An Efficient Approach to Rule Redundancy Reduction in Hierarchical Phrase-Based Translation

Licheng FANG

Institute of Automation, CAS

Beijing, China

lcfang@nlpr.ia.ac.cn

Chengqing ZONG

Institute of Automation, CAS

Beijing, China

cqzong@nlpr.ia.ac.cn

Abstract:

Hierarchical phrase-based machine translation model is a popular syntax model that makes use of the expressive power of Synchronous Context-Free Grammars (SCFG) to address the reordering problem in statistical machine translation. The model, however, generally suffers from a great amount of redundancy in the extracted translation rules. In this paper, we re-introduce the concept of rift into the rule extraction procedure to force the rules with reordering power to concentrate on where reordering has actually happened. Our approach brings a dramatic reduction in the training time and the number of the rules, with only minor sacrifice in translation quality.

Keywords:

Statistical machine translation; hierarchical phrase; redundancy; rift

1. Introduction

Statistical machine translation research has evolved from the ground-breaking word based model developed at IBM [1] to phrase-based models to incorporate contextual information for word disambiguation, idioms, and local word reordering. Franz Och's phrase extraction based on bi-directional word alignment [2] has also become a standard practice. In the recent years, different syntax-based translation models are proposed to address the long distance reordering and dependencies that often haunt the translation task between linguistically distant language pairs such as Chinese-English and Arab-English. These models generally view translation as a parsing task and sometimes use a supervised parser to parse the training corpus before learning translation rules.

Among the syntax-based translation models David Chiang's Hiero system [3] is particularly well received because of its inherent simplicity: it only assumes that natural languages possess a hierarchical structure, and without the help of parsers, extracts synchronous context-free grammar (SCFG) rules from a bilingual corpus. With the reordering ability of SCFG rules, the translation quality is significantly improved. However, typically the SCFG rules greatly outnumber the rules of a conventional phrase-based system, which lays a heavy burden on both the training and decoding processes.

Not much work was put into alleviating this problem, but there are generally two directions pursued. From a better parameter estimating perspective, we can expect an EM procedure would extract rules with improved accuracy, and Viterbi derivations on a training corpus can be used as a smaller yet more accurate rule set. [4] showed such an algorithm on ITG [5]. And [6] applied EM to Galley's tree-to-string model [7] with improved translation results. Yet EM typically comes with a huge computational cost, and almost always suffers from overfitting. Insofar it has not seen successful application to synchronous context-free grammars.

On the other hand, the rule redundancy problem can be attacked with better constraint in the rule extraction procedure. [8] tried to base the rule extraction on linguistically well-formed chunks. And [9] proposed to use mutual information- measured bigram collocations to find anchor points in the source language sentence, which served to constrain Och's phrase extraction algorithm in a conventional phrase-based translation system. Her method in effect also introduced some information about the source language side's well-formedness.

Our approach is based on the assumption that conventional phrases and SCFG rules with nonterminals play different roles in the translation task. Instead of measuring the well-formedness of the rules, we use the definition of *rift* to identify parts of the training sentence pairs where reordering matters most, and let the effort of learning reordering concentrates there. Experiments have shown that our method reduces the training time by an order of magnitude, extracts far less rules, with only less than 2% relative sacrifice of translation quality.

The rest of this paper is organized as follows. In Section 2 we briefly introduce the hierarchical phrase-based model and its rule extraction procedure. In Section 3 an analysis of the usefulness of the rules are given. In Section 4 we describe our approach to combating the redundancy and experimental results are shown in Section 5. The conclusions are drawn in Section 6.

978-1-4244-2780-2/08/$25.00 ©2008 IEEE

2. The Hierarchical Phrase-Based Model

Hierarchical phrase-based translation extracts from bitext rules of the form

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle$$

where X is the single SCFG nonterminal that represents a phrase, and γ and α are strings of terminals and nonterminals. ~ is an alignment between the nonterminals in γ and α.

The rules are learned with the following steps:

1. Bilingual corpus is aligned with GIZA++ bi-directionally.
2. *Initial phrases* are extracted with Och's algorithm[2]. The initial phrases are simply rules with no terminals.
3. If both the source side and the target side of a rule's spans completely cover another rule, we can subtract the second rule from the first to obtain a rule with nonterminals. All possible rules with nonterminals are obtained from the initial phrases using a *phrase-subtract* algorithm, with a few heuristic restrictions to limit the number of rules (see Section 5).

The rules are then sorted by four features $P(\alpha \mid \gamma)$, $P(\gamma \mid \alpha)$, $lex(\alpha \mid \gamma)$, and $lex(\gamma \mid \alpha)$, and then fed into a CYK-style decoder, which combines the feature scores in a log-linear fashion and parses unseen sentences.

