Reordering Model for Forest-to-String Machine Translation

Martin Čmejrek
IBM Watson Group
Prague, Czech Republic
martin.cmejrek@us.ibm.com

Abstract

In this paper, we present a novel extension of a forest-to-string machine translation system with a reordering model. We predict reordering probabilities for every pair of source words with a model using features observed from the input parse forest. Our approach naturally deals with the ambiguity present in the input parse forest, but, at the same time, takes into account only the parts of the input forest used by the current translation hypothesis. The method provides improvement from 0.6 up to 1.0 point measured by (T − B)/2 metric.

1 Introduction

Various commonly adopted statistical machine translation (SMT) approaches differ in the amount of linguistic knowledge present in the rules they employ.

Phrase-based (Koehn et al., 2003) models are strong in lexical coverage in local contexts, and use external models to score reordering options (Tillman, 2004; Koehn et al., 2005).

Hierarchical models (Chiang, 2005) use lexicalized synchronous context-free grammar rules to produce local reorderings. The grammaticality of their output can be improved by additional reordering models scoring permutations of the source words. Reordering model can be either used for source pre-ordering (Tromble and Eisner,), integrated into decoding via translation rules extension (Hayashi et al., 2010), additional lexical features (He et al.,), or using external sources of information, such as source syntactic features observed from a parse tree (Huang et al., 2013).

Tree-to-string (T2S) models (Liu et al., 2006; Galley et al., 2006) use rules with syntactic structures, aiming at even more grammatically appropriate reorderings.

Forest-to-string (F2S) systems (Mi et al., 2008; Mi and Huang, 2008) use source syntactic forest as the input to overcome parsing errors, and to alleviate sparseness of translation rules.

The parse forest may often represent several meanings for an ambiguous input that may need to be transtated differently using different word orderings. The following example of an ambiguous Chinese sentence with ambiguous part-of-speech labeling motivates our interest in the reordering model for the F2S translation.

tǎolùn (0)	hùi (1)	zěnmeyàng (2)
discussion/NN	meeting/NN	how/VV
discuss/VV	will/VV	

There are several possible meanings based on the different POS tagging sequences. We present translations for two of them, together with the indices to their original source words:

(a) NN NN VV:
How$_2$ *was*$_2$ *the*$_0$ *discussion*$_0$ *meeting*$_1$*?*

(b) VV VV VV:
Discuss$_0$ *what*$_2$ *will*$_1$ *happen*$_1$.

A T2S system starts from a single parse corresponding to one of the possible POS sequences, the same tree can be used to predict word reorderings. On the other hand, a F2S system deals with the ambiguity through exploring translation hypotheses for all competing parses representing the different meanings. As our example suggests, different meanings also tend to reorder differently

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 227–232,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

id	rule
r_1	NP(tǎolùn/NN) → discussion
r_2	NP(hùi/NN) → meeting
r_3	NP(x_1:NP x_2:NP) → the x_1 x_2
r_4	IP(x_1:NP zěnmeyàng/VV) → how was x_1
r_5	IP(hùi/VV zěnmeyàng/VV) → what will happen
r_6	IP(tǎolùn/VV x_1:IP) → discuss x_1

Table 1: Tree-to-string translation rules (without internal structures).

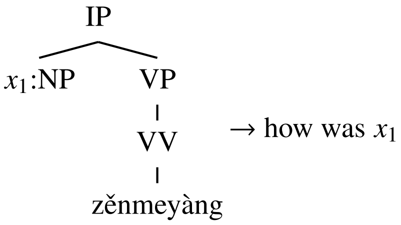

Figure 1: Tree-to-string rule r_4.

during translation. First, the reordering model suitable for F2S translation should allow for translation of all meanings present in the input. Second, as the process of deriving a partial translation hypothesis rules out some of the meanings, the reordering model should restrict itself to features originating in the relevant parts of the input forest. Our work presents a novel technique satisfying both these requirements, while leaving the disambuiguation decision up to the model using global features.

The paper is organized as follows: We briefly overview the F2S and Hiero translation models in Section 2, present the proposed forest reordering model in Section 3, describe our experiment and present results in Section 4.

