

Reordering by Parsing

Jakob Elming and Martin Haulrich

Department of International Language Studies and Computational Linguistics
Copenhagen Business School
Copenhagen, Denmark
{jel, mwh}.isv@cbs.dk

Abstract

We present a new discriminative reordering model for statistical machine translation. The model employs a standard data-driven dependency parser to predict reorderings based on syntactic information. This is made possible through the introduction of a reordering structure, which is a word alignment structure where the target word order is transposed onto the source sentence as a path. The approach is integrated in a phrase-based system. Experiments show a large increase in long distance reorderings. Both automatic and human evaluations show substantial increases in translation quality on an English to German task.

1 Introduction

Handling word order differences between languages is one of the main challenges of statistical machine translation (SMT) today. These differences are often most naturally handled at a syntactic level, since they pertain to entire syntactic constituents.

We present a syntactically motivated discriminative reordering model. The model exploits a *reordering structure*, which is a word alignment where the target sentence is unknown. This structure allows us to treat the reordering problem as a dependency parsing problem. We use a standard data-driven dependency parser to predict reorderings instead of dependencies. This is integrated into a phrase-based SMT (PSMT) framework (Koehn et al., 2003).

2 Reordering Structure

Word alignments are often used to display the relation between a translation and its source by linking up equivalent words. Here we transpose the word alignment information to a representation over a single sentence. This can be done by representing the

corresponding order of the words of the opposite sentences as a path over the words of the current sentence. In this work we will focus on transposing the word alignment onto the source sentence by annotating it with the order in which the aligned target words occur. This is done in the form of a reordering structure, which is a word alignment, where the target sentence is unknown. The idea of a reordering structure is similar to the underlying concept of *source position target order* (Elming, 2008) or *visit sequence* (Ge, 2010), but the extraction algorithm and conceptual representation is different.

Figure 1 gives a simple example of how a reordering structure is created. The figure contains a source and a target sentence with a word alignment in between and the corresponding reordering structure on the source sentence. The numbers are merely used to explain the correlation between links and edges. They are not part of the structure. The reordering structure is created by following the target words from left to right. The first target word is linked to the first source word, and the graph therefore starts by going to the first word. Then the second target word links to the third source word, so the graph proceeds to this word, and so on. The resulting representation consists of the source sentence annotated with a reordering structure which reflects the word order of the corresponding target sentence.

One requirement for the structure is that all source words partake in an edge. The role of null-linked source words in the structure cannot be uniquely determined from a word alignment. It can either be inserted after the previous word or before the following word in the structure.

We employ a *syntactic closeness measure* to decide between left and right attachment. The distance from the null-linked word up to the first common

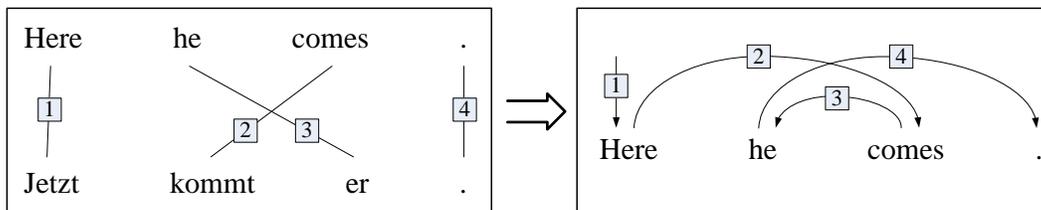


Figure 1: Example of a reordering structure and its underlying word alignment. Numbers are added for explanatory reasons indicating correlation between links and edges. They are not part of the structure.

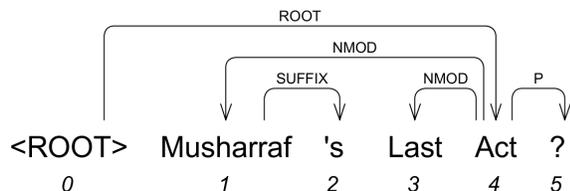


Figure 2: Syntactic dependency structure illustrating syntactic closeness.

node with the left and right neighbor is measured as the number of edges passed on the way. If this is the same for both neighbors, we choose the neighbor with the shortest distance up to the common node. If this is also the same, we attach right. As an example, if we assume *Last* in figure 2 is null-linked, we need to decide whether to insert relative to the left word *'s* or right word *Act*. Common ancestor node is 4 with both neighbors, so distance up from *Last* is the same. We therefore rely on distance up from neighbors, which is 2 passed edges for *'s* and 0 for *Act*. *Act* is therefore syntactically closer, and we attach right. Algorithm 1 illustrates the construction of the reordering structure formally.

