USING WORDNETS AND ONTOLOGIES FOR TEXT-MEANING ASSIGNMENT
Implementation Details of the KYOTO Project First Phase

Aleš Horák and Adam Rambousek
Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic

Keywords: Wordnet, Semantic network, Ontology, Fact extraction.

Abstract: The vision of Semantic Web introduced ontologies as the main unifying tool for management of the knowledge and semantic structure of text documents. However, linking the real text documents with the ontologies (of various kinds and various degree of complexity) is still a matter of current research in knowledge representation projects.
In this paper, we are presenting the work results of the KYOTO project database implementation. The goal of the project is to provide a complex system for automatic processing of documents in order to extract known facts, link them with shared ontology and use this knowledge for Question Answering about the document topic.
We give details about the design and implementation of the KYOTO database, which interlinks national WordNet semantic networks with the general SUMO ontology to offer the basis of the future shared ontology.

1 INTRODUCTION

The standardization of the techniques of knowledge representation and reasoning is driven by designing and incorporating ontologies into the text processing approaches (Mars, 1995). In the process of the design of a knowledge processing system, one of the first decisions must be the choice of the level of complexity of the applied ontological system. Current general ontological systems range from an encyclopaedia-like system Cyc (Lenat, 1995), through the predicate logic based SUMO (Niles and Pease, 2001) to easily exploitable semantic networks based on the Princeton WordNet (Fellbaum, 1998). The number of applications that are using these ontologies for the processing of textual knowledge is proportional to the level of the ontology complexity – the more straightforward the ontology is, the more projects make use of it.

In the following text, we describe the KYOTO project (Vossen, 2008), which aims at a straightforward application of the WordNet like ontologies in the multilingual form (denoted as the *Global WordNet Grid*) and a shared common ontology corresponding to the level of the *Suggested Upper Merged Ontology* (SUMO) as the central knowledge backbone. The ontology here serves as a meaning description tool for all the terms and facts that are extracted, compared and stored within the KYOTO system.

2 THE KYOTO PROJECT – WORDNETS, ONTOLOGIES AND TEXT

WordNet semantic networks allow to express basic language relations[1] in a multigraph structure directly processable by computer systems in many useful ways.[2] However, description of more complicated structured knowledge, e.g. relations with more than one participants, cannot be encoded in a WordNet-standard way that could be further analysed and used by computers.

In the KYOTO system, this (potential) drawback of WordNet is solved by the idea of extending the WordNet into a *Global WordNet Grid* of multiple languages with a shared ontology in the center. Interlink-

[1] hyperonymy/hyponymy, synonymy/antonymy, holonymy/meronymy, etc.

[2] deriving sets of similar objects, classes of more general objects or objects with opposite meaning

Horák A. and Rambousek A. (2009).
USING WORDNETS AND ONTOLOGIES FOR TEXT-MEANING ASSIGNMENT - Implementation Details of the KYOTO Project First Phase.
In *Proceedings of the 4th International Conference on Software and Data Technologies*, pages 303-307
DOI: 10.5220/0002243403030307

