

Source reordering using MaxEnt classifiers and supertags

Maxim Khalilov and Khalil Sima'an
 Institute for Logic, Language and Computation
 University of Amsterdam
 P.O. Box 94242
 1090 GE Amsterdam, The Netherlands
 {m.khalilov,k.simaan}@uva.nl

Abstract

Source language reordering can be seen as the preprocessing task of permuting the order of the source words in such a way that the resulting permutation allows as monotone a translation process as possible. We explore a simple but effective source reordering algorithm that works as a cascade of source string transforms, each consisting of swapping the positions of a single pair of adjacent words in order to unfold a candidate pair of crossing alignments. The decision to swap a pair of words is modelled as a binary classification task formulated as a log-linear model and trained under maximum entropy (MaxEnt). We experiment with features that consist of the local neighborhood of both words as well as lexico-syntactic representations known as supertags. Our experiments on the English-to-Dutch EuroParl translation task show that the cascaded alignment unfolding slightly improves the performance of a state-of-the-art phrase translation system that uses distance-based and lexicalized block-oriented reordering.

1 Introduction

The word-order divergence (also called distortion) between source-target sentence pairs is a major research topic in statistical machine translation (SMT). The problem of reordering in SMT has been attacked from different angles. Standard phrase-based translation models (Och and Ney, 2004) search for the best reordering option during decoding within a limited distortion space de-

finied using a local window of phrases, e.g., (Tillman, 2004). Hierarchical approaches, based on Inversion Transduction Grammar (ITG), e.g., (Wu and Wong, 1998; Chiang, 2005), explore yet a wider range of reorderings defined by the choice of swapping the order of sibling subtrees under each node in a binary parse-tree of the source/target sentence. Finally, source sentence reordering is a preprocessing task that aims at finding a permutation of the words of the source sentence that contains the least number of crossing alignments between source words and their target sentence counterparts. This paper is concerned with the task of source language reordering as a preprocessing task. Figure 1 depicts the translation from source string S to target string T with alignment a (solid line) and the alternative of source reordering S into S' followed by the translation $S' \rightarrow T$ with alignment a' (in dashed lines). Source reordering of S is as successful as much as it will yield a permutation S' that has as monotone an alignment a' as possible with T .

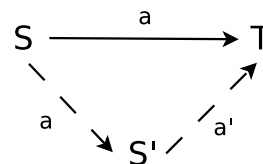


Figure 1: *Translation schemes with and without a reordering step.*

This problem can be seen as the task of learning from a word-aligned parallel corpus a model of source permutation from S to S' , where the latter has monotone alignment with T . The learning task aims at learning a model that minimizes the number of non-monotone alignments in the train-

ing data and expected future data of the same sort. However, this learning problem is hampered with the intractable complexity of computing the most probable permutation under a reasonable probabilistic model of the permutations (see (Tromble and Eisner, 2009) for the Linear Ordering Problem).

We look at an alternative view of source reordering where we view this as a simple cascade of string transforms. At the i^{th} step of the cascade, a single choice is made on the reordering of words/phrases in the source string and this lead to a permutation S'_i . While this general framework does not define a priori a constrained form of the reordering graph (as, e.g., the ITG trees), it has two attractive properties. Firstly, it is efficient enough to allow exploring a variety of such constraints on the order of reordering actions. For example, if a binary tree of the source is given, it is possible to define an ITG-constrained cascade of transforms (but such a binary tree is usually not available unless one commits to a certain source language parser). And secondly, it allows the conditioning of the i^{th} step in the cascade on aspects of previous permutation S'_{i-1} . In this paper we consider Swapping/Monotone as the main transform, and in the reported experiments we limit the swapping to individual words. The swapping decisions are formulated as a binary classification task trained under MaxEnt (Berger et al., 1996). We explore a variety of lexical features of the pair of words, including surrounding words, POS tags, and supertags (Clark and Curran, 2003).

The present exploration is driven by the observation that the existing phrase-based models are quite strong in local word reordering within a fixed window. In light of this observation, it becomes attractive to attempt bridging long distance reorderings as those found in English-Dutch translation.

Figure 2 shows an example of typical difference in word order between Dutch and English: in contrast to English, the Dutch verb normally appears in the end of the relative clause.



Figure 2: Example of long-distance reordering for Dutch-to-English translation.