The measure is linguistically motivated. The common ancestor defines the smallest spanning constituent containing both words. The shorter the path up, the smaller the span, and the syntactically closer the words. If we simply measured the total path length, we might get fooled by a long path down. An example is a noun preceded by a preposition and followed by a relative clause. Here, the noun itself is the common ancestor with the relative pronoun, i.e. it has a 0 distance up, but the down distance may be long. The total path would not classify the noun closest to the relative clause, since the distance to the preceding preposition is 1 up and 0 down.

The advantage of the reordering structure representation is that it is a word alignment representation without explicit reference to the target sentence. Re-

```

sourcePosition previous = rootPosition;
foreach targetPosition t = 0, ..., T do
  foreach sourcePosition s linkedTo t do
    add edge from previous to s;
    previous ← s;
  end
end
foreach sourcePosition s ∈ nullLinked do
  if (s - 1) is syntactically closest to s then
    insert s after (s - 1);
  else
    insert s before (s + 1);
  end
end

```

Algorithm 1: Algorithm for creating reordering structure from word aligned sentence positions.

ordering in machine translation can be viewed as a similar challenge, where we want to find the word alignment knowing only the source sentence. The reordering structure provides us with a focus on this problem, since it refers only to the source sentence and therefore may be predicted from this.

The relation between the reordering structure and the word alignment is not reversible. Whereas all reordering structures correspond to a unique word alignment, the reverse is not the case. Certain word alignments are not representable by a reordering structure. In particular, structures where a source word is linked to target words that are separated by target words linking to a different source word cannot be represented without introducing recursion into the structure as exemplified by figure 3. As a consequence, the structure becomes ambiguous, since a single word would have more out-going edges that could be traversed in different orders.

Crego & Yvon (2009) face similar challenges

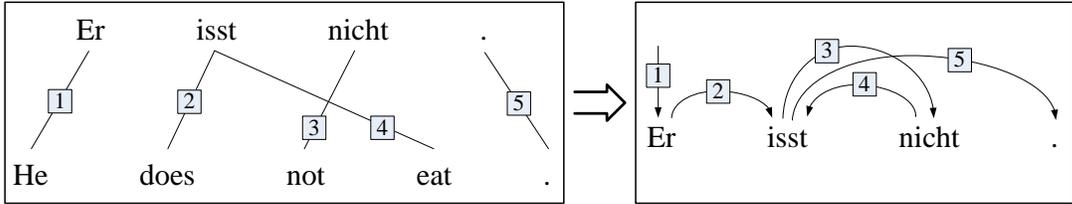


Figure 3: Example of a word alignment that cannot be represented *unrecursively* in a reordering structure. The language pair is here reversed, since we did not find these structures in the direction we are working with.

when monotonicizing the parallel sentences. They handle the problem by making a source word clone for each discontinuous unit it is linked to. We do not adopt this approach here, since these structures are not a major concern with PSMT. PSMT has two means for handling word order differences between languages; phrase-internal reordering, where the equivalent words of a phrase pair appear in different orders, and phrase-external reordering, where the phrases are combined in a different order than they appeared in the source sentence. Only phrase-internal reordering can lead to this problematic word alignment in application, since a single source word token cannot participate in more phrases in the same translation. Since phrase-internal reordering is very reliable, the main purpose of the reordering model is to guide the phrase-external reordering, which will not produce these link constellations.

3 Reordering Structure Modelling

In this work, we will pursue the idea that the reordering structure is conceptually similar to an unlabeled syntactic dependency structure. We therefore use the MSTParser (McDonald et al., 2005), a state-of-the-art data-driven dependency parser, to model the reordering structure.

The basic idea is that the parser predicts the most favorable word alignment to the target sentence based on the source sentence. These predictions are made before translation and passed to the decoder. The level of information included in the reordering structure model therefore only depends on what features we are able to design for the parser, and is fully independent of the PSMT system.

