Syntax Based Reordering with Automatically Derived Rules for Improved Statistical Machine Translation

Karthik Visweswariah
IBM Research
v-karthik@in.ibm.com

Jiri Navratil
IBM Research
jiri@us.ibm.com

Jeffrey Sorensen
Google, Inc.
sorenj@google.com

Vijil Chenthamarakshan
IBM Research
vijil.e.c@in.ibm.com

Nanda Kambhatla
IBM Research
kambhatla@in.ibm.com

Abstract

Syntax based reordering has been shown to be an effective way of handling word order differences between source and target languages in Statistical Machine Translation (SMT) systems. We present a simple, automatic method to learn rules that reorder source sentences to more closely match the target language word order using only a source side parse tree and automatically generated alignments. The resulting rules are applied to source language inputs as a pre-processing step and demonstrate significant improvements in SMT systems across a variety of languages pairs including English to Hindi, English to Spanish and English to French as measured on a variety of internal test sets as well as a public test set.

1 Introduction

Different languages arrange words in different orders, whether due to grammatical constraints or other conventions. Dealing with these word order permutations is one of the fundamental challenges of machine translation. Given an exceptionally large training corpus, a phrase-based system can learn these reordering on a case by case basis. But, if our systems are to generalize to phrases not seen in the training data, they must explicitly capture and model these reorderings. However, permutations are difficult to model and impractical to search.

Presently, approaches that handle reorderings typically model word and phrase movements via a *distortion model* and rely on the target language model to produce words in the right order. Early distortion models simply penalized longer jumps more than shorter jumps (Koehn et al., 2003) independent of the source or target phrases in question. Other models (Tillman, 2004), (Al-Onaizan and Papineni, 2006) generalize this to include lexical dependencies on the source.

Another approach is to incorporate features, based on the target syntax, during modeling and decoding, and this is shown to be effective for various language pairs (Yamada and Knight, 2001), (Zollmann and Venugopal, 2006). Hierarchical phrase-based decoding (Chiang, 2005) also allows for long range reordering without explicitly modeling syntax. While these approaches have been shown to improve machine translation performance (Zollmann et al., 2008) they usually combine chart parsing with the decoding process, and are significantly more computationally intensive than phrase-based systems.

A third approach, one that has proved to be useful for phrase-based SMT systems, is to reorder each source-side sentence using a set of rules applied to a parse tree of the source sentence. The goal of these rules is to make the word order of the source sentence more similar to the expected target sentence word order. With this approach, the reordering rules are applied before training and testing with an SMT system. The efficacy of these methods has been shown on various language pairs including: French to English (Xia and McCord, 2004), German to English (Collins et al., 2005), English to

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1119–1127,
Beijing, August 2010

Chinese, (Wang et al., 2007) and Hindi to English (Ramanathan et al., 2008).

In this paper, we propose a simple model for reordering conditioned on the source side parse tree. The model is learned using a parallel corpus of source-target sentence pairs, machine generated word alignments, and source side parses. We apply the reordering model to both training and test data, for four different language pairs: English → Spanish, English → French, English → Hindi, and English → German. We show improvements in machine translation performance for all of the language pairs we consider except for English → German. We use this negative result to propose extensions to our reordering model. We note that the syntax based reordering we propose can be combined with other approaches to handling reordering and does not have to be followed by an assumption of monotonicity. In fact, our phrase-based model, trained upon reordered data, retains its reordering models and search, but we expect that these facilities are employed much more sparingly with reordered inputs.

2 Related work

There is a significant quantity of work in syntax based reordering employed to improve machine translation systems. We summarize our contributions to be:

- Learning the reordering rules based on training data (without relying on linguistic knowledge of the language pair)
- Requiring only source side parse trees
- Experimental results showing the efficacy for multiple language pairs
- Using a lexicalized distortion model for our baseline decoder

There have been several studies that have demonstrated improvements with syntax based reordering based upon hand-written rules. There have also been studies investigating the sources of these improvements (Zwarts and Dras, 2007). Hand-written rules depend upon expert knowledge of the linguistic properties of the particular language pair. Initial efforts (Niessen and Ney, 2001) were made at improving German-English translation by handling two phenomena: question inversion and detachable verb prefixes in German. In (Collins et al., 2005), (Wang et al., 2007), (Ramanathan et al., 2008), (Badr et al., 2009) rules are developed for translation from German to English, Chinese to English, English to Hindi, and English to Arabic respectively. (Xu et al., 2009) develop reordering rules based upon a linguistic analysis of English and Korean sentences and then apply those rules to translation from English into Korean and four other languages: Japanese, Hindi, Urdu and Turkish. Unlike this body of work, we automatically learn the rules from the training data and show efficacy on multiple language pairs.

