The Global WordNet Grid Software Design

Aleš Horák, Karel Pala, and Adam Rambousek

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
{hales,pala,xrambous}@fi.muni.cz

Abstract. In the presented paper we show how the Global WordNet Grid software is designed. The goal of the Grid is to provide a free network of WordNets linked together through interlingual indexes. We have set as our goal to work on the Grid preparation in the Masaryk University NLP Centre and design its software background. All participating wordnets will be encapsulated by a DEB (Dictionary Editor and Browser) server established for this purpose.
The following text presents design details of the new DEBGrid application with possibilities of three types of public and authenticated user access to the Grid WordNet data.

Key words: WordNet; DEB platform; DEBVisDic; Global WordNet Grid

1 Introduction

In June 2000, the Global WordNet Association (GWA [1]) was established by Piek Vossen and Christiane Fellbaum. The purpose of this association is to "provide a platform for discussing, sharing and connecting WordNets for all languages in the world." One of the most important actions of GWA is the Global WordNet Conference (GWC) that is being held every two years on different places all over the world. The second GWC was organized by the MU NLP Centre in Brno and the NLP Centre members are actively participating in GWA plans and activities. A new idea that was born during the third GWC in Korea is called the Global WordNet Grid with the purpose of providing a free network of smaller (at the beginning) WordNets linked together through ILI. The Grid preparation is currently just starting and the MU NLP Centre is going to secure its software background.

The idea of connecting wordnets has been suggested during the Balkanet project (2001–2004 [2]) in which Patras team developed the core of the WordNet Management System designed to link all the WordNets developed in the course of the project (Deliverable 9.1.04, September 2004).

It was tested successfully on Greek and Czech WordNets. However, the Patras team did not proceed with it and the system remained only as a partial

result of the research that was not pursued further. Before the end of Balkanet project, the Czech team decided to re-implement the local version of the VisDic browser and editor using client/server architecture. This was the origin of the DEBVisDic tool that was fully implemented only after finishing Balkanet project. Fully operational version of DEBVisDic was presented at the 3rd Global WordNet Conference 2006 in Korea [3]. In our view this client/server tool will become a software background for the Grid preparation mentioned above (see below in the Section 3.2).

2 The Global WordNet Grid

Since the first publicly available WordNet, the Princeton WordNet [4], more than fifty national wordnets have been developed all over the world. However, the availability of the wordnets is limited – that is why the idea of a completely free Global WordNet Grid has appeared.

It is a known fact that, for instance, the results of the EuroWordNet are not freely accessible though the participants of the project have developed (and are developing) more complete and larger WordNets for the individual languages. Practically the same can be said also about the results of the Balkanet project. If one wants to exploit WordNets for different languages it is always necessary to get in touch with the developers and ask them for the permission to use the wordnet data.

Another reason for building and having the completely free Global WordNet Grid is the fact that the particular WordNets can be linked to the selected ontologies (e.g. Sumo/Milo) and domains. This has already took place with the WordNets developed in the Balkanet project. The links to the ontologies should be provided for all WordNets included in the Global WordNet Grid.

The Grid also provides a common core of 4.689 synsets serving as a shared set of concepts for all the Grid's languages. These synsets are selected from the EuroWordNet Common Base Concepts used in many wordnet projects.

3 DEBGrid – the DEB Application for the Global WordNet Grid

The DEBGrid application will be built over the DEBVisDic application with the DEB server either set up at the NLP Centre of Masaryk University in Brno or it will be set up by the Global WordNet Association. The DEB platform provides important backgrounds for the WordNet Grid universal features.

