# Alignment-based reordering for SMT

**Maria Holmqvist, Sara Stymne, Lars Ahrenberg and Magnus Merkel**

Department of Computer and Information Science
Linköping University, Sweden
firstname.lastname@liu.se

**Abstract**

We present a method for improving word alignment quality for phrase-based SMT by reordering the source text according to the target word order suggested by an initial word alignment. The reordered text is used to create a second word alignment which can be an improvement of the first alignment, since the word order is more similar. The method requires no other pre-processing such as part-of-speech tagging or parsing. We report improved Bleu scores for English–German and English–Swedish translation. We also examined the effect on word alignment quality and found that the reordering method increased recall while lowering precision, which partly can explain the improved Bleu scores. A manual evaluation of the translation output was also performed to understand what effect our reordering method has on the translation system. We found that where the system employing reordering differed from the baseline in terms of having more words, or a different word order, this generally led to an improvement in translation quality.

**Keywords:** statistical machine translation, reordering, evaluation

## 1. Introduction

Word order differences between languages create several problems for statistical machine translation systems. They present a challenge in translation decoding, where translated phrases must be rearranged correctly, but also during word alignment with statistical methods. For example, the placement of finite verbs in German at the end of a clause makes English and German verbs notoriously difficult to align because of their different positions in the sentence.

In this paper we present a pre-processing method that reorders source words according to the corresponding target word order suggested by an initial word alignment. By making the two texts more similar we hope to address some of the difficulty that word order differences pose to word alignment. A second word alignment is performed on the reordered source and target text when the word order is more similar.

## 2. Word order and SMT

In phrase-based SMT (PBSMT) the decoder tries to find the most probable translation of a sentence by combining translated phrase segments into a well formed target sentence. The final choice of phrases and the order in which they are placed are based on a number of weighted probabilistic features. The phrase translation model and reordering model are estimated from a word aligned parallel corpus. Word alignment is an important step in training a SMT systems since it determines the probabilities of phrase translations and reordering.

During training, state-of-the-art statistical word alignment methods may have difficulty finding the correct alignment of words that are placed at considerably different positions in the source and target sentence. Errors or missing alignments will add incorrect phrase translations to the translation model, and produce a less accurate reordering model as well as less accurate estimations in the reordering model.

## 3. Related work

The challenges of word order differences have been approached in different ways. Since the original word-based distortion models of Brown et al. (1993) reordering models learnt in training and employed by the decoder has become more and more sophisticated, often using both lexical and syntactic features (Koehn et al., 2005; Xiang et al., 2011).

Another approach is to modify the source text before training by making the order of words and phrases more similar to the target language. The most successful of these approaches employ some form of syntactic analysis and the reordering rules can be handwritten as in Collins et al. (2005), or automatically extracted from parallel text as in Xia and McCord (2004); Elming (2008). Language specific reordering rules are applied to the source text and a system is built that translates from reordered source text to target text. This means that a source text must first be reordered using the same reordering rules before it can be translated by the system.

The pre-processing approach has two possible benefits. First, the most obvious benefit is that some of the difficulty of reordering is removed from the translation step. Since the bulk of reordering has already been performed on the source text the translation system will only need to find appropriate phrase translations and do minor changes in word order.

The second benefit appears during the training of the translation system since statistical word alignment methods perform better on translations with similar word order. Improved word alignment quality may also have a positive effect on the translation model and thereby improve translation quality.

Pre-processing does not produce consistent improvements on both translation reordering and word alignment quality for all language pairs. Experiments with German–English (Holmqvist et al., 2009) and English–Arabic (Carpuat et al., 2010) found improvements on translation quality from the improved word alignment rather

than from its effect on reordering during decoding. The effect on alignment quality was isolated by reordering the source text before word alignment, translating alignments back to match the words of the original text and then training the final system on the original text, but with the new (improved) alignment.

## 4. Alignment-based reordering

In this paper, we present a simple, language-independent reordering method to improve word alignment quality and apply it to English–German and English–Swedish translation. After reordering we perform statistical word alignment on the reordered corpus. The hypothesis is that the reordering will result in improved word alignments which in turn will result in a better translation model and better translation quality. Our reordering algorithm is simple, yet effective. It is based on the alignments created by an initial word alignment on the original texts. It does not require any handcrafted or automatically acquired grammatical reordering rules and the process is completely language-independent. The following steps are performed:

(a) perform statistical word alignment with Giza++ (Och and Ney, 2003) on the original texts

(b) reorder one of the texts according to the word alignments

(c) perform statistical word alignment on the preprocessed texts

(d) keep the new word alignments but transfer them back to the original texts to connect words in their original order

The result is a parallel text with potentially improved word alignment from which we build a standard phrase-based SMT system that translates from source to target text.

