Coupling hierarchical word reordering and decoding in phrase-based statistical machine translation

Maxim Khalilov and **José A.R. Fonollosa**
Universitat Politècnica de Catalunya
Campus Nord UPC, 08034,
Barcelona, Spain
{khalilov,adrian}@gps.tsc.upc.edu

Mark Dras
Macquarie University
North Ryde NSW 2109,
Sydney, Australia
madras@ics.mq.edu.au

Abstract

In this paper, we start with the existing idea of taking reordering rules automatically derived from syntactic representations, and applying them in a preprocessing step before translation to make the source sentence structurally more like the target; and we propose a new approach to hierarchically extracting these rules. We evaluate this, combined with a lattice-based decoding, and show improvements over state-of-the-art distortion models.

1 Introduction

One of the big challenges for the MT community is the problem of placing translated words in a natural order. This issue originates from the fact that different languages are characterized by different word order requirements. The problem is especially important if the distance between words which should be reordered is high (global reordering); in this case the reordering decision is very difficult to take based on statistical information due to dramatic expansion of the search space with the increase in number of words involved in the search process.

Classically, statistical machine translation (SMT) systems do not incorporate any linguistic analysis and work at the surface level of word forms. However, more recently MT systems are moving towards including additional linguistic and syntactic informative sources (for example, source- and/or target-side syntax) into word reordering process. In this paper we propose using a syntactic reordering system operating with fully, partially and non- lexicalized reordering patterns, which are applied on the step

prior to translation; the novel idea in this paper is in the derivation of these rules in a hierarchical manner, inspired by Imamura et al (2005). Furthermore, we propose generating a word lattice from the bilingual corpus with the reordered source side, extending the search space on the decoding step. A thorough study of the combination of syntactical and word lattice reordering approaches is another novelty of the paper.

2 Related work

Many reordering algorithms have appeared over the past few years. Word class-based reordering was a part of Och's Alignment Template system (Och et al., 2004); the main criticism of this approach is that it shows bad performance for the pair of languages with very distinct word order. The state-of-the-art SMT system Moses implements a distance-based reordering model (Koehn et al., 2003) and a distortion model, operating with rewrite patterns extracted from a phrase alignment table (Tillman, 2004).

Many SMT models implement the brute force approach, introducing several constrains for the reordering search as described in Kanthak et al. (2005) and Crego et al. (2005). The main criticism of such systems is that the constraints are not lexicalized. Recently there has been interest in SMT exploiting non-monotonic decoding which allow for extension of the search space and linguistic information involvement. The variety of such models includes a constrained distance-based reordering (Costa-jussà et al., 2006); and a constrained version of distortion model where the reordering search problem is tackled through a set of linguistically motivated rules used during decoding (Crego and Mariño, 2007).

A quite popular class of reordering algorithms is a monotonization of the source part of the parallel corpus prior to translation. The first work on this approach is described in Nießen and Ney (2004), where morpho-syntactic information was used to account for the reorderings needed. A representative set of similar systems includes: a set of hand-crafted reordering patterns for German-to-English (Collins et al., 2005) and Chinese-English (Wang et al., 2007) translations, emphasizing the distinction between German/Chinese and English clause structure; and statistical machine reordering (SMR) technique where a monotonization of the source words sequence is performed by translating them into the reordered one using well established SMT mechanism (Costa-jussà and Fonollosa, 2006). Coupling of SMR algorithm and the search space extension via generating a set of weighted reordering hypotheses has demonstrated a significant improvement, as shown in Costa-jussà and Fonollosa (2008).

The technique proposed in this study is most similar to the one proposed for French-to-English translation task in Xia and McCord (2004), where the authors present a hybrid system for French-English translation based on the principle of automatic rewrite patterns extraction using a parse tree and phrase alignments. We propose using a word distortion model not only to monotonize the source part of the corpus (using a different approach to rewrite rule organization from Xia and McCord), but also to extend the search space during decoding.

3 Baseline phrase-based SMT systems

The reference system which was used as a translation mechanism is the state-of-the-art Moses-based SMT (Koehn et al., 2007). The training and weights tuning procedures can be found on the Moses web page[1].

