Key words – Wordnet, TouchGraph, Graph-based semantic editing

Konrad DUSZA, Łukasz BYCZKOWSKI, Julian SZYMANSKI*

COOPERATIVE EDITING APPROACH FOR BUILDING WORDNET DATABASE

The paper presents a approach for cooperative development of WordNet database using graphical component for graph visualizations with interactive navigation. The architecture and a policy for wikipedia-like editing distributed platform as well as the prerequisites for applying such approach in a real-life scenario are discussed. Furthermore, requirements on the tool with details on its implementation are given with some insight on future plans regarding the tool.

1. INTRODUCTION

The challenge of effective and accurate natural language processing remains unsolved. However, there is a a common understanding that proper structuralization of language-related knowledge is a prerequisite for achieving a successful language processing method. There are two main approaches to language structuralization for computational linguistics:

- building hand crafted dictionaries e.g.: WordNet [1], ontologies e.g.: SUMO/MILO [2] or knowledge bases CYC [3]. This approach requires large amount of human resources generally groped in one place. In our approach we would like to exploit power of the internet and give to open community a set of tools for cooperative developing linguistic projects.
- the second approach is automatic or semi-automatic text processing, e.g.: Microsoft Mindnet [4], Open Mind Common Sense Project - ConceptNet [5]. This approach gives very interesting results, however data obtained in that way are very noisy.

WordNet is one of the most widely known linguistic projects. It is developed and maintained by Cognitive Science Laboratory at Princeton University, USA. This lexical database comprises of different types of entities that are related to each other and stored in a flat-file based format. Most important concepts in WordNet structure are: word, synset - as the terms meanings, sense – as the relation between words and synsets, semlink – as the relation between meanings. Other WordNet entities include: word category, example sentences of usage for the certain word.

* Gdansk University of Technology, ul. Narutowicza 11/12, 80-952 Gdansk, Poland. wordteam@vega.eti.pg.gda.pl, julian.szymanski@eti.pg.gda.pl

WordNet is hand-crafted dictionary developed by linguistic engineers at the Princeton University. Even though tools to support further WordNet development are available, it can be observed that cooperative potential of the Internet has not been applied to WordNet yet. In our paper we would like to present a cooperative approach to WordNet edition along with its implementation. Furthermore, we would like to suggest an extrapolation of this idea to facilitate generic approach to building semantic dictionaries in the future.

Success of an editing platform relies on effective and easy-to-use graphical user interface. In order to achieve that, we decided to use a interactive visualization engine that would be able to render graph-like structures and allow to implement editing features. One of such engines is TouchGraph [6] – an open source Java application for graphs visualizations. This light-weight system enables convenient navigation in graph-like structures and provides basic support for graph editing. Our team modified the engine and adapted it for WordNet's data rendering and editing tasks.

The remainder of this paper is organized as follows. The next section presents the cooperative editing approach, which was developed to meet WordNet database requirements. Section 3 describes system architecture and technical details of the developed system. Section 4 provides insight in client application features for WordNet editing. The concluding section presents future plans regarding the presented approach and application.

2. THE COOPERATIVE EDITING APPROACH

2.1. THE APPROACH

Current implementations of WordNet web based applications, are limited to database exploration, moreover they resemble the standard, dictionary alike, web interface for WordNet [7].

Cooperative approach to editing content on the Internet is gaining increasing recognition in many IT fields. The main goal of our project was to create a system that would enable Web users free access and easy-to-use interface for WordNet content 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. Fig.1 presents the overview of the our cooperative editing approach for WordNet.

