Clustered Word Classes for Preordering in Statistical Machine Translation

Sara Stymne
Linköping University, Sweden
sara.stymne@liu.se

Abstract

Clustered word classes have been used in connection with statistical machine translation, for instance for improving word alignments. In this work we investigate if clustered word classes can be used in a preordering strategy, where the source language is reordered prior to training and translation. Part-of-speech tagging has previously been successfully used for learning reordering rules that can be applied before training and translation. We show that we can use word clusters for learning rules, and significantly improve on a baseline with only slightly worse performance than for standard POS-tags on an English–German translation task. We also show the usefulness of the approach for the less-resourced language Haitian Creole, for translation into English, where the suggested approach is significantly better than the baseline.

1 Introduction

Word order differences between languages are problematic for statistical machine translation (SMT). If the word orders of two languages have large differences, the standard methods do not tend to work well, with difficulties in many steps such as word alignment and modelling of reordering in the decoder. This can be addressed by applying a preordering method, that is, to reorder the source side of the corpus to become similar to the target side, prior to training and translation. The rules used for reordering are generally based on some kind of linguistic annotation, such as part-of-speech tags (POS-tags).

For many languages in the world, so called less-resourced languages, however, part-of-speech taggers, or part-of-speech tagged corpora that can be used for training a tagger, are not available. In this study we investigate if it is possible to use unsupervised POS-tags, in the form of clustered word classes, as a basis for learning reordering rules for SMT. Unsupervised tagging methods can be used for any language where a corpus is available. This means that we can potentially benefit from preordering even for languages where taggers are available.

We present experiments on two data sets. First an English–German test set, where we can compare the results of clustered word classes with standard tags. We show that both types of tags beat a baseline without preordering, and that clustered tags perform nearly as well as standard tags. English and German is an interesting case for reordering experiments, since there are both long distance movement of verbs and local word order differences, for instance due to differences in adverb placements. We also apply the method to translation from the less-resourced language Haitian Creole into English, and show that it leads to an improvement over a baseline. The differences in word order between these two languages are smaller than for English–German.

Besides potentially improving SMT for less-resourced languages, the presented approach can also be used as an extrinsic evaluation method for unsupervised POS-tagging methods. This is especially useful for the task of word class clustering which is hard to evaluate.

2 Unsupervised POS-tagging

There have been several suggestions of clustering methods for obtaining word classes that are completely unsupervised, and induce classes from raw

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 28–34,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

text. Brown et al. (1992) described a hierarchical word clustering method which maximizes the mutual information of bigrams. Schütze (1995) described a distributional clustering algorithm that uses global context vectors as a basis for clustering. Biemann (2006) described a graph-based clustering methods for word classes. Goldwater and Griffiths (2007) used Bayesian reasoning for word class induction. Och (1999) described a method for determining bilingual word classes, used to improve the extraction of alignment templates through alignments between classes, not only between words. He also described a monolingual word clustering method, which is based on a maximum likelihood approach, using the frequencies of unigrams and bigrams in the training corpus.

The above methods are fully unsupervised, and produce unlabelled classes. There has also been work on what Goldwater and Griffiths (2007) call POS disambiguation, where the learning of classes is constrained by a dictionary of the allowable tags for each word. Such work has for instance been based on hidden Markov models (Merialdo, 1994), log-linear models (Smith and Eisner, 2005), and Bayesian reasoning (Goldwater and Griffiths, 2007).

Word clusters have previously been used for SMT for improving word alignment (Och, 1999), in a class-based language model (Costa-jussà et al., 2007) or for extracting gappy patterns (Gimpel and Smith, 2011). To the best of our knowledge this is the first study of applying clustered word classes for creating pre-translation reordering rules. The most similar work we are aware of is Costa-jussà and Fonollosa (2006) who used clustered word classes in a strategy they call statistical machine reordering, where the corpus is translated into a reordered language using standard SMT techniques in a pre-processing step. The addition of word classes led to improvements over just using surface form, but no comparison to using POS-tags were shown. Clustered word classes have also been used in a discriminate reordering model (Zens and Ney, 2006), and were shown to reduce the classification error rate.