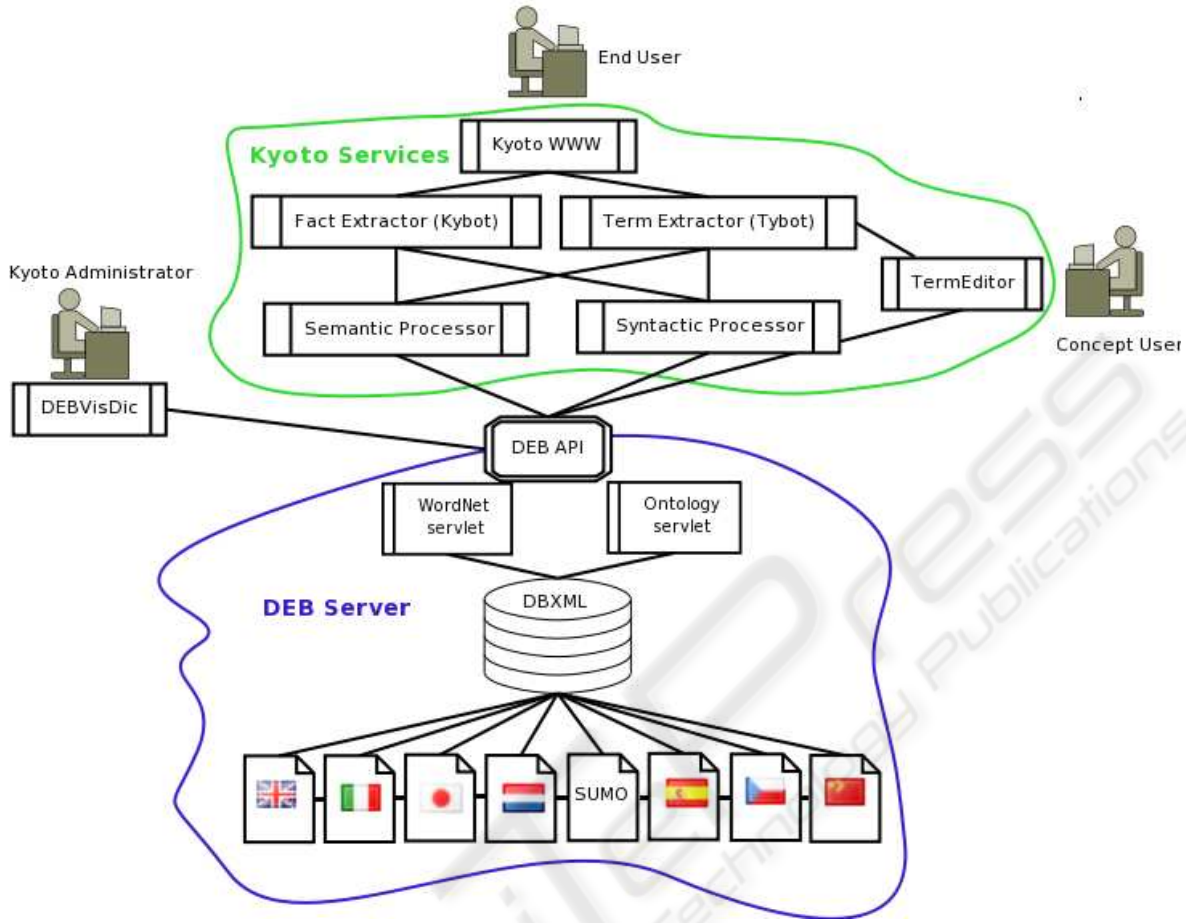

Figure 1: The schema of the KYOTO database within the KYOTO system.

ing of national wordnets is not a new idea, it was introduced e.g. in the EuroWordNet (Vossen, 1998) and Balkanet (Christodoulakis, 2004) projects. In these projects the "pivot," i.e. the *interlingual index*, was represented directly by the English WordNet. This solution had several advantages and several disadvantages. From the point of view of the knowledge analysis, the biggest disadvantage was that the lexical knowledge structure was "hidden" in the English lexicon without the possibility to really extract it for the purpose of further computer processing.

Since the first publicly available WordNet, the Princeton WordNet (Miller, 1990), more than fifty national wordnets have been developed all over the world. However, the availability of the wordnets is limited – that is also a reason why the idea of a completely free Global WordNet Grid has appeared.

It is a known fact that, for instance, the results of 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 KYOTO project will incorporate and expand the Global WordNet Grid and will be the first system that exploits the benefits of storing the definitions of terms and facts in a computer processable logical system using the Grid's shared ontology.