2 Translation Models

Forest-to-string translation (Mi et al., 2008) is an extension of the tree-to-string model (Liu et al., 2006; Huang et al., 2006) allowing it to use a packed parse forest as the input instead of a single parse tree.

Figure 1 shows a tree-to-string **translation rule** (Huang et al., 2006), which is a tuple $\langle \mathit{lhs}(r), \mathit{rhs}(r), \psi(r) \rangle$, where $\mathit{lhs}(r)$ is the source-side tree fragment, whose internal nodes are labeled by nonterminal symbols (like NP), and whose frontier nodes are labeled by source-language words (like "zěnmeyàng") or variables from a finite set $\mathcal{X} = \{x_1, x_2, \ldots\}$; $\mathit{rhs}(r)$ is the target-side string expressed in target-language words (like "how was") and variables; and $\psi(r)$ is a mapping from $\mathcal{X}$ to nonterminals. Each variable $x_i \in \mathcal{X}$ occurs *exactly once* in $\mathit{lhs}(r)$ and *exactly once* in $\mathit{rhs}(r)$.

The Table 1 lists all rules necessary to derive translations (a) and (b), with their internal structure removed for simplicity.

Typically, an F2S system translates in two steps (shown in Figure 2): parsing and decoding. In the parsing step, the source language input is converted into a *parse forest* (A). In the decoding step, we first convert the parse forest into a *translation forest* F^t in (B) by using the fast pattern-matching technique (Zhang et al., 2009). Then the decoder uses dynamic programing with beam search and cube pruning to find the approximation to the best scoring derivation in the translation forest, and outputs the target string.

3 Forest Reordering Model

In this section, we describe the process of applying the reordering model scores. We score pairwise translation reorderings for every pair of source words similarly as described by Huang et al. (2013). In their approach, an external model of ordering distributions of *sibling* constituent pairs predicts the reordering of word pairs. Our approach deals with parse forests rather than with single trees, thus we have to model the scores differently. We model ordering distributions for every pair of *close relatives*–nodes in the parse forest that may occur together as frontier nodes of a single matching rule. We further condition the distribution on a third node–a common ancestor of the node pair that corresponds to the root node of the matching rule. This way our external model takes into acount the syntactic context of the hypothesis. For example, nodes NP$_{0,1}$ and NP$_{1,2}$ are close relatives, NP$_{0,2}$ and IP$_{0,3}$ are their common ancestors; NP$_{0,1}$ and VV$_{2,3}$ are close relatives, IP$_{0,3}$ is their common ancestor; NP$_{0,1}$ and VV$_{1,2}$ are not close relatives.

More formally, let us have an input sentence $(w_0, ..., w_n)$ and its translation hypothesis h. For every i and j such that $0 \le i < j \le n$ we assume that the translations of w_i and w_j are in the hypothesis h either in the same or inverted ordering $o_{ij} \in \{Inorder, Reorder\}$, with a probability $P_{order}(o_{ij}|h)$. Conditioning on h signifies that the probabilistic model takes the current hypothesis as a parameter. The reordering score of the entire hy-

