Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation

Teresa Herrmann, Jochen Weiner, Jan Niehues, Alex Waibel

Institute for Anthropomatics
Karlsruhe Institute of Technology
{teresa.herrmann,jan.niehues,alexander.waibel}@kit.edu, jochen.weiner@student.kit.edu

Abstract

We analyze the performance of source sentence reordering, a common reordering approach, using oracle experiments on German-English and English-German translation. First, we show that the potential of this approach is very promising. Compared to a monotone translation, the optimally reordered source sentence leads to improvements of up to 4.6 and 6.2 BLEU points, depending on the language. Furthermore, we perform a detailed evaluation of the different aspects of the approach. We analyze the impact of the restriction of the search space by reordering lattices and we can show that using more complex rule types for reordering results in better approximation of the optimally reordered source. However, a gap of about 3 to 3.8 BLEU points remains, presenting a promising perspective for research on extending the search space through better reordering rules. When evaluating the ranking of different reordering variants, the results reveal that the search for the best path in the lattice performs very well for German-English translation. For English-German translation there is potential for an improvement of up to 1.4 BLEU points through a better ranking of the different reordering possibilities in the reordering lattice.

1. Introduction

The reordering problem is commonly acknowledged to be one of the main difficulties in machine translation. One widely used approach is to perform reordering as a preprocessing step before translation. The idea is to synthesize a sentence in the source language that simulates the word order of the target language. Reordering the source text results either in a deterministically reordered sentence or multiple reordering variants are generated and stored in a lattice. Then monotone translation can be performed either on the reordered source sentence or the machine translation decoder searches for the best sequence of words in the reordering lattice.

We want to assess the benefits of this common approach of reordering the source before translation and investigate whether it really helps improve the translation quality. For one, we want to determine lower and upper bounds for the translation quality that can be reached by this approach and to identify potential of further development. Furthermore, we want to assess the performance of the reordering model on two levels: The restriction of the search space of possible reorderings and the ranking of different reordering variants.

We designed oracle experiments that address the following questions:

- How good is the translation of the optimally reordered source sentence?
- How beneficial is the restriction of the search space through reordering lattices for translation quality?
- How accurate is the search for the best path in the reordering lattice?

The paper is structured as follows: First, we present related work dealing with the reordering problem, mainly focussing on reordering as preprocessing and the judgement of reordering quality. In Section 3 we explain the reordering approaches applied in this work in detail. Then we describe the setup for the oracle experiments, which include an oracle reordering of the source sentence and the oracle path in the input lattices which is closest to the oracle reordering. We show the results of the experiments in Section 5 and then draw conclusions about future development of the reordering approach in the final section.

2. Related Work

In our work we investigate the benefits of a pre-reordering approach for machine translation by performing oracle experiments. We first present related work regarding reordering methods in machine translation and reference work on judging the quality of a given reordering. Then we mention work using oracles for the analysis of machine translation systems.

Word reordering has been addressed by many approaches in statistical systems. In a state-of-the-art phrase-based machine translation system, the decoder processes the source sentence left to right, but allows changes in the order of source words while the translation hypothesis is generated. Many phrase-based systems also include a lexicalized reordering model [1] which provides additional reordering information for phrase pairs. It stores statistics on the orientation of adjacent phrase pairs on the lexical level.

A very popular approach is to detach the reordering from the decoding procedure and to perform the reordering on the source sentence before translation. Such pre-reordering approaches use linguistic information about the source and or target language, such as parts-of-speech, dependency or constituency tree structure. They apply hand-crafted rules or automatically learn rules that change the order of the source sentence. Then monotone translation is performed.