### 4.1. Reordering algorithm

The reordering algorithm puts the words in one text in the order of the corresponding words in the other text. The initial word-to-word correspondences are created using Giza++ which produces two word alignments one in each translation direction. We then apply a standard algorithm for combining both alignments into one bidirectional alignment. The result is an unconstrained alignment which may contain incomplete alignments where an aligned phrase has not been fully linked as the lines show in Figure 4.1.. Aligned phrases may also contain gaps that consist of words that connect to a phrase in a different position (dashed line in Figure 4.1.) or words that have no alignment.

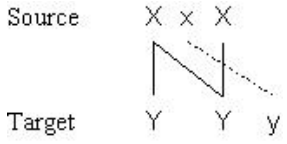


Figure 1: Incomplete alignment with gap.

A correctly unaligned word has no counterpart in the target sentence and by removing it we would make source and target sentences more similar which is the goal of the reordering. However, if the null-alignment is an error (which is often the case) we want to keep the word in the reordered sentence so it can be correctly aligned in the second alignment pass. We therefore keep all words from the source, and move all gap words (unaligned or not) to the right of the containing phrase. The reordering is performed in the following steps, illustrated in Figure 2:

1. Reorder discontinuities by placing the gap words to the right of the containing phrase
2. Add dummy target words for unlinked source words
3. Identify all word aligned groups (phrase alignments)
4. Reorder the source phrases according to the alignment to target phrases.

## 5. Reordering experiments

We have performed experiments on English-German and English-Swedish PBSMT. Systems are built using Moses (Koehn et al., 2007). We report results in Bleu (Papineni et al., 2002) and Meteor ranking scores (Agarwal and Lavie, 2008).

### 5.1. Experiment corpora

Table 1 presents an overview of the corpora used in the experiments. The German–English data was released as shared task data in WMT2009 and WMT2010 workshops (Callison-Burch et al., 2009). This dataset contains both in-domain (news text) and out-of-domain data (Europarl) with a limited amount of in-domain parallel data. The English–Swedish corpora were extracted from the Europarl data and comes in two sizes.

| | | Parallel | | Monolingual | |
|---|---|---|---|---|---|
| | Name | ep | News | ep | News |
| En–De | wmt09 | 1,3M | 81141 | - | de 9,6M<br>en 21,2M |
| | wmt10 | 1,5M | 100K | | de 17,5M<br>en 48,6M |
| En–Sv | ep700K | 700K | - | 700K | - |
| | ep100K | 100K | - | 100K | - |

Table 1: News and Europarl (ep) corpora used in experiments. Size in number of sentences.

### 5.2. English–German translation

The German–English and English–German translation systems consist of two translation models, one from each parallel data set, a reordering model trained on the Europarl data and sequence models on surface form and part-of-speech from all news data. The system is described in (Holmqvist et al., 2009).

The reordered system contains the same components as the baseline system but the parallel corpora have been word aligned using the reordering method described in Section 4.1. A word alignment was created by combining

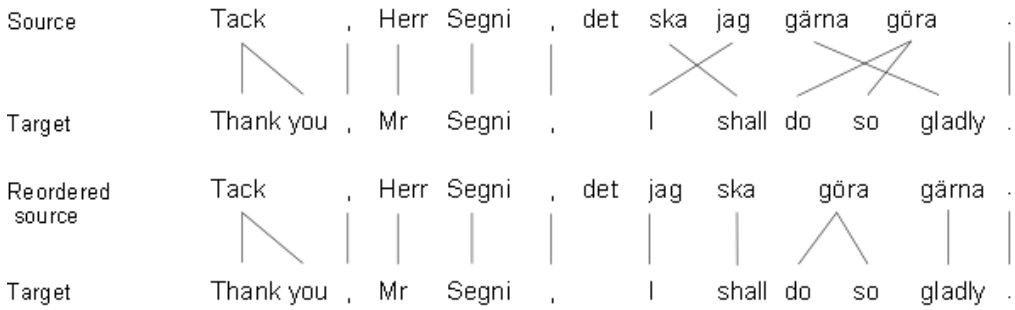


Figure 2: Source text reordered according to alignment with target sentence.

two directed Giza++ alignments using the grow-diag-final-and (gdfa) symmetrization heuristic which gives a good precision-recall trade-off suitable for PBSMT. The results on test data (1025 sentences) are shown in Table 2 and 3.

