ILLC-UvA translation system for EMNLP-WMT 2011

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

In this paper we describe the Institute for Logic, Language and Computation (University of Amsterdam) phrase-based statistical machine translation system for English-to-German translation proposed within the EMNLP-WMT 2011 shared task. The main novelty of the submitted system is a syntax-driven pre-translation reordering algorithm implemented as source string permutation via transfer of the source-side syntax tree.

1 Introduction

For the WMT 2011 shared task, ILLC-UvA submitted two translations (primary and secondary) for the English-to-German translation task. This year, we directed our research toward addressing the word order problem for statistical machine translation (SMT) and discover its impact on output translation quality. We reorder the words of a sentence of the source language with respect to the word order of the target language and a given source-side parse tree. The difference from the baseline Moses-based translation system lies in the pre-translation step, in which we introduce a discriminative source string permutation model based on probabilistic parse tree transduction.

The idea here is to permute the order of the source words in such a way that the resulting permutation allows as monotone a translation process as possible is not new. This approach to enhance SMT by using a reordering step prior to translation has proved to be successful in improving translation quality for many translation tasks, see (Genzel, 2010; Costa-jussà and Fonollosa, 2006; Collins et al., 2005), for example.

The general problem of source-side reordering is that the number of permutations is factorial in n, and learning a sequence of transductions for explaining a source permutation can be computationally rather challenging. We propose to address this problem by defining the source-side permutation process as the learning problem of how to transfer a given source parse tree into a parse tree that minimizes the divergence from target word order.

Our reordering system is inspired by the direction taken in (Tromble and Eisner, 2009), but differs in defining the space of permutations, using local probabilistic tree transductions, as well as in the learning objective aiming at scoring permutations based on a log-linear interpolation of a local syntax-based model with a global string-based (language) model.

The reordering (novel) and translation (standard) components are described in the following sections. The rest of this paper is structured as follows. After a brief description of the phrase-based translation system in Section 2, we present the architecture and details of our reordering system (Section 3), Section 4 reviews related work, Section 5 reports the experimental setup, details the submissions and discusses the results, while Section 6 concludes the article.

2 Baseline system

2.1 Statistical machine translation

In SMT the translation problem is formulated as selecting the target translation *t* with the highest probability from a set of target hypothesis sentences for

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 413–419,
Edinburgh, Scotland, UK, July 30–31, 2011.

the source sentence s: $\hat{t} = \arg\max_t \{ p(t|s) \} = \arg\max_t \{ p(s|t) \cdot p(t) \}$.

2.2 Phrase-based translation

While first systems following this approach performed translation on the word level, modern state-of-the-art phrase-based SMT systems (Och and Ney, 2002; Koehn et al., 2003) start-out from a word-aligned parallel corpus working with (in principle) arbitrarily large phrase pairs (also called blocks) acquired from word-aligned parallel data under a simple definition of translational equivalence (Zens et al., 2002).

The conditional probabilities of one phrase given its counterpart is estimated as the relative frequency ratio of the phrases in the multiset of phrase-pairs extracted from the parallel corpus and are interpolated log-linearly together with a set of other model estimates:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (1)$$

where a feature function h_m refer to a system model, and the corresponding λ_m refers to the relative weight given to this model.

A phrase-based system employs feature functions for a phrase pair translation model, a language model, a reordering model, and a model to score translation hypothesis according to length. The weights λ_m are optimized for system performance (Och, 2003) as measured by BLEU (Papineni et al., 2002).

Apart from the novel syntax-based reordering model, we consider two reordering methods that are widely used in phrase-based systems: a simple distance-based reordering and a lexicalized block-oriented data-driven reordering model (Tillman, 2004).

3 Architecture of the reordering system

We approach the word order challenge by including syntactic information in a pre-translation reordering framework. This section details the general idea of our approach and details the reordering model that was used in English-to-German experiments.

