Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model

Kay Rottmann
InterACT
Interactive System Labs
University of Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
rottmann@ira.uka.de

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

Abstract

In this paper we describe a word reordering strategy for statistical machine translation that reorders the source side based on Part of Speech (POS) information. Reordering rules are learned from the word aligned corpus. Reordering is integrated into the decoding process by constructing a lattice, which contains all word reorderings according to the reordering rules. Probabilities are assigned to the different reorderings. On this lattice monotone decoding is performed. This reordering strategy is compared with our previous reordering strategy, which looks at all permutations within a sliding window. We extend reordering rules by adding context information. Phrase translation pairs are learned from the original corpus and from a reordered source corpus to better capture the reordered word sequences at decoding time. Results are presented for English → Spanish and German ↔ English translations, using the European Parliament Plenary Sessions corpus.

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, 98), (Och and Ney., 2000), (Yamada and Knight, 2000), (Vogel et al., 2003). 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 reordering of words in machine translation still remains one of the hardest problems. Here we will describe our approach using syntax-based reordering rules to create a lattice structure for test sentences that encodes all word reorderings consistent with the reordering rules learned from a word aligned training corpus.

2 Modeling Word Reordering

Different languages differ in their syntactic structure. These differences in word order can be local or global. Local reorderings are for example the swapping of adjective and noun in language pairs like Spanish and English:

Example: ADJ NN → NN ADJ
An important agreement
Un acuerto importante

Word order changes which span across the entire sentence pose a much tougher problem. For example, in the translation from German to English especially verbs participate in long range reorderings.

Example: auxiliary verb and infinite verb
Ich *werde* morgen nachmittag ... *ankommen*
I *will arrive* tomorrow afternoon ...

The '...' indicates that other information (eg. 'mit dem Zug' → 'by train') could be embedded, pushing the auxiliary verb and the infinite verb even apart.

Another example of long-distance reordering is the detached verb prefix in German.

Example: detached verb prefix
Ich *komme* morgen nachmittag ... *an*. I will *arrive* tommorow afternoon ...

The verb prefix 'an' is detached from the main verb 'komme' and moved to the end of the sentence. It is difficult to generate 'arrive' from 'komme' in a phrase-based system. Even more difficult is the translation from English into German, where arrive needs to generate both 'arrive' and 'an' at different positions in the target sentence.

To generate the correct word sequence the translation system needs to have strong, restricting evidence of how to rearrange the words, this is the approach taken in grammar-based systems, or it has to have weak evidence in the form of probabilities, and then test all (or at least a large number) of reorderings, as is the strategy in typical phrase-based statistical translation systems.

The well-known IBM and HMM word alignment models (Brown et al., 1993) and (Vogel et al., 1996) contain as one component a so-called distortion model to capture the different word orders in different languages. These distortion models can be formulated in terms of absolute positions, as in the IBM2 model, or in terms of relative positions, as in the HMM and IBM4 alignment models. These distortion models are rather weak. They essentially boil down to saying that long distance reorderings are less likely then short distance reorderings.

It is important to notice that these distortion models do not pose any restrictions as to which reorderings are possible. At decoding time all permutations need to be considered, which is impossible for any but very short sentences. A restriction to word reordering was introduced in (Wu, 95). The ITG (inverse transduction grammar) constraint allows only reorderings, which can be generated by swapping subtrees in a binary branching tree. Still, for longer sentences the number of possible reorderings is too large to be enumerated; severe pruning is necessary.

To make the distortion models more informative the aligned positions can be conditioned on the length of the sentences, on the words (lexicalized distortion models), or on word classes (parts-of-speech) or automatically generated word classes, using clustering techniques (Al-Onaizan and Papineno, 2006).

State-of-the-art SMT systems use phrases. One advantage is that phrases can capture some of the local reordering patterns. However, this is rather limited as the average length of matching phrases is typically less then two words. To capture longer ranging word reorderings these phrases need to be reordered, which brings us back to the central questions:

- How to model word reordering?
- How to estimate the parameters of the model?
- How to apply the model at translation (decoding) time?