Classical phrase-based translation is considered as a three step algorithm: (1) the source sequence of words is segmented into phrases, (2) each phrase is translated into the target language using a translation table, (3) the target phrases are reordered to fit the target language. The probabilities of the phrases are estimated by relative frequencies of their appearance in the training corpus.

In baseline experiments we used a phrase dependent lexicalized reordering model, as proposed in Tillmann (2004). According to this model, monotonic or reordered local orientations enriched with probabilities are learned from training data. During decoding, translation is viewed as a monotone block sequence generation process with the possibility to swap a pair of neighbor blocks.

4 Syntax-based reordering coupled with word graph

Our syntax-based reordering system requires access to source and target language parse trees and word alignments intersections.

4.1 Notation

Syntax-based reordering (SBR) operates with source and target parse trees that represent the syntactic structure of a string in source and target languages according to a Context-Free Grammar (CFG).

We call this representation *"CFG form"*. We formally define a CFG in the usual way as $G = \langle N, T, R, S \rangle$, where N is a set of nonterminal symbols (corresponding to source-side phrase and part-of-speech tags); T is a set of source-side terminals (the lexicon), R is a set of production rules of the form $\eta \rightarrow \gamma$, with $\eta \in N$ and γ, which is a sequence of terminal and nonterminal symbols; and $S \in N$ is the distinguished symbol.

The reordering rules then have the form

$$\eta_0@0 \ldots \eta_k@k \rightarrow$$
$$\eta_{d_0}@d_0 \ldots \eta_{d_k}@d_k | Lexicon | p_1 \quad (1)$$

where $\eta_i \in N$ for all $0 \leq i \leq k$; $(d_o \ldots d_k)$ is a permutation of $(0 \ldots k)$; *Lexicon* comes from the source-side set of words for each η_i; and p_1 is a probability associated with the rule. Figure 1 gives two examples of the rule format.

4.2 Rules extraction

Concept. Inspired by the ideas presented in Imamura et al. (2005), where monolingual correspondences of syntactic nodes are used during decoding, we extract a set of bilingual patterns allowing for reordering as described below:

(1) align the monotone bilingual corpus with GIZA++ (Och and Ney, 2003) and find the intersection of direct and inverse word alignments, resulting in the construction of the projection matrix P (see below));

(2) parse the source and the target parts of the parallel corpus;

(3) extract reordering patterns from the parallel non-isomorphic CFG-trees based on the word alignment intersection.

Step 2 is straightforward; we explain aspects of Steps 1 and 3 in more detail below. Figures 1 and 2 show an example of the extraction of two lexicalized rules for a parallel Arabic-English sentence:

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

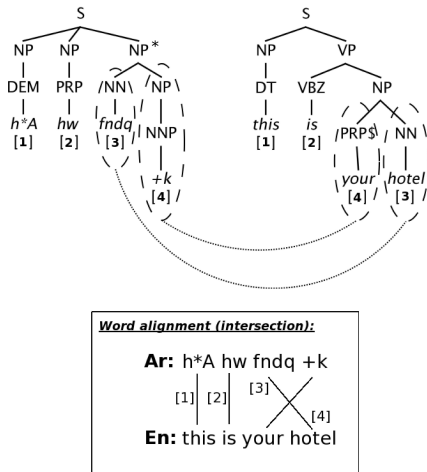We use this below in our explanations.

Figure 2: Example of subtree transfer and reordering rules extraction.

Projection matrix. Bilingual content can be represented in the form of words or sequences of words depending on the syntactic role of the corresponding grammatical element (constituent or POS).

Given two parse trees and a word alignment intersection, a projection matrix P is defined as an $M \times N$ matrix such that M is the number of words in the target phrase; N is the number of words in the source phrase; and a cell (i, j) has a value based on the alignment intersection — this value is zero if word i and word j do not align, and is a unique non-zero link number if they do.

