Cedit – Semantic Networks Manual Annotation Tool

Václav Novák
Institute of Formal and Applied Linguistics
Charles University
Malostranské nám. 25, 11800 Praha, Czech Republic
novak@ufal.mff.cuni.cz

Abstract

We present a demonstration of an annotation tool designed to annotate texts into a semantic network formalism called MultiNet. The tool is based on a Java Swing GUI and allows the annotators to edit nodes and relations in the network, as well as links between the nodes in the network and the nodes from the previous layer of annotation. The data processed by the tool in this presentation are from the English version of the Wall Street Journal.

1 Introduction

Cedit is a part of a project to create a rich resource of manually annotated semantic structures (Novák, 2007) as a new layer of the Prague Dependency Treebank (Sgall et al., 2004). The new layer is based on the MultiNet paradigm described in (Helbig, 2006).

1.1 Prague Dependency Treebank

The Prague Dependency Treebank is a language resource containing a deep manual analysis of text (Sgall et al., 2004). PDT contains three layers of annotation, namely *morphological*, *analytical* (shallow dependency syntax) and *tectogrammatical* (deep dependency syntax). The units of each annotation level are linked to corresponding units from the shallower level. The morphological units are linked directly to the original text.

The theoretical basis of the treebank is described by the Functional Generative Description of language system (Sgall et al., 1986).

1.2 MultiNet

Multilayered Extended Semantic Networks (MultiNet), described in (Helbig, 2006), provide a universally applicable formalism for treatment of semantic phenomena of natural language. They offer distinct advantages over classical predicate calculus and its derivatives. Moreover, semantic networks are convenient for manual annotation because they are more intuitive.

MultiNet's semantic representation of natural language is independent of the language being annotated. However, syntax obviously varies across languages. To bridge the gap between different languages we can the deep syntactico-semantic representation available in the Functional Generative Description framework.

2 Project Goals

The main goals of the project are:

- Test the completeness and intuitiveness of MultiNet specification
- Measure differences in semantic networks of parallel texts
- Enrich the Prague Dependency Treebank with a new layer of annotation
- Provide data for supervised training of text-to-semantic-network transformation

NAACL HLT Demonstration Program, pages 11–12,
Rochester, New York, USA, April 2007.

- Test the extensibility of MultiNet to other languages than German

3 Cedit

The presented tool has two key components described in this section.

3.1 Input/Output processing

The input module of the tool loads XML files in Prague Markup Language (PML) and creates an internal representation of the semantic network, tectogrammatical layer, analytical layer, and the surface text (Pajas and Štěpánek, 2005). There is also an option to use files with named entity annotations. The sentences in this demo are all annotated with named entities.

The XML schema for the semantic network is an application of the Prague Markup Language.

3.2 Network GUI

The annotation GUI is implemented using Java Swing (Elliott et al., 2002). The key features of the tool presented in the demonstration are:

- Editing links between the semantic network and the tectogrammatical layer
- Adding and removing nodes
- Connecting nodes with directed edges
- Connecting edges with directed edges (i.e., creating relations on the metalevel)
- Editing attributes of both nodes and edges
- Undoing and redoing operations
- Reusing concepts from previous sentences

4 Related Work

There are various tools for annotation of the Prague Dependency Treebank. The Tred tool (Hajič et al., 2001), for example, allows users to edit many PML applications, even those that have never been seen before. This functionality is enabled by *roles* in PML specification (Pajas and Štěpánek, 2005). MultiNet structures can be edited using MWR tool (Gnrlich, 2000), but this tool is not primarily intended for annotation; it serves more as an interface to tools automatically transforming German sentences into MultiNet.

Acknowledgement

This work is supported by Czech Academy of Science grant 1ET201120505 and Czech Ministry of Education, Youth and Sports project LC536. The views expressed are not necessarily endorsed by the sponsors.

References

James Elliott, Robert Eckstein, Marc Loy, David Wood, and Brian Cole. 2002. *Java Swing*. O'Reilly.

Carsten Gnrlich. 2000. MultiNet/WR: A Knowledge Engineering Toolkit for Natural Language Information. Technical Report 278, University Hagen, Hagen, Germany.

Jan Hajič, Barbora Vidová-Hladká, and Petr Pajas. 2001. The Prague Dependency Treebank: Annotation Structure and Support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 105—114, Philadelphia, USA. University of Pennsylvania.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer-Verlag, Berlin Heidelberg.

Václav Novák. 2007. Large Semantic Network Manual Annotation. In *Proceedings of 7th International Workshop on Computational Semantics*, pages 355–358, Tilburg, Netherlands.

Petr Pajas and Jan Štěpánek. 2005. A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0. Technical Report 29, UFAL MFF UK, Praha.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing company, Dodrecht, Boston, London.

Petr Sgall, Jarmila Panevová, and Eva Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In A. Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, Massachusetts, USA. Association for Computational Linguistics.