A Comparison of Syntactic Reordering Methods for English-German Machine Translation

Jiří Navrátil Karthik Visweswariah Ananthakrishnan Ramanathan
IBM Research
jiri@us.ibm.com, v-karthik@in.ibm.com, aramana2@in.ibm.com

ABSTRACT
We describe two methods for syntactic source reordering developed for English-German SMT. Both methods learn from bilingual data accompanied by automatic word alignments to reorder the source such that it resembles that of the target. While the first method is an extension of a parse-based algorithm and accommodates contextual triggers in the parse, the second method uses a linear feature-based cost model along with a Traveling Salesman Problem (TSP) solver to perform the reordering. Our results indicate that both methods lead to improvements in BLEU scores in both directions, English→German and German→English. Significant gains in human translation quality assessment are observed for German→English, however, no significant changes are observed in the human assessment for English→German.

KEYWORDS: Parse-based reordering, TSP, English-German SMT.

Proceedings of COLING 2012: Technical Papers, pages 2043–2058,
COLING 2012, Mumbai, December 2012.

1 Introduction

Language-specific word order differences have presented a long-standing challenge in the development of statistical machine translation (SMT) systems. Most mainstream SMT approaches do incorporate word order information, either implicitly (e.g., as part of word phrases), or explicitly, e.g., by means of lexicalized distortion models. However, challenges remain, particularly in modeling long-range word movement. Furthermore, certain language pairs, more than others, exhibit particularly demanding reordering patterns. One such pair is English-German - the focus of this work. With its long-range verb movement and its use of separated verb prefixes, German and English word ordering can induce movements spanning an entire sentence. Section 1.1 describes some of these phenomena more in detail.

Recent improvements in phrase-based SMT algorithms have included source word reordering so as to resemble its target word order. A reordering typically uses rules obtained either manually, e.g., (Niessen and Ney, 2001; Collins et al., 2005), or derived from data by automatic means, e.g., (Xia and McCord, 2004; Rottmann and Vogel, 2007; Zhang et al., 2007; Crego and Habash, 2008; Niehues and Kolss, 2009).

In this paper we further develop two methods first introduced in (Visweswariah et al., 2010, 2011) with the aim to address specific issues arising in English↔German SMT. In particular, we extend the parse-based reordering introduced for a variety of language pairs in (Visweswariah et al., 2010) by creating rules capable of capturing essential contextual triggers in the parse tree hierarchy. For the second method we describe refinements of the feature-based reordering (Visweswariah et al., 2011), which corresponds to solving a Traveling Salesman Problem (TSP). While the method described in (Visweswariah et al., 2011) carries over and performs well for German→English, for English→German we propose a method to integrate the English parse into the reordering model focusing on the issues in English→German reordering. We compare the performance of both methods against various baselines and across multiple evaluation domains for both English→German and German→English directions. Our results indicate that both methods achieve significant improvements in translation quality, as measured by automatic metrics, and some improvements when judged by human reviewers.

Following the description of the methods and refinements in Section 2 we discuss related work in Section 3. Experimental evaluation, results, and discussion are presented in Section 4.

1.1 German Word Order

Despite their common heritage, German and English word order can differ rather dramatically. Most frequently such differences relate to verbs, particularly to verb groups, including auxiliary, main verbs, and their participles. Verb movements tend to span large portions of the sentence, thus presenting a considerable challenge to basic reordering models in standard SMT systems.

Word ordering can often be identical to that in English, as shown in an example in Figure 1. At the same time, an addition of a modal verb in the same example sentence triggers a movement of the main verb to the end of the clause, as shown in Figure 2. Similarly distant verb movements also occur in subordinate clauses. Besides verb movement, word order may vary in other parts of the sentence, including negation, adverbial phrases, etc.

