

Wordventure – developing WordNet in Wikipedia-like style

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 way. The idea of gathering lexical data has been proposed, as well as the system for linguistic data acquisition and management.

1 Introduction

WordNet (Fellbaum and others, 1998) is one of the most popular digital semantic lexicons of English. Its main advantage is that it is made by hand, so data stored within its semantic network are high quality. On the other hand these data cover only a small part of the relations between lexical elements, so there is a need to scale-up the project. Creating a large scale semantic dictionary in a manual way is labor-consuming and relatively slow. Alternative approaches for building semantic networks have been proposed, eg: Microsoft MindNet (Vanderwende et al., 2005), built from text documents parsing, or MIT ConceptNet (Liu and Singh, 2004) built from parsing simple sentences contained common sense knowledge, aquired through web page. Methodology used in this projects allows to build large scale semantic networks, although their quality isn't as high as hand crafted data. The other issue is that they operate only on words, not as WordNet on word meanings (synsets).

WordNet is being built as a research project in Princeton by a group of linguists. The WordNet team has been working on a semantic dictionary for over 22 years. Because of the limited human resources the speed of development of the project is limited. Our goal is to deliver a generally available tools for cooperative development of semantic networks. Building semantic dictionaries by hand requires a large amount of human resources, generally grouped in one place. In our approach we would like to exploit the power of the Internet and give open community a set of tools which

would allow a cooperative modification of WordNet.

The rest of this paper is organized as follows. The next section presents the idea of cooperative editing paradigm, which was applied to WordNet dictionary development. Section 3 describes the architecture and technical details of the Wordventure system. The subsections of this paragraph provide insight into server and client application features of the system. The concluding section presents the future plans regarding the presented approach and application.

2 Cooperative approach for editing WordNet

The best known application of a cooperative approach to gathering textual data is Wikipedia. The project received a 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.

Current implementations of WordNet web based applications are limited to database exploration, moreover they resemble the standard, dictionary-like, web interface for WordNet. Lack of tools for cooperative editing of semantic dictionary databases is the main barrier for rapid WordNet development. Our aim is to deliver a tool enabling a cooperative editing approach for many users placed in distributed Internet environment (Szymański et al., 2007).

Cooperative approach to editing content on the Internet is gaining increasing recognition in many IT fields. The main goal of our project is to create a system that would enable Web users free access and easy-to-use interface for WordNet con-

tent navigation and editing in an interactive, dynamic way. Moreover, the functionalities and the look and feel of the system should encourage web users to feed WordNet database with data.

The editing process in presented scheme consists of the following steps:

1. Users input data on their clients, which communicate changes to the server.
2. Server logs the operation and executes suitable procedures on the database.
3. Periodically, a moderator that has direct access to the server log and the database analyses logs and decides whether any of the user's modifications should be rolled back.

After several editing steps the original database is enriched with the content chosen from users contributions. This procedure is supported with regular database backups. Described editing process is similar to Wikipedia procedures which include regular content checks for vandalism and disrupting activities. If our approach proves successful in presented scenario it could be extended for building semantic databases in general. The example of Wikipedia gives reason for hope that with a proper system design we could achieve satisfactory results in this field at least.

Cooperative editing is connected with publishing the WordNet database and making it open to the Internet community. This might bring advantages for faster WordNet development. However some problems may arise:

- **Vandalism** – may cause loss of important data, kept in current release of lexical database. It can also affect the data structure e.g. creation of pointless connections between words and synsets. Because of that, it is important to deliver tools for moderating the users activities, which will reduce the risk of the above-mentioned.
- **Simultaneous** work on the same part of the database by many users may cause some conflicts resulting from concurrent work of many users at the same time. In the worst case one user can add the connection to an element of the WordNet dictionary that was deleted by another.

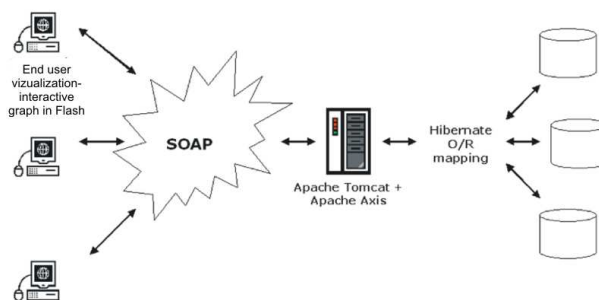


Figure 1: Basic concept of the WordVenture architecture and its elements.