It seems reasonable to aim at bridging the long distance reorderings by attempting resolving them or simply bringing the crossing words within range for phrase-based models. While our work aims at the general problem of learning how to untangle all crossing alignments, we do not a priori exclude the more pragmatic option of narrowing the distance between crossing alignments.

2 Related work

The idea of augmenting SMT by using a reordering step prior to translation has proved to be successful in improving translation quality. Clause restructuring performed with hand-crafted reordering rules for German-to-English and Chinese-to-English tasks are presented in (Collins et al., 2005) and (Wang et al., 2007), respectively. In (Khalilov, 2009; Xia and McCord, 2004) the fundamental problem of word ordering is addressed exploiting syntactic representations of source and target texts. In (Costa-jussà and Fonollosa, 2006) source and target word order harmonization is done using well-established SMT techniques and without the use of syntactic knowledge.

Other reordering models operate in a non-deterministic way providing the decoder with multiple word orders. For example, the MaxEnt reordering model described in (Xiong et al., 2006) provides a hierarchical phrasal reordering system integrated within a CKY-style decoder. In (Galley and Manning, 2008) the authors present an extension of the famous MSD model (Tillman, 2004) able to handle long-distance word-block permutations. One more example of a system performing reordering in this way can be found in (Crego and Mariño, 2007), where syntactic structure on the source side is exploited to reorder the input of a word lattice in an unweighted manner, slightly expanding the monotonic search space. In (Tromble and Eisner, 2009) a $O(n^3)$ chart parsing algorithm aimed to find the best reordering of possible word permutations is described and applied for the German-English language pair.

3 Source reordering by cascaded transforms

For a translation system that employs source reordering as preprocessing step, two training stages are needed given a word-aligned parallel corpus of source and target sentences $\{S - T\}$:

- Conceptually, creating a monolingual word-aligned parallel corpus $\{S - a - S'\}$ from $\{S - T\}$ where S' is obtained by unfolding the crossing brackets between S and T and replacing every word t_i in T with the word it is aligned with $s_{a(i)}$ in S .¹ Some heuristics are needed to fully conduct this stage. We describe this step in subsection 3.1. The word-aligned monotonized parallel corpus is used for training our source reordering cascaded system of transforms. We describe the pre-processing system in subsection 3.2. Notice that $\{S'\}$ features are not used in the present version of the system.
- Once the cascaded system of transforms is available it is used to transform all training source sentences in the original parallel corpus $\{S - to - T\}$ into $\{S'_g - to - T\}$, where S'_g is the guessed permutation of S chosen by the source reordering system. Word-alignment is performed on the latter parallel corpus and that corpus is used for training a phrase-based SMT system.

In this section we describe our cascaded source reordering system that employs MaxEnt classifiers.

3.1 Creating the monotonized parallel corpus

Monotonization of the parallel training corpus is done on the basis of the "grow-diag-final" many-to-many alignment. It is modified to one-to-one in the way described in the next lines. If a source word is aligned to two or more target words, the most probable link given lexical probability model (Brown et al., 1993) is chosen, while the others are omitted. Source side words are permuted within the scope of a sentence such that all the crossing links in the alignment are unfolded. The resulting training corpus is called $\{S'\}$. It is explicitly implied that the number and the set of words in $\{S'\}$ and in $\{S\}$ coincide.

3.2 Cascaded source reordering by classification

Initially, we formulate the problem of defining the set of permutations and selecting the most likely permutation given a source sentence as a conditional probability that we break down into a cascade of transforms where the final permutation S'

is the result of a sequence of permutations $S'_0 = S, S'_1, \dots, S'_n = S'$:

$$\arg \max_{S'} P(S' | S) = \arg \max_{S'} \prod_{i=1}^n P(S'_i | S'_{i-1}) \quad (1)$$

Three issues arise in this definition: (1) How to define the transform S'_{i-1} to S'_i for every i in this cascade, (2) How to define the cascade order, and (3) When to stop the cascade of transforms. For defining each transform $P(S'_i | S'_{i-1})$ in the cascade we narrow this down to a decision on swapping only two words in S'_{i-1} . The decision to swap words is defined as a classification task using MaxEnt which we discuss next (see Section 3.3). This avoids calculating the probability $P(S'_i | S'_{i-1})$ explicitly by focusing on calculating the single decision probability (word pair probability) that leads to S'_i from S'_{i-1} as we will explain in more detail next (subsection 3.3). For the second issue we choose for a simple yet effective strategy by scanning the current source sentence S'_{i-1} from the last position where a word was swapped in the previous step in the cascade. In the model application step, a pair of words belonging to "Swapped" class swaps their order, while words marked with "Monotone" label keep the original positions. The algorithm stops once it reaches the end of the string or there are no word swapping needed according to the MaxEnt classifier.