2 Data-Driven Syntactic Reordering

We use data-driven syntactic reordering to mitigate some of the differences in word order mentioned above. Specifically, we investigate two methods of syntax-based reordering: (1) an

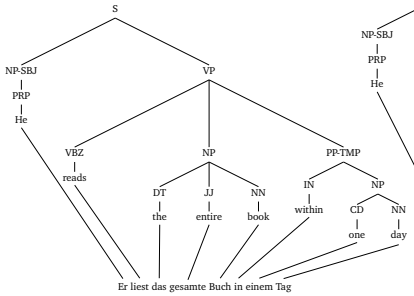

Figure 1: A single verb sentence with monotone ordering pattern

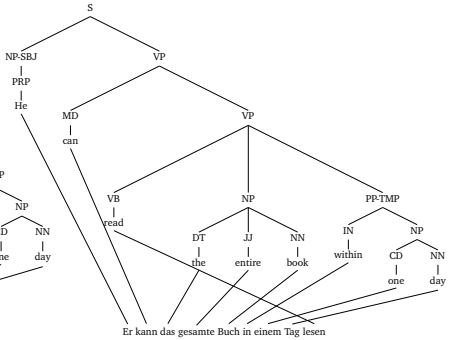

Figure 2: A modal verb triggering a reordering pattern

extension of the parse-based word reordering introduced in (Visweswariah et al., 2010), and (2) feature-based reordering model described in (Visweswariah et al., 2011). Both methods use a bilingual corpus, automatic word alignments, and some source syntax information (parse tree in the first case, POS in the second) to train a reordering model. In both cases the objective is to minimize a measure of an overall distortion observed in the word alignments. We apply the reordering to source sentences in both the training and the test of the system as part of preprocessing.

2.1 Parse-Based Reordering

Visweswariah et al. (Visweswariah et al., 2010) described reordering of each sentence using a set of rules applied to the parse tree of the source sentence. The goal of these rules is to make the source word order resemble the expected target order. Given automatic word alignments and source parses, the rules are inferred from a training corpus.

The parse-based model is probabilistic in nature. Given a source constituency tree, S, it aims to assign the highest probability $P(T|S)$, among all possible reordered trees T, to a tree with its constituents reordered so as to reflect the expected target word order. The model restricts itself to trees that can be obtained only by permuting child nodes of any of its non-terminal constituents, and it makes a simplifying assumption that the children of a node are permuted independently of any other node in the tree. Thus, the overall probability simplifies to a product of constituent-level permutations:

$$P(T|S) = \prod_{n \in I(S)} P(\pi(d_n)|d_n) \quad (1)$$

where n denotes a node, $I(S)$ the set of all non-terminal (interior) nodes of the tree, d_n are the children of the node n, and π is the permutation function. Given the word alignment information, each node in d_n, is permuted based on the individual average *target* position. The average target position is calculated using the alignments $a(w)$ as follows:

$$tpos(n) = \frac{1}{|D(n)|} \sum_{w \in D(n)} a(w) \quad (2)$$

whereby $D(n)$ denotes the set of all descendant leaves (words) of a node n, and $a(w)$ is a function returning the position index of a source word w in the target translation. Each such permutation is recorded over the entire training corpus $\mathscr{C}$ and its permutation probabilities are then estimated to maximize the likelihood $\prod_{S \in \mathscr{C}} P(T|S)$:

$$P(\pi(d_n)|d_n) = \frac{count(\pi(d_n))}{count(d_n)} \tag{3}$$

As already pointed out in (Visweswariah et al., 2010), the above model bears several weaknesses. Besides generic issues regarding parser and alignments accuracy, the assumed independence of a permutation from other nodes (context) can be harmful. This turns out to be particularly true for German. As in the example in Figure 2, a full verb following a modal verb is typically parsed as a VP child node of a VP modal node. In this example the main verb should be placed at the end of the clause due to the presence of the modal, while in the absence thereof typically no reordering occurs (Figure 1). The lack of such distinction in the simplified $\pi(d_n)$ renders extracted rules indiscriminative. The authors reported improvements in translation quality for several languages (French, Spanish, Hindi) with the exception of German where no gains over an unreordered baseline were observed highlighting this weakness as a probable cause.

We propose extending the permutation probability function to include contextual information from S. In this extension each permutation probability is now a function of the permutation node set d_n and a context node subset $\phi_n \in S$:

$$P(T|S) = \prod_{n \in I(S)} P(\pi(d_n)|d_n, \phi_n) \tag{4}$$

Although the subset ϕ can be arbitrary we choose ϕ_n to constitute subtrees of S that are related to d_n via its parental lineage as well as its siblings. For instance, in the Figure 2 this relation would be the subtree including the non-terminals (top-down): S, VP, MD, VP, VB, NP, and PP-TMP. More specifically, in our experiments we investigated ϕ to include the parent of d_n, i.e n, siblings of n (both to its left and right), as well as k levels of grand parents (with k varying between 1 and 5). We relaxed the subtree matching by assuming "wildcards" between any of the siblings. This increases the rule recall for parses with minor variations from the observed patterns (in other words, we apply a *tree grep* instead of a strict tree match).