3 System architecture

A WordVenture portal¹ 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 a WordNet lexical database. The Princeton Cognitive Science Laboratory approach to WordNet development requires a huge amount of resources: e.g. people, time, money (Miller et al., 1990). With WordVenture lexical databases development becomes common and cheap. Our system offers functionalities to browse a WordNet dictionary and display its content on the screen with a graphical user interface based on an interactive graph. (The example is in Figure 3). It gives a user-friendly way for visualizing very large sets of contextual data.

The system supports cooperative editing approach for the WordNet database development. It has been implemented in a standard client-server architecture presented in Figure 1: with database and WordNet logic tier residing on the server and the visualization engine querying the server as a client application.

The success of a platform for cooperative editing depends on effective and easy-to-use graphical user interface. In order to achieve that we decided to use an interactive visualization engine that would be able to render graph-like structures and allow to implement editing features. In our implementation light-weight component for graph visualization enables convenient navigation in graph-like structures and provides basic support for graph editing.

¹<http://wordventure.eti.pg.gda.pl>

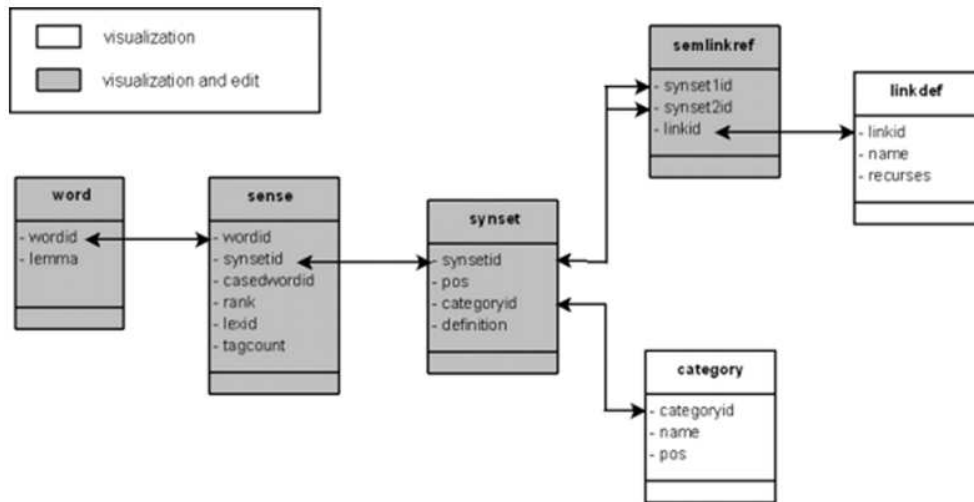


Figure 2: WordNet entities supported by the tool. Grayed out entities have support for both visualization and editing, white entities have only visualization support. Arrows represent relationships between entities.

3.1 Server side and Database

The server-side of the WordVenture application makes its functionalities available through web services. According to communication interoperability requirement it is possible to connect different client applications that can be implemented in different technologies.

Implemented functionalities allow a user to perform four different groups of actions depending on the role that the user has:

- **Functionalities for browsing WordNet lexical database.** Are available to every user (anonymous and logged-in) and gives an opportunity to look through WordNet with interactive interface.
- **Functionalities that allow a user to edit WordNet lexical database.** Are only available to registered users. After performing an edit action on the client-side of an application the proper change proposition is created. Subsequently, this proposition is sent to the server to be added to the database.
- **Functionalities for managing the new data.** A privileged user (moderator) can view all change propositions and select data to commit or cancel. After committing, a proposition is permanently added to database and can be seen by other users.
- **Administrative functionalities connected with user management.** Are available only

for privileged users – administrators. They allow to perform user deletion or user rights editing in WordVenture system. Every administrator can give administrative rights to another user.

The original implementation of a WordNet database uses text files. Because of their structure, modification is available only with dedicated tools. This type of storage doesn't support synchronous access for modification, nor allows to perform efficiently large amount of queries.

