Comparative Study on Hierarchical Phrase Structures and Linguistic Phrase Structures

Tiejun Zhao, Yongliang Ma, Dequan Zheng and Sheng Li

MOE-MS Key Laboratory of Natural Language Processing and Speech
Harbin Institute of Technology, Nangang Xidazhijie 92, 150001 Harbin, China

Abstract. This paper proposes a framework for analysis of SMT translations output from a hierarchical phrase decoder. The tree display tool will show the translation process of the SMT model. An interactive operation tool will provide an adjusting mechanism for translation quality improvement. The work will explore automatic or semi-automatic identification and correction of some translation errors based on comparison between hierarchical phrase structures and linguistic phrase structures. Parts of the framework are implemented and primary results introduced.

1 Motivation

Automatic translation with high quality is the goal pursued by human for a long time. Statistical machine translation (SMT)[1, 2] inspired again people's hope to overcome language barriers between different nations from 1990s. The meaning of any language is expressed in the constraints of grammatical structures; machine translation between two languages cannot be excluded. So, introducing syntactic information to popular phrase-based models or designing syntax-based models becomes one of focuses in the research of SMT currently. While people expect more precise translations from syntax-based SMT model, there is not its steady outperformance than phrase-based SMT model. To what extent, the grammatical structures grasped by syntax model have improved the translation quality? Little in-depth investigation has been published although some work on linguistically motivated analysis had been done [3, 4].

There are two classes of syntactic approaches in recent SMT research - using syntax in formal sense and in linguistic sense [5, 6]. For the approach of using linguistic syntactic structures, there are three issues to limit its wide studies:

- the acquisition of a large size training corpus with syntax annotations in source language side or target language side, or with bilingual linguistic annotations – that demands lots of human labor;
- the lack of universal accepted linguistic syntax formalism;
- impractical precision of parsing techniques–generally, parsing result is the input of linguistic syntax-based translation models.

Zhao T., Ma Y., Zheng D. and Li S. (2009).
Comparative Study on Hierarchical Phrase Structures and Linguistic Phrase Structures.
In *Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science* , pages 97-102
DOI: 10.5220/0002155200970102

On the contrary, the hierarchical phrase translation model[7, 8], one of promising formal syntax approach does not need any manual annotation and has no worry on parsing precision: it uses only one nonterminal X to present all of possible syntactic structures of a sentence. The X-nonterminal is such a formal variable from phrases that extracted by phrase-based SMT model and may express some non-linguistic structures that is useful to translation between two languages. Actually, the set of structures expressed by hierarchical phrase model is a SMT-style syntactic formalism. This formalism carries more frequency information rather than linguistic information for those phrase structures. The study about the respective effect of syntactic structures and phrases to the translation quality will benefit from intensive exploring on the formalism.

Sparse research on the relation between the details of syntactic structures and translation quality is partly imputed to the lack of automatic translation evaluation on sentence-level. BLEU [9], the most popular method for automatic evaluation of machine translation system does not provide a sentence-level evaluation to identify which sentence is better or worse, and only gives a whole evaluation to the translation quality of a system. Many efforts are made to improve the approaches for automatic evaluation of machine translation in recent years[10]. The work has been implemented to automatically generate not only the quality score of a translated sentence but check-points for diagnostic evaluation[11]. We think the display and analysis for syntactic structures of translation output is one alternative for diagnostic evaluation of SMT system performance. It will reveal the reasons for the translation errors.

2 Hierarchical Phrase-based Translation(HPBT) Model[7, 8]

Formally, HPBT model is a weighted synchronous context-free grammar which learned from a parallel text without any syntactic annotations. Rules have the form $X \Rightarrow < \bar{e}||\bar{f} >$ where $\bar{e}$ and $\bar{f}$ are phrases consisting of terminal words and nonterminal symbol X which presents phrase hierarchically, so HPBT model employs a generalization of the conventional phrase-based translation model which does not allow hierarchical phrases. Briefly, decoding of HPBT model is a CKY style parsing process. Given a French sentence f, it finds the English yield of the single best derivation that has French yield f.

3 The Framework of our Study

In our work, we will try to tickle the following problems:

- What is the decoding process of hierarchical phrase translation? How the process affects the output of decoder (translation system)?
- What differences are there between hierarchical phrase structures and linguistic phrase structures, especially those frequent phrase structures used in decoding process?
- Whether these differences make mistakes for translation output? If yes, what are the key positions for those mistakes in a translation sentence?

A toolkit is necessary to solve these problems and it includes three components in shadow rectangles: a tree display tool for demonstration of decoding process while a sentence is translated; a phrase filter with frequency sorting to retain those frequent hierarchical phrases; a structure comparison and analysis tool based on the tree display and human-computer interactive operations. At last, the goal of our research is to build a kind of translation improvement strategy (in shadow ellipse) by adjusting the parameters of the decoder. Figure 1 shows the framework of our research. The core work is to compare the hierarchical phrase structures and linguistic phrase structures (in the mind of operators), to analyze errors and to propose corresponding improvement strategy based on the comparison and other information.