These questions will –at least to some extent– be dealt with in subsequent sections.

2.1 Related Work

Different approaches have been developed to deal with the word order problem. First approaches worked by constraining reorderings at decoding time (Berger et al., 1996). In (Wu, 1996) the alignment model already introduces restrictions in word order, which leads also to restrictions at decoding time. A comparison of these two approaches can be found in (Zens and Ney, 2003). They have in common that they do not use any syntactic or lexical information, therefore they rely on a strong language model or on long phrases to get the right word order. Other approaches were introduced that use more linguistic knowledge, for example the use of bitext grammars that allow parsing the source and target language (Wu, 1997). In (Shen et al., 2004) and (Och et al., 2004) syntactic information was used to rerank the output of a translation system with the idea of accounting for different reordering at this stage. In (Tillmann and Zhang, 2005) and (Koehn et al., 2005) a lexicalised block-oriented reordering model is proposed that decides for a given phrase whether the next phrase should be oriented to its left or right.

The most recent and very promising approaches that have been demonstrated, reorder the source sentences based on rules learned from an aligned training corpus with a POS-tagged source side (Chen et al., 2006), (Popovic and Ney, 2006) and (Crego and Marino, 2006). These rules are then used to reorder the word sequence in the most likely way.

3 Syntactic Reordering Rules

In our approach we follow the idea proposed in (Crego and Marino, 2006) of using a parallel training corpus with a tagged source side to extract rules which allow a reordering before the translation task. By doing it this way we are able to keep the translation process in the decoder monotone and make it significantly faster compared to allowing reorderings in the decoder. To avoid making any hard decisions in reordering the source side we use a lattice structure as input (Crego and Marino, 2006), (Zhang et al., 2007) for our decoder. Lattices are created for the source sentences and contain all the possible reorderings and of course also the original word sequence. As a new feature we use the context in which a reordering pattern is seen in the training data. Context refers to the words or tags to the left or to the right of the sequence for which a reordering has been observed. By doing this we hope to differentiate between reorderings that are dependent on their context.

3.1 Learning Reordering Rules

The rules that are later applied to the source sentences are learned via an aligned corpus for which the POS information of the source sentences is available. Given a sentence pair with source words f_1^J and target words e_1^I, and the alignment a_1^J a reordering rule is extracted whenever the alignment contains a crossing, i.e. whenever there is i and j with $i < j$ and $a_i > a_j$. Within one sentence pair we always extract the longest reordering sequences only. A rule, which is observed as part of a longer reordering, is only stored if it also occurs as the longest reordering sequence in some other sentence pair. The motivation for this is that only those reorderings get learned, which really exist for themselves. This restriction allows us to extract longer reordering patterns and still keeping the number of reordering patterns manageable. This will also restrict the application of rules in wrong place in the later reordering approach.

In a second step of learning, relative frequencies are computed for every rule that has been observed more than a given number of times in the training corpus (we observed good results with more than 5 times). Because the number of rules is very high, a Suffix-Array (Zhang and Vogel, 2006) is used for faster computation of the occurrence-counts for the observed sequences that triggered a reordering.

By the above described mechanisms, we are able to extract rules using as a trigger for the reordering of the words the following types.

- Tag sequence
- Word sequence
- Context of one or two tags before and / or after the Tag sequence
- One or two words before and / or after the Tag sequence

Table 1 shows examples for rules consisting of the plain tag sequence and rules that use an additional (left) context separated by the '::'. The final reordering rule consists of the source side sequence of POS tags or words that trigger a reordering, the permutation of this sequence (given as the numbers indicating the reordering) and the relative frequency of this reordering given the source sequence in the training corpus.

source sequence	rule	freq.
PDAT NN VVINF	3 1 2	0.60
VAFIN :: PDAT NN VVINF	3 1 2	0.63
KOUI :: PDAT NN VVINF	3 2 2	0.88
moechte :: PDAT NN VVINF	3 1 2	0.92

Table 1: Example rules for German to English translation with no context, with one tag of context to the left and one word of context to the left

All four rules in Table 1 reorder the same sequence (moving the infinite Verb to the front), with different relative frequencies assigned to them. The first entry uses no context information, while the other 3 lines show the rules with context information – in this case a left context only. For this POS pattern the strongest evidence for a reordering comes from the tag sequence with one source word in front of the reordering.

