Semi-automatic coreference enrichment for information structure research

Erwin R. Komen
Centre for Language Studies
Radboud University Nijmegen
E.Komen@Let.ru.nl

Knowing how the interplay between information ordering rules and syntax rules has functioned in English will help us understand better how languages in general work. Our project aims to find out how this interplay has shaped English throughout its development. The main direction of our project is to (a) start with available syntactically annotated treebank texts from different time-periods, (b) add coreference information, and (c) find out which word orders or constructions are used to fulfill information-structure related tasks, such as topic intro, maintenance, and shift, and the expression of different kinds of focus.

Our first approach was to add coreference information to the existing texts manually, albeit with the help of a specially developed editor called "CESAC". This approach had several shortcomings. There were inconsistencies in coreference tasks that could have been solved more mechanically, and since our system did not require all NP constituents to be tagged (we only indicated coreference links and assumed information) a certain number of constituents that should have been tagged, were left out. The manual labour was tedious, and motivation to continue marking texts decreased. There also were some technical difficulties connected with the treebank encoding of the texts we worked with.

Our current approach has switched from the treebank encoding to an xml one. What we set out to do is add coreference information semi-automatically for *all* noun phrases. We have developed an algorithm called CESAX, which is based on existing natural-language ones. The algorithm makes links that can be done automatically (about 60%), and makes suggestions for the ambiguous or otherwise suspicious situations it encounters. Its suggestions are correct in almost half the cases. A small amount of the automatically made links (about 6%) are incorrect, but Cesax allows for more fine-tuning.

Several texts have been annotated using CESAX, and we have started developing corpus research projects written in Xquery for a few tasks that combine information structure with syntax.

The results are promising. Cesax is a real improvement and the test cases have shown that the enriched xml texts can effectively be used to answer the kind of questions we ask. What is more: the enriched texts offer the potential for research that makes use of the coreference chains they contain.