model scores are the average of the model scores of both phrase tables. *Union* is a superset of intersection. Additionally to the phrases of intersection, we keep the phrases which belong to only one phrase table. The combination method *intersection + 4 features* is based on intersection. We keep all phrases which occur in both phrase tables and instead of taking the average of each model, we keep the normal scores and add the reverse scores of both phrase translation probabilities and both lexical smoothing probabilities to the phrase tables. For *intersection + 4 features*, we have four additional models. For both the phrase table and the alignment combination, we first reverse the reverse trained alignment as well as the reverse trained phrase table to the normal word order.

4 Translation Systems

We used a phrase-based SMT system, a hierarchical phrase-based SMT system and system combination in our experiments. In this section we describe the three system engines.

4.1 Phrase-based System

The phrase-based translation (PBT) system used in this work is an in-house implementation which is similar to the state-of-the-art PBT decoder described in (Zens and Ney, 2008). We took a standard set of models with phrase translation probabilities and lexical smoothing in both translation directions, word and phrase penalty, distance-based distortion model, an n-gram target language model and four binary count features. The model weights were optimized with MERT (Och, 2003).

4.2 Hierarchical Phrase-based System

For our hierarchical phrase-based (HPBT) setups, we employed the open source translation toolkit Jane (Vilar et al., 2010). The HPBT implementation is similar to (Chiang, 2007) in which a weighted synchronous context-free grammar is induced from parallel corpora. In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. We utilized the cube pruning algorithm (Huang and Chiang, 2007) for decoding. The models integrated into our hierarchical systems were: phrase translation probabilities and lexical smoothing in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four binary count features and an n-gram language model. We optimized the model weights with MERT (Och, 2003).

4.3 System Combination

System combination is used to produce consensus translations from multiple hypotheses generated with different translation engines. The basic concept of our approach to machine translation system combination is similar to the system described in (Matusov et al., 2008). This approach includes an enhanced alignment and reordering framework. Alignments between the system outputs are learned using GIZA++ (Och and Ney, 2000). A confusion network (CN) is then built using one of the hypotheses as "skeleton" or "primary" hypothesis. A hard decision on which of the hypotheses to use for that is not made, but instead combine all possible CNs into one single lattice. Majority voting on the generated lattice is performed using the prior probabilities for each system as well as other statistical models, e.g. a special n-gram language model (LM) which is learned on the input hypotheses. Scaling factors of the models were optimized using MERT. The translation with the best total score within the lattice was selected as consensus translation.

5 NTCIR-9 System Setup

All our experiments were conducted on the NTCIR-9 PatentMT. NTCIR-9 [1] is a machine translation evaluation task for patent domain. We did our experiments on both Japanese-to-English

[1] http://research.nii.ac.jp/ntcir/ntcir-9/index.html

(Jp-En) and Chinese-to-English (Ch-En) subtasks. Table 1 shows the corpus statistics of the bilingual data used for the NTCIR-9 Jp-En task. The

bilingual corpora	Japanese	English
Sentences	3 172 464	
Running Words	109 064 806	109 920 763
Vocabulary	122 295	112 214

Table 1: NTCIR-9 Jp-En bilingual training corpus statistics.

monolingual corpora	running words
us2003	1 486 878 644
us2005	1 295 478 799

Table 2: Corpus statistics of the preprocessed NTCIR-9 English monolingual training data.

segmentation of the Japanese text was done using the publicly available MeCab toolkit[2]. We used the provided *pat-dev-2006-2007* data as tuning set ("*dev*") to optimize the model weights. As unseen test set ("*test*") we used the NTCIR-8 intrinsic evaluation data set. The language model is a 4-gram trained on the bilingual data and the monolingual data sets *us2003* and *us2005* (Table 2).

For the second language pair, we took the Ch-En corpus from the NTCIR-9 evaluation. Table 3 shows the statistics of the bilingual data used. This corpus is smaller compared to the previous one.

bilingual corpora	Chinese	English
Sentences	992 519	
Running Words	41 249 103	42 651 202
Vocabulary	95 320	315 953

Table 3: Corpus statistics of the preprocessed bilingual training data for the NTCIR-9 Chinese-English corpus.

We used the English side of the bilingual data to build our language model. For the phrase-based decoder, we used a 6-gram LM, for the hierarchical system a 4-gram LM. Language models were created with the SRILM toolkit (Stolcke, 2002) using modified Kneser-Ney discounting.