3.3 A MaxEnt Classifier for Word Swapping

The MaxEnt classifier considers every adjacent pair of words w_1 and w_2 in source sentence S and assigns a conditional probability to conduct an operation $O \in \{Swapped, Monotone\}$, as follows:

$$p(O | \phi_1(S), \phi_2(S)) \quad (2)$$

where $\phi_1(\{S\})$ and $\phi_2(\{S\})$ are feature functions of the context of w_1 and w_2 , respectively.

The classification problem here is how to separate pairs of words into "Swapped" or "Monotone" categories, given a set of features describing word preference to stay in the current position or swap it with its counterpart. MaxEnt classifier operates with conditional probabilities $p(O | \phi_1(S), \phi_2(S))$, using reordering instances extracted from the training corpus.

The MaxEnt classifier was trained and applied using open-source Maximum Entropy Modeling

¹Hereafter, we use brackets to refer to corpora and notations without brackets to refer to sentences.

Toolkit².MaxEnt model is trained using 90 iterations of the Limited-Memory Variable Metric method.

Features Feature functions $\phi_1(S)$ and $\phi_2(S)$, apart from the word instances themselves, include:

- *Context-based features.* Source-side context 3-grams, 2-grams, and 1-grams, describing left- and right-hand side neighborhood of the first and the second word. These features can be seen as a contextual predictor describing word preference to change its current location.
- *Morpho-syntactic features.* Linguistic syntax is believed to be helpful in MT, thoroughly handling word order dependencies and accurately modeling many systematic differences between word orders of languages (Bonnie, 1994). Syntax is introduced into MaxEnt reordering system using POS tags and supertags which are assigned for each word according to Lexicalized Tree-Adjoining Grammar (Abeillè and Rambow, 2000) and Combinatory Categorical Grammar (Clark and Curran, 2003), as described in (Hassan et al., 2007). In other words, supertags in compact form describe the way to the highest node of the parse tree.

On the one hand, introducing of syntax into MaxEnt reordering system as a feature contributes to consistency of reordering decisions with grammatical representation of the natural language. On the other hand, it reduces data sparseness by means of clustering words according to their syntactic categories.

3.4 Ambiguous alignment

A high-precision alignment intersection matrix is used to find word swapping examples in the training corpus. The use of intersection matrix allows finding disambiguous crossings indicating that their unfolding leads to monotonization of the alignment matrix (see Figure 3).

However, there is a number of alignments in which one or both words subject to swapping are also aligned to a third word beyond the target-side reordering limits. In some situations it shows that proposed alignment unfolding can destroy a monolithic bilingual unit.

²http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

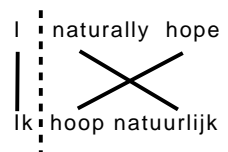


Figure 3: Example of a disambiguous alignment crossing.

The extraction step involves taking a decision on assigning swapping probability for those units. For example, in Figure 4, swapping of “allow” and “us” will possibly break a construction which should be seen as a single and correct translation block (namely, “allows us/maakt het ons mogelijk”)³.

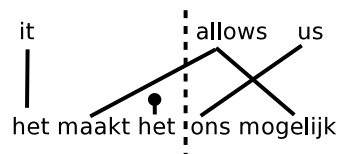


Figure 4: Example of an ambiguous alignment crossing.

We propose two alternative techniques to handle this alignment ambiguity:

1. *Exclude ambiguous crossings.* The source side words which are considered candidates for swapping are always marked with “Monotone” label.
2. *Redistribution of probability mass.* The “strength” of alignment links is estimated with lexical probabilities (Brown et al., 1993) that shows how often a certain source-side word is translated into a particular target-side word. Total reordering probability mass is redistributed between “Swapped” and “Monotone” depending on the values of lexical probabilities.