For the trees in Figure 2,

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 4 & 0 \end{pmatrix}$$

Unary chains. Given an unary chain of the form $X \rightarrow Y$, rules are extracted for each level in this chain. For example given a rule

$$NP@0\, ADVP@1 \rightarrow ADVP@1\, NP@0$$

and a unary chain "$ADVP \rightarrow AD$", a following equivalent rule will be generated

$$NP@0\, AD@1 \rightarrow AD@1\, NP@0.$$

The role of target-side parse tree. Although reordering is performed on the source side only, the target-side tree is of great importance: the reordering rules can be only extracted if the words covered by the rule are entirely covered by both a node in the source and in the target trees. It allows the more accurate determination of the covering and limits of the extracted rules.

4.3 Rules organization

Once the list of fully lexicalized reordering patterns is extracted, all the rules are progressively processed reducing the amount of lexical information. These *initial rules* are iteratively expanded such that each element of the pattern is generalized until all the lexical elements of the rule are represented in the form of fully unlexicalized categories. Hence, from each

NN@0 NP@1 → NP@1 NN@0 | NN@0 << fndq >> NP@1 << +k >> | p
NN@0 NNP@1 → NNP@1 NN@0 | NN@0 << fndq >> NNP@1 << +k >> | p

Figure 1: Directly extracted rules.

initial pattern with N lexical elements, $2^N - 2$ partially lexicalized rules and 1 general rule are generated. An example of the process of delexicalization can be found in Figure 3.

Thus, finally three types of rules are available: (1) fully lexicalized (initial) rules, (2) partially lexicalized rules and (3) unlexicalized (general) rules.

On the next step, the sets are processed separately: patterns are pruned and ambiguous rules are removed. All the rules from the fully lexicalized, partially lexicalized and general sets that appear fewer than k times are directly discarded (k is a shorthand for k_{ful}, k_{part} and k_{gener}). The probability of a pattern is estimated based on relative frequency of their appearance in the training corpus. Only one the most probable rule is stored. Fully lexicalized rules are not pruned ($k_{ful} = 0$); partially lexicalized rules that have been seen only once were discarded ($k_{part} = 1$); the thresholds k_{gener} was set to 3: it limits the number of general patterns capturing rare grammatical exceptions which can be easily found in any language.

Only the one-best reordering is used in other stages of the algorithm, so the rule output functioning as an input to the next rule can lead to situations reverting the change of word order that the previously applied rule made. Therefore, the rules that can be ambiguous when applied sequentially during decoding are pruned according to the higher probability principle. For example, for the pair of patterns with the same lexicon (which is empty for a general rule leading to a recurring contradiction *NP@0 VP@1 → VP@1 NP@0 p1*, *VP@0 NP@1 → NP@1 VP@0 p2*), the less probable rule is removed.

Finally, there are three resulting parameter tables analogous to the "r-table" as stated in (Yamada and Knight, 2001), consisting of POS- and constituent-based patterns allowing for reordering and mono-

tone distortion (examples can be found in Table 5).

4.4 Source-side monotonization

Rule application is performed as a bottom-up parse tree traversal following two principles:

(1) the longest possible rule is applied, i.e. among a set of nested rules, the rule with a longest left-side covering is selected. For example, in the case of the appearance of an *NN JJ RB* sequence and presence of the two reordering rules

$$NN@0 \ JJ@1 \rightarrow ... \ \text{and}$$

$$NN@0 \ JJ@1 \ RB@2 \rightarrow ...$$

the latter pattern will be applied.

(2) the rule containing the maximum lexical information is applied, i.e. in case there is more than one alternative pattern from different groups, the lexicalized rules have preference over the partially lexicalized, and partially lexicalized over general ones.
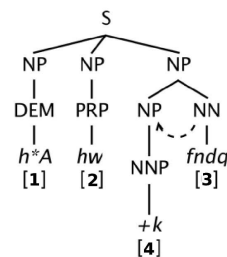

Figure 4: *Reordered source-side parse tree.*

Once the reordering of the training corpus is ready, it is realigned and new more monotonic alignment is passed to the SMT system. In theory, the word links from the original alignment can be used, however, due to our experience, running GIZA++ again results in a better word alignment since it is easier to learn on the modified training example.