3.1 Pre-translation reordering framework

Given a word-aligned parallel corpus, we define the source string permutation as the task of learning to unfold the crossing alignments between sentence pairs in the parallel corpus. Let be given a source-target sentence pair $s \rightarrow t$ with word alignment set a between their words. Unfolding the crossing instances in a should lead to as monotone an alignment a' as possible between a permutation s' of s and the target string t. Conducting such a "monotonization" on the parallel corpus gives two parallel corpora: (1) a source-to-permutation parallel corpus ($s \rightarrow s'$) and (2) a source permutation-to-target parallel corpus ($s' \rightarrow t$). The latter corpus is word-aligned automatically again and used for training a phrase-based translation system, while the former corpus is used for training our model for pre-translation source permutation via parse tree transductions.

In itself, the problem of permuting the source string to unfold the crossing alignments is computationally intractable (see (Tromble and Eisner, 2009)). However, different kinds of constraints can be made on unfolding the crossing alignments in a. A common approach in hierarchical SMT is to assume that the source string has a binary parse tree, and the set of eligible permutations is defined by binary ITG transductions on this tree. This defines permutations that can be obtained only by at most inverting pairs of children under nodes of the source tree.

3.2 Conditional tree reordering model

Given a parallel corpus with string pairs $s \rightarrow t$ with word alignment a, the source strings s are parsed, leading to a single parse tree τ_s per source string. We create a *source permuted* parallel corpus $s \rightarrow s'$ by unfolding the crossing alignments in a without/with syntactic tree to provide constraints on the unfolding.

Our model aims at learning from the source permuted parallel corpus $s \rightarrow s'$ a probabilistic optimization $\arg\max_{\pi(s)} P(\pi(s) \mid s, \tau_s)$. We assume that the set of permutations $\{\pi(s)\}$ is defined through a finite set of local transductions over the tree τ_s. Hence, we view the permutations leading from s to s' as a sequence of local tree transduc-

tions $\tau_{s_0^{'}} \rightarrow \ldots \rightarrow \tau_{s_n^{'}}$, where $s_0^{'} = s$ and $s_n^{'} = s^{'}$, and each transduction $\tau_{s_{i-1}^{'}} \rightarrow \tau_{s_i^{'}}$ is defined using a tree transduction operation that *at most permutes the children of a single node in* $\tau_{s_{i-1}^{'}}$ *as defined next.*

A local transduction $\tau_{s_{i-1}^{'}} \rightarrow \tau_{s_i^{'}}$ is modelled by an operation that applies to a single node with address x in $\tau_{s_{i-1}^{'}}$, labeled N_x, and may permute the ordered sequence of children α_x dominated by node x. This constitutes a direct generalization of the ITG binary inversion transduction operation. We assign a conditional probability to each such local transduction:

$$P(\tau_{s_i^{'}} \mid \tau_{s_{i-1}^{'}}) \approx P(\pi(\alpha_x) \mid N_x \rightarrow \alpha_x, C_x) \qquad (2)$$

where $\pi(\alpha_x)$ is a permutation of α_x (the ordered sequence of node labels under x) and C_x is a local tree context of node x in tree $\tau_{s_{i-1}^{'}}$. One wrinkle in this definition is that the number of possible permutations of α_x is factorial in the length of α_x. Fortunately, the source permuted training data exhibits only a fraction of possible permutations even for longer α_x sequences. Furthermore, by conditioning the probability on local context, the general applicability of the permutation is restrained.

In principle, if we would disregard the computational cost, we could define the probability of the sequence of local tree transductions $\tau_{s_0^{'}} \rightarrow \ldots \rightarrow \tau_{s_n^{'}}$ as

$$P(\tau_{s_0^{'}} \rightarrow \ldots \rightarrow \tau_{s_n^{'}}) = \prod_{i=1}^{n} P(\tau_{s_i^{'}} \mid \tau_{s_{i-1}^{'}}) \qquad (3)$$

The problem of calculating the most likely permutation under this kind of transduction probability is intractable because every local transduction conditions on local context of an intermediate tree[1]. Hence, we disregard this formulation and in practice we take a pragmatic approach and greedily select at every intermediate point $\tau_{s_{i-1}^{'}} \rightarrow \tau_{s_i^{'}}$ the single most likely local transduction that can be conducted on any node of the current intermediate tree $\tau_{s_{i-1}^{'}}$. The individual steps are made more effective by interpolating the term in Equation 2 with string probability ratios:

$$P(\pi(\alpha_x) \mid N_x \rightarrow \alpha_x, C_x) \times \left(\frac{P(s_{i-1}^{'})}{P(s_i^{'})} \right) \qquad (4)$$