The probability of a permutation is estimated as in Eq. (3) with observation counts now collected with respect to the context ϕ of each observed permutation. A permutation rule pruning is performed based on an absolute observation count (in our case any rule with less than 20 observations) as well as a significance threshold (count must be 20% higher relative to the next competing permutation). We only retain the best permutation (in the maximum likelihood sense) for a unique pattern (ϕ, d_n), thus, a final rule now consists of a left- and right-hand side: $(\phi, d_n) \rightarrow \pi(\phi, d_n)$. Reusing the example in Figure 2 the best applicable rule actually extracted from the training corpus is:

```
[S [VP MD [VP VB NP PP-TMP VP] VP] S] --> 1 2 0
```

which enacts a move of the verb "read" (VB) to the end of the sentence and resolves the crossing alignment shown in Figure 2.

Given a parse tree all matching rules are applied recursively (matched always against the original parse) to create a new reordered tree.

2.2 TSP-Based Reordering

In (Visweswariah et al., 2011) a reordering model was proposed that does not require a parser, and learns to reorder words based on reference reorderings derived from hand alignments. The model assigns pairwise costs $c(m,n)$ for word w_m immediately preceding word w_n. The cost of a reordering permutation π for a sentence $\mathbf{w}$ is the sum of these pairwise costs:

$$C(\pi|\mathbf{w}) = \sum_i c(\pi_i, \pi_{i-1}).$$

A sentence $\mathbf{w}$ is reordered by choosing the permutation that minimizes the cost $C(\pi|\mathbf{w})$. The minimization problem is an asymmetric TSP problem, which is converted to a symmetric TSP problem with double the number of states (Visweswariah et al., 2011) and then is solved using the Lin-Kernighan heuristic. The costs $c(m,n)$ are parametrized as a linear model

$$c(m,n) = \theta^T \Phi(\mathbf{w}, m, n)$$

where Φ is a vector of binary feature vectors as described in (Visweswariah et al., 2011). θ is a weight vector learned from reorderings derived from hand alignments using the MIRA update algorithm.

Although the base reordering model does not require a parser, for English→German reordering we experimented with using the TSP model that works on top of an input parse tree. Since English-German word order differences mainly relate to verb movements, we transform the parse tree so that any internal node which has a verb as a descendant is expanded down to its children. If a node does not have a verb descendant we represent the entire subtree by the constituent label of the node. Thus the parse tree in Figure 2 gets transformed to the following being passed as input to the reordering model: *NP-SBJ can read NP PP-TMP*.

2.2.1 Features

We adopted the same base features as described in (Visweswariah et al., 2011) and also experimented with additional features specific to word order differences in English-German. The basic features $\Phi(\mathbf{w}, m, n)$ are binary features that fire based on the identity of the words w_m and w_n and the POS tags of these words. Additionally there are features that examine the identities of words and POS tags one word to the left and right of the positions m and n (see (Visweswariah et al., 2011) for details of feature templates used).

Specific to the English→German direction reordering we use transformed parses (see above) as input and use the constituent labels instead of the words. Additionally, to handle the fact that verbs move to the end of the clause in subordinate clauses, we mark if a verb is a descendant of an SBAR node in the parse tree on the POS tag of the word.

Specific to the German→English direction we added an extra feature template that fires if word w_m and w_n are verbs and there is no verb between positions m and n.

2.3 Manual Rules

We do not have a suitable German parser. Therefore, as an alternative method to the TSP-based method (in the German→English direction only) we consider reordering using hand-crafted rules. The rules use German-side POS sequence as features and focus on verbs:

- Move full verb behind its closest auxiliary to the left
- Move negation behind closest auxiliary to the left
- In a subordinate clause, move auxiliary and full verbs behind an estimated subject
- Void any of the above rules that would cross a barrier token (e.g., a comma)

The rules are applied to a sentence repeatedly until there is no word movement. As will be shown, this small set of heuristic rules improves the output quality. We will compare this method with an unreordered baseline as well with a system pre-ordered via the TSP method.

