DERIVING BENEFIT FROM A GENERALIZED SYNTAX-BASED REORDERING

*Maxim Khalilov, José A.R. Fonollosa**

TALP Research Center
Universitat Politècnica de Catalunya
Campus Nord UPC, 08034,
Barcelona, Spain

Mark Dras

Centre for Language Technology
Macquarie University
North Ryde NSW 2109,
Sydney, Australia

RESUMEN

In this study we describe a syntax-based word reordering technique for n-gram-based statistical machine translation (SMT). The proposed distortion model operates with generalized unlexicalized rules and aims to order source language words so that translation is close to monotonic, simplifying the translation process. In the final step, we apply a translation units blending strategy, combining bilingual tuples extracted from the parallel corpora with monotone and reordered source parts.

Experiments are reported on the BTEC corpus from tourist domain for the Arabic-English translation task, the proposed tuples blending technique significantly outperformes the monotone system.

1. INTRODUCTION

The word disparity problem between source and target languages is a crucial point for many modern SMT systems. Several researchers [1, 2] consider the reordering model to hold great scope for translation quality improvement, and even as a bottleneck bounding further SMT progress. At the same time, there is a controversy about whether a statistical system can benefit from syntactic information, expressed in form of Part-of-Speech (POS) tags, shallow or dependency parse trees.

Though, the word class-based reordering patterns are part of Och's Alignment Template system [1], the classical phrase-based approach does not entirely solve the reordering problem. This problem leads to particularly bad translation when dealing with languages having distinct word orders and linguistic typology. An example of such language pair is Arabic and English: apart from a difference in verbal morphology and the presence of enclitics, they have distinct language topology schemes (VSO for Arabic and SVO for English). Where a monotone translation approach in many cases is not able to deal with such a reordering disparity, a constituent tree structure can be used.

*This work has been funded by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project) and by Generalitat de Catalunya under project TECNOPARLA.

There have already been some efforts to solve this problem both in purely statistical way or involving additional informational sources. The state-of-the-art phrase-based SMT system Moses[1] implements a distance based distortion model [3] as does a word alignment-based MSD (Monotone, Swap and Discontinuous) reordering model as shown in [4].

A linguistically motivated reordering model employing a monotonic search graph extension was proposed in [5]. In [2] another method of word reordering for N-gram-based MT systems was introduced: a monotone sequence of source words is translated into the reordered sequence using the well established mechanism of SMT.

A set of hand-crafted reordering rules demonstrated a significant improvement for German to English translation as shown in [6]. In [7] the authors present a hybrid system for French-English translation, based on the automatically deriving rewrite patterns extraction from a parse tree and phrase alignments. Inspired by this idea we intend to apply a subtree target-to-source mapping as was done in [8], where a two-side subtree transfer was introduced as a part of a syntax-driven SMT. Afterwards, the translation task, realized by a n-gram-based system is reformulated to translate from the reordered source language, that lead to a mutual word order monotonization, shorter translation units and improved translation.

The rest of the paper is organized as follows: Section 2 outlines the n-gram-based SMT system. Section 3 introduces the syntax-based reordering. In Section 4 we present the results and contrast them with an alternative reordering techniques and Section 5 presents the conclusions.

2. NGRAM-BASED SMT

The n-gram-based approach regards translation as a stochastic process maximizing the joint probability $p(f, e)$, leading to a decomposition based on bilingual n-grams, which we call *tuples*, that are extracted from a word-to-word alignment (performed with GIZA++ tool[2]). Tuples are extracted according to the following constraints [9]:

[1] www.statmt.org/moses/
[2] http://code.google.com/p/giza-pp/

- a monotonic segmentation of each bilingual sentence pair is produced
- no word in a tuple is aligned to words outside of it
- no smaller tuples can be extracted without violating the previous constraints

Figure 1 shows an example of tuple monotonic extraction (*regular* technique resulting in one tuple), contrasted with the *unfolding* technique (resulting in three tuples), that allow producing a different bilingual n-gram language model with reordered source words.