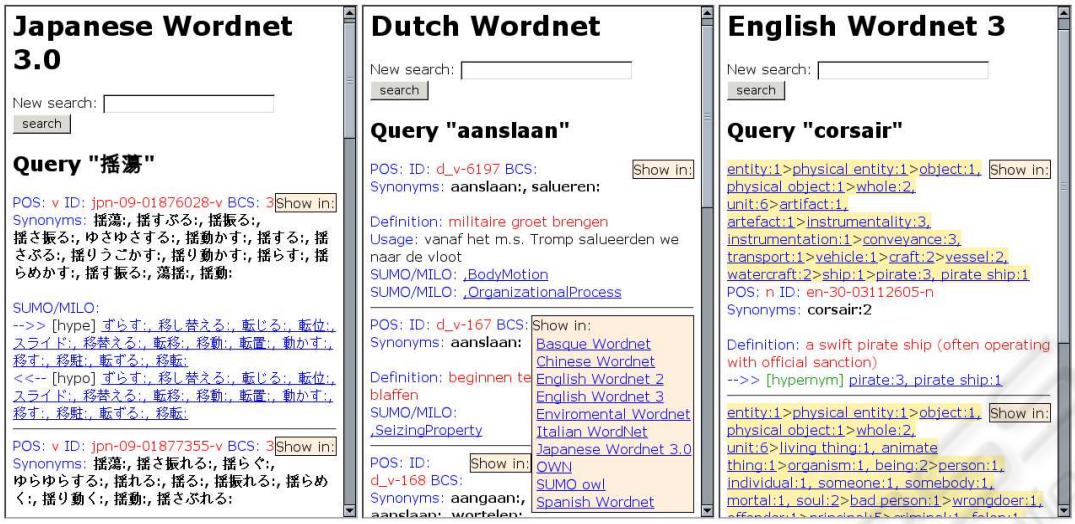

Figure 2: Three national wordnets in the KYOTO Database Viewer.

3 THE KYOTO DATABASES

The KYOTO database is built over the DEBVisDic application with the DEB server either set up at one central locality or it can be set up by several KYOTO partners. The DEB platform provides important backgrounds for the KYOTO project universal features (see Figure 1).

3.1 The DEB Architecture

The Dictionary Editor and Browser (DEB) platform (Horák et al., 2006; Horák and Rambousek, 2007; Horák et al., 2008) 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. 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 (Chaudhri et al., 2003; DB XML, 2007) database, which provides a flexible XML database with standard XPath and XQuery interfaces. The DEB applications are not limited to DB XML, because the database layer can be replaced transparently without the need to change the application itself.

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 KYOTO multilingual knowledge base.

3.2 The Database Implementation

In the DEB platform environment, all the wordnets are usually stored on single DEBVisDic server. In the KYOTO project, each WordNet is provided by different project partner and each of them may have different requirements (for example licensing issues). Thanks to the client-server nature of the DEB platform, KYOTO database can offer three 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.

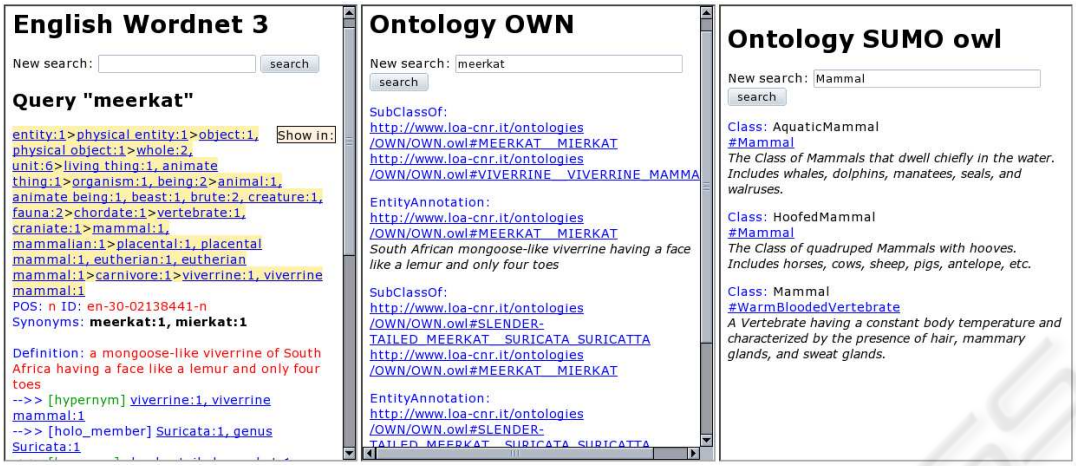

Figure 3: SUMO and OWLWN ontology with the English WordNet.