There have been some studies that try to learn rules from the data. (Habash, 2007) learns reordering rules based on a dependency parse and they report a negative result for Arabic to English translation. (Zhang et al., 2007) learn reordering rules on chunks and part of speech tags, but the rules they learn are not hierarchical and would require large amounts of training data to learn rules for long sentences. Additionally, we only keep a single best reordering (instead of a lattice with possible reorderings) which makes the decoding significantly more efficient. (Xia and McCord, 2004) uses source and target side parse trees to automatically learn rules to reorder French sentences to match English order. The requirement to have both source and target side parse trees makes this method inapplicable to any language that does not have adequate tree bank resources. In addition, this work reports results using monotone decoding, since their experiments using non-monotone decoding without a distortion model were actually worse.

3 Reordering issues in specific languages

In this section we discuss the reordering issues typical of translating between English and Hindi, French, Spanish and German which are the four language pairs we experiment on in this paper.

3.1 Spanish and French

Typical word ordering patterns common to these two European languages relate to noun phrases including groups of nouns and adjectives. In contrast to English, French and Spanish adjectives and adjunct nouns follow the main noun, i.e. we typically observe a reversal of word order in noun phrases, e.g., *"A beautiful red car"* translates into French as *"Une voiture rouge beau"*, and as *"Un coche rojo bonito"* into Spanish. Phrase-based MT systems are capable of capturing these patterns provided they occur with sufficient frequency for each example in the training data. For rare noun phrases, however, the MT may produce erroneous word order that can lead to serious distortions in the meaning. Particularly difficult are nominal phrases from specialized domains that involve challenging terminology, for example: "*group reference attribute*" and "*validation checking code*". In both instances, the baseline MT system generated translations with an incorrect word order and, consequently, possibly a different meaning. We will return to these two examples in Section 5.1 to compare the output of a MT system with and without reordering.

3.2 German

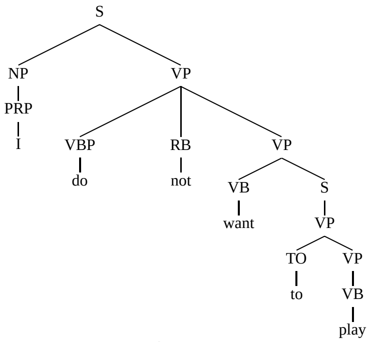
Unlike French and Spanish, German poses a considerably different challenge with respect to word ordering. The most frequent reordering in German relates to verbs, particularly verb groups consisting of auxiliary and main verbs, as well as verbs in relative clauses. Moreover, reordering patterns between German and English tend to span large portions of the sentence. We included German in our investigations to determine whether our automated rule extraction procedure can capture such long distance patterns.

3.3 Hindi

Hindi word order is significantly different than English word order; the typical order followed is Subject Object Verb (although Object Subject Verb order can be used if nouns are followed by appropriate case markers). This is in contrast to English which has a Subject Verb Object order. This can result in words that are close in English moving arbitrarily far apart in Hindi depending on the length of the noun phrase representing the object and the length of the verb phrase. These long range reorderings are generally hard for a phrase based system to capture. Another way Hindi and English differ is that prepositions in English become postpositions in Hindi and appear after the noun phrase. Again, this reordering can lead to long distance movements of words. We include Hindi in our investigation since it has significantly different structure as compared to English.