3.2 Applying Reordering Rules

We begin with a lattice that contains only the monotone path of the sentence that has to be translated. First, the POS tagging is done. Then, for every sequence of POS up to a maximum length (20 in our experiments) it is tested if it occurs as the left-hand side of any reordering rule. If a match is found, then for each right-hand side a new path is added to the lattice with the words now in the reordered sequence. Similarly, for POS sequences plus left/right context, which can be POS tags or words, if a match is found then a new path is added to the lattice. This also covers the reordered part only and ignores the context positions.

To guide the decoder through the lattice by favoring often seen reorderings the relative frequency of every reordering rule is applied to the first edge after a node where the path splits up. In this case it is important to know how the scores are applied to the edges. Since we used different type of rules the relative frequencies do not sum up to 1 over all rules, but only over the rules of one type.

Another problem is introduced by the fact that the reorderings are of different lengths, and only reorderings over the same length are comparable in their scores.

So we decided to score at the outgoing edges of a node, first scoring the longer reorderings and then using the remaining probability mass for the shorter reorderings. That means for one type of rule the score of a reordering in the lattice is its relative frequency seen in the training corpus weighted with the remaining probability mass of the monotone subpath where it takes place. In detail, for reordering subpath p via the m'th of n applied rules from node l to node r for this subpath, the scores are modified and the sum over all scores of edges going out of a node sums up to 1. In the following $P(p_m)$ denotes the relative frequency for the reordering p_m.

$$Score(p_m^{l,r}) = ProbabilityMass^{l,r} \cdot P(p_m)$$

where $ProbabilityMass^{l,r}$ is the probability mass that is remaining for the monotone subsequence from node l to node r. The effective score for the monotone path then computes

$$Score(monotone^{l,r}) =$$

$$ProbabilityMass^{l,r} - \sum_{i=1}^{n} Score(p_i^{l,r})$$

so that the $ProbabilityMass$ left on the subpath from l to $r-1$ is the $Score(monotone^{l,r})$. Figure 1 shows a small example lattice with only one applied rule, and Figure 2 a lattice with more applied rules.

The next step is to combine the scores of rules with different types of context. Those rules all have different relative frequencies, that are not comparable. A high relative frequency however means that this kind of reordering was seen very often during training. So we decided to compute the scores for the rules of different context by their own, only using rules of the same context. Then we applied to a reordering that was seen by more than one ruletype, that score which was the maximum for that rule. This ensures, that those reorderings that are triggered because they occur in a special context are favored. The monotone path however, gets the minimum of all scores computed for the monotone path over the different context rules.

4 Experiments

To study the effect of the POS-based distortion model we did a number of experiments on German-to-English, English-to-German, and English-to-Spanish translation tasks. We used the European Parliament Speeches Corpus as used in the TC-Star[1] project and the SMT-Workshop evaluations. Some details of the corpus are given in Table 2.

Here train-xx is the complete training corpus, dev-xx denotes the development test set used for the MER-training (Och, 2003), and eval-xx is the unseen test set used for evaluation. In the case of