| | En–De | | De–En | |
|---|---|---|---|---|
| | BLEU | Meteor-r | BLEU | Meteor-r |
| Baseline | 14.62 | 49.48 | 19.92 | 38.18 |
| Reorder (src) | 14.63 | 49.80 | 20.54 | 38.86 |
| Reorder (trg) | 14.51 | 48.62 | 20.48 | 38.73 |

Table 2: Results of WMT09 experiments.

| | En–De | | De–En | |
|---|---|---|---|---|
| | BLEU | Meteor-r | BLEU | Meteor-r |
| Baseline | 14.24 | 49.41 | 18.50 | 38.47 |
| Reorder | 14.32 | 49.58 | 18.77 | 38.53 |

Table 3: Results of WMT10 experiments.

We compared the effects of reordering the source text versus the target text and found that reordering the source resulted in better Bleu scores. Reordering improved translation from German into English more than in the other direction. Table 2 shows the most notable improvements on both metrics, +0.6 in Bleu and +0.7 in Meteor-ranking. A possible reason for this result is that alignment quality might be more important in the German–English direction.

### 5.3. English–Swedish translation

In the English-Swedish experiments we compared the effect of reordering on two datasets, a small set of 100K sentences and a larger set of 700K sentences. The results in Table 4 show that the reordered system outperformed baseline in terms of Bleu for both datasets and in both translation directions. However, the improvement was only statistically significant for the large corpus and in translation into Swedish. In terms of word alignment quality, both reordered alignments had higher recall than the baseline alignment, at the expense of lower precision.

#### 5.3.1. Symmetrization heuristic

Creating a word alignment consists of three steps (1) use Giza++ to create 1-to-n alignments from source to target (2) use Giza++ to create 1-to-n alignments in the other direction, and (3) apply a symmetrization heuristic to create a bidirectional m-to-n word alignment.

The symmetrization heuristic determines precision and recall of the word alignment. By keeping only links that the two alignments have in common (intersection) we get a high precision/low recall alignment. Most useful heuristics start from the intersection and add links from the union using the intersection as a guide. The grow-diag (gd) heuristics adds links that are adjacent to previously aligned words. The grow-diag-final-and (gdfa) heuristic also adds links that connect previously unaligned words. The gdfa heuristic has higher recall than gd and is often the preferred heuristic for building PBSMT systems.

When creating a word-alignment in a reordered system we perform two separate alignments. The first alignment is the basis of our reordering. The second alignment is performed on the reordered corpus and it is from this alignment that we extract the phrase table for our translation model.

In the experiments reported above, both word alignments have been performed with the gdfa heuristic. However, there is reason to believe that the reordering algorithm may perform better if it bases the reordering on an alignment with higher precision, i.e., the reorderings that take place will be more accurate while fewer words will be reordered. To test this hypothesis we built systems using different combinations of gd and gdfa alignments and measured word alignment and translation quality. The results are shown in Table 5 where First denotes the word alignment performed before reordering and Final the alignment that the translation model is based on. Only one alignment is performed in the baseline systems.

We found that using gd for the first alignment gave equal or better results for en-sv translation but worse results for sv-en. Word alignment precision and recall for this setup (gd-gdfa) were worse than gdfa-gdfa. An alignment combination of gd-gd showed improvements in Bleu comparable to gdfa-gdfa for en-sv but not for sv-en.

## 6. Manual Evaluation

We found that alignment-based reordering improves Bleu score for translation between German–English and Swedish–English. Since Bleu scores are difficult to interpret we also performed manual analysis to find out what effect alignment-based reordering has on the system and on translation.

| System | | Precision | Recall | F | BLEU | |
|---|---|---|---|---|---|---|
| | | | | | En–Sv | Sv–En |
| 100k | Base | 81.65 | 75.07 | 78.22 | 23.41 | 28.35 |
| | Reo | 80.22 | 75.49 | 77.78 | 23.54 | 28.60 |
| 700k | Base | 83.82 | 77.78 | 80.69 | 24.62 | 30.86 |
| | Reo | 82.78 | 78.54 | 80.61 | 24.96* | 30.98 |

Table 4: Translation and alignment quality for English–Swedish (*Significant at $p < 0.05$ using approximate randomization (Riezler and Maxwell, 2005))

| First | Final | Precision | Recall | F | BLEU | |
|---|---|---|---|---|---|---|
| | | | | | En–Sv | Sv–En |
| - | gd | 85.31 | 76.86 | 80.86 | 24.71 | 30.73 |
| gd | gd | 83.46 | 77.70 | 80.48 | 24.93 | 30.88 |
| gdfa | gd | 83.33 | 78.20 | 80.68 | 24.70 | 30.49 |
| - | gdfa | 83.82 | 77.78 | 80.69 | 24.62 | 30.86 |
| gdfa | gdfa | 82.78 | 78.54 | 80.61 | 24.96* | 30.98 |
| gd | gdfa | 82.69 | 78.28 | 80.43 | 25.12* | 30.73 |

Table 5: A comparison of symmetrization heuristics (ep700k).

