Prague Dependency Treebank of Spoken English

The Prague Dependency Treebank of Spoken English is a collection of English spoken dialogs about personal photograph collections. The original corpus consisted of three interlinked representations:
- ASR output aligned to audio;
- manual transcription;
- reconstructed text.

The data is annotated with formats based on the Prague Markup Language (PML) which is a backbone for the family of XML schema for rich linguistic annotations of texts, such as morphological tagging and dependency trees. The corpus can serve as a training and testing material for machine learning experiments in both intelligent editing as well as in dialog language understanding.

Morphological layer enhancement

We converted the morphological layer of the corpus into a treebank in the standard Penn Treebank bracketing style and enhanced it with:
- part of speech tags;
- named entity labels;
- WordNet hypernyms;
- links to the lower layers of annotation.

The pre-processed corpus data was given as an input into state-of-the-art NLP tools such as the Stanford parser and named entity recognizer, and the WordNet API to obtain the additional analyses. These annotations were added in such a way as to preserve the original PML format.

Browsing, editing and querying possibilities

The main motivation for corpus enrichment is its preparation for information extraction task and linguistic research. The corpus in PML format enables its browsing and editing in PML-tree editor TrEd and querying with a powerful search engine PML Tree Query (PML-TQ).

Figure below shows a query “find the tokens that are nouns and have a hypernym “anniversary” in the PML-TQ environment. The system outputs all the sentences that contain tokens consistent with the query (such as “birthday”).

Concluding remark

A new layer of annotation has been added to multi-layered corpus data in a complex format by combining several tools and merging their partial outputs. The augmented corpus contains interesting strata of linguistic knowledge, is compatible with a specialized open-source query engine and is suitable for extensive information extraction.

Acknowledgment

I would like to thank Dr. Silvie Cinková for her supervision of the project and providing me with access to the data (Institute of Formal and Applied Linguistics, Charles University in Prague).

References