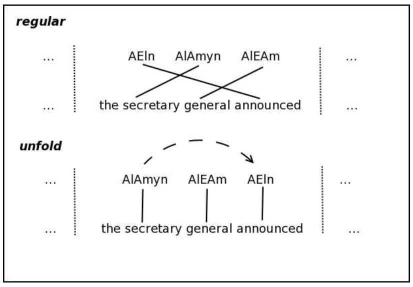

Figura 1. *Example of tuples extraction.*

The Ngram-based translation system implements a log-linear model in which a foreign language sentence $f_1^J = f_1, f_2, ..., f_J$ is translated into another language $e_1^I = e_1, e_2, ..., e_I$ by searching for the translation hypothesis $\hat{e}_1^I$ maximizing a log-linear combination of several feature models [9].

A *translation model (TM)* approximates the joint probability between source and target languages capturing bilingual context, as shown in equation 1:

$$p(S,T) = \prod_{k=1}^{K} p((\tilde{s},\tilde{t})_k | (\tilde{s},\tilde{t})_{k-N+1}, ..., (\tilde{s},\tilde{t})_{k-1}) \quad (1)$$

where s refers to source, t to target, and $(\tilde{s},\tilde{t})_k$ to the k^{th} tuple of a given bilingual sentence pair segmented in K tuples.

The rest of the system models are: a *target language model*, a *POS target language model*, a *word bonus model*, a *source-to-target lexicon model* and a *target-to-source lexicon model*. For more details refer to [9].

We used the MARIE beam-search decoder [10] allowing for efficient pruning of the search space, threshold pruning, histogram pruning and hypothesis recombination. Given the development set and references, the log-linear combination of weights was adjusted using a simplex optimization method (with the optimization criteria of the highest BLEU score) and an n-best re-ranking.

3. SYNTAX-BASED REORDERING

In this study we simulate a situation when the reordering system has access to both the source and target language shallow parsers using word alignment intersection as a ‘bridge’ between two languages. We used the Stanford Parser as a parsing engine[3] [11] and the Arabic and English Penn Treebank sets (26 POS/23 constituent categories for Arabic Treebank and 48 POS and 14 syntactic tags for English Treebank).

Syntax-based reordering as described in this paper operates with a Context-Free Grammar (CFG), where each branch of the parse tree is represented as follows:

$$X \rightarrow \langle N, T, R, S \rangle \quad (2)$$

where N refers to a set of constituents and POS tags, T is a set of terminals (lexicon), R stands for a mapping from N to $(T \bigcup N)^*$ of the form $N_i \rightarrow \gamma$ (γ is a sequence of terminals and non-terminals) and S is the start variable.

Reordering patterns are expressed in the form *NP@0 VP@1* $\rightarrow$ *VP@1 NP@0 p1*, that means that a sequence of constituents *NP@0 VP@1* should be reordered like *VP@1 NP@0* with probability `p1`. Note that here the number of constituents indicates the order of their appearance in the source part of the pattern.

3.1. Rules extraction

The reordering rule extraction procedure consists of the following steps:

Step 1 align the monotone corpus and find the intersection of src-to-trg and trg-to-src word alignments (construct the projection matrix P);

Step 2 parse the source and the target parts of the parallel corpus;

Step 3 convert the parse trees to the CFG form;

Step 4 extract reordering patterns from the parallel non-isomorphic CFG-trees basing on the word alignment intersection and considering POS and constituents equally;

Step 5 estimate and normalize the number of reordering pattern instances.

Figure 2 shows an example of the rule extraction procedure (Step 4) for a parallel sentence

Arabic:	*h*A*	*hW*	*fndq*	*+k*
English:	this	is	your	hotel

Given two parse trees and word alignment intersection expressed in form of projection matrix

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

[3] Generally speaking, the source and targets formal grammars, as well as the parsing mechanisms can differ.