The rationale behind this interpolation is that our source permutation approach aims at finding the optimal permutation $s^{'}$ of s that can serve as input for a subsequent translation model. Hence, we aim at tree transductions that are syntactically motivated that also lead to improved string permutations. In this sense, the tree transduction definitions can be seen as an efficient and syntactically informed way to define the space of possible permutations.

We estimate the string probabilities $P(s_i^{'})$ using 5-gram language models trained on the $s^{'}$ side of the source permuted parallel corpus $s \rightarrow s^{'}$. We estimate the conditional probability $P(\pi(\alpha_x) \mid N_x \rightarrow \alpha_x, C_x)$ using a Maximum-Entropy framework, where feature functions are defined to capture the permutation as a class, the node label N_x and its head POS tag, the child sequence α_x together with the corresponding sequence of head POS tags and other features corresponding to different contextual information.

We were particularly interested in those linguistic features that motivate reordering phenomena from the syntactic and linguistic perspective. The features that were used for training the permutation system are extracted for every internal node of the source tree that has more than one child:

- *Local tree topology.* Sub-tree instances that include parent node and the ordered sequence of child node labels.
- *Dependency features.* Features that determine the POS tag of the head word of the current node, together with the sequence of POS tags of the head words of its child nodes.
- *Syntactic features.* Two binary features from this class describe: (1) whether the parent node is a child of the node annotated with the same syntactic category, (2) whether the parent node is a descendant of a node annotated with the same syntactic category.

[1]Note that a single transduction step on the current tree $\tau_{s_{i-1}^{'}}$ leads to a forest of trees $\tau_{s_i^{'}}$ because there can be multiple alternative transduction rules. Hence, this kind of a model demands optimization over many possible sequences of trees, which can be packed into a sequence of parse-forests with transduction links between them.

4 Related work

The integration of linguistic syntax into SMT systems offers a potential solution to reordering problem. For example, syntax is successfully integrated into hierarchical SMT (Zollmann and Venugopal, 2006). In (Yamada and Knight, 2001), a set of tree-string channel operations is defined over the parse tree nodes, while reordering is modeled by permutations of children nodes. Similarly, the tree-to-string syntax-based transduction approach offers a complete translation framework (Galley et al., 2006).

The idea of augmenting SMT by a reordering step prior to translation has often been shown to improve 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 (Xia and McCord, 2004; Khalilov, 2009) word reordering is addressed by 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 provide 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. Coming up-to-date, in (PVS, 2010) an effective application of data mining techniques to syntax-driven source reordering for MT is presented.

Different syntax-based reordering systems can be found in (Genzel, 2010). In this system, reordering rules capable to capture many important word order transformations are automatically learned and applied in the preprocessing step.

Recently, Tromble and Eisner (Tromble and Eisner, 2009) define source permutation as the word-ordering learning problem; the model works with a preference matrix for word pairs, expressing preference for their two alternative orders, and a corresponding weight matrix that is fit to the parallel data. The huge space of permutations is then structured using a binary synchronous context-free grammar (Binary ITG) with $O(n^3)$ parsing complexity, and the permutation score is calculated recursively over the tree at every node as the accumulation of the relative differences between the word-pair scores taken from the preference matrix. Application to German-to-English translation exhibits some performance improvement.

5 Experiments and submissions

Design, architecture and configuration of the translation system that we used in experimentation coincides with the Moses-based translation system (Baseline system) described in details on the WMT 2011 web page[2].

This section details the experiments carried out to evaluate the proposed reordering model, experimental set-up and data.

5.1 Data

In our experiments we used EuroParl v6.0 German-English parallel corpus provided by the organizers of the evaluation campaign.

A detailed statistics of the training, development, internal (*test int.*) and official (*test of.*) test datasets can be found in Table 1. The development corpus coincides with the 2009 test set and for internal testing we used the test data proposed to the participants of WMT 2010.