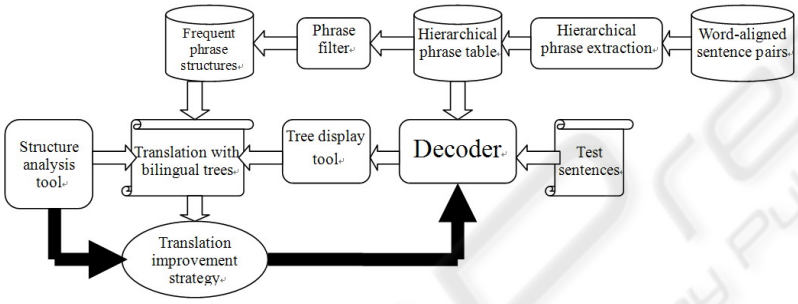

Fig. 1. The framework for improvement of hierarchical phrase decoding.

4 Primary Results

We implemented some modules of hierarchical phrase translation model and used them as the platform of our work which is illustrated in Figure 1.

The IWSLT (International Workshop on Spoken Language Translation) 2008 training set of Chinese-to-English translation evaluation is selected as the corpus of our research. The data set includes 629,101 sentence pairs and the average length of these sentence pairs is short, 12.54 words per Chinese sentence and 12.73 words per English sentence. Using the data, we can pay attention to the basic structures of bilingual mappings between Chinese and English and avoid much noise from long sentences.

Two modules in the framework are implemented: the filter of hierarchical phrases and the tree display tool. Statistics on all items of hierarchical phrase table is got by the phrase filter. More than 104,000 phrases with and without nonterminals were counted, but 80% of them only appeared one time. Table 1 shows the top 20 hierarchical phrases which contain X structures and most are too common to be processed separately. The symbol || is the delimiter of source and target sides in the same synchronous structure rule.

The tree display tool is used to detect the hierarchical phrase structure of decoding process. Figure 2 gives an example. When we want to translate Chinese sentence ”今天晚上的演出是民歌。”, the tree display tool gives a tree presenting the hierarchical phrase structure determined by the model during decoding, and the corresponding English translation is ”Fork song is today evening performance .”. Note that there is a

Table 1. Top 20 frequent hierarchical phrase structures.

Hierarchical phrase structure	Frequency	Hierarchical phrase structure	Frequency
$< X 。 \|\| X . >$	1364	$< X \|\| sorry\ X >$	134
$< X_1 。 X_2 \|\| X_1 . X_2 >$	298	$< X_1$ 可以 $X_2 \|\| X_1\ can\ X_2 >$	125
$< X_1 ， X_2 \|\| X_1 , X_2 >$	246	$< X_1$ 你 $X_2 \|\| X_1$you$X_2 >$	118
$< X_1$ 的 $X_2 \|\| X_1 X_2 >$	240	$<$ 我 $X \|\| I\ X >$	115
$< ， X \|\| , X >$	236	$< X_1$ 我 $X_2 \|\| X_1\ i\ X_2 >$	107
$< 。 X \|\| . X >$	209	$<$ 我 $X \|\| i\ X >$	98
$< X_1 ， X_2 \|\| X_1 . X_2 >$	208	$< X ， \|\| X , >$	93
$< X ？ \|\| X ？ >$	208	$< X$ 你$\|\| X\ you >$	90
$< ， X \|\| . X >$	189	$<$ 可以 $X \|\| can\ X >$	90
$<$ 你 $X \|\| you\ X >$	188	$< X_1$ 是 $X_2 \|\| X_1\ is\ X_2 >$	89

reordering at the node $[X, 0, 6]$, making the monotone Chinese sentence be reordered sentence ”民歌 是 今天 晚上 的 演出 。” which can be mapped to the English translation directly. Despite the reordering, the tree shown in Figure 2 is just the same as the parse tree of Chinese sentence with extracted SCFG from corpus.

We also parse the Chinese sentence with the parser of Stanford[12]. The parse tree from the Stanford parser is showed in Figure 3. Comparing the structure from the tree display tool and the Stanford parse tree, we find that ”的” and ”是” in the SCFG parse tree are encoded in the rule, and they are generated from nonterminals. This means that the SCFG can use these words as lexical information of which model takes advantage. We believe that there are a lot of significant differences between hierarchical phrase structrues and linguistic phrase structures, which could be used to improve the translation quality.

5 Future

We plan to implement the framework shown in Figure 1 in the future. Our studies will be focused on the following aspects:

- Statistics and classification on the distribution of different frequency ranges of hierarchical phrases; And long sentence-pair set (e.g. NIST corpus) is under consideration.
- Improvement on the tree display tool for bilingual trees in both sides to show the difference between source and target syntactic trees.
- Implementation of a platform for interactive operations on comparison between the structures of hierarchical phrases and linguistic phrases, on error analysis of translation output, and on realization of adjusting strategies for decoding process. A series of experiment results about comparison and adjustment will be reported.
- Exploration on automatic learning mechanism for classification and correction of translation errors.