Word clusters have also been used for unsupervised and semi-supervised parsing. Klein and Manning (2004) used POS-tags as the basis of a fully unsupervised parsing method, both for dependency and constituency parsing. They showed that clustered word classes can be used instead of conventional POS-tags, with some result degradation, but that it is better than several baseline systems. Koo et al. (2008) used features based on clustered word classes for semi-supervised dependency parsing and showed that using word class features together with POS-based features led to improvements, but using word class features instead of POS-based features only degraded results somewhat.

3 Reordering for SMT

There is a large amount of work on reordering for statistical machine translation. One way to approach reordering is by extending the translation model, either by adding extra models, such as lexicalized (Koehn et al., 2005) or discriminative (Zens and Ney, 2006) reordering models or by directly modelling reordering in hierarchical (Chiang, 2007) or syntactical translation models (Yamada and Knight, 2002).

Preordering is another common strategy for handling reordering. Here the source side of the corpus is transformed in a preprocessing step to become more similar to the target side. There have been many suggestions of preordering strategies. Transformation rules can be handwritten rules targeting known syntactic differences (Collins et al., 2005; Popović and Ney, 2006), or they can be learnt automatically (Xia and McCord, 2004; Habash, 2007). In these studies the reordering decision was taken deterministically on the source side. This decision can be delayed to decoding time by presenting several reordering options to the decoder as a lattice (Zhang et al., 2007; Niehues and Kolss, 2009) or as an n-best list (Li et al., 2007).

Generally reordering rules are applied to the source language, but there have been attempts at target side reordering as well (Na et al., 2009). Reordering rules can be based on different levels of linguistic annotation, such as POS-tags (Niehues and Kolss, 2009), chunks (Zhang et al., 2007) or parse trees (Xia and McCord, 2004). Common for all these levels is that a tool like a tagger or parser is needed for them to work.

In all the above studies, the reordering rules are applied to the translation input, but they are only applied to the training data in a few cases, for instance in Popović and Ney (2006). Rottmann and Vogel (2007) compared two strategies for reordering the training corpus, by using alignments, and by applying the reordering rules to create a lattice from which they extracted the 1-best reordering. They found that it was better to use the latter option, to reorder the training data based on the rules, than to use the original order in the training data. Using alignment-based reordering was not successful, however. Another option for using reorderings in the training data was presented by Niehues et al. (2009), who directly extracted phrase pairs from reordering lattices, and showed a small gain over non-reordered training data.

3.1 POS-based Preordering

Our work is based on the POS-based reordering model described by Niehues and Kolss (2009), in which POS-based rules are extracted from a word aligned corpus, where the source side is part-of-speech tagged. There are two types of rules. Short-range rules (Rottmann and Vogel, 2007) contain a pattern of POS-tags, and a possible reordering to resemble the target language, such as `VVIMP VMFIN PPER` → `PPER VMFIN VVIMP`, which moves a personal pronoun to a position in front of a verb group. Long-range rules were designed to cover movements over large spans, and also contain gaps that can match one or several words, such as `VAFIN` $\star$ `VVPP` → `VAFIN VVPP` $\star$, which moves the two parts of a German verbs together past an object of any size, so as to resemble English.

Short-range rules are extracted by identifying POS-sequences in the training corpus where there are crossing alignments. The rules are stored as the part-of-speech pattern of the source on the left hand side of the rule, and the pattern corresponding to the target side word order on the right hand side.

Long-range rules are extracted in a similar way, by identifying two neighboring POS-sequences on the source side that have crossed alignments. Gaps are introduced into the rules by replacing either the right hand side or the left hand side by a wild card. In order to constrain the application of these rules, the POS-tag to the left of the rule is included in the rule. Depending on the language pair it might be advantageous to use rules that have wildcards either on the left or right hand side. For German-to-English translation, the main long distance movement is that verbs move to the left, and, as shown by Niehues and Kolss (2009), it is advantageous to use only long-range rules with left-wildcards, as in the example rule above. For the other translation direction, it is important to move verbs to the right, and thus right-wildcard rules were better.

The probability of both short and long range rules is calculated by relative frequencies as the number of times a rule occurs divided by the number of times the source side occurs in the training data.