3. Motivation

3.1 Analysis

Compared with a conventional phrase-based translation system, the rules with nonterminals endow the hierarchical phrase-based system with a greater reordering ability. But this power comes with major increase in computational cost. See the left column of Table 5 for a decomposition of the rules extracted from our experimental data (described in Section 5) according to the number of nonterminals contained.

Apparently the rules with nonterminals and the phrase-subtract algorithm dominated the computational cost, leaving us to wonder if we actually have to devote that much computational resource for the reordering task. [10] proposed a reordering model based on a maximum entropy classifier which estimated the probability of reordering two adjacent phrases. He made the observation that good reordering ability could be achieved by only feeding the maximum entropy classifier with features like the first word of the phrases. This is a strong argument that the reordering task can be fulfilled in a more light-weighted way.

We classify the rules with nonterminals used in the hierarchical phrase-based model by their source and target side patterns, and the percentage of each kind of rule is shown in table 1 and table 2. A w in the pattern represents a string of terminals, and X is the nonterminal.

Table 1. Distribution of rules with one nonterminal

	wXw	*wX*	*Xw*
wXw	51%	2%	3%
wX	2%	18%	1%
Xw	3%	1%	17%

From the tables we can see that the vast majority of the rules are of the forms such as $X \rightarrow \langle wX, wX \rangle$ and $X \rightarrow \langle wX_1wX_2w, wX_1wX_2w \rangle$, which contains no reordering information. A quick examination of the extracted rules will reveal that many rules of this kind simply states sequential translation:

$$X \rightarrow \langle \text{我 } X, \text{I } X \rangle$$

$$X \rightarrow \langle X \text{ 是}, X \text{ is} \rangle$$

$$X \rightarrow \langle X_1 \text{ 这 } X_2, X_1 \text{ this } X_2 \rangle$$

This is a major redundancy of the rules, because such sequential translation has already been formulated by the *glue rule* in the model:

$$S \rightarrow \langle SX, SX \rangle \quad S \rightarrow \langle X, X \rangle$$

3.2 Assumption

To fight the redundancy, we first make the assumption that initial phrases are *meant* to capture the contextual information, and rules with nonterminals are *meant* to capture the reordering information. We have to note that this is only partly true, because sometimes rules with nonterminals do imply a translation context, e.g.

$$X \rightarrow \langle \text{这 是 } X \text{ 吗 ?}, \text{is this } X \text{ ?} \rangle$$

However, most of the times the context can as well be handled by the initial phrases and the glue rules.

So, if we can separate the contextual and reordering information that is contained in the training sentence pairs. We can possibly make the initial phrases and the rules with nonterminals play only their respective roles. We introduce *reordering rift* to fulfill this job.

4. The Rift Constraint

The concept of *rift* is originally proposed by IBM [11] with just decoding efficiency in mind. The original application is that given a unseen French sentence, we want a maximum entropy classifier to predict "safe" points to split the sentence into parts so that the decoder can translate the parts sequentially and efficiently. So, by "safe" it means where reordering doesn't happen.

Given a training sentence pair and their alignment $\langle E, F, A \rangle$, where $a_j = i$ ψin A ψdenotes that the

Table 2. Distribution of rules with two nonterminals

	wXwXw		*XwXw*		*wXwX*		*XwX*	
	straight	inverted	straight	inverted	straight	inverted	straight	inverted
wXwXw	11.4%	0.5%	0.6%	0.3%	0.6%	0.2%	0.1%	0.1%
XwXw	1.8%	1.0%	24.8%	1.5%	0.4%	0.5%	1.1%	0.4%
wXwX	1.8%	0.8%	0.5%	0.4%	28.1%	1.3%	1.4%	0.3%
XwX	0.1%	0.3%	1.8%	1.0%	2.6%	0.9%	13.2%	0.3%

j th French word f_i is aligned to the i th English word e_i. [11] defines rifts as positions j in F such that for all $k < j$, $a_k \le a_j$, and for all $k > j$, $a_k \ge a_j$. e.g., the words to the left of the French word f_j are generated by words to the left of e_{a_j}, and the words to the right of f_j are generated by words to the right of e_{a_j}. This definition reflects the unidirectional alignment at the time and we extend the definition to accommodate the bi-directional alignment used in current translation systems.

Given a sentence pair and the alignment $\langle E, F, A \rangle$, where $A(i, j) = 1$ if e_i and f_j is aligned and $A(i, j) = 0$ otherwise. We define a rift as a tuple $\langle k, l \rangle$ such that for all i, j where $A(i, j) = 1$, if $i < k$, then $j < l$, and if $i > k$, then $j > l$. It can be seen from Figure 1 that our definition of rift is visually just a line simultaneously splitting the source and target side of the sentence pair, without any alignment link penetrates it. It should be noted that a rift is different from an initial phrase boundary. In Figure 1, the dotted link is an initial phrase boundary, while it is not a rift.