Example of correct local reordering done with the SBR model can be found in Figure 4.

| Initial rule: | $NN@0 \ NP@1 \rightarrow NP@1 \ NN@0 \mid NN@0 << fndq >> NP@1 << +k >> \mid p_1$ |

| Part. lexic. rules: | $NN@0 \ NP@1 \rightarrow NP@1 \ NN@0 \mid NN@0 << fndq >> NP@1 << - >> \mid p_2$ |
| | $NN@0 \ NP@1 \rightarrow NP@1 \ NN@0 \mid NN@0 << - >> NP@1 << +k >> \mid p_3$ |

| General rule: | $NN@0 \ NP@1 \rightarrow NP@1 \ NN@0 \mid p_4$ |

Figure 3: Example of a lexical rule expansion.

4.5 Coupling with decoding

In order to improve reordering power of the translation system, we implemented an additional reordering as described in Crego and Mariño (2006).

Multiple word segmentations is encoded in a lattice, which is then passed to the input of the decoder, containing reordering alternatives consistent with the previously extracted rules. The decoder takes the n-best reordering of a source sentence coded in the form of a word lattice. This approach is in line with recent research tendencies in SMT, as described for example in (Hildebrand et al., 2008; Xu et al., 2005). Originally, word lattice algorithms do not involve syntax into reordering process, there-fore their reordering power is limited at representing long-distance reordering. Our approach is designed in the spirit of hybrid MT, integrating syntax trans-fer approach and statistical word lattice methods to achieve better MT performance on the basis of the standard state-of-the-art models.

During training a set of word permutation patterns is automatically learned following given word-to-word alignment. Since the original and monotonized (reordered) alignments may vary, different sets of reordering patterns are generated. Note that no in-formation about the syntax of the sentence is used: the reordering permutations are motivated by the crossed links found in the word alignment and, con-

(a) *Monotonic search, plain text: >n +h mTEm *w tAryx Eryq*
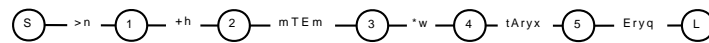
(b) *Word lattice, plain text: >n +h mTEm *w tAryx Eryq*
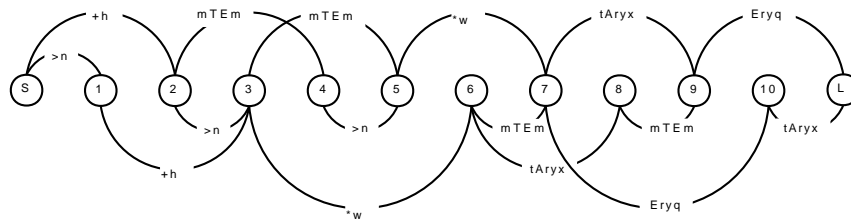
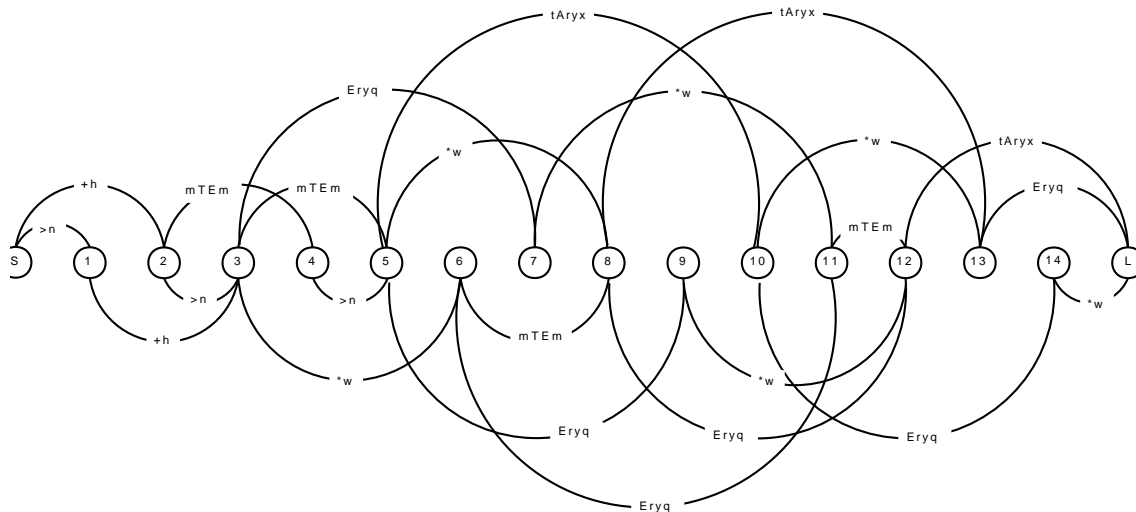(c) *Word lattice, reordered text: >n +h mTEm *w Eryq tAryx*