[1] http://www.tc-star.org

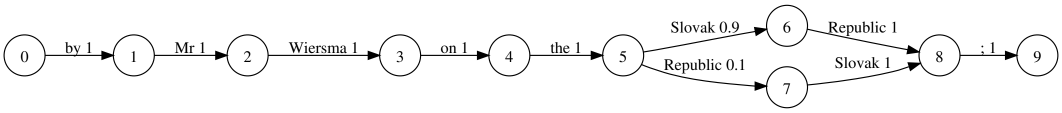

Figure 1: Example for a very small reordering lattice

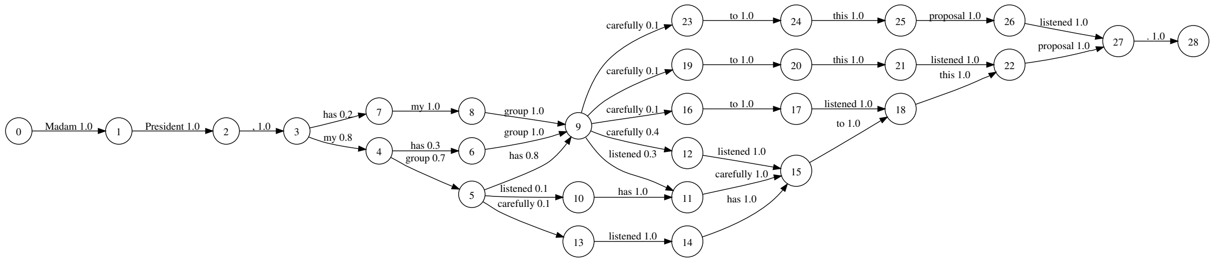

Figure 2: a larger lattice example

	Sentences	Words	Voc/OOV
train-en	1.2M	35M	97K
train-de	1.2M	33M	298K
dev-en	2K	58K	6103 / 62
dev-de	2K	54K	8762 / 306
eval-en	2K	58K	6246 / 250
eval-de	2K	55K	9008 / 551
train-en	1.2M	33M	94K
train-es	1.2M	34M	135K
dev-en	1.2K	30K	4084 / 79
eval-en	1.1K	30K	4100 / 105

Table 2: Corpus statistics EPPS training and test corpora.

German ↔ English translation the evaluation is based on 1 reference, for English → Spanish on 2 references.

For the alignment and the phrase extraction we used the Pharaoh training package (Koehn et al., 2005). To tag the corpora we used the following taggers: for English the Brill tagger (Brill, 1995) with a tag set size of 36 and for German the Stuttgart tree-tagger with a tag set size of 57 tags (Schmid, 1994). From the training corpora and the POS tagged source side we extracted the reordering rules according to the method described in Section 3.1. For the experiments reported in this paper we only learned rules up to a length of 15, since longer rules do not occur often enough in the training corpus. Table 3 displays the counts of rules that consist only of the tag sequence and those that use additional context with the tag to the left and the tag to the right learned from the training data as well as the number of rule usage on the test sentences.

4.1 Threshold and Context

In the first series of experiments we wanted to study two questions: how does the threshold value for the relative frequencies of the rules affect the translation quality, and is using context for the reordering patterns helpful. For the influence of the context we used only those rules that used the tags to the left and to the right of a reordered tag sequence. We chose that kind of context for this task because although it would probably perform worse than no context, it would indicate, which threshold is best for both types of context, those only before the reordering sequence and those after the sequence. Higher threshold, i.e. fewer rules should eventually hurt the performance. On the other side, allowing unreliable reordering rules to be used could also lead to a degradation. The results for those experiments can be seen in Table 4 and in Table 5.

The systems named *POS no Context* are those that only use the tag sequence for triggering reorderings, while those named *POS + Context* use only rules with left and right tags as context. The value behind the system name indicates the relative frequency threshold for the rules. All BLEU scores are for case sensitive evaluation. As a base-

System		# en → es		# en → de		# de → en	
Context	Threshold	Rules Learned	Rule Matches	Rules Learned	Rule Matches	Rules Learned	Rule Matches
no	0.05	21388	12715	7929	60692	13396	72728
	0.1	6848	7740	4061	27809	8528	32233
	0.2	2321	4247	1291	8192	3738	14615
	0.3	1136	3369	469	3879	1601	7076
yes	0.01	72772	21119	32380	89225	38858	88549
	0.05	46014	6888	22836	36765	28485	37608
	0.1	25962	4924	15941	19319	21469	17148
	0.2	15304	3461	8462	8574	14466	9534