In a preprocessing step to decoding, all rules are applied to each input sentence, and when a rule applies, the alternative word order is added to a word lattice. To keep lattices of a reasonable size, Niehues and Kolss (2009) suggested using a threshold of 0.2 for the probability of short-range rules, of 0.05 for the probability of long range rules, and blocked rules that could be applied more than 5 times to the same sentence. We adopt these threshold values.

In this work we use the short-range reordering rules of Rottmann and Vogel (2007) and the long-range rules of Niehues and Kolss (2009). As suggested we use only right-wildcard rules for English–German translation. For Haitian Creole, we have no prior knowledge of the reordering direction, and thus choose to use both left and right long-range rules. In previous work only one standard POS-tagset was explored. In this work we investigate the effect of different type of annotation schemes, besides only POS-tags. We use several types of tags from a parser, and compare them to using unsupervised tags in the form of clustered word classes. We also apply the reordering techniques to translation from Haitian Creole, a less-resourced language for which no POS-tagger is available.

4 Experimental Setup

We conducted experiments for two language pairs, English–German and Haitian Creole–English. We always applied the reordering rules to the translation input, creating a lattice of possible reorderings as input to the decoder. For the training data we applied two strategies. As the first option we used training data from the baseline system with original word order. As the second option we reordered the training data as well, using the learnt reordering rules to create reordering lattices for the training data, from which we

ID	Form	Lemma	Dependency	Functional tag	Syntax	POS	Morphology
1	Resumption	resumption	main:>0	@NH	%NH	N	NOM SG
2	of	of	mod:>1	@<NOM-OF	%N<	PREP	
3	the	the	attr:>4	@A>	%>N	DET	
4	session	session	pcomp:>2	@<P	%NH	N	NOM SG

Table 1: Parser output

extracted the 1-best reordering, as suggested by Rottmann and Vogel (2007).

For the supervised tagging of the English source side we use a commercial functional dependency parser.[1] The main reason for using a parser instead of a tagger was that we wanted to explore the effect of different tagging schemes, which was available from this parser. An example of a tagged English text can be seen in Table 1. In this work we used four types of tags extracted from the parser output, part-of-speech tags (pos), dependency tags (dep), functional tags (func) and shallow syntax tags (syntax). The dependency tags consist of the dependency label of the word and the POS-tag of its dependent. For the example in Table 1, the sequence of dependency tags is: `main_TOP mod_N attr_N pcomp_PREP`. The other tag types are directly exemplified in Table 1. The tagsets have different sizes, as shown in Table 2.

For the unsupervised tags, we used clustered word classes obtained using the mkcls software,[2] which implements the approach of Och (1999). We explored three different numbers of clusters, 50, 125, and 625. The clustering was performed on the same corpus as the SMT training.

The translation system used is a standard phrase-based SMT system. The translation model was trained by first creating unidirectional word alignments in both directions using GIZA++ (Och and Ney, 2003), which are then symmetrized by the grow-diag-final-and method (Koehn et al., 2005). From this many-to-many alignment, consistent phrases of up to length 7 were extracted. A 5-gram language model was used, produced by SRILM (Stolcke, 2002). For training and decoding we used the Moses toolkit (Koehn et al., 2007) and the feature weights were optimized using minimum error rate training (Och, 2003). The baseline systems were trained using no additional preordering, only a distance-based reordering penalty for modelling reordering. For the Haitian Creole–English experiments we also added a lexicalized reordering model (Koehn et al., 2005), both to the baseline and to the reordered systems.

Tagset	Classes	Rules	Paths
pos	23	319147	2.1e09
dep	523	328415	2.8e09
func	49	325091	1.5e10
syntax	20	315407	4.5e11
class50	50	303292	6.2e09
class125	125	271348	1.3e07
class625	625	211606	31654

Table 2: Number of tags for each tagset in the English training corpus, number of rules extracted for each tagset, and average numbers of paths per sentence in the testset lattice using each tagset to create rules

For the English–German experiments, the translation system was trained and tested using a part of the Europarl corpus (Koehn, 2005). The training part contained 439513 sentences and 9.4 million words. Sentences longer than 40 words were filtered out. The test set has 2000 sentences and the development set has 500 sentences.