6 Analysis of Reverse and Normal Systems

In the following section, we point out the differences between a reverse and normal system. We analyze the alignment and language model training. For both decoders, we analyze the phrase extraction and the decoding process.

6.1 Language Model

The procedure of LM training for both standard and reverse system is the same. We used the same amount of training data and utilize the same smoothing method. Language models were created with the SRILM toolkit (Stolcke, 2002) using modified Kneser-Ney discounting.

LM	perplexity
Ch-En standard	59.12
Ch-En reverse	58.93
Jp-En standard	37.83
Jp-En reverse	37.78

Table 4: LM perplexities, all LMs are 4-grams.

Perplexity (ppl) information are given in Table 4. There is no evidence that the reverse LM training can be reason for a translation quality difference between a normal and an reverse system.

6.2 Alignment

For the alignment training we used IBM model 1 (IBM1) (Brown et al., 1993), HMM (Vogel et al., 1996) and IBM model 4 (IBM4) (Brown et al., 1993). The word order does not affect alignment probability if we use IBM model 1. Hence for the reverse and normal systems IBM1 produces the same result. However, IBM model 4 as well as the HMM model are effected as they depend on the specific word positions which calculate a penalty for the distance of two words. Two examples are listed in Table 5 and Table 6. Table 5 is a German-English sentence pair. For the normal direction the source position of the full stop is 9, the target position of the full stop is 6. The reordering model penalizes this distance of 3. For the reverse direction both full stops are at position 1 and no penalty is calculated. Table 6 is a Chinese-English sentence pair. For the reverse sentence pair the English word 'increase' has much higher probability to be aligned to the Chinese word 'zengzhang'. The alignment training for the reverse and for the normal system head to different results. However, it is not clear which alignment training yields better translation quality. As we are going to show later, it is better to utilize both alignments. In Figure 1 an example alignment for a normal system

[2]http://mecab.sourceforge.net/

is given. Compared to the alignment in Figure 2 for an reverse system, the normal alignment differs slightlv.

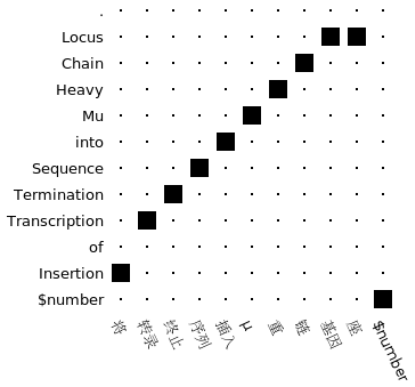

Figure 1: Example of normal alignment

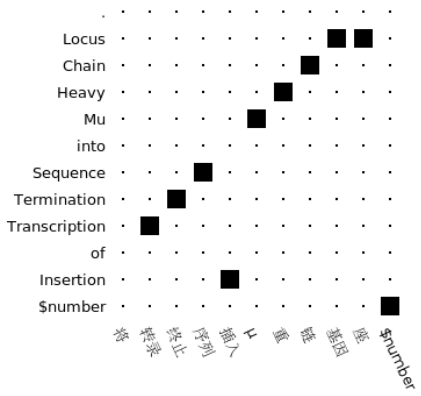

Figure 2: Example of reverse alignment

6.3 Phrase Extraction and Decoding

The phrase extraction settings for lexical as well as for hierarchical phrases are the same for the normal and the reverse direction. Formally, for lexical phrases we get the following criterion for a given sentence pair (f_1^J, e_1^I) with alignment A:

$$P(f_1^J, e_1^I, A) = \left\{ (f_{j_1}^{j_2}, e_{i_1}^{i_2}) : \forall (j,i) \in A : j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i_2 \wedge \exists (j,i) \in A : j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2 \right\} \quad (1)$$

Followed from this equation, the source and target word order direction is not relevant for the lexical phrases. The hierarchical phrases are built from the lexical phrases. All heuristics for generating these are independent from the phrase direction. To verify this with our phrase extraction, we compared the phrase table of the reverse system and the alignment-reverse system. Both systems used the same alignment, but the word order differs. In Table 9 the number of lexical as well as the number of hierarchical phrases are listed. If we reverse the source and target part of each reverse phrase pair back to normal word order both phrase tables are the same.

Another point we discover from Table 7 and Table 8 is that the normal system has more unaligned words which explains that the normal phrase table size is bigger than the reverse one, as shown in Table 9 and Table 10.

system	source	target
reverse	17 074 676	20 278 170
normal	18 039 699	20 598 882
total words	109 064 806	109 920 763

Table 7: Amount of unaligned words for Jp-En.

system	source	target
reverse	4 787 877	7 129 962
normal	4 874 689	7 233 732
total words	41 249 103	42 651 202

Table 8: Amount of unaligned words for Ch-En.