In the first pre-reordering approach, reordering rules for English-French translation are automatically learned from source and target language dependency trees [2]. Since then many adopted this method. In the beginning manually crafted reordering rules based on syntactic or dependency parse trees or part-of-speech tags were designed for particular languages [3, 4, 5, 6]. Later data-driven methods followed, learning reordering rules automatically based on part-of-speech tags or syntactic chunks [7, 8, 9, 10]. Alternatively, word class information may be used to perform a translation of the original source sentence into a reordered source sentence [11]. More recent work includes reordering rules learned from source and target side syntax trees [12], automatically learned reordering rules from IBM1 alignments and source side dependency trees [13] and using a classifier to predict source-sentence reordering [14]. An approach presenting automatically learned reordering rules based on syntactic parse tree constituents [15] further combines the tree-based rules with two types of part-of-speech-based rules [7, 10]. This produces complementary reordering variants which result in an improved translation quality. While some of the presented approaches perform a deterministic reordering of the source sentence, others store reordering variants in a word lattice leaving the selection of the reordering path to the decoder.

Related work regarding reordering metrics and reordering quality includes the first description of reorderings as permutations [16]. Later, the use of permutation distance metrics to measure reordering quality [17] leveraged research into distance functions for ordered encodings. An approach to transform alignments into permutations [18] takes the particular characteristics of alignment functions into account.

Oracle experiments have shown to be a valuable method for analyzing different aspects of machine translation. While an oracle BLEU score may serve for identifying translation errors in the phrase table [19], another approach uses oracles for punctuation and segmentation prediction in speech translation [20]. Efficient methods for finding the best translation hypothesis in a decoding lattice have been proposed [21]. Furthermore, research on oracles regarding the reordering problem have been conducted [22, 23]. The first uses linear programming to compare the best achievable BLEU scores when using different reordering constraints [22]. The latter presents a reordering method for translations from English to Spanish, Dutch and Chinese where deterministic reordering decisions are conditioned on source tree features and compared to several oracles [23].

Rule Type	Example Rule	
Short	*VVIMP VMFIN PPER*	$\rightarrow$ 2 1 0
Long	*VAFIN * VVPP*	$\rightarrow$ 0 2 1
Tree	*VP PTNEG NP VVPP*	$\rightarrow$ 0 2 1

Figure 1: *Rule Types*

Our work differs in three ways: First, we investigate a reordering approach where reordering decisions are not deterministic. Instead, reordering variants produced by both part-of-speech-based and tree-based reordering rules are stored in a lattice and the final order of the source sentence is decided during decoding. Second, we perform a separate analysis of two different aspects: the quality of the restriction of the search space through reordering lattices and the accuracy of the search. Third, we perform translations from English to German and German to English for 2 different translation tasks.

3. Reordering Approach

We first describe the reordering methods applied in the systems used in our oracle experiments. We use two approaches based on continuous and discontinuous sequences of parts-of-speech of the words in the sentence [7, 10]. In addition we perform reordering based on constituents of syntactic parse trees [15] and we combine the different types of rules. Thus, we cover both short-range and long-range reordering phenomena between source and target language.

3.1. Rule Types

In our experiments we distinguish between short-range, long-range and tree-based rules. Examples for each of the rule types are presented in Figure 1.

3.1.1. Short-range Rules

Short-range rules consist of a sequence of part-of-speech (POS) tags on the left hand side and an indexed representation of the target order of those POS tags on the right hand side of the rule. Each rule comes with an associated probability which is the relative frequency of the occurrence of this reordering in the training corpus.

3.1.2. Long-range Rules

A long-range rule consists of a sequence of POS tags with placeholders on the left hand side. Placeholders can match arbitrary types and numbers of POS tags. The right hand side of the rule contains the reordered indices where the tags matched by the placeholder are assigned one index as a whole. Again, a probability is assigned to each rule.

3.1.3. Tree-based Rules

The tree-based rules address reordering within one constituent of a syntactic tree. The rule consists of the head category as well as the the child categories of the constituent on the left hand side of the rule. The right hand side represents the reordered sequence of the children where each child constituent is assigned one index and the words covered by it are moved as a whole. An additional type of rules called partial rules need not cover all the children in the constituent, but consecutive sequences of children.