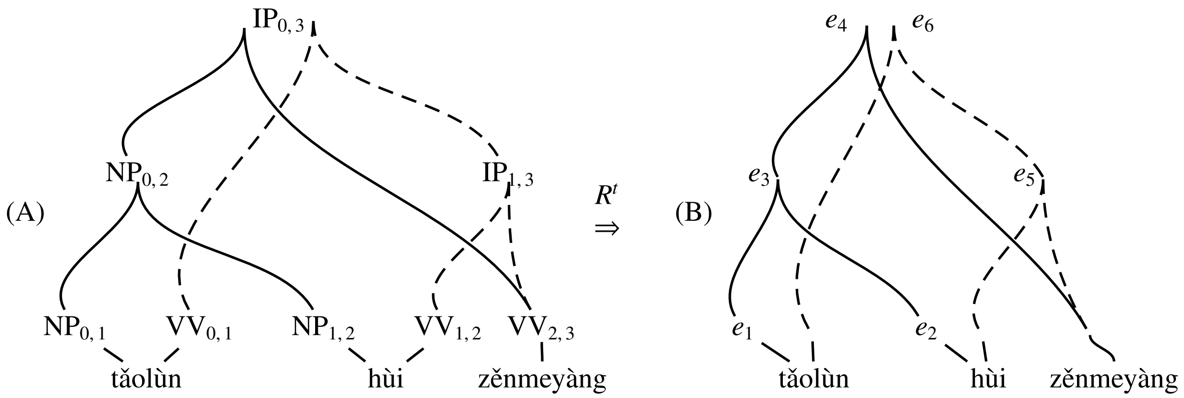

Figure 2: Parse and translation hypergraphs. (A) The parse forest of the example sentence. Solid hyperedges denote the best parse, dashed hyperedges denote the second best parse. Unary edges were collapsed. (B) The corresponding translation forest F^t after applying the tree-to-string translation rule set R^t. Each translation hyperedge (e.g. e_4) has the same index as the corresponding rule (r_4). The forest-to-string system can produce the example translation (a) (solid derivation: r_1, r_2, r_3, and r_4) and (b) (dashed derivation: r_5, r_6).

pothesis $f_{order}(h)$ is then computed as

$$f_{order} = \sum_{0 \leq i < j \leq n} - \log P_{order}(o_{ij} = o^h_{ij} \mid h), \quad (1)$$

where o^h_{ij} denotes the actual ordering used in h.

The score f_{order} can be computed recursively by dynamic programing during the decoding. As an example, we show in Table 2 reordering probabilities retrieved in decoding of our sample sentence.

(a) If h is a hypothesis formed by a single translation rule r with no frontier nonterminals, we evaluate all word pairs w_i and w_j covered by h such that $i < j$. For each such pair we find the frontier nodes x and y matched by r such that x spans exactly w_i and y spans exactly w_j. (In this case, x and y match preterminal nodes, each spanning one position). We also find the node z matching the root of r. Then we directly use the Equation 1 to compute the score using an external model $P_{order}(o_{ij}|xyz)$ to estimate the probability of reordering the relative nodes. For example, when applying rule r_5, we use the ordering distribution $P_{order}(o_{1,2}|\mathrm{VV}_{1,2}, \mathrm{VV}_{2,3}, \mathrm{IP}_{1,3})$ to score reorderings of hùi and zěnmeyàng.

(b) If h is a hypothesis formed by a T2S rule with one or more frontier nonterminals, we evaluate all word pairs as follows: If both w_i and w_j are spanned by the same frontier nonterminal (e.g., tǎolùn and hùi when applying the rule r_4), the score f_{order} had been already computed for the underlying subhypothesis, and therefore was already included in the total score. Otherwise, we compute the word pair ordering cost. We find the close relatives x and y representing each w_i and w_j. If w_i is matched by a terminal in r, we select x as the node matching r and spanning exactly w_i. If w_i is spanned by a frontier nonterminal in r (meaning that it was translated in a subhypothesis), we select x as the node matching that nonterminal. We proceed identically for w_j and y. For example, when applying the rule r_4, the word zěnmeyàng will be represented by the node $\mathrm{VV}_{2,3}$, while tǎolùn and hùi will be represented by the node $\mathrm{NP}_{0,2}$.

Note that the ordering o^h_{ij} cannot be determined in some cases, sometimes a source word does not produce any translation, or the translation of one word is entirely surrounded by the translations of another word. A weight corresponding to the binary discount feature $f_{o_{unknown}}$ is added to the score for each such case.