The default output of the parser is the most probable reordering structure given its model. This is too restrictive for our purpose. The model would often not be relevant, if it expected a single word order

		To position			
		1	2	3	4
From position	0	-0.16	-1.03	-1.21	-1.39
	1		0.50	1.01	0.51
	2	0.91		1.16	1.48
	3	1.22	1.34		1.14
	4	0.23	0.12	-0.07	

Table 1: Illustration of the edge scores that the parser provides for the English sentence in figure 1. The highest scoring structure in bold.

during translation. Especially for longer sentences it would be unlikely to get this exact word order.

One of the characteristics of first-order MST parsing is that the score of each edge is independent of the rest of the structure. The parser therefore calculates a matrix of scores for each possible edge in the sentence before searching for the most probable combined structure. We exploit this behavior by emitting the matrix of edge scores instead of the best structure. This way, we can provide the decoder with scores for each possible reordering it can produce.

Table 1 gives an example of such an edge score matrix with the scores that the parser provided for the English sentence in figure 1. As an example, an edge from word 2 to word 4 has a cost of 1.48. Higher scores are better. Position 0 is the root position, which can only have out-going edges. The bold scores mark the most probable structure, which is the structure represented in figure 1.

4 Integration in PSMT

As described in the previous section, the decoder receives an edge score matrix in addition to the source sentence. This extra information is only used by a *word alignment scoring model*. This model returns a score each time a phrase is added to a translation

System	Lexical reordering	Tune <i>newstest2008</i>	Development <i>newstest2009</i>	Test <i>newstest2010</i>
Baseline	-	13.69	13.20	14.18
	+	13.98	13.74	14.80
Reordering Structure	-	14.04	13.76	14.69
	+	14.48	14.11	14.93
Oracle Reordering Structure	-	16.99	16.34	17.66
	+	17.17	16.67	18.06

Table 2: BLEU evaluation for the systems on different data sets.

hypothesis. Since the model returns a single score, it only introduces one additional parameter to system optimization. The score is calculated as the sum of scores for alignment links of adjacent target word positions within the phrase:

$$s_{wa} = \sum_{i=1}^n s(a_{i-1}, a_i) \quad (1)$$

where i is the target word position in a sentence of length n , a_i is the source word position it links to, and $s(a_{i-1}, a_i)$ is the score for target word positions $i - 1$ and i being aligned to source word positions a_{i-1} and a_i respectively. That is, the score of an edge going from a_{i-1} to a_i .

The scoring process is exemplified in figure 4. The left box illustrates the end of a translation hypothesis, and the right box illustrates the new phrase being added to this hypothesis. Only the reordering structure being scored at this stage is shown above the translation. Again we indicate the relationship between the word alignment and the edges using indexes. The index 0 shows the final link of the translation hypothesis, which decides where the new phrase links up. That is, the score for adding a new phrase is the sum of the score for connecting to the previous phrase (edge 1) and the scores for the phrase-internal edges (edges 2 and 3). These phrase-internal edges can be computed at phrase retrieval to save computation.

5 Experiments

Experiments were conducted from English to German, a language pair which exhibits substantial word order challenges.

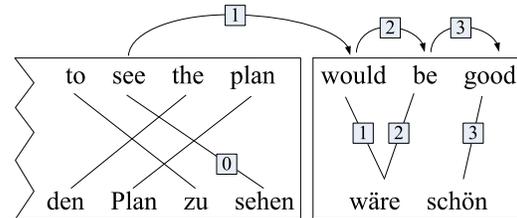


Figure 4: Illustration of the scoring done by the word alignment scoring model when extending the translation hypothesis with a new phrase.

5.1 Data

We use the English-German data from the Workshop on Statistical Machine Translation 2011 (WMT11)¹. This consists of 3.4/3.3 million words of parallel news data, 46.0/43.7 million words of parallel Europarl data, and 309 million monolingual words of europarl and news. We only use unique sentences from the monolingual data. We use *newstest2008* for tuning, *newstest2009* for development, and *newstest2010* for testing.