For the Haitian Creole–English experiments we used the SMS corpus released for WMT11 (Callison-Burch et al., 2011). The corpus contains 17192 sentences and 352326 words. The test and development data both contain 900 sentences each. Since we know of no POS-tagger for Haitian Creole, we only compare the clustered result to a baseline system.

Reordering rules were extracted from the same corpora that were used for training the SMT system. The word alignments needed for reordering were created using GIZA++ (Och and Ney, 2003), an implementation of the IBM models (Brown et al., 1993) of alignment, which is trained in a fully unsupervised manner based on the EM algorithm (Dempster et al., 1977).

[1] `http://www.connexor.eu/technology/machinese/machinesesyntax/`

[2] `http://www-i6.informatik.rwth-aachen.de/web/Software/mkcls.html`

5 Results

Table 2 shows the number of rules, and the average number of paths for each sentence in the test data lattice, using each tagset. For the standard tagsets the number of rules is relatively constant, despite the fact that the number of tags in the tagsets are quite different. For the clustered word classes, there are slightly fewer rules with 50 classes than for the standard tags, and the number of rules decreases with a higher number of classes. For the average number of lattice paths per sentence, there are some differences for the standard tags, but it is not related to tagset size. Again, the clustering with 50 classes has a similar number as the standard classes, but here there is a sharp decrease of lattice paths with a higher number of classes.

The translation results for the English–German experiments are shown in Table 3. We report translation results for two metrics, Bleu (Papineni et al., 2002) and NIST (Doddington, 2002), and significance testing is performed using approximate randomization (Riezler and Maxwell, 2005), with 10,000 iterations. All the systems with reordering have higher scores than the baseline on both metrics. This difference is always significant for NIST, and significant for Bleu in all cases except for two systems, one with standard tags and one with clustered tags. Between most of the systems with reordering the differences are small and most of them are not significant. Overall the systems with standard word classes perform slightly better than the clustered systems, especially the func tagset gives consistently high results, and is significantly better than four of the clustered systems on Bleu, and than one system on NIST. The fact that the number of paths were much smaller for a high number of clustered classes than for the other tagsets does not seem to have influenced the translation results.

Clustering of word classes is nondeterministic, and several runs of the cluster methods give different results, which could influence the translation results as well. To investigate this, we reran the experiment with 50 classes and baseline training data three times. The differences of the results between these runs were small, Bleu varied between 20.08–20.19 and NIST varied between 5.99–6.01. This variation is smaller than the difference between the baseline and the reordering systems, and should not influence the overall conclusions.

	Baseline training		Reordered training	
Tagset	Bleu	NIST	Bleu	NIST
Baseline	19.84	5.92	–	–
pos	20.34**	6.05**	20.26**	5.98*
dep	20.11	6.03**	20.25**	6.06**
func	20.40**	6.05**	20.40**	6.06**
syntax	20.29**	6.07**	20.32**	6.06**
class50	20.15*	6.05**	20.15*	5.99**
class125	20.15*	6.03**	20.17*	6.02**
class625	20.19**	6.05**	20.07	6.05**

Table 3: Translation results for English–German. Statistically significant differences from baseline scores are marked * ($p < 0.05$), ** ($p < 0.01$).

Tagset	Classes	Rules	Paths
class50	50	4588	3.70
class125	125	3554	1.46
class625	625	2388	1.42

Table 4: Number of classes for Haitian Creole, number of rules extracted for each tagset, and average numbers of paths per sentence in the testset lattice using each tagset to create rules

For the Haitian Creole testset both the average number of reorderings per sentence, and the number of rules, are substantially lower than for the English testset. As shown in Table 4, the trends are the same, however. With a higher number of classes there are both fewer rules and fewer rule applications. That there are few rules and paths can both depend on the fact that there are fewer word order differences between these languages, that the corpus is smaller, and that the sentence length is shorter.