The external model $P_{order}(o_{ij}|xyz)$ is implemented as a maximum entropy model. Features of the model are observed from paths connecting node z with nodes x and y as follows: First, we pick paths $z \rightarrow x$ and $z \rightarrow y$. Let z' be the last node shared by both paths (the closest common ancestor of x and y). Then we distinguish three types of path: (1) The *common prefix* $z \rightarrow z'$ (it may have zero length), the *left path* $z \rightarrow x$, and the *right path* $z \rightarrow y$. We observe the following features on each path: the syntactic labels of the nodes, the production rules, the spans of nodes, a list of stop words immediately preceding and following the span of the node. We merge the features observed from different paths $z \rightarrow x$ and $z \rightarrow y$. This approach

rule	word pair	order	probability
a) how$_2$ was$_2$ the discussion$_0$ meeting$_1$			
r_3	(tǎolùn,hùi)	Inorder	$P_{order}(o_{0,1}\|\mathrm{NP}_{0,1},\mathrm{NP}_{1,2},\mathrm{NP}_{0,2})$
r_4	(tǎolùn,zěnmeyàng)	Reorder	$P_{order}(o_{0,2}\|\mathrm{NP}_{0,2},\mathrm{VV}_{2,3},\mathrm{IP}_{0,3})$
	(hùi,zěnmeyàng)	Reorder	$P_{order}(o_{1,2}\|\mathrm{NP}_{0,2},\mathrm{VV}_{2,3},\mathrm{IP}_{0,3})$
b) discuss$_0$ what$_2$ will$_1$ happen$_1$			
r_5	(hùi, zěnmeyàng)	Reorder	$P_{order}(o_{1,2}\|\mathrm{VV}_{1,2},\mathrm{VV}_{2,3},\mathrm{IP}_{1,3})$
r_6	(tǎolùn, hùi)	Inorder	$P_{order}(o_{0,1}\|\mathrm{VV}_{0,1},\mathrm{IP}_{1,3},\mathrm{IP}_{0,3})$
	(tǎolùn, zěnmeyàng	Inorder	$P_{order}(o_{0,2}\|\mathrm{VV}_{0,1},\mathrm{IP}_{1,3},\mathrm{IP}_{0,3})$

Table 2: Example of reordering scores computed for derivations (a) and (b).

ignores the internal structure of each rule[1], relying on frontier node annotation. On the other hand it is still feasible to precompute the reordering probabilities for all combinations of *xyz*.

4 Experiment

In this section we describe the setup of the experiment, and present results. Finally, we propose future directions of research.

4.1 Setup

Our baseline is a strong F2S system (Čmejrek et al., 2013) built on large data with the full set of model features including rule translation probabilities, general lexical and provenance translation probabilities, language model, and a variety of sparse features. We build it as follows. The training corpus consists of 16 million sentence pairs available within the DARPA BOLT Chinese-English task. The corpus includes a mix of newswire, broadcast news, webblog data coming from various sources such as LDC, HK Law, HK Hansard and UN data. The Chinese text is segmented with a segmenter trained on CTB data using conditional random fields (CRF).

Bilingual word alignments are trained and combined from two sources: GIZA (Och, 2003) and maximum entropy word aligner (Ittycheriah and Roukos, 2005).

Language models are trained on the English side of the parallel corpus, and on monolingual corpora, such as Gigaword (LDC2011T07) and Google News, altogether comprising around 10 billion words.

We parse the Chinese part of the training data with a modified version of the Berkeley parser (Petrov and Klein, 2007), then prune the obtained parse forests for each training sentence with the marginal probability-based inside-outside algorithm to contain only $3n$ CFG nodes, where n is the sentence length.

We extract tree-to-string translation rules from forest-string sentence pairs using the forest-based GHKM algorithm (Mi and Huang, 2008; Galley et al., 2004).

In the decoding step, we use larger input parse forests than in training, we prune them to contain $10n$ nodes. Then we use fast pattern-matching (Zhang et al., 2009) to convert the parse forest into the translation forest.

The proposed reordering model is trained on $100,000$ automatically aligned forest-string sentence pairs from the parallel training data. These sentences provide 110M reordering events that are used by `megam` (Daumé III, 2004) to train the maximum entropy model.

The current implementation of the reordering model requires offline preprocessing of the input hypergraphs to precompute reordering probabilities for applicable triples of nodes (x, y, z). Since the number of levels in the syntactic trees in T2S rules is limited to 4, we only need to consider such triples, where z is up to 4 levels above x or y.