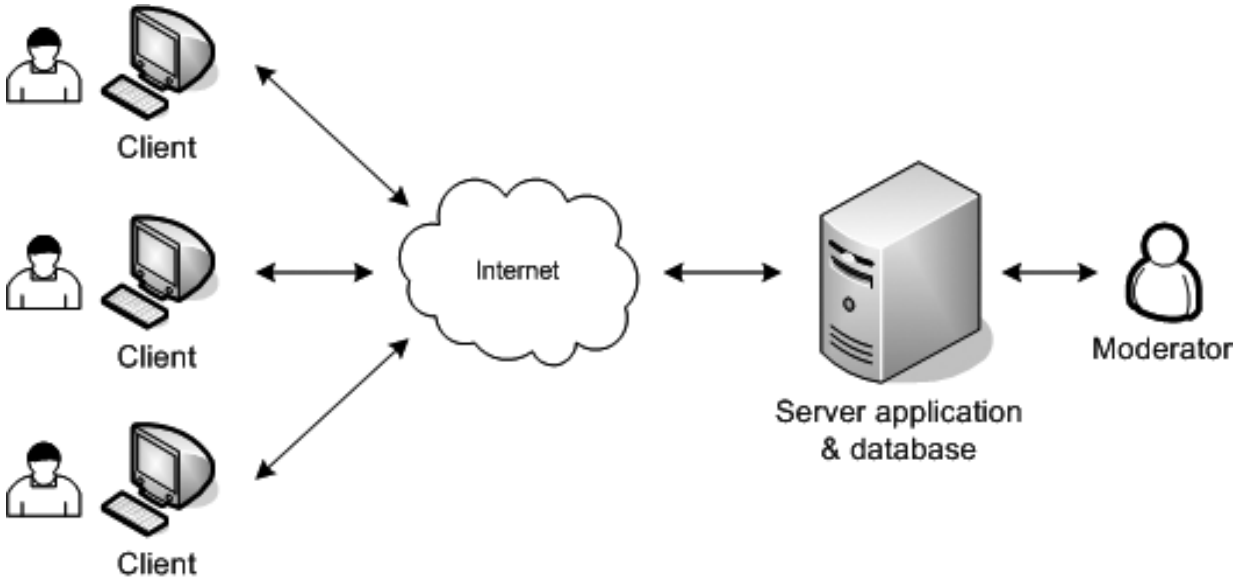

Fig. 1. Cooperative editing scheme.

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

- Users input data on their clients, which communicate changes to the server.
- Server logs the operation and executes suitable procedures on the database.
- 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 edit steps, the original database is enriched with content chosen from users contributions. This procedure is supported with regular database backups. Described editing process is similar to Wikipedia's [8] 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 at least satisfactory results in this field.

2.2. VISUALIZATION AND EDITING

In our project's initialization phase, it was decided to start with the latest WordNet version available at the time – WordNet 2.1. Analysis of the WordNet database schema made it clear that both the visualization and the editing features of the system to be created, have to be limited to the most important elements of the schema. It was argumented that overloading the system with concepts like word positioning or morphological definition that are not essential for cooperative edition of the database, would discourage users from inputing data to the system. The final structure used in the system, derived from the original WordNet 2.1, is presented on Fig.2 (the diagram is based on MySQL WordNet port by Bernard Bou [9]).

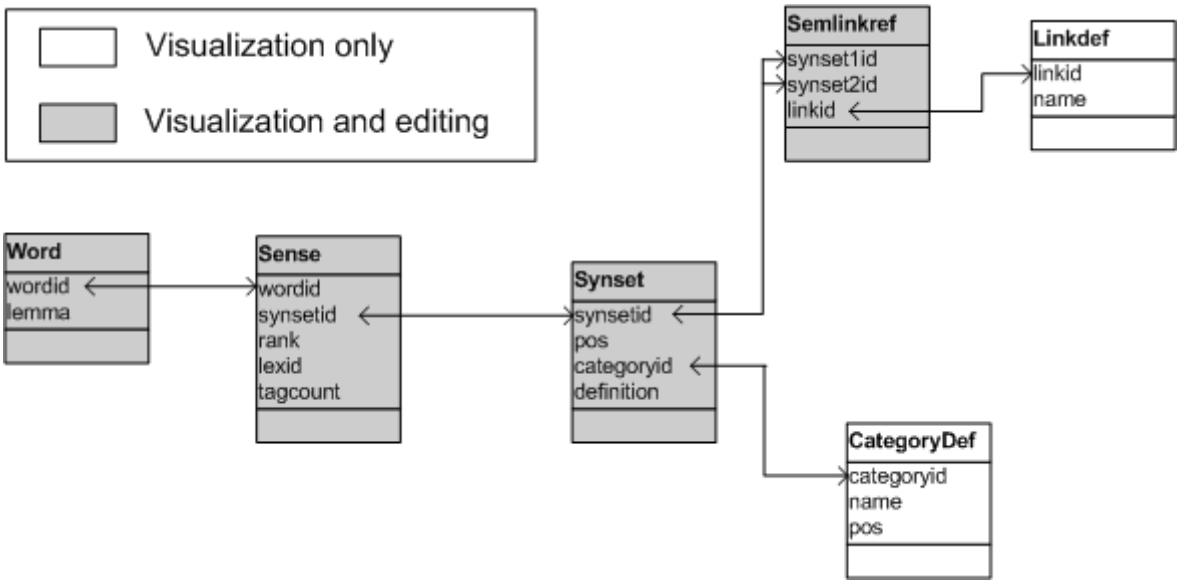

Fig. 2. WordNet entities to be supported by the tool. Greyed out entities will have support for both visualization and editing, white entities will have only visualization support. Arrows represent relationships between entities.