Algorithm 1 is a fast algorithm to determine all the rifts given the alignments of a sentence pair. It runs in $O(n)$ time suppose both the sentences have length n. According to the above definition of rifts, many rifts will be found around consecutive null-aligned words. In practice we only set rifts at the boundaries of these words.

We can easily see that rifts separate the sentence pair into bilingual chunks, and reordering can happen only inside these chunks. We change the rule extraction to the following procedure, which extracts the same initial phrases for any sentence pair while the phrase-subtract procedure is only allowed to work inside rift-separated chunks.

1. Initial phrases are extracted from a sentence pair.
2. Rifts from the sentence pair are determined.
3. Any initial phrase that spans over a rift is marked.
4. Run the phrase-subtract algorithm only on the unmarked initial phrases.

The set of rules generated by the above procedure thus contains the same set of initial rules as previous, but contains a much smaller set of rules with nonterminals.

Algorithm 1. Determine all rifts given source length *m*, target length *n*, and alignment *A*

```
S = NULL array of length m
T = NULL array of length n
for all i such that S[i] is not null-aligned do
    S[i] = max{j | A(k,j) = 1, 0 <= k <= i}
end for
for all j such that T[j] is not null-aligned do
    T[j] = max{i | A(i,k) = 1, 0 <= k <= j}
end for
rift = [], i = 0, j = 0
while i < m and j < n do
    append (i, j) to rift
    if = NULL then
        i = i + 1
        continue
    end if
    if T[j] = NULL then
        j = j + 1
        continue
    end if
    while T[j] != i or S[i] != j do
        i, j = T[j], S[i]
    end while
    i = i + 1, j = j + 1
end while
return rift
```

5. Experiments

5.1 Experimental Settings

Table 3 gives a summary of the experimental data of a Chinese-English translation task. The IWSLT 2006 evaluation campaign corpus consists of spoken language related to traveling scenarios such as hotel booking, ticket booking, restaurant ordering, etc. We used the devset2 and devset3 from IWSLT 2006 as our development data (for tuning decoder parameters) and test data, respectively.

We tested the rift constraint on our implementation

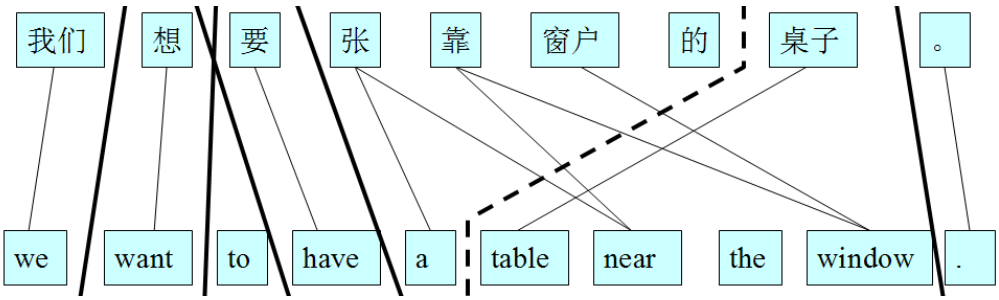

Figure 1. The bold line segments represent rifts in this aligned pair of sentences. The dotted line is an initial phrase boundary but not a rift

Table 3. The IWSLT 2006 Chinese-English translation experimental data. The development and test sets are accompanied by 16 reference translations

	Sentences
Training (C-E)	39,953
Development (devset2)	500
Test (devset3)	506

of the hierarchical phrase-based translation engine. The results from a conventional phrase-based translation engine with a beam search decoder developed in our group is also presented for comparison. Both systems used phrases with possibly null-aligned words at the boundaries (soft boundary). Aside from that, all phrases or rules followed the restriction that no initial phrase should exceed 10 words in length, and no rule with nonterminal should contain more than 2 nonterminals, or more than 5 terminals and nonterminals in total. Immediately adjacent nonterminals in rules were also forbidden.

Both the conventional and hierarchical phrase-based decoders used an ngram language model trained on the English side of the training corpus. The model was generated by the SRILM toolkit [12].

The parameters to log-linearly combine the feature scores were tuned by minimum error rate training [13] on the development set, maximizing BLEU scores [14]. The translation results were scored with case-insensitive BLEU against 16 references for each test Chinese sentence. The results are shown in Table 4.

5.2 Discussions

The results are encouraging. The rift constraint brought the BLEU score down to 0.5715 from 0.5791, but note that compared with average machine translation results these scores are high and this is in fact just a 1.3% decrease in accuracy. And what is important is that the rift constraint has not deprived the hierarchical phrase-based model of its definitive advantage over the conventional phrase-based model.