"ASL" stands for average sentence length. All the sets were provided with one reference translation.

Data		Sent.	Words	Voc.	ASL
train	En	1.7M	46.0M	121.3K	27.0
train	Ge	1.7M	43.7M	368.5K	25.7
dev	En	2.5K	57.6K	13.2K	22.8
test int.	En	2.5K	53.2K	15.9K	21.4
test of.	En	3.0K	74.8K	11.1K	24.9

Table 1: *German-English EuroParl corpus (version 6.0).*

Apart from the German portion of the EuroParl parallel corpus, two additional monolingual corpora from news domain (the News Commentary corpus (NC) and the News Crawl Corpus 2011 (NS)) were

[2] http://www.statmt.org/wmt11/baseline.html

used to train a language model for German. The characteristics of these datasets can be found in Table 2. Notice that the data were not de-duplicated.

Data		Sent.	Words	Voc.	ASL
NC	Ge	161.8M	3.9G	136.7M	23.9
NS	Ge	45.3M	799.4M	3.0M	17.7

Table 2: *Monolingual German corpora used for target-side language modeling.*

5.2 Experimental setup

Moses toolkit (Koehn et al., 2007) in its standard setting was used to build the SMT systems:

- GIZA++/mkcls (Och, 2003; Och, 1999) for word alignment.
- SRI LM (Stolcke, 2002) for language modeling. A 3-gram target language model was estimated and smoothed with modified Kneser-Ney discounting.
- MOSES (Koehn et al., 2007) to build an unfactored translation system.
- the Stanford parser (Klein and Manning, 2003) was used as a source-side parsing engine[3].
- For maximum entropy modeling we used the maxent toolkit[4].

The discriminative syntactic reordering model is applied to reorder training, development, and test corpora. A Moses-based translation system (corpus realignment included[5]) is then trained using the reordered input.

5.3 Internal results and submissions

The outputs of two translation system were submitted. First, we piled up all feature functions into a single model as described in Section 3. It was our "secondary" submission. However, our experience tells that the system performance can increase if the set of patterns is split into partial classes conditioned on the current node label (Khalilov and Sima'an, 2010). Hence, we trained three separate MaxEnt models for the categories with potentially high reordering requirements, namely NP, $SENT$ and $SBAR(Q)$. It was defines as our "primary" submission.

The ranking of submission was done according to the results shown on internal testing, shown in Table 3.

System	BLEU dev	BLEU test	NIST test
Baseline	11.03	9.78	3.78
Primary	11.07	10.00	3.79
Secondary	10.92	9.91	3.78

Table 3: *Internal testing results.*

5.4 Official results and discussion

Unfortunately, the results of our participation this year were discouraging. The primary submission was ranked 30th (12.6 uncased BLEU-4) and the secondary 31th (11.2) out of 32 submitted systems.

It turned out that our preliminary idea to extrapolate the positive results of English-to-Dutch translation reported in (Khalilov and Sima'an, 2010) to the WMT English-to-German translation task was not right.

Analyzing the reasons of negative results during the post-evaluation period, we discovered that translation into German differs from English-to-Dutch task in many cases. In contrast to English-to-Dutch translation, the difference in terms of automatic scores between the internal baseline system (without external reordering) and the system enhanced with the pre-translation reordering is minimal. It turns out that translating into German is more complex in general and discriminative reordering is more advantageous for English-to-Dutch than for English-to-German translation.

A negative aspect influencing is the way how the rules are extracted and applied according to our approach. Syntax-driven reordering, as described in this paper, involves large contextual information applied cumulatively. Under conditions of scarce data, alignment and parsing errors, it introduces noise to the reordering system and distorts the feature prob-

[3]The parser was trained on the English treebank set provided with 14 syntactic categories and 48 POS tags.

[4]http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[5]Some studies show that word re-alignment of a monotonized corpus gives better results than unfolding of alignment crossings (Costa-jussà and Fonollosa, 2006).

ability space. At the same time, many reorderings can be performed more efficiently based on fixed (hand-crafted) rules (as it is done in (Collins et al., 2005)). A possible remedy to this problem is to combine automatically extracted features with fixed (hand-crafted) rules. Our last claims are supported by the observations described in (Visweswariah et al., 2010).