4 Learning reordering rules

In this section we describe how we learn rules that transform source parse trees so the leaf word order is more like the target language. We restrict ourselves to reorderings that can be obtained by permuting child nodes at various interior nodes in a parse tree. With many reordering phenomena discussed in Section 3 this is a fairly strong assumption about pairs of languages, and there are examples in English→Hindi where such an assumption will not allow us to generate the right reordering. As an example consider the English sentence "*I do not want to play*". The sentence has a parse:



The correct word order of the translation in Hindi is "*I to play not want*" In this case, the word *not* breaks up the verb phrase *want to play* and hence the right Hindi word order cannot be obtained by the reordering allowed by our model. We found such examples to be rare in English→Hindi, and we impose this restriction for the simplicity of the model. Experimental results on several languages show benefits of reordering in spite of this simplifying assumption.

Consider a source sentence s and its corresponding constituency parse tree S[1]. We set up the problem in a probabilistic framework, i.e. we would like to build a probabilistic model $P(T|S)$ that assigns probabilities to trees such that the

[1] In this paper we work with constituency parse trees. Initial experiments, applying similar techniques to dependency parse trees did not yield improvements.

word order in trees T which are assigned higher probability match the order of words in the target language. A parse tree, S is a set of nodes. Interior nodes have an ordered list of children. Leaf nodes in the tree are the words in the sentence s, and interior nodes are labeled by the linguistic constituent that they represent. Each word has a parent node (with only one child) labeled by the part-of-speech tag of the word.

Our model assigns non-zero probabilities to trees that can be obtained by permuting the child nodes at various interior nodes of the tree S. We assume that children of a node are ordered independently of all other nodes in the tree. Thus

$$P(T|S) = \prod_{n \in I(S)} P(\pi(c_n)|S, n, c_n),$$

where $I(S)$ is the set of interior nodes in the tree S, c_n is the list of children of node n and π is a permutation. We further assume that the reordering at a particular node is dependent only on the labels of its children:

$$P(T|S) = \prod_{n \in I(S)} P(\pi(c_n)|c_n).$$

We parameterize our model using a log-linear model:

$$P(\pi(c_n)|c_n) = \frac{1}{Z(c_n)} \exp(\lambda^T f(\pi, c_n)). \quad (1)$$

We choose the simplest possible set of feature functions: for each observed sequence of non-terminals we have one boolean feature per permutation of the sequence of non-terminals, with the feature firing iff that particular sequence is observed. Assuming, we have a training corpus C of (T, S) tree pairs, we could optimize the parameters of our model to maximize : $\prod_{S \in C} P(T|S)$. With the simple choice of feature functions described above, this amounts to:

$$P(\pi(c_n)|c_n) = \frac{\mathrm{count}(\pi(c_n))}{\mathrm{count}(c_n)},$$

where $\mathrm{count}(c_n)$ is the number of times the sequences of nodes c_n is observed in the training data and $\mathrm{count}(\pi(c_n))$ is the number of times that c_n in S is permuted to $\pi(c_n)$ in T. In Section 6, we show considering more general feature functions and relaxing some of the independence might yield improvements on certain language pairs.

For each source sentence s with parse S we find the tree T that makes the given alignment for that sentence pair most monotone. For each node n in the source tree S let D_n be the set of words that are descendants of n. Let us denote by $tpos(n)$ the average position of words in the target sentence that are aligned to words in D_n. Then

$$tpos(n) = \frac{1}{|D_n|} \sum_{w \in D_n} a(w),$$

where $a(w)$ is the index of the word on the target side that w is aligned with. If a word w is not aligned to any target word, we leave it out from the mean position calculation above. If a word w is aligned to many words we let $a(w)$ be the mean position of the words that w is aligned to. For each node n in the tree we transform the tree by sorting the list of children of n according to $tpos$. The pairs of parse trees that we obtain (S, T) in this manner form our training corpus to estimate our parameters.

In using our model, we once again go for the simplest choice, we simply reorder the source side sentences by choosing $\arg\max_T P(T|S)$ both in training and in testing; this amounts to reordering each interior node based on the most frequent reordering of the constituents seen in training. To reduce the effect of noise in training alignments we apply the reordering, only if we have seen the constituent sequence often enough in our training data (a *count threshold* parameter) and if the most frequent reordering is sufficiently more frequent than the next most frequent reordering (a *significance threshold*).