2.4 Sentence Pre-Selection

The quality of reordering rules, i.e. their precision, depends on several factors: (1) the alignment, tagging, and parsing accuracy, and (2) the bilingual training data quality. The latter is considered here from a point of view of suitability for extraction of meaningful reordering patterns. Bilingual corpora typically contain sentence pairs with a varying degree of mutual parallelism. While a sentence pair may be considered an acceptable translation, the individual sides may differ significantly in terms of their syntactic and clausal structure. Non-literal translations may introduce considerable noise into the rule extraction. Consider this example taken from the Europarl (Koehn, 2005) corpus:

- English: *A ban would have less serious repercussions in this respect if it applied throughout the EU.*
- German: *Im Falle eines europäischen Verbots von Nachtflügen sind diese Folgen weniger stark ausgeprägt.*
- Gloss: *In case of a European ban of night flights are these consequences less strongly contrastful.*

Obviously, issues will arise for rule extraction given that these two sentences are composed quite differently. In particular the verb "applied" remains unaligned in this case thus introducing an error to the estimated target position for the corresponding set of node descendants, as described in Section 2.1.

To mitigate the impact of noise in bilingual data we propose an automatic sentence preselection algorithm. The goal here is to assign a quality score to each sentence pair given its automatic alignment and to select a subset of higher quality (i.e. literally translated) sentences for the rule extraction. We borrow the confidence measure proposed originally for block extraction in (Huang, 2009) and adopt it as an indicator for sentence parallelism. A description of the algorithm follows.

Let (V,W) denote a bilingual sentence pair where $V = \{v_1,...,v_I\}$ and $W = \{w_1,...,w_J\}$ is the source and the target sentence, respectively. The sentences are word-aligned via a set of alignment links $A = \{a_{ij}\}$. The alignment A represents the best way to relate the content of V to W and is obtained by automatic means (e.g. Giza or Maxent methods). The quality measure involves calculating two quantities:

$$P(W|V,A) = \epsilon \prod_{j=1}^{J} \sum_{\forall i:a_{ji} \in A} p(w_j|v_i) \tag{5}$$

which is the lexical probability of the target sentence given the source words and their alignment, and

$$P(W|V) = \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=1}^{I} p(w_j|v_i) \tag{6}$$

the lexical probability of W given V, independent of A. The terms $p(w_j|v_i)$ are Model 1 word translation probabilities as estimated during training. The constants ϵ and $\frac{1}{(I+1)^J}$ are due to the simplifying assumptions of Model 1 as described in (Brown et al., 1993), namely sentence lengths, and alignments being all equally probable. These constants will later cancel out. Similar to the above, transposed quantities $P(V|W)$ and $P(V|W,A')$ are also obtained using Model 1 probabilities in in the reverse direction along with their corresponding alignment A'. Noting that

$$P(W,A|V) = P(A)P(W|V,A) = \frac{1}{(I+1)^J} P(W|V,A) \tag{7}$$

we now define the bilingual sentence confidence score as follows:

$$F(V,W) = \frac{1}{J} \log \frac{P(A,W|V)}{P(W|V)} + \frac{1}{I} \log \frac{P(A',V|W)}{P(V|W)} \tag{8}$$

Since $F(V,W)$ is an average of alignment posteriors $P(A|V,W)$ and $P(A'|V,W)$, we expect it to produce large values for translations with sharp posterior distributions which, in turn, can be viewed as relatively literal and complete. For sentence pairs with a lower degree of parallelism, for incomplete, or perhaps incorrect translations $F(V,W)$ should tend to be small.

The overall sentence pre-selection step consists of calculating $F(V,W)$ values for all sentence pairs in the reordering training corpus followed by thresholding and selecting a certain proportion of best scoring sentences for the rule training. In our case the proportion was set to be around 50%.

In our preliminary experiments we observed translation quality gains of up to 1 point BLEU due to this selection method.