It was required to create special mechanisms for editing, including synchronization and file structure refactoring after any operation. To enable editing a WordNet lexical database through web we had to perform mappings between WordNet text files and a relational database. Transformation from text files to its relational representation was performed by the WordNet SQL Builder tool². Data access routines were implemented with Hibernate ORM engine³. Manipulating the database content is made via implemented server API exposed as Web Services, which fulfills requirements of Service Oriented Architecture (SOA) (Erl, 2005) paradigm. The Web Services has been deployed on Apache Log4j on a Tomcat server. All the server components reside on a Debian Linux OS.

The elements of the original WordNet like a word position or morphological definitions are not

²<http://wnsqlbuilder.sourceforge.net>

³<http://www.hibernate.org>

as necessary as lemmas and synsets. To simplify the editing process it was decided to allow only for modification of the semantic network structure. The database structure for handling data provided by WordVenture is presented in Figure 2, where editable and dictionary tables of the system are shown.

3.2 Client side and visualization

WordVenture has been developed in rich-client architecture (Boudreau et al., 2007). Because of that, some logic connected with data visualization can be executed on the client-side of application. Because of ease-of-use requirement it was decided that the client application will be developed as a flash rich client application. The client is a modified gossamer component⁴ for interactive graph visualization, where graph elements represent WordNet entities. The visualization allows a user to:

- **Browse WordNet lexical database.** It enables the user to navigate over the WordNet semantic network in a user-friendly way. Words and synsets are visualized as graph nodes, connections between them are presented as graph edges. Additionally, the user can filter graph nodes and edges to obtain required content (according to a selected relation or part of speech type), which makes user interface clean and readable.
- **Perform modifications on WordNet lexical database** – the tool enables a user to change graph content by adding, editing, or deleting its elements: nodes and edges. Modification of above-mentioned elements of WordNet lexicon (see Figure 2) does not cover all the components of WordNet. It includes only the four most desired, from the user point of view, elements of the semantic network: words, synsets, senses and relations.

Furthermore, the application offers additional features: manipulating the visible plane via zoom, rotating and moving, hiding selected nodes, etc. Currently, the application editing capabilities are as follows:

- adding new words and synsets,
- adding new links by dragging an edge between two nodes,

- editing existing relations, words, synsets.

Described tool functionalities allow WordNet database to edit according to the approach presented in section 2. Our team has tested the tool in scenarios of extending the existing WordNet database and building a semantic network from scratch (only schema with no data). User's feedback on the approach and the support provided by the tool has been positive. Some users pointed out that using the tool for WordNet dictionary browsing actually supports extending English vocabulary. This is achieved by the eye-catching visualization of database exploration in the client and discovering word synonyms and other related words.

Graph-based visualization in a WordVenture system depicted in Figure 3 allows a user to work efficiently, and keep clean and readable a large amount of lexical data. In every moment a user can enable or disable required elements of the visualization, which makes his workspace personalized. Additionally, it is possible to zoom in or zoom out a view of graph, so a user is able to keep a lot of graph nodes on his workspace.

4 Conclusions and future work

The system for cooperative WordNet editing has reached the end of its third iteration. Since deployment, we have received positive feedback and feature proposals for extending the application. In general, future improvements in the system can be classified in one of the following categories:

- server-side API extensions (allow more types of WordNet data to be visualized and edited),
- UI improvements (tabbed viewing, more filtering capabilities, improved rendering, etc.)
- miscellaneous (server administration console, client-side action history, etc.).

At present we are evaluating future proposals for the system, gathering more feedback from users via our web-based forum system, prioritizing future goals, and evaluating the applied solution as a base for generic approach to semantic data editing tasks. We believe that our approach and the system can be used for effective management of WordNet-based dictionaries and that it is important to support ontology-based systems with editors similar to the one presented in this paper.