- a WordNet can be stored on a DEBVisDic server located at the WordNet owner's institution. All servers can then communicate with each other (depending on the server setup). The central server 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 system 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.

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

- 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 synsets in three WordNets is displayed in Figure 2. This type of WordNet view is probably the best for public anonymous access to the KYOTO knowledge base, since it does not need any installation of user software or packages.
- using the full DEBVisDic application. This application needs to be installed as an extension of the freely available Firefox web browser, but it offers much more 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).
- by means of a defined interface of the DEBVisDic server, the *Application Programming Interface* (API). This way any external application[3] may query the server and receive WordNet entries (in XML or other format) for a subsequent processing.
- using the Term Editor – a Wiki-based WordNet browser and editor developed within the KYOTO project.

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

All the national WordNets are provided in Lexical Markup Framework (LMF) format (Francopoulo et al., 2008). The DEBVisDic server is optimized for its own WordNet format, so all the data are converted from and to LMF using XSLT stylesheets. For batch operations (importing and exporting the whole WordNet), a special application based on `libxml` (Veillard, 2002) is used, because this solution offers fast conversion. For example, 80MB XML file takes two days to

[3]including DEBVisDic or the Term Editor

convert using XSLT, and only 40 minutes using the special conversion application.

3.3 Interlinking Wordnets and Ontologies

All wordnets in the KYOTO database are interlinked using the common central ontology. The solution is not limited to one ontology only. At the current state, SUMO and OWL-WN ontologies are used, both of them are stored in the OWL format.

An ontology is either referenced from a synset, or a user can browse it independently using the DEB HTML interface (similar to the WordNet HTML interface, see Figure 3). However, the ontology browser is not based on the DEBVisDic WordNet browser, because of the differences in structure and format. It is a standalone module integrated to the KYOTO database.

The ontology application allows the user to search for classes, properties, descriptions and relations within a single query.

4 CONCLUSIONS

This paper has presented the main ideas of developing the multilingual Global WordNet Grid with a shared knowledge ontology within the KYOTO project.

We have described the design and implementation of the KYOTO database storing the wordnets and ontologies in a versatile DEB (Dictionary Editor and Browser) server, which allows to abstract the actual data structures and provides the requested high level functionality to the system.

ACKNOWLEDGEMENTS

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 102/09/1842.

REFERENCES

Chaudhri, A. B., Rashid, A., and Zicari, R., editors (2003). *XML Data Management: Native XML and XML-Enabled Database Systems*. Addison Wesley Professional.

Christodoulakis, D. (2004). *Balkanet Final Report*. University of Patras, DBLAB. No. IST-2000-29388.

DB XML (2007). Oracle Berkeley DB XML web. http://www.oracle.com/database/berkeley-db/xml.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2008). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation Journal*.

Horák, A., Pala, K., and Rambousek, A. (2008). The Global WordNet Grid Software Design. In *Proceedings of the Fourth Global WordNet Conference*, Szegéd, Hungary. University of Szegéd.

Horák, A., Pala, K., Rambousek, A., and Rychlý, P. (2006). New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems*, pages 17–23, Italy. Lexical Computing Ltd., U.K.

Horák, A. and Rambousek, A. (2007). 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)*, pages 129–138, Funchal, Portugal. INSTICC PRESS.

Lenat, D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Mars, N. (1995). *Towards very large knowledge bases*. Ios Press.

Miller, G. (1990). Five Papers on WordNet. *International Journal of Lexicography*, 3(4). Special Issue.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, pages 2–9. ACM New York, NY, USA.

Veillard, D. (2002). The XML C library for Gnome (libxml). http://xmlsoft.org/.

Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.

Vossen, P. (2008). KYOTO Project (ICT-211423), Knowledge Yielding Ontologies for Transition-based Organization. http://www.kyoto-project.eu/.