5 Experiments

5.1 Results for French, Spanish, and German

In each language, the rule extraction was performed using approximately 1.2M sentence pairs aligned using a maxent aligner (Ittycheriah and Roukos, 2005) trained using a variety of domains (Europarl, computer manuals) and a maximum entropy parser for English (Ratnaparkhi, 1999). With a significance threshold of 1.2, we obtain about 1000 rules in the eventual reordering process.

Phrase-based systems were trained for each language pair using 11M sentence pairs spanning a variety of publicly available (e.g. Europarl, UN speeches) and internal corpora (IT technical and news domains). The system phrase blocks were extracted based on a union of HMM and maxent alignments with corpus-selective count pruning. The lexicalized distortion model was used as described in (Al-Onaizan and Papineni, 2006) with a window width of up to 5 and a maximum number of skipped (not covered) words during decoding of 2. The distortion model assigns a probability to a particular word to be observed with a specific jump. The decoder uses a 5-gram interpolated language model spanning the various domains mentioned above. The baseline system without reordering and a system with reordering was trained and evaluated in contrastive experiments. The evaluation was performed utilizing the following (single-reference) test sets:

- *News*: 541 sentences from the news domain.
- *TechA:* 600 sentences from a computer-related technical domain, this has been used as a dev set.
- *TechB:* 1038 sentences from a similar domain as *TechA* used as a blind test.
- *Dev09:* 1026 sentences defined as the *news-dev2009b* development set of the Workshop on Statistical Machine Translation 2009 [2]. This set provides a reference measurement using a public data set. Previously published results on this set can be found, for example, in (Popovic et al., 2009).

In order to assess changes in word ordering patterns prior to and after an application of the reordering, we created histograms of word jumps in the alignments obtained in the baseline as well as in the reordered system. Given a source word s_i at index i and the target word t_j it is aligned to at index j, a jump of 1 would correspond to s_{i+1} aligning to target word t_{j+1}, while an alignment to t_{j-1} corresponds to a jump of -1, etc. A histogram over the jump values gives us a summary of word order distortion. If all of the jumps were one, then there is no reordering between the two languages. To gain insight into changes introduced by our reordering we look at differences of the two histograms i.e., counts after reordering minus counts before reordering. We would hope that after reordering most of the jumps are small and concentrated around one. Figure 1 shows such *difference*-histograms for the language pairs English→Spanish and English→French, respectively, on a sample of about 15k sentence pairs held out of the system training data. Here, a positive difference value indicates an increased number *after* reordering. In both cases a consistent trend toward monotonicity is observed, i.e more jumps of size one and two, and fewer large jumps. This confirms the intended reordering effect and indicates that the reordering rules extracted generalize well.

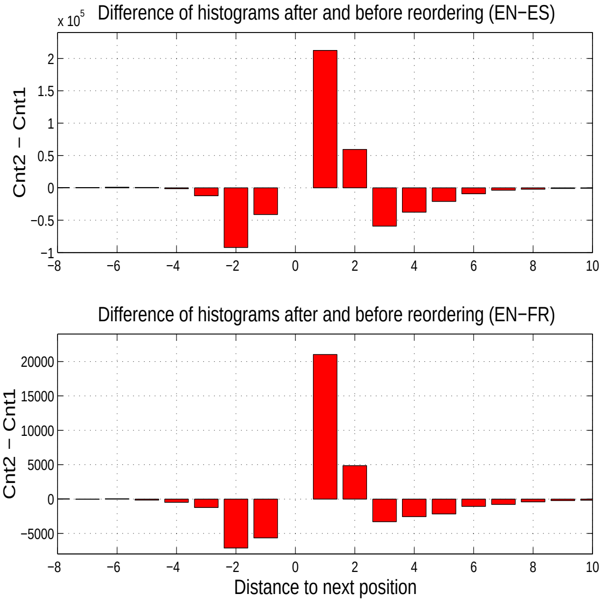


Figure 1: Difference-histogram of word order distortions for English→Spanish (upper), and English→French (lower).

Table 1 shows the resulting uncased BLEU scores for English-Spanish and English-French.