Even though the number of reorderings is relatively small, there are consistent significant improvements for all reordered options on both Bleu and NIST compared to the baseline, as shown in Table 5. Between the clustered systems the differences are relatively small, and the only significant differences are that the system with 50 classes and reordered training data is worse on Bleu than 50 classes with baseline reordering and 125 classes with reordered training data, at the 0.05-level. The trend for the systems with 125 and 625 classes is in the other direction with slightly higher results with reordered data. There is hardly any difference between these two systems, which is not surprising, seeing that the number of ap-

	Baseline training		Reordered training	
Tagset	Bleu	NIST	Bleu	NIST
Baseline	29.04	5.58	–	–
class50	29.59**	5.73**	29.60**	5.69**
class125	29.52**	5.70**	29.78**	5.73**
class625	29.55**	5.70**	29.75**	5.74**

Table 5: Translation results for Haitian Creole–English. Statistically significant differences from baseline BLEU score are marked ** ($p < 0.01$).

plied rules is very similar.

6 Conclusion and Future Work

We have presented experiments of using clustered word classes as input to a preordering method for SMT. We showed that the proposed method perform better than a baseline and nearly on par with using standard tags for an English–German translation task. We also showed that it can improve results over a baseline when translating from the less-resourced language Haitian Creole into English, even though the word order differences between these languages are relatively small.

The suggested preordering algorithm with word classes is fully unsupervised, since unsupervised methods are used both for word classes and word alignments that are the basis of the preordering algorithm. This means that the method can be applied to less-resourced languages where no taggers or parsers are available, which is not the case for the many preordering methods which are based on POS-tags or parse trees.

This initial study is quite small, and in the future we plan to extend it to larger corpora and other language pairs. We would also like to compare the performance of different unsupervised word clustering and POS-tagging methods on this task.

Acknowledgments

I would like to thank Jan Niehues for sharing his code, and for his help on the POS-based reordering, and Joakim Nivre and the anonymous reviewers for their insightful comments.

References

Chris Biemann. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, pages 7–12, Sydney, Australia.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Peter F. Brown, Stephen 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.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT*, pages 22–64, Edinburgh, Scotland.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):202–228.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pages 531–540, Ann Arbor, Michigan, USA.

Marta R. Costa-jussà and José A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of EMNLP*, pages 70–76, Sydney, Australia.

Marta R. Costa-jussà, Josep M. Crego, Patrik Lambert, Maxim Khalilov, José A. R. Fonollosa, José B. Mariño, and Rafael E. Banchs. 2007. Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses. In *Proceedings of WMT*, pages 167–170, Prague, Czech Republic.

Arthur E. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurence statistics. In *Proceedings of the Second International Conference on Human Language Technology*, pages 228–231, San Diego, California, USA.

Kevin Gimpel and Noah A. Smith. 2011. Generative models of monolingual and bilingual gappy patterns. In *Proceedings of WMT*, pages 512–522, Edinburgh, Scotland.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, pages 744–751, Prague, Czech Republic.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of MT Summit XI*, pages 215–222, Copenhagen, Denmark.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, pages 478–485, Barcelona, Spain.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and

David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

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 *Proceedings of ACL, demonstration session*, pages 177–180, Prague, Czech Republic.

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

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL*, pages 595–603, Columbus, Ohio.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 720–727, Prague, Czech Republic.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Hwidong Na, Jin-Ji Li, Jungi Kim, and Jong-Hyeok Lee. 2009. Improving fluency by reordering target constituents using MST parser in English-to-Japanese phrase-based SMT. In *Proceedings of MT Summit XII*, pages 276–283, Ottawa, Ontario, Canada.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Proceedings of WMT*, pages 206–214, Athens, Greece.

Jan Niehues, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe translation system for the EACL-WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 80–84, Athens, Greece.

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

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of EACL*, pages 71–76, Bergen, Norway.

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

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

Maja Popović and Hermann Ney. 2006. POS-based reorderings for statistical machine translation. In *Proceedings of LREC*, pages 1278–1283, Genoa, Italy.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*, pages 57–64, Ann Arbor, Michigan, USA.

Kay Rottmann and Stephan Vogel. 2007. Word reordering in statistical machine translation with a POS-based distortion model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 171–180, Skövde, Sweden.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In *Proceedings of EACL*, pages 141–148, Dublin, Ireland.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of ACL*, pages 354–362, Ann Arbor, Michigan, USA.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 901–904, Denver, Colorado, USA.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of CoLing*, pages 508–514, Geneva, Switzerland.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of ACL*, pages 303–310, Philadelphia, Pennsylvania, USA.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of WMT*, pages 55–63, New York City, USA.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 21–28, Trento, Italy.