3.2. Learning Reordering Rules

For the training of the reordering rules a parallel corpus and a word alignment is required. In addition, we need the POS tags for the source side of the corpus for training the POS-based reordering rules. For the tree-based rules we need syntactic parse trees for the source side. For each sentence in the training corpus we search for changes of word order between the source and target language sentence. When we find a crossing alignment indicating a different order of source and target language words, we monotonize the alignment and extract a rule that rearranges the source words in the order of the aligned target words. For more details refer to the descriptions of POS-based rules [7, 10] and tree-based rules [15].

3.3. Applying Reordering Rules

Before translation, a word lattice is created that includes the original source sentence as the monotone translation path. Initially all edges of the monotone path are assigned a transition probability of 1. Then the reordering rules are applied to the source text. For each sentence all applicable rules are applied where the tree rules might be applied recursively to reordered paths. The resulting reordering variants are stored in the word lattice. The edges of the reordered path are assigned transition probabilities according to the probability of the applied reordering rule. An edge branching from the monotone path receives the probability of the rule. The following edges in the reordered path are assigned a probability of 1. The edge on the monotone path where the branching started receives an update such that the probability of the applied rule is subtracted from the current transition probability of this edge. Finally, the word lattice including all reordering variants is used as input to the decoder.

3.4. Judging Reordered Paths

The probability of a given path in a reordering lattice is calculated as the product of the individual transition probabilities of the traversed edges. Since the transition probabilities are based on the occurrences of the reordering in the training data, the highest scoring path in the lattice should represent the best reordering for the sentence. The reordering lattice is one model in the log-linear model combination of the translation system. Its weight is set during optimization of the whole system together with the weights of the other models in the translation system.

4. Oracle Reordering

We want to investigate the impact of the reordering on the translation quality. We compare the actual system performance against two different oracle reorderings of the input sentence. With these experiments we want to address the questions raised in the introduction.

The first oracle is the optimally reordered source sentence which presents the source words according to the target language word order. With this experiment we analyze the usefulness of the pre-reordering approach. By reordering the source sentence according to the target language word order we estimate an upper bound for translation quality using this strategy.

Then we investigate how the reordering lattices produced by our reordering model restrict the search space for translation. Therefore, we compare the aforementioned oracle translation with the translation of the oracle path. It corresponds to the path in the lattice that is closest to the oracle reordering of the source sentence. We perform this experiment for each of the different rule types.

In a third experiment we evaluate how good our models are at determining the best path in the lattice. In order to evaluate this aspect, we compare the translation of the oracle path with the actual translation.

4.1. Optimally Reordered Sentence

In order to measure the oracle performance of the pre-reordering approach, we use an optimally reordered sentence as input to the translation system and do not allow additional reordering during decoding. In order to create this oracle reordering for the source sentence, we make use of the word alignment between source sentence and reference translation. This alignment is generated by applying the alignment model trained during system development to the test data and its reference translation. After source and reference are aligned, we create a permutation of the source sentence [17].

In the permutation, words are generally assigned the position of word they are aligned with. However, permutations are one-to-one alignments, while word alignments may also contain unaligned words, many-to-one alignments and one-to-many alignments. Therefore, some simplifying assumptions have to be made when transforming alignments to permutations [18]: *unaligned source words* are aligned to the word after its predecessor or to the first word if it has no predecessor; *unaligned target words* are irrelevant to the source sentence order and are therefore ignored; for *many-to-one source-to-target alignments* the ordering is assumed to be monotone; in *one-to-many source-to-target alignments* the word is assumed to be aligned to the first target word. We will refer to this reordered source sentence as the oracle reordering of the input sentence.

4.2. Oracle Path

With our reordering model we generate many reordering variants by applying reordering rules to the source sentence and store these variants in a lattice. In order to know the upper bound of the restriction of the search space by the lattice we want to identify the best reordering variant in the reordering lattice. We define it as the path in the lattice which has the smallest distance to the oracle reordering as described above.

