Abstract

Parsing natural language is an attempt to discover some structure in a text (or textual representation) generated by a person. This structure can be put to a variety of uses, including machine translation, grammar conformance checking, and determination of prosody in text-to-speech tasks.

Recent theories of Syntax use Unification to better describe the intricacies of natural language [137]. For parsing systems, unification techniques have been either added to a context-free base system [152, 40, 4, 23], or replaced the context-free base entirely [118, 135, 45] (possibly putting it back later [136]). The seemingly small step of adding unification has opened a Pandora's Box of computational complexity, increasing the difficulty of the problem from polynomial [48] to somewhere between NP-complete and intractable, depending on the details of the unification system and how it was added [10]. Worse, unification on a context-free base parser can break the packing technique used to address the problem of ambiguity, leading to exponential blow-ups of the parser's performance in both space and time in practice.

I propose the use of a multi-pass strategy to avoid these problems in practice. I describe a parser which combines the use of shallow, simple value unification with some approximation techniques in order to find a covering packed parse-forest. This parse-forest is then searched for a single-best fully-unifying value; the scoring system which drives the heuristic search encodes linguistically-based disambiguation preferences.

The resulting two-pass parser is compared to an ordinary single-pass parser in the context of a heavy-weight knowledge-based machine translation system. The two-pass parser is shown to be competitive with the single-pass parser on average data, both in terms of time and space. It is also shown to be able to avoid a common class of ambiguity blow-up that the single-pass parser is subject to. These results indicate that the multi-pass technique, interleaving *some* of the unification equations in the parse, is the superior approach for heavy-weight unification parsing.