Cooperative WordNet Editor for Lexical Semantic Acquisition

Julian Szymański
Gdańsk University of Technology
Narutowicza 11/12, 80-952 Gdańsk, Poland
julian.szymanski@eti.pg.gda.pl

Abstract. The article describes an approach for building WordNet semantic dictionary in a collaborative approach paradigm. The presented system enables functionality for gathering lexical data in a Wikipedia-like style. The core of the system is a user-friendly interface based on component for interactive graph navigation. The component has been used for WordNet semantic network presentation on web page, and it brings functionalities of modification its content by the distributed group of people.

Keywords: Lexical semantic, Acquisition, Semantic dictionaries, Collaborative editing, WordNet.

1 Introduction

WordNet is one of the largest semantic lexicons of English. It has been developed since 1985 by the Cognitive Science Laboratory at Princeton University. Its authors, based on theories of human cognition, try to reflect all linguistic dependencies between concepts in a common lexical database. The WordNet team has been working on a semantic dictionary for over 22 years. Nowadays, the dictionary contains about 155287 words, organized in 117659 synsets (meaning representations), and includes 206941 pair words – meaning. Introduction of all words with their connections, as well as examples of their usage in language, requires a lot of human work, however the WordNet team has only seven members. The WordNet project has been supported by plenty of grants, which brought together 3 millions dollars. Currently the third release of the WordNet lexical database is available at the project website. WordNet develops as a research project in a closed academic environment. The first version of the dictionary appeared in 1993, and now a third version is available. The dictionary is publicly available, but its modification is restricted to internauts. Probably, the reason for that, is the fact that the lexicon is organized as a set of text files in a specific format, which makes it hard to apply cooperative approach for WordNet development. Lack of cooperative editing functionality is the biggest barrier to scale-up semantic database.

The most well known application of a cooperative approach for gathering data is Wikipedia. The project has experienced great interest from the Internet community.
which brought many positive results. Wikipedia has been developed since 2001 by volunteers from all over the world. Currently, the Wikipedia initiative is supported by almost 75000 people, working on over nine million articles written in 125 languages. The largest set of articles is available in English, and contains over 2 million articles.

Nowadays, a lot of projects has been created on the basis of WordNet\textsuperscript{3}. They use semantic dictionary as a core knowledge base about language, what enables to implement elementary linguistic competences in a machines.

Some of the implementations do the mapping from WordNet files to other models, especially relational. This can be used to enable a cooperative editing approach.

\section{Description of WordVenture System}

A WordVenture portal\textsuperscript{4} has been developed at the Gdansk University of Technology at the Faculty of Electronics, Telecommunications and Informatics. It provides mechanisms for simultaneous work on lexical dictionaries for distributed groups of people and enables cooperative work on aWordNet lexical database. The Cognitive Science Laboratory approach to WordNet development required huge amounts of resources e.g human, time, money \cite{7}. With WordVenture, lexical database development becomes common and cheap.

With WordVenture, a user can browse a WordNet dictionary, and display its content on the screen with a graphical user interface based on an interactive graph. It gives a user-friendly way for visualizing very large sets of contextual data. A user can also query WordVenture to find a specified word and display its senses and related concepts. Connections between nodes (words or senses) are illustrated as edges of a given type. To keep graphs clear, a user can set some constraints to visualize only required types of data. There is also the possibility of interactive graph traversing. Selecting one node all elements that are connected with the marked one are displayed (according to given constraints on data selection).

The advantage brought to WordNet development by WordVenture system is a possibility of editing semantic database by the open, Internet community, which fasten lexical data acquisition process. To provide high quality of the acquired data, all changes introduced by users are represented as change propositions, which are approved or rejected by a privileged user – moderator.

\section{System Architecture}

It was decided that the WordVenture system will be implemented in client-server architecture, with the following assumptions:

\begin{itemize}
  \item WordNet database and data access logic resides on the server,
  \item Data visualization mechanisms reside at the client side and provide interfaces to the lexical database in the form of interactive graphs.
\end{itemize}

\textsuperscript{3} See: related projects http://wordnet.princeton.edu/links
\textsuperscript{4} http://wordventure.eti.pg.gda.pl