3 Related Work

A significant quantity of work in syntax based reordering has accumulated in the machine translation literature. Relevant to our topic, there have been studies investigating sources of improvements (Zwarts and Dras, 2007) due to syntax-based reordering. Initial efforts (Niessen and Ney, 2001) were made at improving German-English translation using handwritten rules by handling two phenomena: question inversion and detachable verb prefixes in German. In (Collins et al., 2005; Carl, 2007) rules are developed for translation from German to English based on source POS and manual rules covering a variety of patterns including verb movement.

There have been studies that try to learn rules from the data, among others (Ringger et al., 2004; Rottmann and Vogel, 2007; Zhang et al., 2007; Crego and Habash, 2008; Niehues and Kolss, 2009; Khalilov and Sima'an, 2011). Work by Rottmann and Vogel utilizes sequences of source POS and the corresponding alignments to automatically extract reordering rules along with their left and right context. They then reorder a source sentence using applicable rules thus obtaining a word lattice for decoding. Building on this work, Niehues

and Kolss explore an extension of such rules introducing a variable length gaps to better generalize long range reordering patterns. Both studies report significant gains in BLEU scores for both English→German and German→English translation.

The work in (Rottmann and Vogel, 2007; Niehues and Kolss, 2009) relates to both our parse-based as well as feature-bases reordering methods in that they take the context of the reordered tokens into account - a key to English-German word order. Our method, like (Niehues and Kolss, 2009), allows for gaps in the matching patterns. In contrast to the above mentioned work, the parse-based method operates on the entire source parse, including coarse levels where reordering can induce long range word movements. We use contextual information from the parse tree thus modeling left and right context which, in contrast to (Niehues and Kolss, 2009), is hierarchical. Our second, feature-based method operates on word level and takes into account arbitrary syntactic and lexical source features. The method utilizes machine learning to find appropriate model to capture most beneficial source word order, given a set of automatic alignments.

Recently, reordering models that can learn to reorder source sentences to make them match the target language order without requiring a parser have been proposed. (Tromble and Eisner, 2009) propose a model based on the Linear Ordering Problem, where for each pair of words in the sentence the cost that one of them occurs somewhere before the other is modeled. (DeNero and Uszkoreit, 2011) take a two pronged approach where they first learn to parse the sentence and then learn a reordering model on the resultant parses. (Visweswariah et al., 2011) proposed a model based on the Travelling Salesman Problem to learn to reorder sentences where costs of a word immediately preceding another word in the sentence are learned. This model was shown to be better than (Tromble and Eisner, 2009) in terms of reordering performance. In this work we extend and apply the (Visweswariah et al., 2011) to German-English reordering.

While the focus of our paper is on pre-ordering techniques, there has been considerable work on handling the reordering problem as part of the decoding process (Chiang, 2007; Yamada and Knight, 2002; Galley et al., 2006; Liu et al., 2006; Zollmann and Venugopal, 2006). These approaches are computationally expensive compared with phrase based systems due to the inclusion of bilingual parsing in the decoding process.

4 Experimental Evaluation

4.1 Parse-Based Rules

The rule extraction was performed using approximately 2.6 million sentence pairs obtained through the sentence pre-selection process as described in Section 2.4, operating at a 50% sentence rejection rate. The sentence pairs were aligned by a maximum entropy aligner (Ittycheriah and Roukos, 2005) trained using a subset of the Europarl corpus, and selected computer manual corpora). A maximum entropy parser (Ratnaparkhi, 1999) was used to generate the parse trees for the English side in the English→German direction, and a maximum entropy tagger to generate POS (using the Stuttgart-Tübingen tag set (Schiller et al., 1995)) for the German side in the German→English direction. With a significance threshold of 1.2 and minimum count threshold of 20, we obtain about 3200 rules in the final reordering model.

4.2 TSP-Based Model

For both English→German and German→English reordering we train the TSP-based models on a set of hand alignments (roughly 30K sentences) and a subset of machine alignments selected using the sentence pre-selection method described in Section 2.4 (roughly 300k sentences). For English→German we use the transformed English parse as described in Section 2.2 instead of the original word sequence.

During development stage we used the monolingual BLEU score (mBLEU) to compare the reordered output to a reference reordering derived from hand alignments on a test set of 400 sentences from the news domain. For German→English we obtain larger improvements, going from a mBLEU score of 61.5 for unreordered German to 72.5 for German reordered using the TSP-based model. For English→German we get a relatively modest improvement; going from 64.2 for unreordered English to 67.6 for reordered English.

