From: AAAI-97 Proceedings. Copyright © 1997, AAAI (www.aaai.org). All rights reserved.

Computing Discourse Information with Statistical Methods[1,2]

Kenneth B. Samuel
Department of Computer and Information Science
University of Delaware
Newark, Delaware 19716 USA
samuel@cis.udel.edu

This dissertation research involves implementing a computer system that, given a natural language dialogue, will automatically tag each utterance with a *discourse label* (a concise abstraction of the intentional function of the speaker) and a *discourse pointer* (a focusing mechanism that represents the dialogue context in which an utterance is to be understood). (Samuel 1996)

Since the discourse label of an utterance is dependent on the surrounding dialogue, tagging utterances with discourse labels is similar to the part-of-speech (PoS) tagging problem in syntax. Within the domain of PoS tagging, extensive experimental research has shown that statistical learning algorithms are among the most successful. I will investigate two methods that have been effective in PoS tagging: Hidden Markov Models (HMMs) (Charniak 1993) and Transformation-Based Learning (TBL) (Brill 1995).

Unlike these PoS taggers, which determine a word's tag based on the surrounding words (within a fixed window size), a discourse-tagging system must use the surrounding utterances as input. Thus, the sparse data problem is much more severe for the discourse tagger, since the number of possible utterances is infinite. To alleviate this problem, rather than directly processing each utterance verbatim (which would probably bombard the system with a great deal of extraneous information that is not relevant to the task at hand), I have identified a small set of features that can be extracted from each utterance to provide the relevant information to the learning algorithm.

Since HMMs and TBL deal with contiguous sequences of discourse labels, they are unable to take focus shifts into consideration. But it is crucial to account for the focus shifts that frequently occur in discourse. I have proposed a solution to this problem for both algorithms. For HMMs, this involves modifying the Markov assumption slightly, while still retaining the linear-time efficiency of the HMMs approach. With TBL, the solution is more straightforward.

The TBL algorithm also requires other modifications. Since the tag assigned can depend on information in the future context, TBL is unable to effectively produce discourse labels for a dialogue that has not yet completed. But if a computer system is to participate in a conversation, it must be able to analyze an incomplete dialogue. I have proposed a strategy that involves training the system on preliminary answers (which do not depend on the future context) as well as the final discourse labels.

Also, TBL does not provide a measure of confidence for each answer it produces. Confidence measures can help to decide how to resolve discourse labels and conflicting information from other sources. To compute confidence measures, I have outlined a novel application of the recently-proposed Committee-Based Sampling (Dagan and Engelson 1995) technique.

To improve the reliability of human tagging, I am developing a method for automatically organizing discourse labels into a taxonomic structure and then converting it into a decision tree that will aid in manually annotating dialogues. The human annotator will not be required to determine the discourse labels directly. Instead, he will make judgements about simple features of each utterance.

Space constraints preclude discussing other problems that I am currently addressing.

References

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics.* 21(4): 543–566.

Charniak, Eugene. 1993. *Statistical Language Learning.* Cambridge, Massachusetts: MIT Press.

Dagan, Ido and Engelson, Sean P. 1995. Committee-Based Sampling for Training Probabilistic Classifiers. 150–157. Proceedings of the 12th International Conference on Machine Learning.

Samuel, Ken. 1996. Using Statistical Learning Algorithms to Compute Discourse Information, Technical Report, #97-11, Department of Computer and Information Science, The University of Delaware.

[1] Copyright © 1997, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

[2] This work was partially supported by the NSF Grant #GER-9354869.