### 6.1. System comparison

A comparison between the reordered system and the baseline system based on the large English-Swedish corpus show that the phrase table of the reordered system is almost 10% smaller than the baseline table. One reason could be that higher word alignment recall creates fewer alternative phrases that apparently still produces good translations.

We also compared the tuned weights of the different system components. By comparing the tuned weights of components that rely on the word alignment with the tuned weights of the monolingual language models we wanted to find out if in fact, the improvement in translation quality come from a stronger reliance on the language model which would indicate that alignment-based reordering created a less accurate translation model. Fortunately, this was not the case as the language model weight was slightly higher for the baseline system (0.048 vs. 0.045). On the contrary, it shows that more importance is attributed to the translation model created from alignment-based reordering.

Another difference in the tuned weights is that the reordered system favors slightly shorter output than the baseline system. This is determined by the tuned word penalty parameter which was set to -0.096 and -0.102, respectively.

Another thing to note is that the language model has higher weight in the Swedish–English system than the English–Swedish, which explains the smaller effect of reordering on the Swedish–English systems.

### 6.2. Manual translation evaluation

The English–Swedish reordered system achieved a statistically significant improvement in Bleu over the baseline. To find out what this actually means a manual evaluation was performed on 133 sentences that differed between systems. The two systems were anonymized and three annotators were asked to categorize each difference between translations into one of six categories, using the Blast annotation tool (Stymne, 2011). Annotators also had to judge if the difference was better in one of the systems or similar in quality. The classification of each difference and which system this difference was in favor of is shown in Table 6.

| Category | Reordered | Baseline | Neutral |
|---|---|---|---|
| Word choice | 91 | 111 | 223 |
| Agreement | 29 | 32 | 53 |
| Word order | 18 | 7 | 14 |
| Addition | 90 | 22 | 23 |
| Deletion | 29 | 66 | 27 |
| Other | 2 | 2 | 4 |
| Total | 31% | 28% | 41% |

Table 6: Frequency of judged improvements per system and divergence category.

Three categories have a clear effect in favor of one system: Addition, Deletion and Word order. Added word(s) tend to be in favor of the reordered system and deletions are often favorable to the baseline system. Both systems tend to benefit from having the extra word, but the reordered system has the most additions. Each sentence from the reordered system was labeled as better, worse or neutral compared to the sentence from the baseline system based on a majority vote of the non-neutral differences from each annotator (Table 7). The difference between reordered and baseline was not significant using Wilcoxon signed rank test. The sentence level judgments were fairly consistent between annotators. All three annotators agreed on 54% of the sentences and at least two agreed on 97%.

| | Reordered | Baseline | Neutral |
|---|---|---|---|
| Sentences | 50 | 42 | 43 |

Table 7: Frequency of judged improvements per system at the sentence-level.

## 7. Conclusion

We have presented alignment-based reordering, a language-independent reordering method to improve word alignment

for phrase-based SMT systems. Translation experiments on German–English and Swedish–English showed improvements in translation quality as measured in Bleu. The improvements were larger for German-English translation than English–German, and larger for English–Swedish than Swedish–English Manual evaluation of the differences in translations from reordered and baseline systems revealed that reordered systems are better in cases of additions and word order differences. In terms of word alignment quality, improved Bleu score often correlates with improved word alignment recall. Reordered systems tend to have higher recall which results in smaller phrase translation models.

## 8. References


Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the 48th Annual Meeting of the ACL, Short papers*, pages 178–183, Uppsala, Sweden.

Michael Collins, Philipp Koehn, and Ivona Kucerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540, Ann Arbor, Michigan.

Jakob Elming. 2008. Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, pages 46–54, Columbus, Ohio, USA.

Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 120–124, Athens, Greece.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demonstration session*, Prague, Czech Republic.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania.

Stefan Riezler and John Maxwell. 2005. On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, USA.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the ACL, demonstration session*, Portland, Oregon, USA.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, Geneva, Switzerland.

Bing Xiang, Niyu Ge, and Abraham Ittycheriah. 2011. Improving reordering for statistical machine translation with smoothed priors and syntactic features. In *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 61–69, Portland, Oregon, USA.