For example provided in Figure 4 in case of first strategy application, the probability of swapping or keeping the words monotone are found as follows:

$$p(\text{allows,us,} \text{ "Monotone"}) = 1$$

$$p(\text{allows,us,} \text{ "Swapped"}) = 0$$

For the “redistribution of probability mass” strategy, these probabilities are defined as follows:

³Dutch word “het” is aligned to NULL.

$$\begin{aligned}
p(\textit{allows,us,} \textit{Monotone}) &= \\
&= \frac{p_{LP}(\textit{allows,maakt})}{\{p_{LP}(\textit{allows,maakt}) + p_{LP}(\textit{allows,mogelijk})\}}
\end{aligned}$$

$$\begin{aligned}
p(\textit{allows,us,} \textit{Swapped}) &= \\
&= \frac{p_{LP}(\textit{allows,mogelijk})}{\{p_{LP}(\textit{allows,maakt}) + p_{LP}(\textit{allows,mogelijk})\}}
\end{aligned}$$

where $p_{LP}(word_1, word_2)$ is the lexical probability of translating a source-side $word_1$ by a target-side $word_2$.

In the following sections, we explore the impact of this hard decision in reordering accuracy and in translation quality.

4 Baseline translation system

Rather than translating single words, phrase-based systems (Och and Ney, 2002) work with (in principle) arbitrarily large phrase pairs (also called blocks) acquired from word-aligned parallel data under a set of constraints (Zens et al., 2002). A bilingual phrase (which in the context of SMT do not necessarily coincide with their linguistic analogies) is any aligned pair of m source words and n target words that satisfies two basic constraints: (1) words are consecutive along both sides of the bilingual phrase and (2) no word on either side of the phrase is aligned to a word outside the phrase (Och and Ney, 2004). The probability of the phrases is estimated by counts of their appearance in the training corpus. The finite, fixed inventory of phrases obtained from a parallel corpus is stored in a “phrase-table”.

The translation of a source sentence in phrase-based system proceeds by “tiling” the source sentence with source-side phrases in the phrase table. Every tiling of the source sentence with a sequence of source-side phrases provides a bag of aligned target phrases, which can be reordered using a reordering model and the language model statistics. The different possible tilings of the source sentence lead to a set (represented efficiently as a lattice) of output hypotheses in the target-language. The highest scoring hypothesis is selected using a decoder that performs the optimization under strict pruning regimes.

Two reordering methods are considered: a distance-based distortion model (see 4.1) and lexicalized MSD block-oriented model (see 4.2).

4.1 Distance-based

A simple distance-based reordering model default for Moses system is the first reordering technique under consideration. This model provides the decoder with a cost linear to the distance between words that should be reordered.

4.2 MSD

A lexicalized block-oriented data-driven MSD reordering model (Tillman, 2004) considers three different orientation types: monotone (M), swap (S), and discontinuous (D). MSD model conditions reordering probabilities on the word context of each phrase pair and considers decoding process a block sequence generation process with the possibility of swapping a pair of word blocks. Notice that in the experiments conducted within the framework of this study a MSD model was used together with a distance-based reordering model.

5 Experiments and results

This section describes experiments carried out to evaluate the proposed reordering model.

5.1 Data

The experiment results were obtained using the English-Dutch corpus of the European Parliament Plenary Session transcription (*EuroParl*). Basic training corpus statistics can be found in Table 1.

	Dutch	English
Sentences	1.2 M	1.2 M
Words	32.9 M	33.0 M
Average sentence length	27.20	27.28
Vocabulary	228 K	104 K

Table 1: Basic statistics of the English-Dutch EuroParl training corpus.

Development and test datasets were randomly chosen from the corpus and consisted of 500 and 1,000 sentences, respectively. Both were provided with 1 reference translation.

5.2 Experimental setup

SMT system used in the experiments is implemented within the open-source MOSES toolkit (Koehn et al., 2007). Standard training and weight tuning procedures which were used to build our system are explained in details on the MOSES web page:

<http://www.statmt.org/moses/>. Word alignment was estimated with GIZA++ tool⁴ (Och, 2003), coupled with the mkcls⁵ (Och, 1999) tool, which allows for statistical word clustering for better generalization.

N -gram target language model was estimated using the SRI LM toolkit (Stolcke, 2002) and is a 5-gram model with modified Kneser-Ney discounting.

Evaluation conditions were case-sensitive and included punctuation marks.