2.2. ADDITIONAL REQUIREMENTS

Desired system characteristics defined in p. 2.1 and schema elements for editing described in p.2.2. were picked as a basis for defining feature set for the entire system. Basic visualization and editing features for synset, word, semlink and sense have been complemented with some additional requirements:

- provide logging for the server side to monitor all system activities
- provide means for securing the database from vandalism (esp. deletion activities)
- keep the client application small (below 300kB total)

Aforementioned requirements were introduced to increase effectiveness of the platform and support system's administering activites.

3. IMPLEMENTATION

The system supporting the cooperative editing approach for the WordNet database has been implemented in a standard client-server architecture, with database and WordNet logic tier residing on the server and the visualization engine querying the server as a client application.

Due to the ease-of-use requirement, it has been decided to implement the client application as a J2SE 5 applet. The client is a modified TouchGraph application, in which the communication layer has been introduced along with mappings between TouchGraph shapes and WordNet entities. The TouchGraph code itself has been improved in some parts in terms of UI efficiency and the use of J2SE 5 constructs (the original TouchGraph available on SourceForge was implemented in Java 1.2).

The server consists database migrated from 'flat' WordNet files to MySQL 5 DBMS in order to allow easy data manipulation (esp. deleting and updating operations). The migration has been conducted by Bernard Bou [9], who published his work on the Internet for free use. Data access routines were implemented with Hibernate ORM engine [10]. Manipulating the database content is

done via implemented server API exposed as Web Services and which use is logged with Apache Log4j on a Tomcat server. All of server components reside on a Debian Linux OS.

Resulting application has a very flexible architecture. The encapsulation of WordNet database modification routines in a form of web services residing on a server, produced a wide range of possible future applications that could include the functionality of browsing and modification of a WordNet database. Presented in section 4, the J2SE 5 applet, is only one of the possible solutions. For example, we could reuse the server side of the system, so that the server would gather data both from users and from automatic mechanisms like web robots, etc. Nevertheless, this can be considered as a result of a proper system design and in further paragraphs we will strictly focus on editing capabilities of the system provided by the J2SE applet.

4. CLIENT APPLICATION PRESENTATION

Project has been developed and deployed at the Gdansk University of Technology at Faculty of Electronics, Telecommunications and Informatics. Project's web page and client application is accessible at: http://wordventure.eti.pg.gda.pl/. WordNet's words and synsets are visualised as graph nodes (word – rectangle, synset – rounded rectangle), whereas senses and semlinks are visualized as edges in the rendered graph, each semlink type is identified by name and color (see Fig.3). Currently, the following functionalities are supported in the client application: search word by lemma, filter visible relations by type (acc. to WordNet's sense and semlink linkid feature), WordNet's structure exploration via search and clicking on the nodes (selecting a node renders all its edges by default). Furthermore, the application offers standard TouchGraph features: manipulating the visible plane via zoom, rotate and move, hiding selected nodes, etc.

Currently, the application editing capabilities are as follows:

- adding new words and synsets,
- adding new senses and semlinks by dragging an edge between two nodes,
- editing existing words, synsets and senses.

Semlink edition is limited due to constraints enforced by WordNet structure. Furthermore, users are allowed to mark nodes and edges as deleted, the real deletion process is deferred to limit the consequences of possible vandalism – it is up to the system administrator, when and if the entities mark as deleted in the database will eventually be removed from the database or restored.