4.3 SMT Model

The phrase-based systems were trained in both directions using same amounts of training data. In a first stage, smaller models were created adhering to the WMT2010 constrained training condition (WMT Website, 2010). In a second stage, about 16M sentence pairs spanning a variety of publicly available (e.g. Europarl) as well as internal corpora (IT and news domains) served the training of unconstrained systems. The phrase pairs were extracted based on a union of HMM and maxent alignments with corpus-selective count pruning. The lexicalized distortion model (Al-Onaizan and Papineni, 2006) was used with a window width of up to 5 and a maximum number of 2 skipped (not covered) words during decoding. 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, except in the constrained mode where a single 5-gram language model was trained on the WMT2010 training data (WMT Website, 2010).

4.4 Evaluation Sets

The methods were evaluated in contrastive experiments utilizing the following (single-reference) test sets:

- *News*: 166 sentences (8700 words) 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 (WMT) 2009 (WMT Website, 2009). Results of others on this set can be found, for example, in (Popovic et al., 2009).
- *WMT10:* 2034 sentences from news domain used as the eval set in the WMT 2010. Results of others on this test set can be found in (WMT Website, 2010).

4.5 Evaluation Metrics

The translation quality in our experiments is evaluated using BLEU (Papineni et al., 2002), as well as using human assessment. The latter is carried out by a judge rating the quality of three translations for each source sentence. The defined assessment levels correspond to

following judgements: 0=Exceptionally poor, 1=Poor (difficult to understand the meaning), 2=Not good enough (errors in grammar, vocabulary, and style make understanding difficult), 3=Good enough (there are errors, however one can understand the meaning with a reasonable confidence), 4=Very good (there may be minor errors, but one can understand the meaning with high confidence), 5=Excellent (the information is presented clearly and with appropriate grammar, vocabulary, and style.) When performing the assessment, the reviewer is presented the source and the competing (blind) translations with their order randomized. The ratings are averaged within the test resulting in a single human score per system. We employed a single judge, who was not involved in the related technical work, and who is proficient in both English and German.

4.6 Results

The cased BLEU scores for English→German are shown in Table 1, and for German→English in Table 2. Overall, we make the following observations: (1) the scores behave consistently between the constrained and unconstrained training, (2) all reordering methods in both directions improve the BLEU scores over their unreordered baselines, (3) the parse- and TSP-based methods seem to fare comparably across the testsets with BLEU differences less than 0.3 points, (4) while the manual rules for German→English do improve the translation quality, the automatic TSP reordering outperforms the manual rules by more than 1 BLEU point leading to overall improvements of about 2 points in the unconstrained condition. Tables 3 and 4 show

English→German Test	Baseline (no RO)	Parse-Based RO Cased BLEU	TSP RO	Diff (TSP vs. Bsl.)
TechA	0.170	**0.184**	0.181	+0.019
TechB	0.190	**0.201**	0.199	+0.018
News	0.248	0.253	**0.255**	+0.023
Dev09	0.142	0.144	**0.146**	+0.018
WMT10	0.150	0.156	**0.158**	+0.012
Dev09 (Constrained)	0.137	0.138	**0.140**	+0.003
WMT10 (Constrained)	0.148	0.154	**0.155**	+0.007

Table 1: BLEU scores for English→German phrase-based machine translation with and without reordering (RO). Last column shows score differences between TSP and Baseline.

German→English Test	Baseline (no RO)	Manual RO Cased BLEU	TSP RO	Diff (TSP vs. Bsl.)
TechA	0.297	0.303	**0.316**	+0.019
TechB	0.294	0.298	**0.312**	+0.018
News	0.250	0.262	**0.273**	+0.023
Dev09	0.184	0.194	**0.202**	+0.018
WMT10	0.194	0.197	**0.206**	+0.012
Dev09 (Constrained)	0.180	0.186	**0.190**	+0.010
WMT10 (Constrained)	0.186	0.191	**0.196**	+0.010

Table 2: BLEU scores for English→German phrase-based machine translation with and without reordering (RO). Last column shows score differences between TSP and Baseline.

the human assessments for the two directions, English→German and German→English, respectively. In both cases 50 randomly sampled sentences from the *TechB* testset were used.