In both cases the reordering has a consistent positive effect on the BLEU scores across test sets. In examining the sources of improvement, we noticed that word order in several noun phrases that

[2] http://statmt.org/wmt09/

	System	News	TechA	TechB	Dev09
Spanish	Baseline	0.3849	0.3371	0.3483	0.2244
	Reordered	0.4031	0.3582	0.3605	0.2320
French	Baseline	0.5140	0.2971	0.3035	0.2014
	Reordered	0.5242	0.3152	0.3154	0.2092
German	Baseline	0.2580	0.1582	0.1697	0.1281
	Reordered	0.2544	0.1606	0.1682	0.1271
Hindi	Baseline	20.0			
	Reordered	21.7			

Table 1: Uncased BLEU scores for phrase-based machine translation.

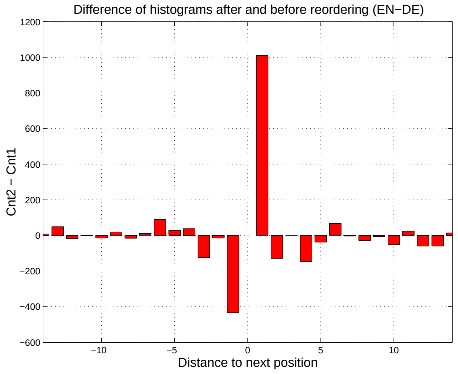


Figure 2: Difference-histogram of word order distortions for English→German.

were not common in the training data were fixed by use of the reordering rules.

Table 1 shows the BLEU scores for the English→German language pair, for which a mixed result is observed. The difference-histogram for English→German, shown in Figure 2, differs from those of the other languages with several increases in jumps of large magnitude, indicating failure of the extracted rules to generalize.

The failure of our simple method to gain consistent improvements comparable to Spanish and French, along with our preliminary finding that a relatively few manually crafted reordering rules (we describe these in Section 6.4) tend to outperform our method, leads us to believe that a more refined approach is needed in this case and will be subject of further discussion below.

5.2 Results for Hindi

Our Hindi-English experiments were run with an internal parallel corpus of roughly 250k sentence pairs (5.5M words) consisting of various domains (including news). To learn reordering rules we used HMM alignments and a maxent parser (Ratnaparkhi, 1999), with a count threshold of 100, and a significance threshold of 1.7 (these settings gave us roughly 200 rules). We also experimented with other values of these thresholds and found that the performance of our systems were not very sensitive to these thresholds. We trained Direct Translation Model 2 (DTM) systems (Ittycheriah and Roukos, 2007) with and without source reordering and evaluated on a test set of 357 sentences from the News domain. We note that the DTM baseline includes features (functions of target words and jump size) that allow it to model lexicalized reordering phenomena. The reordering window size was set to +/- 8 words for the baseline and system with reordered inputs. Table 1 shows the uncased BLEU scores for English-Hindi, showing a gain from using the reordering rules. For the reordered case, the HMM alignments are rederived, but the accuracy of these were no better than those of the unreordered input and experiments showed that the gains in performance were not due to the effect on the alignments.

Figure 3 shows *difference*-histograms for the language pair English→Hindi, on a sample of about 10k sentence pairs held out of the system training data. The histogram indicates that our reordering rules generalize and that the reordered English is far more monotonic with respect to the Hindi.

6 Analysis of errors and future directions

In this section, we analyze some of the sources of errors in reordering rules learned via our model, to better understand directions for further improvement.

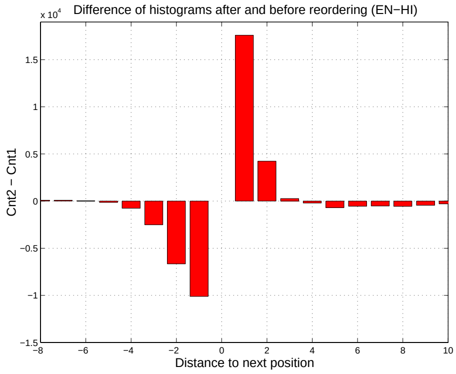


Figure 3: Difference-histogram of word order distortions for English→Hindi.