The decoding step of the hierarchical phrase-based system is independent of the word order direction. In our HPBT, the final translation is build hierarchical, it is irrelevant if the corpus is reverse or normal. The hierarchical decoder proceeds with the search process in the same way for both directions.

For the standard phrase based approach, the search is done by Dynamic Programming Beam Search (Zens and Ney, 2008). As seen in the publications mentioned in Section 2, the Dynamic Programming Beam Search gives different results when changing the decoding direction.

7 Experiments

We compared different methods with both NTCIR-9 corpora. All experiments were evaluated with BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

phrases	HPBT normal	HPBT reverse
hierarchical	20 303 097	19 883 111
lexical	14 473 801	14 464 621
total size	34 776 898	34 347 732

Table 10: Phrase table size for the Ch-En task for the HPBT systems.

Madam	President	,	on	a	point	of	order	.
Frau	Präsidentin	,	zur	Geschäftsordnung	.			
.	order	of	point	a	on	,	President	Madam
.	Geschäftsordnung	zur	,	Präsidentin	Frau			

Table 5: For the reverse sentence pair the source punctuation **full stop** has much higher probability to be aligned to the target punctuation **full stop**.

china	's	foreign	trade	imports	and	exports	continue	to	increase
zhongguo	duiwaimaoyi	jinchukou	jixu	zengzhang					
increase	to	continue	exports	and	imports	trade	foreign	's	china
zengzhang	jixu	jinchukou	duiwaimaoyi	zhongguo					

Table 6: For the reverse sentence pair the English word **increase** has much higher probability to be aligned to the Chinese word **zengzhang** (we use pinyin to represent Chinese words).

7.1 NTCIR-9 Japanese-English

For the Japanese-English corpus, we investigated several different setups. All results are listed in Table 11. First we compared for both translation systems the performance of a normal and a reverse system. For our standard phrase-based translation system, the reverse system performs better then the normal system with 0.7 points in BLEU and 0.8 points in TER. For our hierarchical system , the reverse system outperforms the normal system with 0.7 points in BLEU and 1.9 points in TER.

We first did the system combination with only the normal PBT and normal HPBT systems. We get an improvement of 0.2 points in BLEU compared to the HPBT normal system. Nevertheless, the TER score is 0.3 points worse compared to the normal PBT system. In summary, system combination that only uses the normal PBT and normal HPBT systems gave no improvement.

Secondly, we did the system combination with combining four hypotheses (normal PBT, reverse PBT, normal HPBT and reverse HPBT) . We get an improvement of 1.7 points in BLEU and 3.1 points in TER compared to the best single system HPBT reverse.

Thirdly, we used HPBT to run some experiments with only reversing the source (source-reverse system) or the target corpus (target-reverse system) . For the test set, the source-reverse system performs worse in BLEU compared to both HPBT normal and HPBT reverse systems. For TER, we get an improvement of 0.4 points compared to HPBT normal, but also loose performance compared to HPBT reverse. The target-reverse system performs worse compared to both HPBT systems. Nevertheless, if we added the hypotheses to our system combination we get an additional improvement of 0.8 points in TER compared to the system combination of the first four systems in Table 11.

7.2 NTCIR-9 Chinese-English

For Chinese-English, we focused on the comparison of the reverse and the normal systems and the experiments of combining alignments and phrase tables. The results are listed in Table 12. For PBT, the normal system is slightly better than the reverse system with 0.3 points in BLEU. For HPBT, the reverse system performs better with 0.5 points in BLEU and 1.7 points in TER compared to the normal HPBT system. The HPBT alignment-reverse system performs similar to the HPBT reverse system. As the reverse word order does not affect the decoding, we can see that the reverse trained language model does not change the translation quality. The system combination of all four systems gives us an additional improvement of 0.4 points in BLEU and 0.5 points in TER compared to our best single system. For the combination of the normal and reverse alignments, the best result is given by the union heuristic. We get an improvement of 0.2 points in BLEU, but we loose 0.7 points in TER compared to the best single system. All in all, alignment combination gives us no similar improvement like system combination. The combination of the phrase tables degrades the scores for the intersection heuristic as well as for the union heuristic. Adding the features of the reverse phrase table to the normal one yields an improvement of 0.3 points in BLEU. Compared to the system combination we loose 0.1 points in BLEU and 0.6 points in TER. Nevertheless, adding the four reverse model scores to the normal phrase ta-

phrases	HPBT normal	HPBT reverse	HPBT alignment-reverse	PBT normal	PBT reverse
hierarchical	34 205 034	33 150 034	33 150 034	-	-
lexical	31 345 790	31 137 318	31 137 318	27 620 220	27 433 584
total size	65 550 824	64 287 352	64 287 352	27 620 220	27 433 584