Table 3: Number of reordering rules learned from the training corpus and number of rule matches on the test sentences with respect to the relative frequency threshold, without and with using the context POS tags

System	en → es
Baseline(RO3)	49.98
POS no Context 0.05	50.36
POS no Context 0.1	51.09
POS no Context 0.2	50.66
POS no Context 0.3	50.59
POS + Context 0.01	50.92
POS + Context 0.05	50.90
POS + Context 0.1	50.84
POS + Context 0.2	50.74
unseen Baseline(RO3)	48.51
unseen no Context	49.57
unseen with Context	49.49

Table 4: Case sensitive BLEU scores on English to Spanish development and test sets for the different applied threshold values

System	en→ de	de → en
Baseline(RO3)	18.92	25.64
POS no Context 0.05	19.48	26.69
POS no Context 0.1	19.55	26.46
POS no Context 0.2	19.30	26.01
POS no Context 0.3	19.22	25.73
POS + Context 0.01	19.34	25.85
POS + Context 0.05	19.34	25.86
POS + Context 0.1	19.44	25.79
unseen Baseline(RO3)	17.69	23.70
unseen no Context	17.78	24.79
unseen with Context	17.79	23.87

Table 5: Case sensitive BLEU scores on English and German development sets for the different applied threshold values

line we used our decoder with internal reordering (Vogel, 2003). The internal reordering was deactivated for every other system. So the scores reported for the reordering using the POS information does not use any additional internal reordering.

Although the first series of experiments was conducted on the developement set, it is possible to draw some conclusions from the observed results. Somewhat surprising is the fact that the system that used only the rules with context for the English to Spanish task was nearly as good as the system that did not use any context. The results get even more surprising, if you review the number of rules that were used to generate the lattices (Table:3). With a threshold value of 0.05 the number of rules with context that were applied, were even lower than the number of rules for the best setting without context while achieving nearly the same BLEU score. This means that the rules with context are able to cover as many reorderings as the rules without context although they are more specific. From this it can be seen that the reorderings in the translation from English to Spanish often occur in the same context.

In the English and German translations however, the situation is quite different. Here the score with the rules that make use of context information is below the scores without context information by ≈ 0.2 BLEU points. This is what we expected, since the German language allows a lot of reorderings of the same word sequence, because this type of context of reorderings in the German language varies a lot and it is hard to extract specific rules without omitting others. However the number of rules for the best settings with and without context shows that the system without context applied 50% more rules to the devset, which also shows the more general form of the rules without context.

Nevertheless there are some reorderings in the German language that suggest that some rules require context information. For example in sentences with auxiliary verbs, it is possible to learn a rule that moves the verb to the auxiliary verb which stays in place (e.g. " Er hat ... gesagt."). Without context it is not possible to cover those dependencies without a huge increase of wrong reorderings or the score for such a reordering is much to low to get ever applied.

Using the best system tunded on the developement data for the unseen data provided a nice improvement over the baseline system and even the system that used the context of the left and right tags performed in all three tests on the unseen data better than the internal reordering. This along with the results we observed indicate that while some reordering are better covered when context information is used, there are some reordering for which no context is useful.

In order to utilize this, we built reordering lattices that contained reorderings triggered by all extracted rules, not only just one type (Table 6 and Table 7). One problem which arose was that the rules that only used the source word sequence and no POS information hurt performance. This is obvious, since these rules only get learned if the word sequence appears often enough in the training corpus. The problem is that this however also leads to good phrases for these sequences. By having high probability reorderings for those sequences, those phrases that provide the good translation are not useful anymore and the performance is hurt.

Overall the results show that the approach of using syntactic reordering outperforms the internal reordering. In all tested language pairs we saw an improvement: in the German do English and the English to Spanish task the improvement was more than 1.0 BLEU. Also the combination of rules with different context types can lead to better performance. The improvement achieved over a single type of rule depends on the language pair, but for the translation task from English to Spanish we saw an improvement of more than 0.3 BLEU and for English to German it was more than 0.4 BLEU. In the German to English task the Improvement was only 0.1 BLEU.