Among Hamming distance, Ulam's Distance and Kendall's tau distance, a version of Kendall's tau resulted to be the best distance, being the most reliable and correlating strongly with human fluency judgement [17]. Hence, we calculate the Kendall's tau distance [24] in order to find the path that is closest to the oracle reordering. The Kendall's tau distance is the minimum number of swaps between two adjacent symbols that transforms a permutation σ into another permutation π. This metric measures relative differences and takes both the number and the size of reorderings into account. We use the square root version [18] which corresponds closely with human perception of word order quality:

$$d(\pi, \sigma) = 1 - \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=i}^{n} x_{ij}}{Z}}$$

$$\text{where } x_{ij} = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } Z = \frac{n \cdot (n-1)}{2}$$

If a path with the oracle reordering is in the lattice, this path is the closest path. However, if the oracle reordering is not in the lattice, several paths can have the smallest distance to the oracle reordering. Then we create lattices containing only the best paths and use these as input to the translation system.

Note that the best path or even the oracle reordering need not result in the best possible translation quality for two reasons. First, we rely on the alignment between source and reference for generating the oracle reordering. Errors in the alignment can introduce errors into the oracle reordering and the closest path. Another reason is that we generate an artificial word order which does not match the word order as seen in the training data. Therefore, we might not have well matching phrase pairs for generating the best possible translation.

5. Experiments

In this section we present three experiments designed to address the three questions raised in the introduction. First, we will briefly describe the systems we used to generate the translations. Afterwards, we will analyze the potential of the pre-reordering approach. Then we investigate how the reordering lattices produced by our reordering model restrict the search space for translation. In a third experiment we compare the oracles with the actual performance of a system using the reordering lattices to see how good our models are at ranking different word orders.

5.1. System Description

We perform experiments with four different systems covering two translation directions and two different translation tasks. We translate between German and English in both directions. For each direction we use competitive systems used in WMT and IWSLT evaluations to translate News texts and TED talks in order to cover different domains. For the News systems, the training data includes the European Parliamentary Proceedings and the News Commentary data. The test data is news2011. For details of the WMT system refer to the WMT system description [25]. The systems are optimized once on news2010, but in the experiments described in this paper, no new optimizations were run between system variants using different rule types to reduce the noise to a minimum. The system translating TED talks is trained on European Parliamentary Proceedings, News Commentary data, the Common Crawl corpus and TED talks, while development and test data consist of TED talks only. Again, the systems are only optimized once. A detailed system description can be referred to in [26]. All translations are produced using the input sentence with a word order stated in the given experiment description. No additional reordering in the decoder is allowed.

5.2. Potential of Reordering the Source Sentence

When applying reordering as preprocessing, it is commonly assumed that arranging the source sentence according to target language word order should result in better translation quality. We want to question this assumption and investigate the benefits of the pre-reordering approach in this first experiment that identifies the lower and upper bounds of translation quality with respect to word order. We consider the lower bound of translation quality to be the performance that is obtained by translating the monotone source sentence without allowing any additional reordering. Since the objective of the pre-reordering approach is to obtain the source words in the order of the target language words, we regard the translation of the optimally reordered path to be the upper bound for translation quality. We generate the optimally reordered path using the reference translation and the alignment between source and reference as described in Section 4.1.

5.2.1. German-English

Table 1 presents the results for the translation from German to English in two different domains. The difference between monotone translation and the translation of the oracle reordering is 5.2 and 6.2 BLEU points, respectively. With a system using our lattice-based reordering approach that does not have any oracle information, but the decoder chooses the path, we achieve a performance that is approximately in the middle of that range.

Reordering Type	News	TED
Monotone	20.23	27.18
Lattice Reordering	22.45	30.87
Oracle	25.42	33.39

Table 1: *Oracle Reordering: German-English*

5.2.2. English-German

For the other translation direction, we can see lower absolute BLEU scores, since translation into German is more difficult due to the highly inflective morphology of the German language. Compared to German-English translation, the difference between monotone and oracle translation is smaller, 2.9 and 4.6 BLEU points, respectively. The decoder using lattice reordering performs better than the monotone translation, but the gap towards the oracle translation is bigger. That means that for English to German translation, there is even more potential for improvement through better reordering lattices.