5.2 Reordering Structure Model Setup

The reordering structure model is created with the MSTParser², a dependency parser based on online discriminative learning. Since the reordering structure will contain many crossing edges, it is necessary to use non-projective parsing. There are no algorithmic modification to the parser. The only modification we make is that we make it emit the edge score matrix for each sentence that it parses. We only train the model on 25,000 sentences from the parallel news data to keep down computational costs. The English side of this subcorpus is dependency parsed using an MSTParser trained on the

¹ <http://www.statmt.org/wmt11/translation-task.html>

² <http://sourceforge.net/projects/mstparser/>

Penn Treebank converted to dependency structures, and grow-diag-final-and word alignments from creating the PSMT system are used to extract reordering structures in CoNLL format. The dependency parse is used to connect null-linked source words as described in section 2, and it provides word form, part-of-speech tag, and dependency relation features for the reordering structure parser. We did not do extensive feature selection for the reordering structure model, but excluding either of the three information levels decreased performance on a translation task.

5.3 PSMT Setup

All our PSMT systems are created with the Moses toolkit (Koehn et al., 2007). We use the baseline system from WMT11³ as our baseline with the small modifications that we use truecasing instead of lowercasing and recasing, and allow training sentences of up to 80 words. For our reordering experiments, we expand the baseline Moses system with the word alignment scoring model described in section 4. This is the only change to the baseline system. The baseline system got the best results with a distortion limit of 10, which we used for all experiments. The phrase table and the lexical reordering model is trained on the union of all parallel data with a max phrase length of 7, and the 5-gram language model is trained on the entire monolingual data set.

6 Results

6.1 Automatic Evaluation

Table 2 shows the results from automatic evaluation using the BLEU metric (Papineni et al., 2002). We report on the performance of the baseline and the reordering structure system with and without the lexical reordering model switched on. We use bootstrapping⁴ to test the significance of the results (Zhang et al., 2004). For all the data sets, the reordering structure system significantly outperforms the corresponding baseline system.

An interesting observation is that adding either the lexical reordering model or the reordering structure model to the baseline brings an improvement, and adding both improves performance even further.

³ See a detailed description at <http://www.statmt.org/wmt11/baseline.html>

⁴ <http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm>

	All edges	Non-monotone edges
Tune	69.47	11.81
Development	69.03	11.88
Test	71.32	14.51

Table 3: Unlabeled attachment scores for the reordering structure model on the data sets.

This indicates that the two models target different areas of reordering, and therefore they do not even each other out. Instead we see a cumulative effect where performance is increased even further.

The final system represented in table 2 called Oracle Reordering Structure gives an indication of the performance that is attainable if the predictions of the reordering structure model are improved. Here the gold standard reordering structure was added as a feature, so the parser obtained a 100% unlabeled attachment score on the data sets. The idea of this system is to see how much there is to gain if we have a perfect reordering structure model. However, the gold standard builds on erroneous automatic word alignments, which means that the “correct” structure may mislead the translation. Also this is the oracle best structure, not the oracle best edge score matrix, which is what is actually used by the system.

Table 3 shows the unlabeled attachment scores for the basic reordering structure model on the data sets. The scores are computed based on the most probable parse for each sentence, and they are reported for all edges and for non-monotone edges, i.e. edges going anywhere else than to the right neighboring word. These non-monotone edges are the most interesting edges, since they represent the reordering, and the prediction of these is very poor. We therefore expect that a fair part of the gain indicated by the Oracle Reordering Structure system is attainable through improvement of the reordering structure model.

6.2 Human Evaluation

In addition to the automatic evaluation, we also perform a small human evaluation using sentence translation ranking (Callison-Burch et al., 2010). We have two native German speakers rank the translations from the baseline system and the reordering structure system relative to each other. We evaluate on the first 100 sentences of the test corpus (*new-*

	Equally good	Baseline best	RS best
Evaluator 1	50	17	33
Evaluator 2	34	20	46
Average	42.0	18.5	39.5

Table 4: Human evaluation comparing the baseline and the reordering structure (RS) system on the first 100 sentences of the test set.

stest2010). On this subset, the baseline system gets a BLEU score of 10.55, and the reordering structure system gets 10.70.

The evaluators are presented with the source sentence and the two translations in randomized order. They are told to rank the systems from best to worst. Ties are allowed. The evaluators agreed on their judgements in 67 of the 100 sentences. Compared to an expected chance agreement of 1/3, the kappa coefficient is 0.505, which is much in line with findings from WMT10 (Callison-Burch et al., 2010).

Table 4 shows the results from the human evaluations. The translations from the reordering structure system were chosen as better than the baseline system more than twice as often as the reverse. This indicates that the reorderings introduced by the RS system may improve translation quality more than what the BLEU scores reflect. It has previously been reported that BLEU can be insensitive to word order improvements (Callison-Burch et al., 2007).