English→German Assessment	Average	Counts per rating grade					
		0	1	2	3	4	5
Bsl (no RO)	2.4	0	15	13	12	8	2
Parse RO	2.4	0	14	17	9	5	5
TSP RO	2.4	0	13	16	12	8	1

Table 3: Human assessment results of English→German MT output for systems with and without reordering (RO) based on 50 randomly sampled sentences from the TechB testset.

German→English Assessment	Average	Counts per rating grade					
		0	1	2	3	4	5
Bsl (no RO)	2.3	1	12	22	8	1	6
Manual RO	3.0	0	11	11	10	5	13
TSP RO	3.4	0	4	10	10	14	12

Table 4: Human assessment results of German→English MT output for systems with and without reordering (RO) based on 50 randomly sampled sentences from the TechB testset.

4.6.1 Discussion

In the German→English direction the results let us conclude that the TSP method leads to improvements in the translation quality. A post-review analysis of the human assessment indicates that most gains can be linked to either word order or better phrase match. The latter is a consequence of more monotone alignments and thus an improved phrase extraction. We have investigated the latter point by recording the count of phrase pairs extracted from the same training corpus (WMT2010) with and without reordering. While in the unreordered system, the extraction resulted in 30.0M phrase pairs, in the reordered system this number increased to 36.3M phrase pairs - corresponding to about 20% increase in extraction efficiency. The reordered phrases also tend to be slightly longer, on average. An example from the human assessment illustrates the nature of the improvement:

- Source: *Die Vergabekriterien sind für alle Kategorien klar definiert, wie auch ihre Beurteilung.*
- Baseline: *The award for all categories clearly defined, and their assessment.* (Rating: 2)
- TSP-RO: *The lending criteria are clearly defined for all categories, as well as their assessment.* (Rating: 5)

The baseline output seems to suffer not only from incorrect word order but also from dropping essential words (“criteria” and “are”), both of which are rectified in the reordered output.

We observe a somewhat different picture for the English→German SMT with results mixed between the automatic and the human metrics. The improvements in BLEU fail to produce noticable difference for the human judge. This could be caused by several factors. First, the overall translation quality is low (2.4), therefore word order improvements may not help in presence of dominant errors. An example from the assessment sheet illustrates this issue:

- Source: *The fix is to relax these attributes as optional.*
- Baseline: *Der Fix ist sich zu entspannen diese Attribute optional sein soll.* (Rating: 1)
- TSP-RO: *Der Fix ist diese Attribute als optionale zu lockern.* (Rating: 1)

Clearly, the word choice of the baseline ("sich zu entspannen" - "to rest" or "to unwind") is poor as is its placement. The reordered system has undoubtedly better word order and arguably better word choice, however, the judge still considered the overall translation as "poor."

Second, correctly placing the verbs in the English→German direction may be a harder problem. Elaborating on this conjecture, we may first consider the relevant word patterns in the two language directions: while going into English the reordering moves (mostly) verbs together (e.g., with the auxiliary verb being an anchor), it moves them far apart when going into German. Given that these verbs (e.g., auxiliary and their main verb) tend to form a single semantic unit of the sentence, they will thus be captured by the phrase extraction as a unit. Conversely, when splitting the verb groups and moving them apart, the positive effect of a more monotone word alignment may be out-weighed by the fact that these verbs will not be captured in one unit. Moreover, we hypothesize an accurate verb placement to be a harder problem going into German than into English. In the latter case an unambiguous placement pattern is typical, e.g., a main verb being placed immediately behind its auxiliary. On the other hand, the placement of a main verb to resemble German depends heavily on the parser accuracy (clause boundaries), and it can sometimes be ambiguous (e.g., verbs may or may not move past embedded clauses). One way to mitigate this problem might be combining reordered and unreordered phrase pairs in the training and then rely on word reordering lattices being used inside the decoder to choose the best matching alternative (as proposed in (Rottmann and Vogel, 2007)), which, however, increases the search complexity of the SMT system.

