

Issue Statement – Digital Tool Summit in Linguistics

Marcus Sammer

Corpus driven methods are a staple in natural language processing. Part of speech taggers, syntactic parsers, morphological analyzers, document classifiers, and named entity recognizers are among the tools that depend on high quality tagged corpora. With the amount of textual information available on the web, it is surprising how difficult it can be to obtain large high quality corpora even for major languages. And when looking for corpora tagged with syntactic or semantic information, it becomes all the more difficult. Even when a high quality tagged corpus exists, tools trained on that corpus degrade quickly when applied outside of the domain of the original corpus.

Right now the Linguistic Data Consortium is a major provider of corpora used in natural language processing. One critical problem though, besides expense, is that their corpora are largely static objects: purchased, downloaded, and worked on in isolation. Instead, imagine a Wiki style environment in which anyone can add a new corpus or submit a new tagging of another corpus. This would be a dynamic environment, responsive to the needs of the community. It would necessitate the adoption of standards for formatting and saving data, and it could also spawn the creation of new tools for automating tasks necessary for building and annotating corpora.

Such an environment could also be a destination for the data collected in projects like the StoryCorps project (<http://storycorps.net/>) which is currently creating a snapshot of American oral history by recording people interviewing their friends and families. And so this environment could be used by those recording history, analyzing and preserving languages, or just creating language processing software.