3.1 The DEB Architecture

The Dictionary Editor and Browser (DEB) platform [3, 5, 6] has been developed as a general framework for fast development of wide range of dictionary writing applications. The DEB platform provides several very important foundations

that are common to most of the intended dictionary systems. These foundational features include:

- a strict separation to the *client* and *server* parts in the application design. The server part provides all the necessary data manipulation functions like data storage and retrieval, data indexing and querying, but also various kinds of data presentations using templates. In DEB, the dictionary entries are stored using a common XML format, which allows to design and implement dictionaries and lexicons of all types (monolingual, translational, thesauri, ontologies, encyclopaedias etc.). The client part of the application concentrates on the user interaction with the server part, it does not produce any complicated data manipulation. The client and server parts communicate by means of the standard HTTP (or secured HTTPs) protocol.
- a common *administrative interface* that allows to manage user accounts including user access rights to particular dictionaries and services, dictionary schema definitions, entry locking administration or entry templates definitions.
- *XML database* backend for the actual dictionary data storage. Currently, we are working with the Oracle Berkeley DB XML [7, 8] database, which provides a flexible XML database with standard XPath and XQuery interfaces. The DB XML database is well suited for processing complicated XML structures, however, we (and according to private discussions other DB XML users as well) have encountered efficiency problems when processing certain kinds of queries that result in large lists of answers. Simple processing of the data (like export or import of the whole dictionary) is not a problem as the whole English WordNet export (over 100.000 entries) takes less than 1 minute, but searching for values of specific subtags can take several seconds in such large dictionary even when indexes are used. We are currently working on several solutions for this, which include link caching, specific DB XML indexing and also trying a completely different database backend. The key advantage for all the DEB applications is that a replacement of the DB XML backend with another database will be a completely transparent process which does not need any change in the applications themselves.

Based on these common features several developed and widely used dictionary applications have been implemented, including the well-known wordnet editor DEBVisDic that has been used in several national wordnets development recently (Czech, Polish, Hungarian or South African languages). With this evidence, we believe that DEB is the right concept for the Global WordNet Grid data provision.

3.2 The DEBGrid Design and Implementation

In the DEB platform environment, all the wordnets are usually stored on single DEBVisDic server. In the Grid, most of the wordnets will be also stored in this way, however, since the Grid could be finally composed of large number of

wordnet dictionaries developed by different organizations, this solution may not be always the best option (for example because of licensing issues). Thanks to the client-server nature of the DEB platform, DEBGrid can offer two possible types of encapsulating wordnets in the server:

- a WordNet can be physically stored on the central server. This is the traditional DEBVisDic setup and offers the best performance.
- a WordNet can be stored on a DEBVisDic server located at the wordnet owner's institution. All servers in the Grid can then communicate with each other (depending on the server setup). The Central Grid server for this wordnet has only the knowledge of which server to contact, instead of having the full wordnet database stored locally, and all queries are dynamically resolved over the Internet. This option may be slower as it depends on the quality of connection to different servers and their performance. On the other hand, the WordNet owner has full control over the displayed data and access permissions.
- a mixed solution – some wordnets are stored on central server and some are stored on their respective owners' servers. This is just an extension of the previous option. Again, the performance of the whole Grid depends on the performance of single servers, but the speed can be improved if the most used wordnets are stored on the central server.

The DEB framework provides several possibilities of working with the wordnet data. All the types of the Grid access undergo the same control of service and user management with the option to provide information for public (anonymous) access as well as authenticated access for registered users.

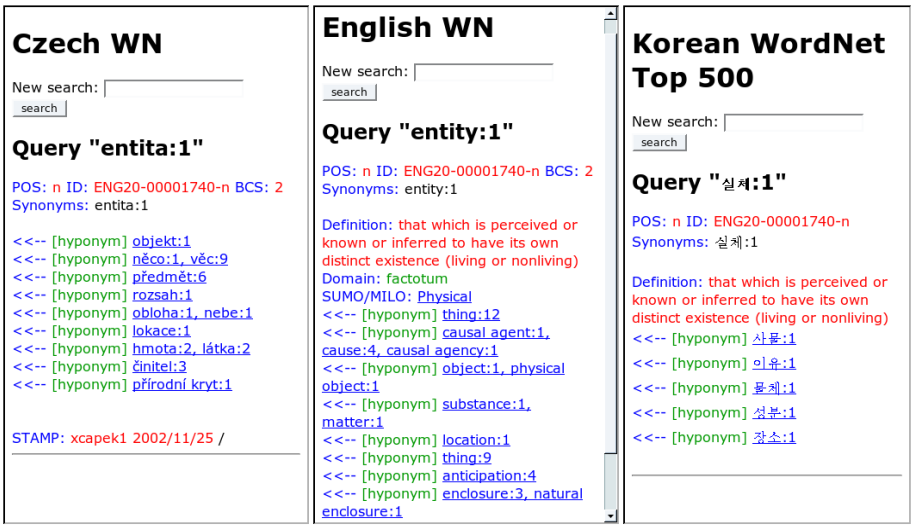

Fig. 1. The web interface of DEBGrid with three interlinked wordnets.