Table 9: Phrase table size for the Jp-En task for the HPBT systems.

system	dev		test	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
PBT normal (*)	27.9	63.5	30.1	61.9
PBT reverse (*)	28.9	62.9	30.8	61.1
HPBT normal (*)	29.1	64.7	30.7	63.9
HPBT reverse (*)	29.6	63.3	31.4	62.0
syscombi of HPBT normal and PBT normal	29.4	63.2	30.9	62.2
syscombi of above 4 (*) systems	30.6	60.4	33.1	58.9
HPBT source-reverse (*)	28.0	64.1	30.0	62.4
HPBT target-reverse (*)	27.9	65.5	29.2	64.3
syscombi of all 6 (*) systems	30.8	59.4	33.1	58.1

Table 11: Results for NTCIR Jp-En. For all reverse systems, source and target language is reverse. For the source-reverse and target-reverse systems, only one language is reverse.

system	dev		test	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
PBT normal (*)	34.8	50.7	33.3	51.9
PBT reverse (*)	34.9	50.6	33.0	51.9
HPBT normal (*)	35.8	50.5	34.1	51.7
HPBT reverse (*)	35.6	49.2	34.6	50.0
syscombi of above 4 (*) systems	36.7	48.1	35.0	49.5
HPBT alignment-reverse	36.0	49.6	34.6	50.1
HPBT alignment merge union	36.2	49.4	34.8	50.7
HPBT alignment merge acl	36.2	49.8	34.7	50.9
HPBT alignment merge intersection	35.7	50.2	34.6	51.3
HPBT alignment merge iu	35.7	50.0	34.5	50.8
HPBT phrase table merge intersection	35.1	50.1	34.0	51.1
HPBT phrase table merge union	33.5	54.8	32.1	55.3
HPBT phrase table merge intersection + 4 features	36.1	49.2	34.9	50.1

Table 12: Results for NTCIR-9 Ch-En. For all reverse systems, source and target language is reverse.

ble gives us the best result without system combination.

8 Conclusion

In this work we revisited the idea of translation from right-to-left instead of the normal direction left-to-right. In order to do so, we did alignment and language model training as well as decoding for the reverse word order. We built up several systems with different translation directions and proposed to apply system combination to take benefit of the strength of each of the setups. Without any changes in preprocessing and with the standard set of models, we achieved an improvement over normal phrase-based and hierarchical phrase-based setups. The improvement is up to 1.7 points in BLEU and 3.1 points in TER on the NTCIR-9 PatentMT task for Japanese-English. For Chinese-English we were 0.4 points in BLEU and 0.5 points in TER better than the best single system. In anal-

ysis of our setups, we came to the conclusion that the trained alignment is the main reason for varying different translation results of systems built with different translation directions. We got improvement with the combination of the reverse and normal alignments. The combination of the reverse and normal trained phrase tables degrades the translation quality. Nevertheless, if we add the feature of the reverse phrase table to the normal one, we get the best results without system combination.

For future work, we could focus on more language pairs with large amount of training data. It could be useful to know on which language pair the reverse system produces better results, as the computational effort for reverse and normal systems is the same. Furthermore, we could try different heuristics to reorder source and target language and learn the alignment with better reordered source and target sides.

Acknowledgments

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

References

Brown, Peter F., Stephan 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, June.

Chiang, D. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Finch, Andrew and Eiichiro Sumita. 2009. Bidirectional phrase-based statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1124–1132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huang, Liang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

Matusov, E., G. Leusch, R.E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J.B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237.

Och, Franz J. and Hermann Ney. 2000. GIZA++: Training of statistical translation models.

Och, Franz Josef and Hermann Ney. 2002. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*. To appear.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Stolcke, Andreas. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

Watanabe, Taro and Eiichiro Sumita. 2002. Bidirectional decoding for statistical machine translation. In *IN PROC. OF COLING 2002*, pages 1079–1085.

Zens, Richard and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 195–205, Honolulu, Hawaii, October.