Figure 5: *Comparative example of a monotone search (a), word lattice for a plain (b) and reordered (c) source sentences.*

82

sequently, the generalization power of this framework is limited to local permutations.

On the step prior to decoding, the system generates word reordering graph for every source sentence, expressed in the form of a word lattice. The decoder processes word lattice instead of only one input hypothesis, extending the monotonic search graph with alternative paths.

Original sentence in Arabic, the English gloss and reference translation are:

Ar.: *>n +h mTEm   *w Eryq   tAryx*
Gl.: *this   restaurant has history illustrious*
Ref: 'this restaurant has an illustrious history'

The monotonic search graph (a) is extended with a word lattice for the monotonic train set (b) and reordered train sets (c). Figure 5 shows an example of the input word graph expressed in the form of a word lattice. Lattice (c) differ from the graph (b) in number of edges and provides more input options to the decoder. The decision about final translation is taken during decoding considering all the possible paths, provided by the word lattice.

5 Experiments and results

5.1 Data

The experiments were performed on two Arabic-English corpora: the BTEC'08 corpus from the tourist domain and the 50K first-lines extraction from the corpus that was provided to the NIST'08 evaluation campaign and belongs to the news domain (NIST50K). The corpora differ mainly in the average sentence length (ASL), which is the key corpus characteristic in global reordering studies.

A training set statistics can be found in Table 1.

	BTEC		NIST50K	
	Ar	En	Ar	En
Sentences	24.9 K	24.9 K	50 K	50 K
Words	225 K	210 K	1.2 M	1.35 M
ASL	9.05	8.46	24.61	26.92
Voc	11.4 K	7.6 K	55.3	36.3

Table 1: Basic statistics of the BTEC training corpus.

The BTEC development dataset consists of 489 sentences and 3.8 K running words, with 6 human-made reference translations per sentence; the dataset used to test the translation quality has 500 sentences, 4.1 K words and is also provided with 6 reference translations.

The NIST50K development set consists of 1353 sentences and 43 K words; the test data contains 1056 sentences and 33 K running words. Both datasets have 4 reference translations per sentence.

5.2 Arabic data preprocessing

We took a similar approach to that shown in Habash and Sadat (2006), using the MADA+TOKAN system for disambiguation and tokenization. For disambiguation only diacritic unigram statistics were employed. For tokenization we used the D3 scheme with -TAGBIES option. The scheme splits the following set of clitics: w+, f+, b+, k+, l+, Al+ and pronominal clitics. The -TAGBIES option produces Bies POS tags on all taggable tokens.

5.3 Experimental setup

We used the Stanford Parser (Klein and Manning, 2003) for both languages, Penn English Treebank (Marcus et al., 1993) and Penn Arabic Treebank set (Kulick et al., 2006). The English Treebank is provided with 48 POS and 14 syntactic tags, the Arabic Treebank has 26 POS and 23 syntactic categories.

As mentioned above, specific rules are not pruned away due to a limited amount of training material we set the thresholds k_{part} and k_{gener} to relatively low values, 1 and 3, respectively.

Evaluation conditions were case-insensitive and with punctuation marks considered. The target-side 4-gram language model was estimated using the SRILM toolkit (Stolcke, 2002) and modified Kneser-Ney discounting with interpolation. The highest BLEU score (Papineni et al., 2002) was chosen as the optimization criterion. Apart from BLEU, a standard automatic measure METEOR (Banerjee and Lavie, 2005) was used for evaluation.