5.3 Systems

We contrast five alternative system configurations differing in feature set and reordering example extraction algorithm, along with two Moses-based baseline systems (Table 2).

Both baseline systems implement non-deterministic reordering algorithm providing the decoder with multiple word order options.

5.4 Results

Source reordering analysis In the first step of system evaluation we estimated the total number of crossings found in the word alignment. Table 3 shows these values found in the last 10,000 lines of the alignment intersection matrix between different variations of the source and the target languages⁶. All the numbers are calculated on the re-aligned corpora.

The first row shows the number of alignment crossings found in the original (unmonotonized) corpus. There are 1.8 - 1.9 swappings per sentence and the maximum reduction of crossings in comparison with the original corpus is relatively small

⁴code.google.com/p/giza-pp/

⁵<http://www.fjoch.com/mkcls.html>

⁶A smaller portion of the corpus is used for analysis in order to reduce evaluation time.

($\approx 4\%$ for the $\{S'_gMER + ST\}$ corpus). However, moving words closer to each other can have a positive impact on translation process, since it can potentially bring them into the scope of the Moses distance-based distortion system.

Source	#of crossings
$\{S\}$	18,898
$\{S'_gMER\}$	18,458
$\{S'_gMER + Ext\}$	18,496
$\{S'_gMER + LexProb\}$	18,590
$\{S'_gMER + ST\}$	18,136
$\{S'_gMER + POS + ST\}$	18,762

Table 3: Number of crossings in word alignment intersection.

As the next step of analysis, we calculate BLEU and NIST scores obtained on the basis of the last 10,000 lines of the training corpus and using monotonized parallel $\{S'_g\}$ corpus as a reference. Table 4 reports the results of this evaluation.

Source	BLEU	NIST
$\{S'_gMER\}$	75.40	15.33
$\{S'_gMER + Ext\}$	75.41	15.33
$\{S'_gMER + LexProb\}$	75.17	15.27
$\{S'_gMER + ST\}$	75.38	15.31
$\{S'_gMER + POS + ST\}$	76.39	15.44

Table 4: Automatic translation scores of $\{S'_g\}'$ estimated with MaxEnt models vs. directly monotonized corpus ($\{S\}$).

This evaluation can be seen as an alternative metric of reordering algorithm effectiveness. $MER+POS+ST$ systems outperforms the simplest MaxEnt configuration by about 1 BLEU point.

It is worth noticing that the system config-

System	Features	Alignment ambiguity
Baseline 1	Moses + Distance-based	
Baseline 2	Moses + Distance-based + MSD	
MER	Lexical	-
MER+Ext	Lexical	Exclude ambiguous crossings
MER+LexProb	Lexical	Lexical probabilities
MER+ST	Lexical + supertags	-
MER+POS+ST	Lexical + POS + supertags	-

Table 2: Translation systems under consideration.

uration that provides better results in terms both of reordering effectiveness (Table 4) and translation performance (Table 5), namely *MER+POS+SuperTags*, is the same that produces a lesser reduction in the number of crossings (Table 3). This observation supports the claim that efficient alignment unfolding leads to a higher-quality translation.

Figure 5 demonstrates an example of the sentences that presumably benefitted from the monotonization of the source part of the parallel corpus. English infinitive “*to lead*” appears in the end of the clause in the Dutch reference translation. Moses-based system coupled with a MSD reordering model translates the verb as “*om het voortouw te nemen*” (“*to take the initiative*”), while *MER+POS+ST* system is able to produce the correct translation, even if the particle “*to*” has not been moved to the end of the clause.

Translation scores Table 5 shows the results of translation, both starting with baseline configurations, and contrasts them with the performance shown by the MaxEnt systems. Best scores are placed in cells filled with grey.

Neither of two algorithms intended to handle alignment ambiguity manages to outperform the baselines. However, according to results, “Exclude ambiguous crossings” strategy is believed to be the best way to resolve alignment ambiguity. Supertag features do not lead to any gain in terms of transla-

tion scores, however when coupled with POS features they allow outperforming the baseline system by about 0.3 BLEU points.

6 Conclusions

In this paper we explored a cascaded approach to source sentence reordering that works by swapping adjacent pairs of words. The decision to swap adjacent words is framed as a classification task using the MaxEnt framework that works with feature functions. The feature functions we employed are lexical as well as lexico-syntactic (POS tags and supertags). Our English-Dutch best system performs as well (or slightly better) than the Moses baseline although our system did not succeed in monotonizing a large majority of the crossing alignments. This hints at the fact that our source reordering is simply bringing long-distance crossing alignments into smaller neighborhoods that can be tackled using standard phrase reordering mechanisms implemented in Moses.