System	en → es
unseen Baseline(RO3)	48.51
unseen no Context	49.52
unseen with Context	49.49
unseen combination	49.58
unseen combination-Lex	49.83

Table 6: Case sensitive BLEU scores on English to Spanish translation with with combination of all rule types and all rules except those that use only source words as trigger

System	en→ de	de → en
unseen Baseline(RO3)	17.69	23.70
unseen no Context	17.78	24.79
unseen with Context	17.79	23.87
unseen combination	18.27	24.85
unseen combination-Lex	18.21	24.88

Table 7: Case sensitive BLEU scores on English and German translation with combination of all rule types and all rules except those that use only source words as trigger

4.2 Reordering the Training Corpus

The next series of experiments we tried examined the influence of reordering in the training corpus (Popovic and Ney, 2006). One main reason why this should lead to further improvement lies in the the observation we made above, that often seen rules may contradict phrases. This effect can be seen most significantly when looking at the performance with and without rules that are only based on the exact word sequence on

Corpus	en → de	de → en
Combination	19.61	26.88
Reordered (Giza)	19.44	26.76
Reordered (Lattice)	20.00	27.06
unseen Baseline(RO3)	17.69	23.70
unseen combination	18.27	24.85
unseen reordered corpus	18.42	25.06

Table 8: Case sensitive BLEU scores using phrases from reordered training corpus

the source side. (Popovic and Ney, 2006) also reported improvements when reordering the training corpus. We conducted experiments on the English to German and German to English translation task and tried two different ways of reordering the training corpus.

The first way was to extract phrases from a corpus that had been reordered based on the existing alignment information. That is to say, the source sentence was reordered to make the alignment between source and target sentence monotone.

The second approach we tested was using the learned reordering rules to create a reordering lattice for every source sentence. Then we used the word sequence on the best path, i.e. the path with the highest score, as new source sentence. The scores we used for the edges were the same as described above. After reordering the source corpus we used this to extract a new phrase table. The results of the tests can be seen in Table 8.

As it can be seen in Table 8, the phrases extracted from the reordered training corpus using the alignment information directly performed worse than those phrases that were obtained from the corpus that was reordered using the reordering lattices.

On the unseen test data, we see an improvement of 0.15 in BLEU score compared to the previously best configuration for English to German and an improvement of 0.2 for German to English. So we were able to reproduce the effect reported by (Popovic and Ney, 2006), that a reordered training Corpus leads to a further improvement of the translation quality. As a result you can say that using the same reordering strategy for the training data as for the test data is preferable over just reordering the training corpus based on the word alignment generated by the word alignment models.

5 Future work

In the future we will try to minimize the rules that are applied to a test set for further reduction of the runtime. We believe the way to achieve this is by a better estimation of the scores for the monotone path and by alternative scoring methods so that effective pruning can be done. Also the effect of smoothing the relative frequencies should be revisited for the reordering rules.

One question that has not been answered yet, is whether additional decoder-internal reordering is still helpful. Some experiments have indicated this, and the effect seems to depend on the language pair. Another field we are working on is the integration of long range reordering rules (e.g. of the form: AUX * VB - 0 2 1, which would allow in German to English translations to move a verb next to the corresponding auxiliary verb). This can be done via the above stated rules, or as a combination with chunk reordering (Zhang et al., 2007). In the experiments described in the paper we relied on existing POS taggers. An alternative would be to use automatic clustering to obtain word classes. This would especially be useful when dealing with languages for which no good POS taggers are available. First experiments on applying word clustering for that task seem to be promising.

6 Conclusions

We presented a reordering model based on rules learned from a tagged aligned corpus. The results we obtain show that this approach outperforms our previous word reordering strategy, which used only distance information. We presented results on English to Spanish translation, which showed improvements of up to 1.3 BLEU points on unseen test data. For German to English and English to German the improvements where 0.6 and 1.1 BLEU point respectively on unseen data.