The reduction in computational cost, however, is obvious. We extracted 73% less rules from the training corpus. Table 5 breaks down the unfiltered rules into different categories according to their number of nonterminals. We can see that without the rift constraint, the vast majority of the rules are those with nonterminals, most of which hardly contribute to the translation. With the rift constraint enabled, the distribution is closer to our ideal: a majority of initial phrases, and an elite set of more informative reordering rules.

The training time is reduced by an order of magnitude. This sharp reduction can be attributed to the fact that phrase-subtract algorithm dominates the training time in extracting hierarchical phrases. The decoding time is halved, but has not experienced as much a reduction because the number of translation hypotheses inside the decoder increases exponentially with the number of rules, and the decoder has to prune its search space in either case.

6. Conclusions and future work

We proposed a very simple, yet effective way for taming the enormous rule redundancy in hierarchical phrase-based translation systems. The positive experimental results confirmed our assumption that rules with nonterminals could do as good a job by only concentrating on segments of the training corpus where reordering happens.

Statistical machine translation often faces the trade-off between translation quality and computational cost, which means that if we can dramatically reduce the computational cost, it would then be possible for us to devote the saved computing power to other factors that affect the translation quality. And in our case, we might extract longer initial phrases, relax the pruning criteria of the decoder, etc. Such effort gives us better understanding and insight of how the information contained in the training corpus is distributed and how we can effectively use it.

But this is only a first encouraging sign that we might be heading toward the right direction. Many further explorations are yet to be made. First, by only training the reordering rules on reordering abound chunks, we run the risk of overestimating the rules

Table 4. The result comparing the conventional phrase-based system, the hierarchical phrase-based system, and the hierarchical phrase-based system with rift constraint. The decoding time of the hierarchical phrase-based system is based on rule sets filtered by the test set. The last row shows the percentage of reduction in each measure after adding the rift constraint.

	BLEU	Rules(filtered/unfiltered)	Training time	Decoding time
Conventional	0.5096	608,006	1'48"	2'37"
Hierarchical	0.5791	484,953/4,333,306	172'11"	17'34"
Hierarchical+rift	0.5715	128,571/1,180,768	8'57"	7'30"
Percentage of change	-1.3%	-73%/-73%	-95%	-57%

Table 5. The rules are classified by the number of nonterminals contained. The rift constraint renders a more reasonable distribution of rules.

	w/o rift constraint	w/ rift constraint
Initial phrases	618.960	618,960
Rules with one nonterminal	1,900,372	336,780
Rules with two nonterminals	1,813,974	225,028
TOTAL	4,333,306	1,180,768

suggesting reordering instead of a sequential translation. Further testing on different corpora is needed. Second, we believe that there is the potential in our method to outperform the original hierarchical phrase-based model, because with more accurate estimating, less rules can mean better rules. Also, the reduction in the number of rules can mean less search errors for the decoder. Less rules also give opportunity to computationally costly procedures like EM to further hone them, considering that the rule extraction procedure is actually a parsing of the training corpus.

Acknowledgements

The research work in this paper is partially funded by the Natural Science Foundation of China under Grant No. 60575043 and 60736014, the National Key Technology R&D Program under Grant No. 2006BAH03B02, the Hi-Tech Research and Development Program (863) of China under Grant No. 2006AA01Z194 and 2006AA010108-4, and Nokia Research Center, Beijing as well.

References

[1] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The Mathematic of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol 19, no. 2, 1994, pp. 263-311.

[2] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, 2004 pp. 417-449.

[3] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, 2007, pp 201-228.

[4] H. Zhang and D. Gildea, "Stochastic lexicalized inversion transduction grammar for alignment," *Proceedings of the 43rd Annual Meetings of Association for Computational Linguistics*, 2005, pp. 475-482.

[5] D. Wu, "Stochastic inversion grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, 1997, pp. 377-403.

[6] J. May and K. Knight, "Syntactic Re-Alignment Models for Machine Translation," *Proceedings of the 2007 Joint Conferences on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), 2007, pp. 360-368.

[7] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a translation rule," *Proceedings of HLT/NAACL*, vol. 4, 2004, pp. 273-280.

[8] W. Wei and B. Xu, "Hierarchical chunking-phrase based translation," *International Conference on Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007*, 2007, pp. 268-273.

[9] Y. Zhou, "Approaches to Bilingual Alignment for Statistical Machine Translation," *Ph.D. Dissertation, Institute of Automation, Chinese Academy of Sciences*, 2008.

[10] D. Xiong, Q. Liu, and S. Lin, "Maximum entropy based phrase reordering model for statistical machine translation," *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the AC*L, 2006, pp. 521-528.

[11] A. Berger, V. Della Pietra, and S. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, 1996, pp. 39-71.

[12] A. Stolcke, "SRILM-an Extensible Language Modeling Toolkit," *Seventh International Conference on Spoken Language Processing*, 2002.
[13] F. Och, "Minimum error rate training in statistical machine translation," *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003, pp. 160-167.
[14] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 311-318