Reordering Type	News	TED
Monotone	15.91	24.22
Lattice Reordering	16.34	24.95
Oracle	18.84	28.77

Table 2: *Oracle Reordering: English-German*

From this experiment we can draw the conclusion that reordering the source text prior to translation indeed holds promising results. Our system using reordering lattices as translation input outperforms the monotone translation in all four translation tasks, and the oracle reordering shows that there is still potential for improvement through better reordering methods. In the following we will investigate how we can best address this potential by analyzing different aspects of the reordering approach in detail.

5.3. Lattice-based Restriction of the Search Space

In the previous experiment we have identified a gap between the actual performance of the system using reordering lattices and the oracle reordered translation. In our reordering approach we restrict the search space of possible reorderings by the reordering lattice. In this second experiment we want to investigate how much this restriction influences the drop in performance. Therefore, we evaluate how much better we could get, if the decoder found the best path in the given reordering lattices. As described in Section 4.2 we define the best path as the one that is closest to the oracle reordered sentence used in the previous experiment.

In order to compare the benefits of individual reordering rule types we apply all the different types of reordering rules and identify the oracle path within the lattices produced by those rules. Then we perform translation of the oracle path and compare the translation quality.

All results tables repeat the scores for the monotone and oracle translation presented above. In addition, they show the translation results for systems using first short and long-range rules based on POS tags. Afterwards follow the tree-based rules, first the plain tree rules, then the tree-based rules with recursive rule application and the third tree rule option includes partial rules. More details on recursive rule application and partial rules are described in [15]. The three final systems combine all rule types.

5.3.1. German-English

Table 3 shows the results for German-to-English translation and the size of the search space by indicating the number of edges in the lattices. As can be seen, the more complex the rule types that are used to generate the reordering lattice and the larger the search space gets, the better the translation of the oracle path in that lattice. Hence, we are able to improve the word order by increasing the search space. The oracle path that is closest to the oracle reordering stems from the lattice produced by applying all rule types.

Reordering Type	News		TED	
	BLEU	Size	BLEU	Size
Monotone	20.23		27.18	
Short	21.37	193K	29.98	68K
Short+Long	21.41	255K	30.66	163K
Tree	21.88	140K	29.74	51K
Tree-rec	22.17	244K	30.11	81K
Tree-rec-partial	22.28	249K	30.22	82K
Short+Long+Tree	22.49	429K	30.97	182K
Short+Long+Tree-rec	22.64	534K	31.10	212K
Short+Long+Tree-rec-part.	22.65	538K	31.12	213K
Oracle	25.42		33.39	

Table 3: *Oracle Path: German-English*

5.3.2. English-German

Table 4 presents the same experiments for English-to-German translation. Again, the more complex rules and bigger search spaces lead to better oracle paths.

Thus, we can confirm the findings in [15], namely that the different rule types produce complementary reordering possibilities which result in the best translation quality if combined in one lattice. We can also see that the translation of the best oracle path is still far from the oracle reordered translation. The lattices generated with the help of our reordering rules restrict the search space in a sensible way to allow for reorderings that are getting closer to the oracle reordered sentence. However, some reordering possibilities are still missing from our lattices. Therefore, research in the area of extending the search space by better rules seems to be promising.