6.1 Model weakness

In our initial experiments, we noticed that for the most frequent reordering rules in English→Hindi (e.g that IN NP or NP PP flips in Hindi) the probability of a reordering was roughly 65%. This was concerning since it meant that on 35% of the data we would be making wrong reordering decisions by choosing the most likely reordering. To get a better feel for whether we needed a stronger model (e.g by lexicalization or by looking at larger context in the tree rather than just the children), we analyzed some of the cases in our training data where (IN,NP), (NP, PP) pairs were left unaltered in Hindi. In doing that analysis, we noticed examples involving negatives that our model does not currently handle. The first issue was mentioned in Section 4, where the assumption that we can achieve the right word order by reordering constituent phrases, is incorrect. The second issue is illustrated by the following sentences: *I have some/no books*, which have similar parse structures, the only difference being the determiner *some* vs the determiner *no*. In Hindi, the order of the fragments *some books* and the fragment *no books* are different (in the first case the words stay in order, in the second the flip). Handling this example would need our model to be lexicalized. These issue of negatives requiring special handling also came up in our analysis of German (Section 6.4). Other than the negatives (which require a lexicalized model), the major reason for the lack of sharpness of the reordering rule probability was alignment errors and parser issues. We look at these topics next.

Aligner	Number of Sentences	fMeasure	BLEU score
HMM	250k	62.4	21.7
MaxEnt	250k	76.6	21.4
Manual	5k	-	21.3

Table 2: Using different alignments

6.2 Alignment accuracy

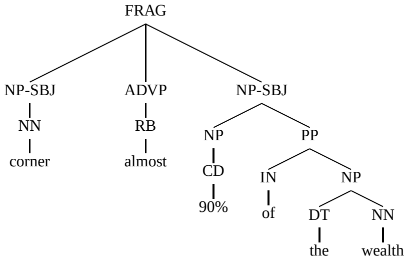
Since we rely on automatically generated alignments to learn the rules, low accuracy of the alignments could impact the quality of the rules learned. This is especially a concern for English→Hindi since the quality of HMM alignments are fairly low. To quantify this effect, we learn reordering rules using three sets of alignments: HMM alignments, alignments from a supervised MaxEnt aligner (Ittycheriah and Roukos, 2005), and hand alignments. Table 2 summarizes our results using aligners with differing alignment qualities for our English→Hindi task and shows that quality of alignments in learning the rules is not the driving factor in affecting rule quality.

6.3 Parser accuracy

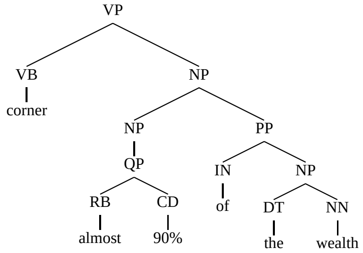
Accuracy of the parser in the source language is a key requirement for our reordering method, because we choose the single best reordering based on the most likely parse of the source sentence. This would especially be an issue in translating *from* languages other than English, where the parser would not be of quality comparable to the English parser.

In examining some of the errors in reordering we did observe a fair fraction attributable to issues in parsing, as seen in the example sentence: *The rich of this country , corner almost 90% of the wealth .*

The second half of the sentence is parsed by the Berkeley parser (Petrov et al., 2006) as:



and by IBM's maximum entropy parser parser (Ratnaparkhi, 1999) as:



With the first parse, we get the right Hindi order for the second part of the sentence which is: *the wealth of almost 90% corner* . To investigate the effect of choice of parser we compared using the Berkeley parser and the IBM parser for reordering, and we found the BLEU score essentially unchanged: 21.6 for the Berkeley parser and 21.7 for the IBM parser. A potential source of improvements might be to use alternative parses (via different parsers or n-best parses) to generate n-best reorderings both in training and at test.