During post-evaluation period we analyzed the reasons why the system performance has slightly improved when separate MaxEnt models are applied. The outline of reordered nodes for each of syntactic categories considered ($SENT$, $SBAR(Q)$ and NP) can be found in Table 4 (the size of the corpus is 1.7 M of sentences).

Category	# of applications
NP	497,186
SBAR(Q)	106,243
SENT	221,568

Table 4: *Application of reorderings for separate syntactic categories.*

It is seen that the reorderings for NP nodes is higher than for $SENT$ and $SBAR(Q)$ categories. While SENT and SBAR(Q) reorderings work analogously for Dutch and German, our intuition is that German has more features that play a role in reordering of NP structures than Dutch and there is a need of more specific features to model NP permutations in an accurate way.

6 Conclusions

This paper presents the ILLC-UvA translation system for English-to-German translation task proposed to the participants of the EMNLP-WMT 2011 evaluation campaign. The novel feature that we present this year is a source reordering model in which the reordering decisions are conditioned on the features from the source parse tree.

Our system has not managed to outperform the majority of the participating systems, possibly due to its generic approach to reordering. We plan to investigate why our approach works well for English-to-Dutch and less well for the English-to-German translation in order to discover more generic ways for learning discriminative reordering rules. One possible explanation of the bad results is a high sparseness of automatically extracted rules that does not allow for sufficient generalization of reordering instances.

In the future, we plan (1) to perform deeper analysis of the dissimilarity between English-to-Dutch and English-to-German translations from SMT perspective, and (2) to investigate linguistically-motivated ideas to extend our model such that we can bring about some improvement to English-to-German translation.

7 Acknowledgements

Both authors are supported by a VIDI grant (nr. 639.022.604) from The Netherlands Organization for Scientific Research (NWO).

References

M. Collins, P. Koehn, and I. Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL'05*, pages 531–540.

M. R. Costa-jussà and J. A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of HLT/EMNLP'06*, pages 70–76.

M. Galley and Ch. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP'08*, pages 848–856.

M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thaye. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of COLING/ACL'06*, pages 961–968.

D. Genzel. 2010. Aumotatically learning source-side reordering rules for large scale machine translation. In *Proc. of COLING'10*, pages 376–384, Beijing, China.

M. Khalilov and K. Sima'an. 2010. A discriminative syntactic model for source permutation via tree transduction. In *Proc. of the Fourth Workshop on Syntax and Structure in Statistical Translation (SSST-4) at COLING'10*, pages 92–100, Beijing (China), August.

M. Khalilov. 2009. *New statistical and syntactic models for machine translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, October.

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the ACL'03*, pages 423–430.

Ph. Koehn, F. Och, and D. Marcu. 2003. Statistical phrase-based machine translation. In *Proceedings of the HLT-NAACL 2003*, pages 48–54.

Ph. 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 2007*, pages 177–180.

F. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL'02*, pages 295–302.

F. Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of ACL 1999*, pages 71–76.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL'03*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318.

A. PVS. 2010. A data mining approach to learn reorder rules for SMT. In *Proceedings of NAACL/HLT'10*, pages 52–57.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of SLP'02*, pages 901–904.

C. Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*, pages 101–104.

R. Tromble and J. Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of EMNLP'09*, pages 1007–1016.

K. Visweswariah, J. Navratil, J. Sorensen, V. Chenthamarakshan, and N. Kambhatla. 2010. Syntax based reordering with automatically derived rules for improved statistical machine translation. In *Proc. of COLING'10*, pages 1119–1127, Beijing, China.

C. Wang, M. Collins, and Ph. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP-CoNLL'07*, pages 737–745.

F. Xia and M. McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of COLING'04*, pages 508–514.

D. Xiong, Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL'06*, pages 521–528.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL'01*, pages 523–530.

R. Zens, F. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Proceedings of KI: Advances in Artificial Intelligence*, pages 18–32.

A. Zollmann and A. Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of NAACL'06*, pages 138–141.