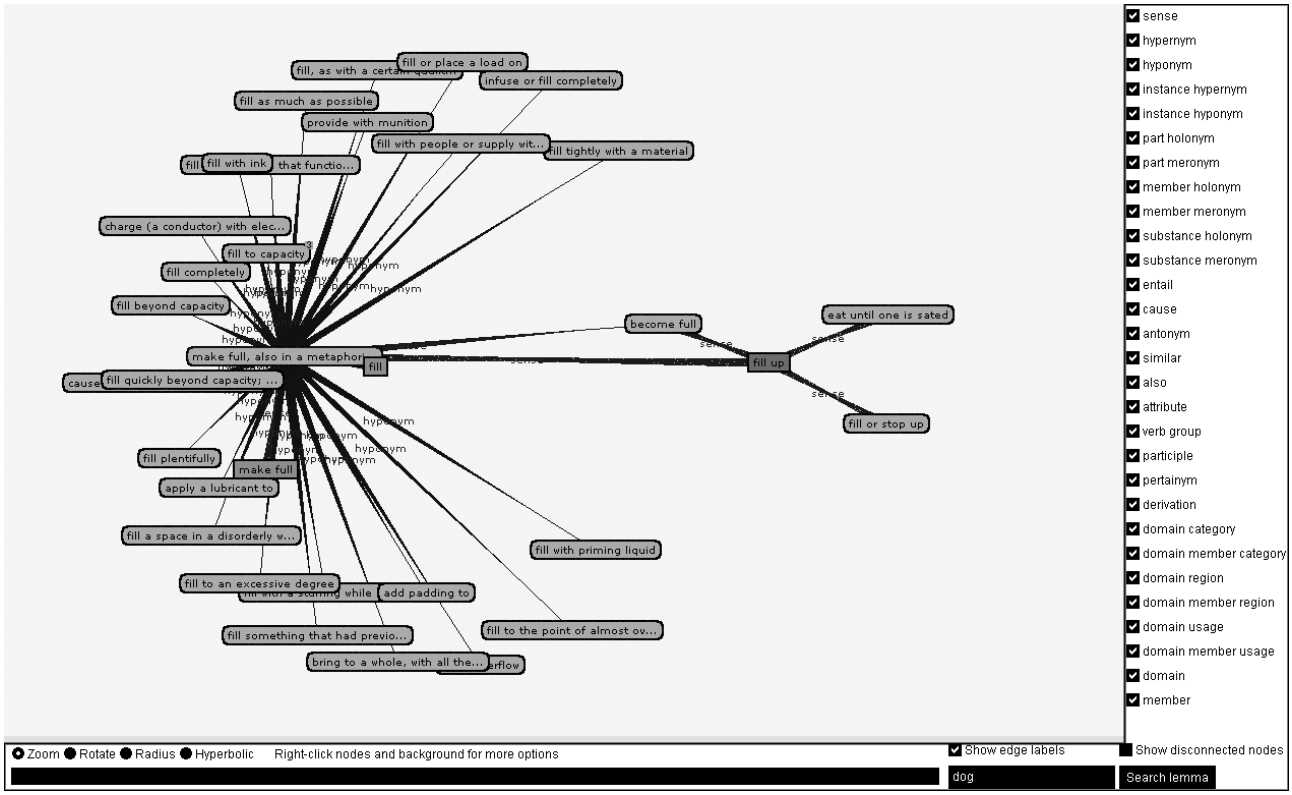

Fig. 3. Screenshot of the client application (colors/shading modified for better printing).

Described tool functionalities allows WordNet database editing 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 WordNet database from scratch (only schema with no data). User 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's synonyms and other related words.

5. CONCLUSION

The system for cooperative WordNet editing has reached the end of its first 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),
- upgrades: we plan to import new data released in version 3.0 of WordNet to our system, also other functionalities offered by WordNet (i.e. sample sentences) will be include in next iteration,
- searching (search by keywords in synset descriptions, etc.),
- UI improvements (tabbed viewing, more filtering capabilities, improved rendering, etc.)
- miscellaneous (server administration console, client-side action history, etc.).

Notably, implementing some of the UI-related improvements (i.e. tabbing) would require in-depth redesign and reimplementation of visualization engine, due to TouchGraph's limitations.

At present, we are evaluating feature 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. During the development of the application we became aware of many problems and restrictions of the initial plan, like TouchGraph's architectural limitations. Nevertheless, 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.

6. ACKNOWLEDGEMENT

Julian Szymański is grateful for the support by the Polish Committee for Scientific Research, research grant 2006-2008.

REFERENCES

[1] MILLER G. A. WordNet: An on-line lexical database, International Journal of Lexicography, 1990 3(4):235–312. Special Issue.

[2] NILES I., TERRY A.: The MILO: A general-purpose, mid-level ontology, In Proceedings of International Conference on Information and Knowledge Engineering (IKE'04). Las Vegas, NV.

[3] VANDERWENDE L., KACMARCIK G., SUZUKI H. and MENEZES A., 2005. MindNet: An Automatically-Created Lexical Resource, In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, Vancouver, British Columbia, Canada, October 2005.

[4] LENAT D.B. CYC: A Large-Scale Investment in Knowledge Infrastructure, Comm. of the ACM 38, 33-38, 1995.

[5] LIU H., SINGH, P. (2004). Commonsense Reasoning in and over Natural Language, Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2004). Wellington, New Zealand. September 22-24. Lecture Notes in Artificial Intelligence, Springer 2004

[6] TouchGraph homepage, http://www.touchgraph.com/

[7] WordNet main site, http://wordnet.princeton.edu/perl/webwn

[8] Wikipedia main site, http://wikipedia.org

[9] Bernard Bou's port of WordNet database to MySQL, http://sourceforge.net/projects/wnsqlbuilder/

[10] Hibernate ORM Engine, http://www.hibernate.org/