Reordering Type	News				TED			
	DecoderPath		OraclePath		DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance	BLEU	Distance	BLEU	Distance
Monotone			20.23				27.18	
Short	21.59	0.290	21.37	0.250	30.00	0.179	29.98	0.124
Long	21.35	0.286	21.41	0.259	30.73	0.181	30.66	0.112
Tree	21.78	0.286	21.88	0.250	29.60	0.180	29.74	0.140
Tree-rec	22.01	0.284	22.17	0.243	29.88	0.179	30.11	0.135
Tree-rec-partial	22.10	0.284	22.28	0.241	29.96	0.179	30.22	0.133
Short+Long+Tree	22.33	0.289	22.49	0.224	30.82	0.182	30.97	0.106
Short+Long+Tree-rec	22.44	0.288	22.64	0.220	30.86	0.182	31.10	0.104
Short+Long+Tree-rec-partial	22.45	0.288	22.65	0.220	30.87	0.182	31.12	0.104
Oracle			25.42				33.39	

Table 5: *Oracle vs. Real: German-English*

Reordering Type	News		TED	
	BLEU	Size	BLEU	Size
Monotone	15.91		24.22	
Short	16.31	186K	25.83	76K
Short+Long	16.70	383K	25.99	170K
Tree	16.48	189K	25.31	71K
Tree-rec	16.60	726K	25.49	237K
Tree-rec-partial	16.60	727K	25.49	237K
Short+Long+Tree	17.00	496K	26.28	208K
Short+Long+Tree-rec	17.07	1M	26.38	373K
Short+Long+Tree-rec-part.	17.07	1M	26.38	373K
Oracle	18.84		28.77	

Table 4: *Oracle Path: English-German*

5.4. Ranking different word orders

The experiments above revealed the best translation that can be produced by using the individual rule types and combinations thereof. Now we want to examine how well we actually perform in finding the best path in the lattices. Again, we tested on all the different rule types, but let the decoder find the best path for translation. It is worth mentioning that the decoder does not only utilize the reordering model described in Section 3 to find the path, but all the models in the log-linear model of the translation system. For reference we include the scores achieved with the oracle paths from the previous experiment. In addition, we present the average distances between the decoder path used for translation and the optimally reordered sentence both for the decoder translation and for the translation of the oracle path. The distances are calculated using the Kendall's tau metric.

5.4.1. German-English

We present the results for German-to-English translation in Table 5. The differences between the oracle path scores and the real performance of the system (decoder path) with the reordering lattices are actually very small. This means that the decoder is already quite good at finding the best path in the reordering lattice. To reach the translation quality of the oracle path, a further increase of 0.2 and 0.3 BLEU points would be possible for the News and the TED task, respectively.

The distances between decoder translation path and oracle reordering are shown in the column to the right of the decoder path, while the distances between the oracle path and the oracle reordering are shown in the column to the right of the scores reached by the oracle path translations. We can see that both the distances and the translation quality for the oracle path systems converge nicely for the News task. The closer the translation quality comes to the translation quality of the oracle reordering, the smaller the distance to the oracle reordering. In the TED task we also observe a good correspondence between translation quality and reordering distance for the oracle path results. The drop in BLEU score when using only tree rules is also obvious in the distance scores, which raise for those systems. For the decoder translation path, the distance to the oracle reordering seems to be not converging at all, it stays about the same both for News and TED translations.

5.4.2. English-German

The results for English-to-German translation are presented in Table 6. For this translation direction, the path in the reordering lattices chosen by the decoder is not very close to the optimal one yet. The decoder performance is 0.7 BLEU points worse than the translation of the oracle path in the best rule type of the News task. For the TED task, the difference between oracle path translation and decoder performance is even 1.4 BLEU points.

The distance scores show a similar behavior as observed in the other translation direction. The distances from oracle path to oracle reordering get smaller as the translation quality increases. The distances from decoder translation path to oracle reordering do not converge. Compared to the other