⁴<http://gossamer.eti.pg.gda.pl/>

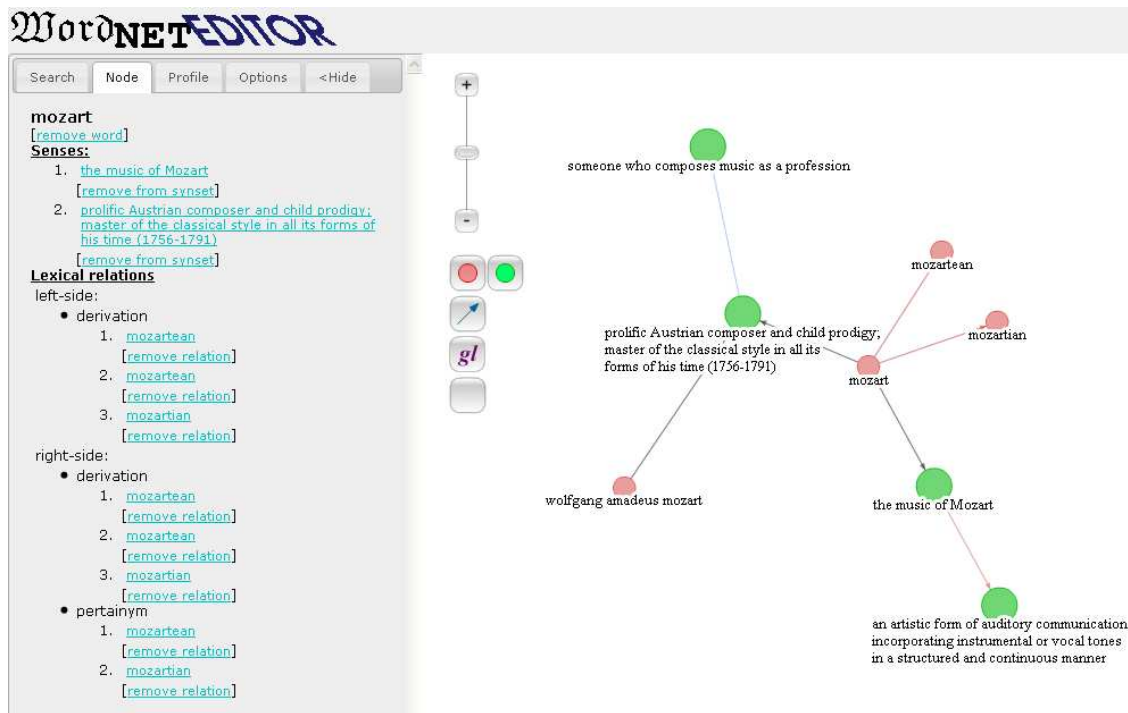


Figure 3: WordVenture visualization interface for WordNet

Wordventure can also be used as an interface for the correction of data obtained in an automated way, as it was in projects MindNet and ConceptNet. We plan mining Wikipedia to obtain new relations between synsets, it is also possible to enrich WordNet with data imported from other ontologies mentioned earlier: MindNet, ConceptNet or Sumo/Milo ontology (Niles and Pease, 2001). One of the most important things is synsets stratification, which will allow to filter data in terms of data importance.

In a few months we plan to integrate Wordventure with the second of our projects for visualization knowledge in Wikipedia⁵ where WordNet stands as ontology for articles categorization system. Long term goal is to join WordNet synsets with Wikipedia articles (Szymaski and Kilanowski, 2009), which will allow to look through Wikipedia knowledge effectively.

References

T. Boudreau, J. Tulach, and G. Wielenga. 2007. Rich client programming: plugging into the netbeans platform.

T. Erl. 2005. *Service-oriented architecture: concepts, technology, and design*. Prentice Hall PTR Upper Saddle River, NJ, USA.

⁵<http://swn.eti.pg.gda.pl>

C. Fellbaum et al. 1998. WordNet: An electronic lexical database.

H. Liu and P. Singh. 2004. ConceptNet – a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to wordnet: An online lexical database. *International Journal of lexicography*, 3(4):235–244.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9.

J. Szymański, K. Dusza, and Ł. Byczkowski. 2007. Cooperative Editing Approach for Building Wordnet Database. *Proceedings of the XVI International conference on system science*, pages 448–457.

J. Szymaski and D. Kilanowski. 2009. Wikipedia and wordnet integration based on words co-occurrences. *Proceedings of International conference on system science and technology*.

L. Vanderwende, G. Kacmarcik, H. Suzuki, and A. Menezes. 2005. MindNet: an automatically-created lexical resource. *HLT/EMNLP. The Association for Computational Linguistics*.