ALIGNING MILDLY CONTEXT-SENSITIVE FORMALISMS FOR DATA-DRIVEN PARSING

ILYA KASHKAREV

ABSTRACT

In this thesis, we work with threee formalisms: head-driven phrase structure grammars (HPSG), abstract categorial grammars (ACG), and linear context-free rewriting systems (LCFRS). The ACG and LCFRS formalisms are weekly equivalent, i.e. they generate the same string languages. They represent the class of mildly context-sensitive formalisms, whereas HPSG is capable of generating more complex languages. We concentrate on practical and theoretical properties of the two mildly context-sensitive formalisms.

On the theoretical level we construct a partial convertion of HPSG into ACG. In fact, we use hyperedge replacement grammars (HRG) as the target of our convertion, but some very tight correspondence of HRG and ACG was already proved before. Moreover, the complexity results suggest that if we try to build such a conversion in a more general fashion, then the second-order ACG will not be sufficient. But higher-order ACGs are not yet well-studied. It seems also clear that there may not in principle exist a fair conversion of HPSG into HRG, basically because head-driven phrase structure grammars do not have context-free derivations. Thus, in this work, we try to establish conditions, possibly not too restrictive, under which a grammar in HPSG formalism may be modeled as an HRG.

The practical side of our investigation concerns LCFRS as there is on-going research on LCFRS applications. We enrich the formalism with additional external information to improve parsing. As the target language we chose German. There are big availiable German corpora which one may use as the training data. German has variable word order patterns and it is very difficult to obtain good results within formalisms restricted to local trees. Frequently occuring discontinuities present lots of material to work on.

We implement an LCFRS parser for German, train and test it. The parser, used with a context-free grammar does the usual probilistic context-free parsing. It provides us with a baseline to compare our results to. We also compare the results to other current state-of-art parsers for German. Our theoretical results and error analysis guide us to introduce modification to the parser.