Reordering Type	News				TED			
	DecoderPath		OraclePath		DecoderPath		OraclePath	
	BLEU	Distance	BLEU	Distance	BLEU	Distance	BLEU	Distance
Monotone			15.91				24.22	
Short	16.27	0.297	16.31	0.249	24.83	0.200	25.83	0.141
Long	16.31	0.311	16.70	0.236	24.87	0.214	25.99	0.129
Tree	16.21	0.306	16.48	0.252	24.47	0.206	25.31	0.163
Tree-rec	16.18	0.312	16.60	0.244	24.51	0.207	25.49	0.158
Tree-rec-partial	16.18	0.312	16.60	0.244	24.50	0.207	25.49	0.158
Short+Long+Tree	16.32	0.318	17.00	0.227	24.94	0.217	26.28	0.123
Short+Long+Tree-rec	16.34	0.321	17.07	0.222	24.95	0.218	26.38	0.120
Short+Long+Tree-rec-partial	16.34	0.321	17.07	0.222	24.95	0.218	26.38	0.120
Oracle			18.84				28.77	

Table 6: *Oracle vs. Real: English-German*

direction they vary even more. It is possible that this is due to the smaller differences in translation quality. In addition, outliers in the paths chosen by the decoder could cause the variations in the distance scores.

From these results on the translation quality we can draw the conclusion that there still lies some potential in the reordering rules and consequently in the reordering lattices that the decoder is not yet able to make use of. The differences in the decoder path translation scores and oracle path translation scores suggest that more complex scoring models for better assessing the quality of different reordering possibilities seem to be a promising research direction for English-German translation.

6. Conclusion

We have analyzed the performance of an approach to reordering as a preprocessing step using oracle experiments. We conducted experiments on German-to-English and English-to-German translation of News texts and TED talks.

In a first series of experiments we could show that source sentence reordering is a very promising approach. By translating an optimally reordered source sentence, we could improve the translation performance by up to 6.2 BLEU points.

Then we translated the optimally reordered source sentence and compared it with the oracle path in reordering lattices produced by different types of reordering rules. This led to the conclusion that the restriction of the search space using our reordering lattices approximates the oracle reordering better when more complex and complementary reordering rules are used. However, the best oracle path and the oracle reordering are still far apart, leaving a lot of potential for finding better reordering rules that approximate the oracle reordering even better. While for German-to-English translation the distance between actual performance and the best possible translation is 2.5 to 3 BLEU points, the gap for English-German is a little bigger. An additional 2.5 to 3.8 BLEU points are missing until the best possible translation result can be reached. As a consequence, one direction of promising research is to extend the search space further to include reordering variants that better approximate the optimally reordered source sentence.

Comparing the decoder path translation with the oracle path showed that the path chosen by the decoder is quite close to the oracle path, both in terms of translation quality and reordering distance for German-to-English translation. The decoder translation path and the oracle path are only 0.2 and 0.3 BLEU points apart. Consequently, the current models used in the machine translation system are able to find almost the best source word order that is in the search space. For English-to-German translation, however, finding the best path in the reordering lattice seems to be more difficult. A gap of 0.7 and 1.4 BLEU remains until the oracle path performance is reached. We can conclude that at least for English-to-German translation a better ranking of the different reordering possibilities in the search space seems to hold a promising perspective for future research.

All in all, our experiments confirmed the usefulness of reordering the source sentence before translation. The approach displayed a good performance with potential for improvement by extending the search space of reordering possibilities. For English-to-German the ranking of reordering quality for finding a better path in the reordering lattice is another promising research direction. In total, the approach has a potential for a further 3 and 3.8 BLEU points of improvements, depending on the language. This potential could be reached by improving the restriction of the search space with better rules and a better ranking of reordering quality.

7. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

8. References

[1] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proceedings of IWSLT 2005*, Pittsburgh, USA, 2005.

[2] F. Xia and M. McCord, "Improving a statistical MT system with automatically learned rewrite patterns," in *Proceedings of COLING 2004*, Geneva, Switzerland, 2004.

[3] M. Collins, P. Koehn, and I. Kučerová, "Clause Restructuring for Statistical Machine Translation," in *Proc. of ACL 2005*, Ann Arbor, Michigan, USA, 2005.

[4] M. Popović and H. Ney, "POS-based Word Reorderings for Statistical Machine Translation," in *Proceedings of LREC 2006*, Genoa, Italy, 2006.

