Syntax-Directed Phrase-based Statistical Machine Translation

Name: Nguyen Thai Phuong
Supervisor: Prof. Akira Shimazu

Keywords: statistical machine translation (SMT), phrase-based SMT, syntactic transformation, word sense disambiguation (WSD)

1. The Aim: Incorporating Syntactic Transformation into the Decoding Phase of Phrase-Based SMT

A number of studies [Collins et al. (2005); Thai and Shimazu (2006)] used syntactic transformation in the preprocessing phase of phrase-based SMT. Though these studies showed that translation quality can be improved significantly, integrating syntactic transformation into the decoding phase is obviously a more natural approach than preprocessing. This is the aim of our study.

2. Approach: Using Syntax-Directed Translation Schemata (SDTS)

The SDTS has been applied in the field of compiler and in transfer-based machine translation. After the parsing step, the syntactic structure of a sentence is identified. The parse tree will be analyzed, augmented, and transformed by later phases in the SMT system. Those phases are controlled by syntax. We use the stochastic SDTS to model such kind of translation process for phrase-based SMT. This approach has a number of advantages as follow:

- Since translation is separated from parsing, parsers of the source language can be exploited.
- Since syntactic information of the source side is made use of, more control over the translation process can be taken.
- Do not require syntactic information of the target side since for many languages good parsers are still not available.
- Do not give up the strength of the baseline approach: phrases.

3. Progress of 2007

We consider a syntax-directed phrase-based SMT approach based on the stochastic SDTS. We propose a tree transformation algorithm and a tree-based decoding algorithm. The transformation algorithm converts a tree with word leaves into a tree with phrase leaves (phrase tree). The decoding algorithm is a dynamic programming algorithm which processes an input tree in a bottom-up manner. The syntactic transformation model is employed to control and score reordering operations. We conducted experiments with English-Vietnamese and English-Japanese language pairs. Experimental results showed a significant improvement in terms of translation quality.

A number of MT applications such as Web translation require high speed. Since full parsing may be slow for such applications, we consider chunking as an alternative. We study a chunking-based reordering method for phrase-based SMT [Thai & Shimazu (2007)]. This is an instance of the syntax-directed phrase-based approach. We employ the syntactic transformation model for phrase reordering within chunks. The transformation probability is also used for scoring

translation hypotheses. Chunk reordering is carried out in the decoding phase. This study shows another way to apply the syntactic transformation model to SMT.

Beside the word order problem, word choice is another obstacle for MT. Though phrase-based SMT has an advantage of word choice based on local context, exploiting larger context is an interesting research topic. We carried out an empirical study of integrating WSD into SMT. We implemented the approach proposed by [Carpuat and Wu (2007)]. Our experiments reconfirmed that WSD can improve SMT significantly. We used two WSD models including MEM and NB while [Carpuat and Wu (2007)] used an ensemble of four combined WSD models (NB, MEM, Boosting, and Kernel PCA-based) and [Chan et al. (2007)] employed SVM. We evaluated WSD accuracy, effect of phrase length, the use of syntactic relation feature for SMT.

We built a SMT system for phrase-based log-linear translation models. This system has three decoders: beam search, chunking-based, and syntax-based. We used the system for our experiments with reordering and WSD.

4. Future direction

There are several ways to extend our frame work of syntax directed phrase-based SMT. First, syntactic parsing is not perfect especially when a parser trained on Penn Treebank comes to analyze texts in a different domain. Using a n-best list of parses instead of 1-best is an extension to improve translation quality. Our decoding algorithm in Chapter 6 should be upgraded to represent an input tree forest and to search over it. A second way to improve translation quality is to deal with the flexible of adjunct attachment. Our decoder should allow a movement of adjuncts without changes the dependency structure of the input syntactic tree. This treatment leads to deal with a set of parses whose dependency structure is the same.

We also intend to apply the syntactic transformation model to improve word alignment. The notable GIZA++ tool is an implementation of IBM translation models (Model 1, 2, 3, 4, and 5). All models are word-based. The input and the output of the noisy channel are just sequences of words. The channel's operations are word duplications (including insertion and deletion), word movements, and word translations. Using a string-to-tree noisy channel model for word alignment, we expect to improve word alignment accuracy for language pairs which are very different in word order such as English and Japanese.

5. Publications

Systems:

[1] A program that generates WSD training data from bitext
[2] A chunk-based SMT decoder and a syntax-directed decoder

Papers:

[1] Nguyen Phuong Thai, Akira Shimazu, 2006. Improving Phrase-Based Statistical Machine Translation with Morphosyntactic Transformation. *Machine Translation*, Vol. 20, No. 3, pp 147-166.

[2] Nguyen Phuong Thai, Akira Shimazu, Le-Minh Nguyen, and Van-Vinh Nguyen, 2007. A Syntactic Transformation Model for Statistical Machine Translation. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, Vol. 20, No. 2, 1-21.