

Research Statement

Hao Zhang

June 5, 2008

My research focuses on statistical machine translation using synchronous grammars. My research interests also include parsing, automata, machine learning, efficient algorithms for language and string processing.

Research Overview

Machine translation was reborn in the early 90s when statistical approach was first introduced for this highly-complex NLP task. In recent years, MT quality has improved significantly by learning phrasal translation pairs from very large data sets of human-translated sentence pairs. The model for the state-of-art MT systems, however, remains to be very simple: flat reordering of phrases.

The most evident drawback of the the phrase-based systems is that they are incapable of handling systematic word or phrase re-orderings across two languages. Hierarchical models answer the call for richer models and are theoretically superior to flat models in terms of computational complexity. But in reality, simple flat models can use simple methods to tame the complexity. For hierarchical models, to control the complexity, we need more sophisticated techniques. Most of the hierarchical systems nowadays can be categorized into the framework of *synchronous context-free grammars* (SCFG). My thesis focuses on handling the complexity of SCFGs, specifically the structural complexity for both search and learning.

Research Contributions

Taming the complexity of synchronous grammars is the key to realize their theoretically promised advantage for the particular problem of re-ordering in MT. I believe that this line of research is leading to a certain type of probabilistic synchronous CFG that admits efficient algorithms and can serve as the backbone of MT as the role of probabilistic finite state transducers in speech recognition.

Factorization of Synchronous Grammars Synchronous CFGs are different from monolingual CFGs on factorization. I worked on binarization of synchronous grammars at USC/ISI. I found out that the majority (99.7%) of the synchronous CFG rules can be factorized into an equivalent set of binary rules, reducing the complexity of the decoding algorithm significantly and enabling the system to surpass a state-of-art flat phrase-based system [11]. This work led to theoretical investigations of the non-collapsing hierarchy of Synchronous CFGs. I developed efficient algorithms for factorizing arbitrary SCFG later [8], including combinatorial analyses for cases that can not be binarized [4], and multi-text grammars dealing with more than two languages [7]. Most recently, I discovered that given an alignment matrix, we can recover the minimal synchronous tree that generates the surface alignment, in time linear to the number of alignment links [10].

Efficient Alignment and Decoding As most practical SCFGs can be binarized, the search efficiency of binary SCFGs, such as ITG becomes more relevant to the advance in this field. My contributions in this line include A* synchronous parsing for ITG which uses IBM model 1 to compute the outside heuristic costs, which speeds up ITG Viterbi alignment by several times over CYK [5]. I also implemented an A* decoder for ITG. My most recent work [9] improved the

efficiency of the A* decoder by using a two-pass decoding approach where the first pass is a bigram-integrated decoding pass that parses inside-outside to produce outside estimates for the trigram-integrated states in the second pass. I also extensively experimented with other techniques to speed up the decoding process, such as the dynamic programming hook trick [1]. To address the mismatch between intrinsic model scores and BLEU scores, I used another decoding pass on the trigram-integrated forest to maximize the expected count of synchronous constituents which improved the translation quality significantly in terms of BLEU score [9].

Learning Synchronous Grammars This part of research is on the unsupervised learning of synchronous grammars. I have been focusing on ITG extensions. I first discovered one-to-one alignment given by EM training of word-based ITG is competitive [2]. I have worked on lexicalized ITG [3], its more efficient variants that are bilexicalized [6]. My latest work done at MSR in the summer of 2007 [12] was on the phrasal extension of ITG where the terminal rules generate multi-word pairs. I applied Variational Bayes to phrasal ITG to substitute EM as the learning algorithm and showed a sparse prior could be successfully enforced to improve both alignment and translation. The system has the minimal external reliance on non-syntactic models such as IBM models and shows significant improvement over the state-of-art in an end-to-end system.

Future Directions

I foresee several promising directions for research on synchronous grammars. Following the line of alignment factorization, I hope to discover the minimal set of synchronous rule patterns that are necessary for covering the majority of word alignment phenomena. I will answer the question on how much expressiveness is required for MT from a practical point-of-view. The first successful hierarchical system of David Chiang starts from the word alignment provided by a non-syntactic, non-hierarchical model. I have the goal of building an end-to-end SCFG-based system that releases the full power of SCFG. My PhD research is leading to the goal. This is not an easy goal. The work I did in the summer of 2007 at MSR is a good starting point. It only needs IBM model 1 for pruning the phrasal alignment space. There are many possible improvements that can be made to the work, such as using non-parametric Bayesian models to favor even sparser solutions thus to reduce the reliance on hard constraints. The SCFG it uses can be augmented to include discontinuous alignment patterns as well.

Research Goals

Broadly speaking, my interests are in building applications that can promote access of information written or spoken in natural languages. Machine translation is one such example. As the Internet expands and enriches in every possible way, natural language processing is facing both opportunities and challenges. I think machine learning for machine translation in the web environment will become more and more important and may bring forth a new learning paradigm for machine translation and other cross-lingual information access applications. At the same time, efficient algorithms that can handle problems of large scale and complex structures will become more and more important. With my background in dealing with complex NLP problems, I hope I can contribute to reaching the goal of bridging human languages and information systems.

References

- [1] Liang Huang, Hao Zhang, and Daniel Gildea. Machine translation as lexicalized parsing with hooks. In *International Workshop on Parsing Technologies (IWPT05)*, Vancouver, BC, 2005.
- [2] Hao Zhang and Daniel Gildea. Syntax-based alignment: Supervised or unsupervised? In *COLING-04*, Geneva, Switzerland, August 2004.

- [3] Hao Zhang and Daniel Gildea. Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of ACL-05*, 2005.
- [4] Hao Zhang and Daniel Gildea. Efficient factorization of synchronous context-free grammars. Technical Report 889, University of Rochester, 2006.
- [5] Hao Zhang and Daniel Gildea. Efficient search for inversion transduction grammar. In *Proceedings of EMNLP*, Sydney, 2006.
- [6] Hao Zhang and Daniel Gildea. Inducing word alignments with bilexical synchronous trees. In *COLING/ACL-06 Poster*, Sydney, 2006.
- [7] Hao Zhang and Daniel Gildea. Enumeration of factorizable multi-dimensional permutations. *Journal of Integer Sequences*, 07(5.8), 2007.
- [8] Hao Zhang and Daniel Gildea. Factorization of synchronous context-free grammars in linear time. In *NAACL Workshop on Syntax and Structure in Statistical Translation (SSST)*, 2007.
- [9] Hao Zhang and Daniel Gildea. Efficient multi-pass decoding for synchronous context free grammars. In *Proceedings of the 46th Annual Conference of the Association for Computational Linguistics: Human Language Technologies (ACL-08:HLT)*, 2008.
- [10] Hao Zhang, Daniel Gildea, and David Chiang. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, 2008.
- [11] Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. Synchronous binarization for machine translation. In *Proceedings of NAACL-06*, pages 256–263, 2006.
- [12] Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proceedings of the 46th Annual Conference of the Association for Computational Linguistics: Human Language Technologies (ACL-08:HLT)*, 2008.