Basically, each wordnet in the Grid can be presented to the Grid users in one of the following forms:

a) by means of a simple purely HTML interface working in any web browser. This interface is able to display one WordNet dictionary or the same synset in several WordNets. Synsets are displayed using XSLT templates – the server can provide several view of the synset data ranging from a terse view up to a detailed view. The view can be even different for each dictionary. An example of such presentation of one synset in three WordNets is displayed in the Figure 1. This type of WordNet view is probably the best for public anonymous access to the Grid, since it does not need any installation of user software or packages.

a) using the full DEBVisDic application. This application needs to be installed as an extension of the freely available Firefox web browsers, but it offers much complex functionality than the web access. Each WordNet is opened in its own window which offers several views of the WordNet data (a textual preview, hypero/hyponymic tree structures, user query lists or XML) and also the possibility to edit the data (for users with the write permissions). With this type of the Grid access, the user would have the most advanced environment for working with the Grid WordNets.

a) by means of a defined interface of the DEBVisDic server. This way any external application may query the server and receive wordnet entries (in XML or other form) for subsequent processing. In this way, local external applications can easily process the Grid data in standard formats.

In all cases, users (or external applications) could authenticate with a login and password over secure HTTP connection. Each user can be given a read-only or read-write access to particular WordNets.

For some applications it is useful to use a visualization tool that allows to view synsets and their links as graphs. Such tool is under development at the MU NLP Centre, it is called Visual Browser [9]. Its important feature is the ability to process wordnet synsets from a DEB server storage and convert them into the RDF notation for visualization. Visual Browser is also suitable for representing ontologies that can and will be integrated within Global WordNet Grid.

4 Conclusions

In this article, we have presented a report of the design and implementation of the Global WordNet Grid software background. The basic idea of the WordNet Grid introduced by P. Vossen, Ch. Fellbaum and A. Pease at GWC 2006 includes establishing of an interlinked network of national wordnets connected by means of the interlingual indexes. In the starting phase the Grid contains only a subset of the EuroWordNet Base Concepts with nearly 5.000 synsets.

The management and intelligent processing of the included wordnets is driven by the DEB development platform tool called DEBGrid. This tool is built on top of the DEBVisDic wordnet editor and allows thus a versatile environment for working with large number of wordnets in one place and style.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Ministry of Education of CR in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 201/05/2781.

References

1. : The Global WordNet Association. (2007) `http://www.globalwordnet.org/`.
2. : Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`. (2002)
3. Horák, A., Pala, K., Rambousek, A., Povolný, M.: First version of new client-server wordnet browsing and editing tool. In: Proceedings of the Third International WordNet Conference - GWC 2006, Jeju, South Korea, Masaryk University, Brno (2006) 325–328
4. Miller, G.: Five Papers on WordNet. International Journal of Lexicography **3**(4) (1990) Special Issue.
5. Horák, A., Pala, K., Rambousek, A., Rychlý, P.: New clients for dictionary writing on the DEB platform. In: DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems, Italy, Lexical Computing Ltd., U.K. (2006) 17–23
6. Horák, A., Rambousek, A.: Dictionary Management System for the DEB Development Platform. In: Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS, aka NLUCS), Funchal, Portugal, INSTICC PRESS (2007) 129–138
7. Chaudhri, A.B., Rashid, A., Zicari, R., eds.: XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional (2003)
8. : Oracle Berkeley DB XML web (2007) `http://www.oracle.com/database/berkeley-db/xml`.
9. Nevěřilová, Z.: The Visual Browser Project. `http://nlp.fi.muni.cz/projects/visualbrowser` (2007)