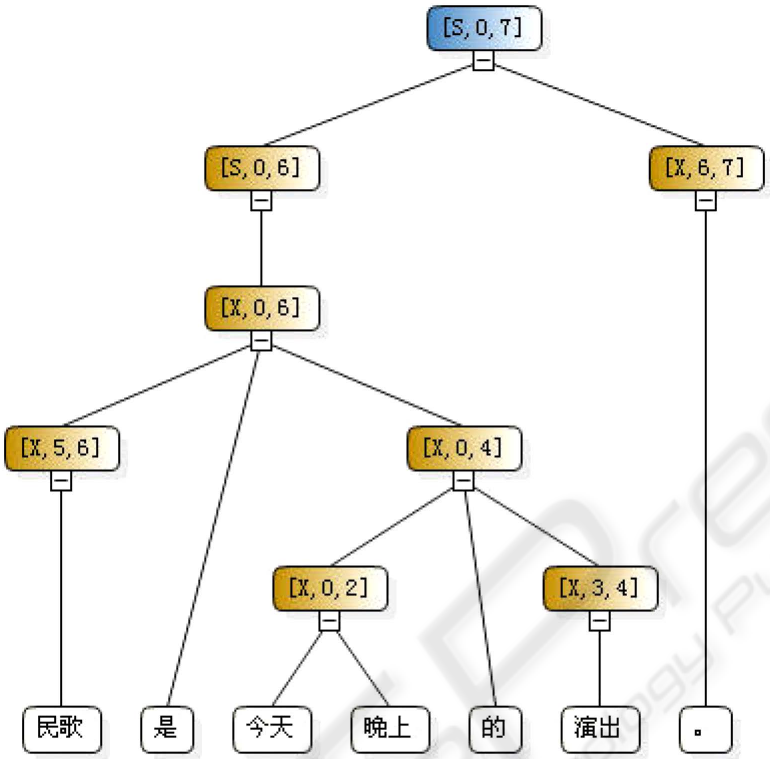

Fig. 2. Tree diagram of hierarchical phrase output from the decoder.

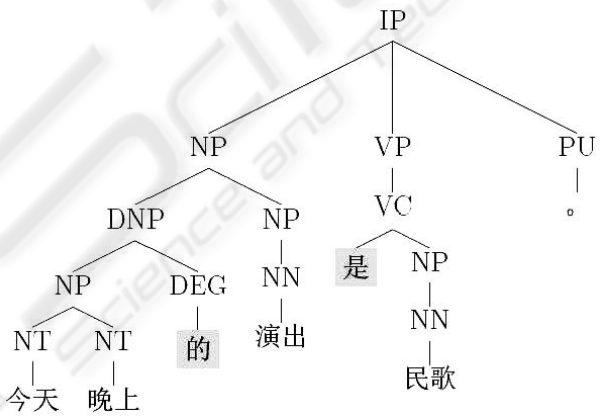

Fig. 3. Parse tree of Stanford parser.

Acknowledgements

The work of this paper is funded by the project of National Natural Science Foundation of China (No. 60736014) and the project of National High Technology Research and Development Program of China (863 Program) (No.2006AA0100108).

References

1. Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin: A statistical approach to machine translation. Computational Linguistics, Vol. 16(2), MIT Press(1990) 79–85
2. Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer: The Mathematics of Statistical Machine translation: Parameter estimation. Computational Linguistics, Vol, 19(2), MIT Press(1993) 263–311
3. Yuval Marton and Philip Resnik: Soft Syntactic Constraints for Hierarchical Phrased-Based Translation. Proceedings of ACL-08:HLT, Association for Computational Linguistics(2008) 1003–1011
4. David Chiang, Yuval Marton, and Philip Resnik: Online Large-margin Training of syntactic and structural translation features. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics(2008)224–233
5. Xiong Deyi, Liu Qun and Lin Shouxun: A Survey of Syntax-based Statistical machine translation. Journal of Chinese Information Processing, Vol. 22(2), The Commerciacl Press(2008) 28–39 (in Chinese)
6. Adam Lopez: Statistical Machine Translation. ACM Computing Surveys, Vol. 40(3), ACM(2008) 1–49
7. David Chiang: A Hierarchical Phrase-based Model for Statistical Machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics(2005) 263–270
8. David Chiang: Hierarchical Phrase-based translation. Computational Linguistics, Vol. 33(2), MIT Press(2007) 201–228
9. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu; Bleu: A Method for Automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics(2001) 311–318
10. http://www.nist.gov/speech/tests/metricsmatr/
11. Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang and Tiejun Zhao: Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points. Proceedings of the 22nd International Conference on Computational Linguistics, Coling Organizing Committee(2008) 1121–1128
12. Dan Klein and Chris D. Manning: Fast exact Inference with a factored Model for natural language parsing. Advances in Neural Information Processing System 15, MIT Press(2002) 3–10