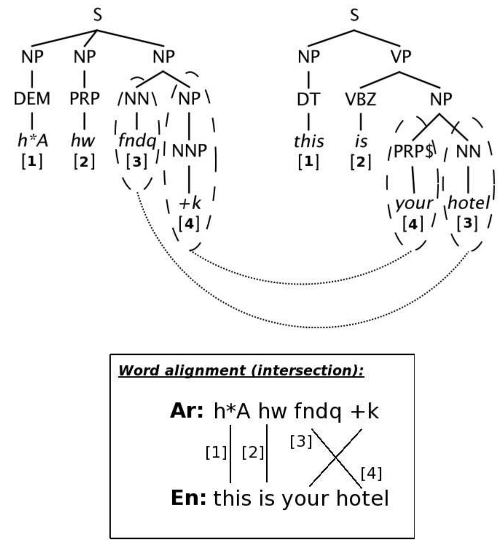

Figura 2. *Rules extraction step.*

the directly extracted reordering rule is *NN@0 NP@1* $\rightarrow$ *NP@1 NN@0* and since the "NP" node leads to the leaf *"+k"* through the "NNP" POS tag, one more unlexicalized rule can be induced: *NN@0 NNP@1* $\rightarrow$ *NNP@1 NN@0*. It is worth noticing that the left side of the reordering pattern is always monotone and the right side can be monotone or reordered.

If a word that is aligned in only one direction (source to target or target to source) appears in the branch that is considered as a candidate to be involved into a reordering pattern, it does not exert influence on the alignment projection matrix.

3.2. Organizing reordering rules

Once the list of reordering patterns is extracted, they are organized following the strategy similar to the one proposed in [7] for generalized rules. All the rules that appear less than k times are directly discarded (in experiments we used the threshold $k = 3$). A probability of alternative patterns is estimated basing on absolute counting of their appearance in the training corpus and the most probable rules are stored.

Ambiguous rules are pruned out according to the higher probability principle, for example, for the pair of patterns *NP@0 VP@1 ->VP@1 NP@0 p1*, *VP@0 NP@1 ->NP@1 VP@0 p2*, leading to the recurring contradiction, one rule will be removed depending on the ratio $p2/p1$).

Finally, the reordering table (analogous to the "r-table" as stated in [8]) is a set of POS- and constituent-based patterns allowing for reordering and monotone distortion.

3.3. Source-side monotonization

Rules application is performed as a bottom-up parse tree traversal applying the longest possible rule, i.e. among a set of nested rules, the rule with a longest left-side covering is selected (e.g. in case of *NN JJ RB* sequence appearance and two reordering rules presence *NN@0 JJ@1 ->...* and *NN@0 JJ@1 RB@2 ->...*, the former pattern will be applied).

Figure 3 shows the example of the reordered source-side parse tree with the applied pattern *NN@0 NNP@1 ->NNP@1 NN@0*. The resulting Arabic sentence is

```
h*A hW +k fndq
```

that more closely matches the order of the target language and reflects *possessive pronoun - noun* typical English word order.

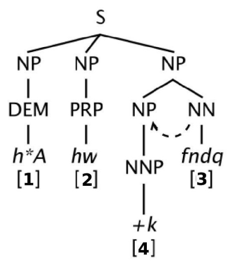

Figura 3. *Reordered source-side parse tree.*

3.4. Tuples blending

In terms of this study, we operate exclusively with generalized (i.e. unlexicalized) reordering rules, that along with improved translation units, cause errors induced by a certain number of grammatical exceptions which can be easily found in any language. Therefore, after the corpus with reordered source part is aligned, two sets of tuples are extracted basing on the reordered and monotone alignment matrices. In the final stage of the translation model construction, the bilingual units from these sets are combined following the criterion of maximizing the number of tuples at the sentence level. This technique entails more tuples involvement into TM contruction that provides better bilingual generalization (shorter translation units have higher probability of appearance in the translanting corpus than the longer ones).

4. EXPERIMENTAL SETTINGS, RESULTS AND COMPARISON WITH UNFOLDING METHOD

The experiments were performed on the BTEC'08 corpus from the tourist domain. A basic corpus statistics can be found in table 1.

