

Student Research Symposium 2004

Kenji Sagae

Using Dependencies for Easy, Fast and Accurate Grammatical/Functional Analysis

Modern statistical syntactic parsers have achieved very high levels of accuracy over the past ten years, and we have begun to see their impact on several areas of language technologies, such as question answering, machine translation, and semantic-role labeling. Because the Penn Treebank (PTB) is widely used for training of such parsers, it is common to associate PTB-style constituent trees with statistical parsing. However, there are instances where other syntactic representations would be easier to use, and just as useful (if not more). One such instance is the assignment of grammatical relations (or even PTB function tags) to words. In this case, dependencies are not only more comfortable to understand and faster to annotate, but also easier to process and largely just as effective.

I will discuss a simple representation based on lexical dependencies, which I have been using in the syntactic analysis of parent-child dialogs. I will present a simple deterministic algorithm for dependency parsing, and show the accuracy of the dependencies it produces is very close to the accuracy of current PTB constituent statistical parsers (91% vs. 93%). Although PTB constituent parsers have a slight edge, they are quite complex. I will show that a dependency parser that performs almost as well can be surprisingly simple and fast.

I will also discuss how these dependencies can be used to determine PTB function tags (such as subject, predicate, temporal, beneficiary, locative, etc). The current state-of-the-art on assigning function tags to text is the work of Blaheta (2000, 2003), and it uses (among other features) PTB parse trees nodes. I will present results that are very similar using no constituent information, only dependencies. Both methods achieve an overall accuracy of about 87% in function tagging (not counting .NULL. tags). Blaheta's method is slightly better on tags classified as .grammatical. (subject, predicate, etc), while the dependency approach is slightly better on .form/function. tags (temporal, locative, manner, etc).

This approach to function tagging can also be used to label all dependency arcs, when training data is available. In fact, a relatively small training corpus (less than 10,000 words) can be used to produce a system that assigns a grammatical relation label to every dependency arc with an accuracy of about 90% in a corpus of parent-child dialogs.