Furthermore we investigated the effect of extracting the phrase table from an reordered training corpus. By doing so we were able to obtain an additional improvement on the tested language pair German to English and English to German. So overall the improvement of the German to English translation added up to 0.8 BLEU points over the baseline result and the total improvement from English to German was 1.3 BLEU points.

It is important to note that there was no further internal reordering applied when translating the lattices - so this can possibly lead to a further performance boost. The translation time we observed was in all settings ≈ 2 times faster than the approach of reordering only in the decoder. This is due to the monotone decoding over the lattice. Some sample translations of the baseline system with internal reordering, the system with POS-reordering without context and the combination of POS-reordering with and without context can be seen in Table 9.

7 Acknowledgements

This work was partly funded by the National Science Foundation under the project STR-DUST (Grant IIS-0325905) and by DARPA under the GALE project.

References

Yaser Al-Onaizan and Kishore Papineno. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 4th annual meeting of the ACL*, pages 529–536, Sydney, Australia.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39.

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565.

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

B. Chen, M. Cettolo, and M. Federico. 2006. Reordering rules for phrase-based statistical machine translation. In *Int. Workshop on Spoken Language Translation Evaluation Campaign on Spoken Language Translation*, pages 1–15.

Josep M. Crego and Jose B. Marino. 2006. Reordering Experiments for N-Gram-based SMT. In *Spoken Language Technology Workshop*, pages 242–245, Palm Beach, Aruba.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *ACL 2000*, pages 440–447.

F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. *Proc. 2004 HLT-NAACL*, page 161.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

M. Popovic and H. Ney. 2006. POS-based word reorderings for statistical machine translation. In *Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC)*, page 1278, Genoa, Italy.

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

L. Shen, A. Sarkar, and F. J. Och. 2004. Discriminative reranking for machine translation. *In HLT-NAACL 2004: Main Proc.*, page 177.

C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 557–564, Ann Arbor, MI.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. In *COLING 16*, pages 836–841.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venogupal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical translation system. In *Proceedings of MT Summit IX*, New Orleans, LA, September.

Stephan Vogel. 2003. SMT decoder dissected: Word reordering. In *Proceedings of the Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 561–566, Beijing, China, October.

System	Translation
English Source	. . . - which we chose to set up -to continue to play a full role in this area .
Baseline	. . . , die wir haben eingerichtet, um weiterhin eine vollwertige Rolle spielen in diesem Bereich.
POS	. . . , die wir haben eingerichtet, um weiterhin eine umfassende Rolle in diesem Bereich spielen.
Combination	. . . , die wir festgelegt haben, weiterhin eine umfassende Rolle in diesem Bereich spielen .
German Source	. . . geschah, bevor das Umweltbewusstsein ausreichend geschaerft war und ehe man wusste , welche Auswirkungen das haben wuerde.
Baseline	. . . happened before the increased environmental awareness sufficient was and before we knew what impact this would have .
POS	. . . happened before the environmental awareness sufficient was and before we knew what the impact of the would have .
Combination	. . . happened before the environmental awareness was sufficient and before we knew what impact this would have .

Table 9: Sample translations of different system types

Yeyi Wang and Alex Waibel. 98. Fast Decoding for Statistical Machine Translation. In *Proc. ICSLP 98*, pages 2775–2778.

D. Wu. 1996. A polynomial-time algorithm for statistical machine translation. *Proc. 34th Annual Meeting of the Assoc. for Computational Linguistics*, page 152.

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

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

R. Zens and H. Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 144–151, Sapporo, Japan.

Ying Zhang and Stephan Vogel. 2006. Suffix array and its applications in empirical natural language processing. Technical Report CMU-LTI-06-010, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Dec.

Y. Zhang, R. Zens, and H. Ney. 2007. Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 1–8, Rochester, NY.