4.7 Future Directions

We described two methods for syntactic source reordering developed for English-German SMT. Our results indicate that both methods, applied as a pre-processing step, lead to improvements, as measure by BLEU. Significant gains in human assessment of the translation quality are observed for German→English, however, no observable changes were achieved for English→German according to the judge. The latter finding poses several questions: (1) why does same reordering method help significantly more in one direction than vice versa? and (2) how can the discrepancy between our automated metric BLEU and the human assessment results be mitigated? Regarding (1) we hypothesized the reordering problem to be harder when going into German due to a higher word "scatter" as well as sensitivity to parser and tagger accuracy. We plan to further investigate this problem and to refine both methods to to increase their robustness. Independently, the challenge of the overall low quality of the English→German remains to be addressed. Improving the baseline could help the reordering gains in BLEU to show up also in the human assessment, as a better word order will arguably tend to make more impact going from a rating of 3 to a rating of 4, or 5, than, say, from 1 to 2. English→German translation is a challenging language pair and we believe significant improvements will be achieved only by addressing several problems beyond just word ordering. These include generating correct inflections, grammatical agreement, and preventing content word loss. Our results also confirm human assessment to be a necessary component of the overall system evaluation helping to avoid over-optimistic conclusions.

References

Al-Onaizan, Y. and Papineni, K. (2006). Distortion models for statistical machine translation. In *Proceedings of ACL*.

Brown, P. F., Pietra, V. J., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Carl, M. (2007). Metis-ii. the german to english mt system. In *In Proceedings of the 11th Machine Translation Summit*, pages 65–72.

Chiang, D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

Crego, J. and Habash, N. (2008). Using shallow syntax information to improve word alignment and reordering for smt. *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61.

DeNero, J. and Uszkoreit, J. (2011). Inducing sentence structure from parallel corpora for reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 193–203, Stroudsburg, PA, USA. Association for Computational Linguistics.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of ACL*.

Huang, F. (2009). Confidence measure for word alignment. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 932–940, Suntec, Singapore. Association for Computational Linguistics.

Ittycheriah, A. and Roukos, S. (2005). A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of HLT/EMNLP*.

Khalilov, M. and Sima'an, K. (2011). Context-sensitive syntactic source-reordering by statistical transduction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 38–46, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Liu, Y., Liu, Q., and Lin, S. (2006). Tree-to-String alignment template for statistical machine translation. In *Proceedings of ACL*.

Niehues, J. and Kolss, M. (2009). A pos-based model for long-range reorderings in smt. In *In Proc. of Fourth ACL Workshop on Statistical Machine Translation*.

Niessen, S. and Ney, H. (2001). Morpho-syntactic analysis for reordering in statistical machine translation. In *Proc. MT Summit VIII*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Popovic, M., Vilar, D., Stein, D., Matusov, E., and Ney, H. (2009). The RWTH machine translation system for WMT 2009. In *Proceedings of WMT 2009*.

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

Ringger, E., Gamon, M., Moore, R. C., Rojas, D., Smets, M., and Corston-Oliver, S. (2004). Linguistically informed statistical models of constituent structure for ordering in sentence realization. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rottmann, K. and Vogel, S. (2007). Word reordering in statistical machine translation with a pos-based distortion model. In *Proc. of the 11th Internat. Conf. on Theoretical and Methodological Issues in Machine Translation*, Sweden.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das tagging deutscher textcorpora mit stts. Technical report, IMS Stuttgart/Seminar f. Sprachwiss. Tübingen.

Tromble, R. and Eisner, J. (2009). Learning linear ordering problems for better translation. In *Proceedings of EMNLP*.

Visweswariah, K., Navratil, J., Sorensen, J., Chenthamarakshan, V., and Kambhatla, N. (2010). Syntax based reordering with automatically derived rules for improved statistical machine translation. In *COLING*, pages 1119–1127.

Visweswariah, K., Rajkumar, R., Gandhe, A., Ramanathan, A., and Navratil, J. (2011). A word reordering model for improved machine translation. In *EMNLP*, pages 486–496.

WMT Website (2009). http://statmt.org/wmt09/.

WMT Website (2010). http://statmt.org/wmt10/.

Xia, F. and McCord, M. (2004). Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling*.

Yamada, K. and Knight, K. (2002). A decoder for syntax-based statistical MT. In *Proceedings of ACL*.

Zhang, Y., Zens, R., and Ney, H. (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, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*.

Zwarts, S. and Dras, M. (2007). Syntax-based word reordering in phrase-based statistical machine translation: Why does it work? In *Proc. MT Summit*.