7 Analysis

An interesting aspect of the effect of the reordering structure model is the amount of word order differences it leads to. This information can be extracted from the word alignment information produced during a translation. Figure 5 shows the amount of reordering created by the baseline and reordering structure model systems on the development data. The figure shows that the reordering structure system introduces a lot more long distance reordering to the translation than the baseline systems. With lexical reordering on, it produces more than twice as many reorderings with a distance of 4 words, and more than 4 times as many reorderings with a distance of 8 words.

Here we also see a cumulative effect of combining the reordering structure model with the lexical

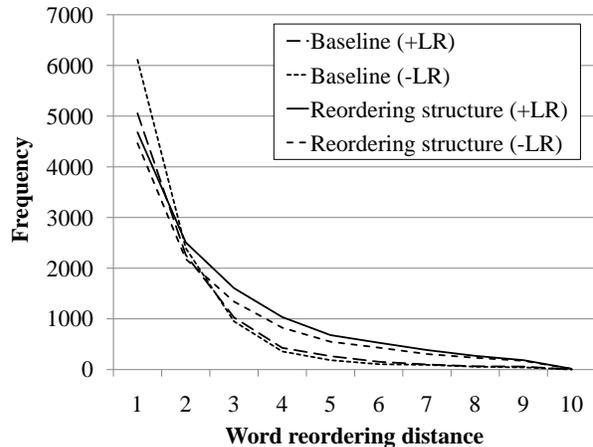


Figure 5: The amount of *non-monotone* decoding done by the baseline and reordering structure systems on the development data. More specifically, the number of target words representing a given reordering distance. LR specifies whether lexical reordering is in use.

reordering model. Together with the baseline system, the lexical reordering model does not introduce much long distance reordering, but combined with the reordering structure model the amount of long distance reordering gets boosted, also compared to the reordering structure model by itself.

8 Related Work

In recent years, the integration of syntactic knowledge into statistical machine translation has received much attention. The main motivation for this has been the need for better reordering of the words during translation. In a framework like synchronous context-free grammars (SCFGs) syntax is incorporated either on the source side (Liu et al., 2006), the target side (Galley et al., 2004), or both sides (Liu et al., 2009), and reordering is handled through the rules that constitute the building blocks for the translation. Such approaches have proven successful especially for language pairs which exhibit much non-local reordering (Zollmann et al., 2008; Birch et al., 2009). The hard constraints within the formalisms of these frameworks may however be too restrictive to handle frequently occurring aspects of parallel languages (Wellington et al., 2006; Søgaard and Kuhn, 2009; Galley and Manning, 2010).

In order to avoid such hard constraints introduced by the formalism, we place reordering information in the model to motivate certain word orders rather than prohibit others. That is, we create a reorder-

ing model that scores translation in parallel to other scoring models. This is much in line with Chiang (2010), who place syntactic correspondence information in the model as a soft constraint, but their approach is heavily tied to the SCFG framework, whereas our approach is framework independent.

A lot of work has been done on reordering in PSMT. The original approach deterred reordering by applying a *distortion penalty* for each word that is moved across (Koehn et al., 2003). Another approach is *lexicalized reordering*, which conditions the probability of moving a phrase in a certain direction on the lexical content of the phrase (Tillmann, 2004; Koehn et al., 2005). A third approach is *pre-translation reordering*, which reorders the source words in an attempt to assimilate the word order of the target language prior to translation. This can be done by supplying the decoder with a single permutation (Xia and McCord, 2004; Collins et al., 2005; Habash, 2007; Xu et al., 2009) or multiple weighted permutations (Zhang et al., 2007; Li et al., 2007; Elming, 2008; Ge, 2010). The present approach relates to the pre-translation reordering approaches in that it tries to predict the target word order from source sentence syntax. However, in these previous approaches, the source words are reordered prior to translation. This is not done in the current approach – instead, we use a decoder-internal model for scoring all generated reorderings.