5.4 Results

The scores considered are: BLEU scores obtained for the development set as the final point of the MERT procedure (*Dev*), and BLEU and METEOR scores obtained on test dataset (*Test*).

We present BTEC results (Tables 2), characterized by relatively short sentence length, and the re-

sults obtained on the NIST corpus (Tables 3) with much longer sentences and much need of global reordering.

	Dev	Test	
	BLEU	BLEU	METEOR
Plain	48.31	45.02	65.98
BL	48.46	47.10	68.10
SBR	48.75	47.52	67.33
SBR+lattice	48.90	48.78	68.85

Table 2: Summary of BTEC experimental results.

	Dev	Test	
	BLEU	BLEU	METEOR
Plain	41.83	43.80	62.03
BL	42.68	43.52	62.17
SBR	42.71	44.01	63.29
SBR+lattice	43.05	44.89	63.30

Table 3: Summary of NIST50K experimental results.

Four SMT systems are contrasted: *BL* refers to the Moses baseline system: the training data is not reordered, lexicalized reordering model (Tillman, 2004) is applied; *SBR* refers to the monotonic system configuration with reordered (SBR) source part; *SBR+lattice* is the run with reordered source part, on the translation step the input is represented as a word lattice.

We also compare the proposed approach with a monotonic system configuration (*Plain*). It shows the effect of source-reordering and lattice input, also decoded monotonically.

Automatic scores obtained on the test dataset evolve similarly when the SBR and word lattice representation applied to BTEC and NIST50K tasks. The combined method coupling two reordering techniques was more effective than the techniques applied independently and shows an improvement in terms of BLEU for both corpora. The METEOR score is only slightly better for the SBR configurations in case of BTEC task; in the case of NIST50K the METEOR improvement is more evident. The general trend is that automatic scores evaluated on the test set increase with the reordering model complexity.

Application of the SBR algorithm only (without a word lattice decoding) does not allow achieving statistical significance threshold for a 95% confidence interval and 1000 resamples (Koehn, 2004) for either of considered corpora. However, the *SBR+lattice* system configuration outperforms the *BL* by about 1.7 BLEU points (3.5%) for BTEC task and about 1.4 BLEU point (3.1%) for NIST task. These differences is statistically significant.

Figure 6 demonstrates how two reordering techniques interact within a sentence with a need for both global and local word permutations.

5.5 Syntax-based rewrite rules

As mentioned above, the SBR operates with three groups of reordering rules, which are the product of complete or partial delexicalization of the originally extracted patterns. The groups are processed and pruned independently. Basic rules statistics for both translation tasks can be found in Table 4.

The major part of reordering rules consists of two or three elements (for BTEC task there are no patterns including more than three nodes). For NIST50K there are a few rules with higher size in words of the move (up to 8). In addition, there are some long lexicalized rules (7-8), generating a high number of partially lexicalized patterns.

Table 5 shows the most frequent reordering rules with non-monotonic right part from each group.

Ar. plain.: **AElnt**   *Ajhzp AlAElAm l bEvp*   *AlAmm AlmtHdp fy syrAlywn An ...*
En. gloss: **announced** press release   by mission <u>nations united</u>   in sierra leone that ...

En. ref.: 'a press release by the <u>united nations</u> mission to sierra leone **announced** that ...'

*Ar. reord.: Ajhzp AlAElAm l bEvp <u>AlmtHdp AlAmm</u> fy syrAlywn **AElnt** An ...*

Figure 6: Example of SBR application (**highlited bold**) and local reordering error corrected with word lattice reordering (<u>underlined</u>).

6   Conclusions

In this study we have shown how the translation quality can be improved, coupling (1) SBR algorithm and (2) word alignment-based reordering framework applied during decoding. The system automatically learns a set of syntactic reordering patterns that exploit systematic differences between word order of source and target languages.

Translation accuracy is clearly higher when allowing for SBR coupled with word lattice input representation than standard Moses SMT with existing (lexicalized) reordering models within the decoder and one input hypothesis condition. We have also compared the reordering model a monotonic system.