The BLEU score obtained on the development set (489 lines, $3{,}7K$ running words and 6 reference translations)

	Arabic	English
Sentences	23.7 K	23.7 K
Words	166.0 K	183.9 K
Average sentence length	7.75	6.99
Vocabulary	10.8 K	6.8 K

Tabla 1. Basic statistics of the BTEC training corpus.

as the final point of the simplex optimization procedure and the translation results done on the test set (500 lines, $4{,}1K$ running words and 16 references) are summarized in table 2. We consider four translation systems: *monotone* and *reordered* configurations that correspond to the systems involving the parallel corpora with monotone and reordered source parts, respectively; a *blending* model as described in subsection 3.4; and the alternative *UC* method, that include the unfold algorithm of tuples extraction and constrained distance-based distortion model used on the decoding step (as described in [12]).

	dev BLEU	test BLEU	# tuples
Monotone	40.55	43.78	135.855
Reordered	41.05	45.15	143.934
Blending	43.20	47.92	170.572
UC	43.61	47.46	163.755

Tabla 2. Summary of the experimental results.

For the tuples *blending* configuration, about 40 % of the tuples came from the system with reordered source part. Curiously, more tuples were generated by this system than by *unfolded* algorithm (the number of bilingual units generated by the former system is the maximum theoretical possible with invariable alignment). We explain this phenomena by several "noisy" tuples generated by the reordered system under conditions of a lack of training material.

In terms of BLEU score, the unfolded and the combined reordered-monotone system demonstrate comparable performance significantly outperforming both the monotone and the syntactically reordered SMT systems.

5. CONCLUSIONS AND FUTURE WORK

The proposed syntactically motivated reordering coupled with the bilingual units blending method shows competitive performance comparing with an alternative reordering method on the small Arabic-English corpus preserving potential power of fully or partially lexicalized reordering rules using. However, more profound analysis of generated bilingual units and their impact on the translation quality is needed and will be done in the near future.

6. BIBLIOGRAFÍA

[1] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, y D. Radev, "A Smorgasbord of Features for Statistical Machine Translation," in *Proceedings of HLT/NAACL04*, 2004, pp. 161–168.

[2] M. R. Costa-jussà y J. A. R. Fonollosa, "Statistical machine reordering," in *Proceedings of the HLT/EMNLP 2006)*, 2006.

[3] Ph. Koehn, F. J. Och, y D. Marcu, "Statistical phrase-based machine translation," in *Proceedings of the HLT-NAACL 2003*, 2003, pp. 48–54.

[4] C. Tillmann y T. Zhang, "A localized prediction model for statistical machine translation," in *Proceedings of the 43rd Annual Meeting on ACL 2005*, 2005, pp. 557–564.

[5] J. M. Crego y J. B. Mariño, "Improving statistical MT by coupling reordering and decoding," *Machine Translation*, vol. 20(3), pp. 199–215, 2007.

[6] M. Collins, Ph. Koehn, y I. Kučerová, "Clause restructuring for statistical machine translation," in *Proceedings of the 43rd Annual Meeting on ACL 2005*, 2005, pp. 531–540.

[7] F. Xia y M. McCord, "Improving a statistical mt system with automatically learned rewrite patterns," in *Proceedings of the COLING 2004*, 2004.

[8] K. Imamura, H. Okuma, y E. Sumita, "Practical approach to syntax-based statistical machine translation," in *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, 2005, pp. 267–274.

[9] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. A. Fonollosa, y M. R. Costa-juss, "N-gram based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[10] J. M. Crego, J. B. Mariño, y A. de Gispert, "An ngram-based statistical machine translation decoder," in *Proceedings of INTERSPEECH05*, 2005.

[11] D. Klein y C. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting of the ACL 2003*, 2003, pp. 423–430.

[12] M. R. Costa-jussà, J. M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J. A. Fonollosa, J. B. Mariño, y R. E. Banchs, "TALP phrase-based system and TALP system combination for IWSLT 2006," in *Proceedings of the IWSLT 2006*, 2006, pp. 123–129.