In this exploratory work we employed a simple cascaded approach that works from left to right over the source sentence, and we limited the attention to swapping pairs of words. Obviously the cascaded framework we present can be implemented in terms of phrases instead of words, and it can incorporate ITG tree-driven inversion transforms that might constrain the reordering space to better motivated reorderings from a linguistic perspective. We intend to explore these ideas within

Src: *it was not enough to reassure me that he would be the right person to lead the Commission during the next five years*

Ref.: *het heeft mij er niet van overtuigd dat hij de juiste persoon is om de komende vijf jaar de Commissie te leiden*

Baseline 2: *het niet genoeg is voor mij verzekeren dat hij de juiste persoon is om het voortouw te nemen in de Commissie tijdens de komende vijf jaar*

Rrd src: *it was not enough to reassure me that he would be the right person to the Commission during the next five years lead*

MER+POS+ST: *het niet genoeg was voor mij verzekeren dat hij de juiste persoon is om de Commissie tijdens de komende vijf jaar te leiden*

Figure 5: Translation example.

System	Internal reordering	External reordering	Dev	Test	
			BLEU	BLEU	NIST
Baseline 1	Distance-based	-	23.88	24.04	6.29
Baseline 2	Distance-based + MSD	-	24.07	24.04	6.28
MER	Distance-based	MER	23.59	24.27	6.30
MER+Ext	Distance-based	MER+external links	23.69	24.04	6.28
MER+LexProb	Distance-based	MER+LexProb	23.11	23.72	6.21
MER+ST	Distance-based	MER+ST	23.68	23.90	6.28
MER+POS+ST	Distance-based	MER+POS+SuperTags	23.89	24.34	6.31

Table 5: English-to-Dutch experimental results.

the next stages of this work and we will expand on employing source-language syntactic structure for this purpose.

A further extension of this approach is to employ language model features from S' to guide the cascaded transforms towards better permutations of source sentence S . We also intend to study algorithms that aim at global optimization (Viterbi-like) of the conditional probability $P(S' | S)$ instead of the current local classification approach that isolated each step in the cascade.

References

- Abeillé, A. and O. Rambow. 2000. *Tree Adjoining Grammars: formalisms, linguistic analysis and processing*. CSLI, Stanford, CA, USA.
- Berger, A., S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1(22):39–72.
- Bonnie, B. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 4(20):597–663.
- Brown, P., V. Della Pietra, S. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311.
- Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.
- Clark, St. and J. R. Curran. 2003. Log-linear models for wide-coverage CCG parsing. In *Proceedings of EMNLP 2003*, pages 97–104.
- Collins, M., P. Koehn, and I. Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL 2005*, pages 531–540.
- Costa-jussà, M. R. and J. A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of HLT/EMNLP'06*, pages 70–76.
- Crego, J. M. and J. B. Mariño. 2007. Syntax-enhanced N-gram-based SMT. In *Proceedings of MT SUMMIT XI*.
- Galley, M. and Ch. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP'08*, pages 848–856.
- Hassan, H., K. Sima'an, and A. Way. 2007. Supertagged phrase-based statistical machine translation. In *Proceedings of ACL 2007*, pages 288–295.
- Khalilov, M. 2009. *New statistical and syntactic models for machine translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, October.
- Koehn, P., 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 2007*, pages 177–180.
- Och, F. and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, pages 295–302.
- Och, F. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.
- Och, F. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the ACL 1999*, pages 71–76.
- Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*, pages 160–167.
- Stolcke, A. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of SLP 2002*, pages 901–904.
- Tillman, C. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004*, pages 101–104.
- Tromble, R. and J. Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of EMNLP'09*, pages 1007–1016.
- Wang, C., M. Collins, and Ph. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP-CoNLL 2007*, pages 737–745.
- Wu, D. and H. Wong. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of ACL-COLING 1998*, pages 1408–1415.
- Xia, F. and M. McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, pages 508–514.
- Xiong, D., Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL'06*, pages 521–528.
- Zens, R., F. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Proceedings of KI: Advances in Artificial Intelligence*, pages 18–32.