We tune on 1275 sentences, each with 4 references, from the LDC2010E30 corpus, initially released under the DARPA GALE program.

We combine two evaluation metrics for tuning and testing: B (Papineni et al., 2002) and T (Snover et al., 2006). Both the baseline and the reordering experiments are optimized with MIRA (Crammer et al., 2006) to maximize (T - B)/2.

We test on three different test sets: GALE Web test set from LDC2010E30 corpus (1239 sentences, 4 references), NIST MT08 Newswire

[1] Only to some extent, the rule still has to match the input forest, but the reordering model decides based on the sum of paths observed between the root and frontier nodes.

System	GALE Web			MT08 Newswire			MT08 Web		
	$\frac{T\ \ -B}{2}$	B	T	$\frac{T\ \ -B}{2}$	B	T	$\frac{T\ \ -B}{2}$	B	T
F2S	8.8	36.1	53.7	5.6	40.6	51.8	12.0	31.3	55.3
+Reordering	8.2	36.4	52.7	4.8	41.7	50.5	11.0	31.7	53.7
Δ	*-0.6*	*+0.3*	*-1.0*	*-0.8*	*+1.1*	*-1.3*	*-1.0*	*+0.4*	*-1.6*

Table 3: Results.

portion (691 sentences, 4 references), and NIST MT08 Web portion (666 sentences, 4 references).

4.2 Results

Table 3 shows all results of the baseline and the system extended with the forest reordering model. The (T − B)/2 score of the baseline system is 12.0 on MT08 Newswire, showing that it is a strong baseline. The system with the proposed reordering model significantly improves the baseline by 0.6, 0.8, and 1.0 (T − B)/2 points on GALE Web, MT08 Newswire, and MT08 Web.

The current approach relies on frontier node annotations, ignoring to some extent the internal structure of the T2S rules. As part of future research, we would like to compare this approach with the one that takes into accout the internal structure as well.

5 Conclusion

We have presented a novel reordering model for the forest-to-string MT system. The model deals with the ambiguity of the input forests, but also predicts specifically to the current parse followed by the translation hypothesis. The reordering probabilities can be precomputed by an offline process, allowing for efficient scoring in runtime. The method provides improvement from 0.6 up to 1.0 point measured by (T − B)/2 metrics.

Acknowledgments

We thank Jiří Havelka for proofreading and helpful suggestions. We would like to acknowledge the support of DARPA under Grant HR0011-12-C-0015 for funding part of this work. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the DARPA.

References

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL.*

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7.

Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the HLT-NAACL.*

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the COLING-ACL.*

Katsuhiko Hayashi, Hajime Tsukada, Katsuhito Sudoh, Kevin Duh, and Seiichi Yamamoto. 2010. Hierarchical Phrase-based Machine Translation with Word-based Reordering Model. In *Proceedings of the COLING.*

Zhongjun He, Yao Meng, and Hao Yu. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proceedings of the EMNLP.*

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the AMTA.*

Zhongqiang Huang, Jacob Devlin, and Rabih Zbib. 2013. Factored soft source syntactic constraints for hierarchical machine translation. In *Proceeedings of the EMNLP.*

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of the HLT and EMNLP.*

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL.*

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of the IWSLT*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL*.

Haitao Mi and Liang Huang. 2008. Forest-based translation rule extraction. In *Proceedings of EMNLP*.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL: HLT*.

Franz Joseph Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the AMTA*.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. Proceedings of the HLT-NAACL.

Roy Tromble and Jason Eisner. Learning linear ordering problems for better translation. In *Proceedings of the EMNLP*.

Martin Čmejrek, Haitao Mi, and Bowen Zhou. 2013. Flexible and efficient hypergraph interactions for joint hierarchical and forest-to-string decoding. In *Proceedings of the EMNLP*.

Hui Zhang, Min Zhang, Haizhou Li, and Chew Lim Tan. 2009. Fast translation rule matching for syntax-based statistical machine translation. In *Proceedings of EMNLP*, pages 1037–1045, Singapore, August.