6.4 Remarks on German reordering

Despite a common heritage, German word order is distinct from English, particularly regarding verb placement. This difference can be dramatic, if an auxiliary (e.g. modal) verb is used in conjunction with a full verb, or the sentence contains a subordinate clause. In addition to our experiments with automatically learned rules, a small set of handcrafted reordering rules was created and evaluated. Our preliminary results indicate that the latter rules tend to outperform the automatically derived ones by 0.5-1.0 BLEU points on average. These rules are summarized as follows:

1. In a VP immediately following an NP, move the negation particle to main verb.
2. Move a verb group away from a modal verb; to the end the of a VP. Negation also moves along with verb.
3. Move verb group to end of an embedded/relative clause.
4. In a VP following a subject, move negation to the end of VP (handling residual cases)

The above hand written rules show several weaknesses of our automatically learned rules for reordering. Since our model is not lexicalized, negations are not handled properly as they are tagged RB (along with other adverbs). Another limitation apparent from the first rule above (the movement of verbs in a verb phrase depends on the previous phrase being a noun phrase) is that the automatic reordering rule for a node's children depends only on the children of that node and not a larger context. For instance, a full verb following a modal verb is typically parsed as a VP child node of the modal VP node, hence the automatic rule, as currently considered, will not take the modal verb (being a sibling of the full-verb VP node) into account. We are currently investigating extensions of the automatic rule extraction alorithm to address these shortcomings.

6.5 Future directions

Based on our analysis of the errors and on the hand designed German rules we would like to extend our model with more general feature functions in Equation 1 by allowing features: that are dependent on the constituent words (or headwords), that examine a large context than just a nodes children (see the first German rule above) and that fire for all permutations when the constituent X is moved to the end (or start). This would allow us to generalize more easily to learn rules of the type "move X to the end of the phrase". Another direction that we feel should be explored, is the use of multiple parses to obtain multiple reorderings and combine these at a later stage.

7 Conclusions

In this paper we presented a simple method to automatically derive rules for reordering source sentences to make it look more like target language sentences. Experiments (on internal and public test sets) indicate performance gains for English→French, English→Spanish, and English→Hindi. For English→German we did not see improvements with automatically learned rules while a few hand designed rules did give improvements, which motivated a few directions to explore.

References

[Al-Onaizan and Papineni2006] Al-Onaizan, Yaser and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of ACL*.

[Badr et al.2009] Badr, Ibrahim, Rabih Zbib, and James Glass. 2009. Syntactic phrase reordering for english-to-arabic statistical machine translation. In *Proceedings of EACL*.

[Chiang2005] Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

[Collins et al.2005] Collins, Michael, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

[Habash2007] Habash, Nizar. 2007. Syntactic preprocessing for statistical machine translation. In *MT Summit*.

[Ittycheriah and Roukos2005] Ittycheriah, Abraham and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of HLT/EMNLP*.

[Ittycheriah and Roukos2007] Ittycheriah, Abraham and Salim Roukos. 2007. Direct translation model 2. In *Proceedings of HLT-NAACL*, pages 57–64.

[Koehn et al.2003] Koehn, Philipp, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*.

[Niessen and Ney2001] Niessen, Sonja and Hermann Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proc. MT Summit VIII*.

[Petrov et al.2006] Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL*.

[Popovic et al.2009] Popovic, Maja, David Vilar, Daniel Stein, Evgeny Matusov, and Hermann Ney. 2009. The RWTH machine translation system for WMT 2009. In *Proceedings of WMT 2009*.

[Ramanathan et al.2008] Ramanathan, A., P. Bhattacharyya, J. Hegde, R. M. Shah, and M. Sasikumar. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *Proceedings of International Joint Conference on Natural Language Processing*.

[Ratnaparkhi1999] Ratnaparkhi, Adwait. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34(1-3).

[Tillman2004] Tillman, Christoph. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*.

[Wang et al.2007] Wang, Chao, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP-CoNLL*.

[Xia and McCord2004] Xia, Fei and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling*.

[Xu et al.2009] Xu, Peng, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for Subject-Object-Verb languages. In *Proceedings of NAACL-HLT*.

[Yamada and Knight2001] Yamada, Kenji and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL*.

[Zhang et al.2007] Zhang, Yuqi, Richard Zens, and Hermann Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *NAACL-HLT AMTA Workshop on Syntax and Structure in Statistical Translation*.

[Zollmann and Venugopal2006] Zollmann, Andreas and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*.

[Zollmann et al.2008] Zollmann, Andreas, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of COLING*.

[Zwarts and Dras2007] Zwarts, Simon and Mark Dras. 2007. Syntax-based word reordering in phrase-based statistical machine translation: why does it work? In *Proc. MT Summit*.