The method was tested translating from Arabic to English. Two corpora and tasks were considered: the BTEC task with much need of local reordering and the NIST50K task requiring long-distance permutations caused by longer sentences.

The reordering approach can be expanded for any other pair of languages with available parse tools. We also expect that the method scale to a large training set, and that the improvement will still be kept, however, we plan to confirm this assumption experimentally in the near future.

Group	# of rules	Voc	2-element	3-element	4-element	[5-8]-element
BTEC experiments						
Specific rules	703	413	406	7	0	0
Partially lexicalized rules	1,306	432	382	50	0	0
General rules	259	5	259	0	0	0
NIST50K experiments						
Specific rules	517	399	193	109	72	25
Partially lexicalized rules	17,897	14,263	374	638	1,010	12,241
General rules	489	372	180	90	72	30

Table 4: Basic reordering rules statistics.

Specific rules
*NN@0 NP@1 -> NP@1 NN@0	NN@0 « Asm » NP@1 « +y »	0.0270*
*DTNN@0 DTJJ@1 -> DTJJ@1 DTNN@0	DTNN@0 « AlAmm »DTJJ@1 « AlmtHdp »	0.0515*
Partially lexicalized rules
*DTNN@0 DTJJ@1 -> DTJJ@1 DTNN@0	DTNN@0 « NON »DTJJ@1 « AlmtHdp »	0.0017*
*NN@0 NNP@1 -> NNP@1 NN@0	NN@0 « NON »NNP@1 « $rm »	0.0017*
General rules
*PP@0 NP@1 -> PP@0 NP@1	0.0432*
NN@0 DTNN@1 DTJJ@2 -> NN@0 DTJJ@2 DTNN@1 |0.0259

Table 5: Examples of Arabic-to-English reordering rules.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

M. Collins, Ph. Koehn, and I. Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on ACL 2005*, pages 531–540.

M.R. Costa-jussà and J.A.R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of the HLT/EMNLP 2006*.

M.R. Costa-jussà and J.A.R. Fonollosa. 2008. Computing multiple weighted reordering hypotheses for a statistical machine translation phrase-based system. In *In Proc. of the AMTA'08*, Honolulu, USA, October.

M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J. A. Fonollosa, J.B. Mari no, and R.E. Banchs. 2006. TALP phrase-based system and TALP system combination for IWSLT 2006. In *Proceedings of the IWSLT 2006*, pages 123–129.

J.M. Crego and J. B Mariño. 2006. Reordering experiments for N-gram-based SMT. In *SLT'06*, pages 242–245.

J.M. Crego and J.B. Mariño. 2007. Syntax-enhanced N-gram-based smt. In *Proceedings of MT SUMMIT XI*.

J.M. Crego, J. B. Mariño, and A. de Gispert. 2005. Reordered search and tuple unfolding for ngram-based smt. In *In Proc. of MT SUMMIT X*, pages 283–289, September.

S. Nießen and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. volume 30, pages 181–204.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 49–52.

A.S. Hildebrand, K. Rottmann, M. Noamany, Q. Gao, S. Hewavitharana, N. Bach, and S. Vogel. 2008. Recent improvements in the cmu large scale chinese-english smt system. In *Proceedings of ACL-08: HLT (Companion Volume)*, pages 77–80.

K. Imamura, H. Okuma, and E. Sumita. 2005. Practical approach to syntax-based statistical machine translation. In *Proceedings of MT Summit X*, pages 267–274.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *In Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174, June.

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

Ph. Koehn, F. J. 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.

Ph. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395.

S. Kulick, R. Gabbard, and M. Marcus. 2006. Parsing the Arabic Treebank: Analysis and improvements. *Treebanks and Linguistic Theories*.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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

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, and D. Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of HLT/NAACL04*, pages 161–168.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904.

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

C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the Joint Conference on EMNLP*.

F. Xia and M. McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the COLING 2004*.

J. Xu, E. Matusov, R. Zens, and H. Ney. 2005. Integrated chinese word segmentation in statistical machine translation. In *Proc. of IWSLT 2005*.

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