[5] N. Habash, "Syntactic preprocessing for statistical machine translation," *Proceedings of MT Summit*, 2007.

[6] C. Wang, M. Collins, and P. Koehn, "Chinese syntactic reordering for statistical machine translation," in *Proceedings of EMNLP 2007*, Prague, Czech Republic, 2007.

[7] K. Rottmann and S. Vogel, "Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model," in *TMI*, Skövde, Sweden, 2007.

[8] Y. Zhang, R. Zens, and H. Ney, "Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation," in *Proceedings of SSST 2007*, Rochester, NY, USA, 2007.

[9] J. M. Crego and N. Habash, "Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT," in *Proceedings of ACL-HLT 2008*, Columbus, Ohio, USA, 2008.

[10] J. Niehues and M. Kolss, "A POS-Based Model for Long-Range Reorderings in SMT," in *Proceedings of WMT 2009*, Athens, Greece, 2009.

[11] M. R. Costa-jussà and J. A. R. Fonollosa, "Statistical Machine Reordering," in *Proceedings of EMNLP 2006*, Sydney, Australia, 2006.

[12] M. Khalilov, J. Fonollosa, and M. Dras, "A new subtree-transfer approach to syntax-based reordering for statistical machine translation," in *Proceedings of EAMT 2009*, Barcelona, Spain, 2009.

[13] D. Genzel, "Automatically learning source-side reordering rules for large scale machine translation," in *Proceedings of COLING 2010*, Beijing, China, 2010.

[14] U. Lerner and S. Petrov, "Source-side classifier preordering for machine translation," in *Proceedings of EMNLP 2013*, Seattle, Washington, USA, 2013.

[15] T. Herrmann, J. Niehues, and A. Waibel, "Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation," in *Proceedings of SSST 2013*, Altanta, Georgia, USA, 2013.

[16] J. Eisner and R. W. Tromble, "Local Search with Very Large-Scale Neighborhoods for Optimal Permutations in Machine Translation," in *Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, 2006.

[17] A. Birch, M. Osborne, and P. Blunsom, "Metrics for MT Evaluation: Evaluating Reordering," *Machine Translation*, vol. 24, no. 1, 2010.

[18] A. Birch, "Reordering Metrics for Statistical Machine Translation," Ph.D. dissertation, University of Edinburgh, 2011.

[19] G. Wisniewski, A. Allauzen, and F. Yvon, "Assessing phrase-based translation models with oracle decoding," in *Proceedings of EMNLP 2010*, Cambridge, Massachusetts, USA, 2010.

[20] E. Cho, J. Niehues, and A. Waibel, "Segmentation and Punctuation Prediction in Speech Language Translation Using a Monolingual Translation System," in *Proceedings of IWSLT 2012*, Hong Kong, 2012.

[21] A. Sokolov, G. Wisniewski, and F. Yvon, "Computing lattice BLEU oracle scores for machine translation," in *Proceedings of EACL 2012*, Avignon, France, 2012.

[22] M. Dreyer, K. Hall, and S. Khudanpur, "Comparing Reordering Constraints for SMT Using Efficient BLEU Oracle Computation." in *Proc. of SSST 2007*, Rochester, USA, 2007.

[23] M. Khalilov and K. Sima'an, "Context-sensitive syntactic source-reordering by statistical transduction," in *Proceedings of IJCNLP 2011*, Chiang Mai, Thailand, 2011.

[24] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed. A Charles Griffin Title, September 1990.

[25] J. Niehues, Y. Zhang, M. Mediani, T. Herrmann, E. Cho, and A. Waibel, "The Karlsruhe Institute of Technology Translation Systems for the WMT 2012," in *Proceedings of WMT 2012*, Montreal, Canada, 2012.

[26] T.-L. Ha, J. Niehues, T. Herrmann, M. Mediani, E. Cho, Y. Zhang, I. Slawik, and A. Waibel, "The KIT Translation Systems for IWSLT 2013," in *Proceedings of IWSLT 2013*, Heidelberg, Germany, 2013.