The approach utilizes syntactic dependency relations to predict reorderings. This has previously been suggested to provide a better basis for reordering in machine translation due to higher inter-lingual phrasal cohesion than phrase structure (Fox, 2002). Much previous work has included dependency structure information in an SMT system. Quirk et al. (2005) use a source side dependency structure in their treelet SMT system, which translates from subtrees to strings. Galley & Manning (2009) use a dependency parser in a phrase-based setup for assigning a dependency structure to the target side during translation. This allows for the integration of a dependency language model directly into the system. Gimpel & Smith (2009; 2011) treat translation as a monolingual dependency parsing problem, creating a dependency structure over the translation during decoding. No syntactic structure is created during decoding in our approach. Instead the dependency

parser is used for the sole purpose of scoring the word order of the target sentence.

9 Conclusion and Future Work

We have introduced a new syntactically motivated discriminative reordering model. The model employs a standard data-driven dependency parser to predict reorderings. This is made possible by introducing a reordering structure. Within the framework of PSMT, we obtain substantial increases in translation quality both measured automatically and by human evaluators on an English to German task.

In the present work, we did very little feature selection and only provided word form, part-of-speech, and dependency relation information for the parser. In the future, we will experiment with additional features to improve the reordering structure model. In particular, we expect that more syntactic features will be beneficial. Also approaches such as second-order and stacked parsing may be helpful, since first-order parsing may be too weak to handle the complexities of the reordering structure. We also want to look closer at the features exploited by the standard MTSParser. These features are optimized to learn dependency structures, and they may not be optimal for learning the reordering structure.

One concern with the approach is that the model is trained against a gold standard which was extracted from automatic word alignment. This means that there will be a lot of noise in the training material. Also when training against the gold standard, all edges are considered equally important, but this may in fact not be the case for translation. Certain reorderings should always apply, while other may be stylistic and optional. A better way of training the model might be to train it as part of optimizing the PSMT system. This way, the system would be optimizing directly towards improving the word order of the translation. Due to the discriminative model’s large number of weights, using a discriminative algorithm to optimize the system (Watanabe et al., 2007; Chiang et al., 2008) would be an interesting option. This could either be done by learning from only the data set used for tuning the PSMT system, or by taking the model trained in the present work as a point of departure and revising the weights in the context of optimizing a PSMT system. We expect to pursue this direction in future work.

References

- A. Birch, P. Blunsom, and M. Osborne. 2009. A quantitative analysis of reordering phenomena. In *Proceedings of the WMT*.
- C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the WMT*.
- C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of WMT*.
- D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of EMNLP*.
- D. Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of ACL*.
- M. Collins, P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*.
- J. M. Crego and F. Yvon. 2009. Gappy translation units under left-to-right smt decoding. In *Proceedings of EAMT*.
- J. Elming. 2008. Syntactic reordering integrated with phrase-based smt. In *Proceedings of COLING*.
- H. J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP*.
- M. Galley and C. D. Manning. 2009. Quadratic-time dependency parsing for machine translation. In *Proceedings of ACL*.
- M. Galley and C. D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Proceedings of HLT-NAACL*.
- M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In *Proceedings of NAACL*, Boston, Massachusetts, USA.
- N. Ge. 2010. A direct syntax-driven reordering model for phrase-based machine translation. In *Proceedings of HLT-NAACL*.
- K. Gimpel and N. A. Smith. 2009. Feature-rich translation by quasi-synchronous lattice parsing. In *Proceedings of EMNLP*.
- K. Gimpel and N. A. Smith. 2011. Quasi-synchronous phrase dependency grammars for machine translation. In *Proceedings of EMNLP*.
- N. Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of MTSummit*.
- P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.
- P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of IWSLT*.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL: Demo and Poster Sessions*, Prague, Czech Republic.
- C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic.
- Y. Liu, Q. Liu, and S. Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL*.
- Y. Liu, Y. Lü, and Q. Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of ACL*.
- R. T. McDonald, F. Pereira, K. Ribarov, and J. Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP*.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, PA.
- C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of ACL*.
- A. Søgaard and J. Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of SSST*.
- C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*.
- T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of EMNLP-CoNLL*.
- B. Wellington, S. Waxmonsky, and I. D. Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of ACL*.
- F. Xia and M. McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING*. COLING.
- P. Xu, J. Kang, M. Ringgaard, and F. Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of NAACL*.
- Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of LREC*.
- Y. Zhang, R. Zens, and H. Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proceedings of IWSLT*, Trento, Italy.
- A. Zollmann, A. Venugopal, F. J. Och, and J. M. Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of COLING*.