BALKANET: A Multilingual Semantic Network for Balkan Languages

Stamou Sofia	*Oflazer Kemal*	*Pala Karel*	*Christodoulakis Dimitris*	*Cristea Dan*	*Tufis Dan*
Koeva Svetla	*Totkov George*	*Dutoit Dominique*	*Grigoriadou Maria*		

Abstract

BalkaNet aims at building a multilingual lexical database consisting of WordNets in several Central and Eastern European languages. Even though it will be built in a similar way with EuroWordNet, new features will be implemented ranging from structuring the Inter-Lingual-Index to ensure linking of conceptual equivalencies across WordNets to the development of an inter-networked WordNet Management so that each partner retains full responsibility and independence of his local WordNet whereas at the same time they will be able to view other WordNets and check their compatibility.

1 Introduction

BalkaNet is a funded project (IST-2000-29388) that aims at building a multilingual lexical database consisting of WordNets in Central and Eastern European languages. Each monolingual WordNet will be structured along the same lines as EuroWordNet (EWN), (Vossen, 1998) i.e., synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations such as hyponymy, meronymy, antonymy, etc. Equivalence relations between synsets in different languages will be made explicit in the so-called Inter-Lingual-Index (ILI) adopted from EWN. ILI is an unstructured collection of concepts with the only purpose to provide an efficient mapping across languages. However, ILI will be modified in order to reflect the lexicalization patterns of Balkan languages and will be structured to enable efficient mapping of senses in the BalkaNet database. BalkaNet aims at developing a multilingual resource representing semantic relations among basic concepts of the following Balkan languages: Greek, Turkish, Romanian, Bulgarian, Czech and Serbian. BalkaNet includes semantic relations existing in each of the above languages, as language internal relations, as well as among them, as equivalence relations to the ILI. BalkaNet will as much as possible be built from available lexical resources so that it will be possible to combine information from independently created resources, making the final database more consistent and reliable while keeping at the same time the richness and diversity of the vocabularies of the languages involved. The main resources of information are going to be the individual monolingual WordNets that have already been developed or are currently under development for most of the participant languages. Where a monolingual WordNet is not available dictionaries, thesauri or corpora of the respective languages will be used for the terminology extraction. For the development of BalkaNet a merge model approach will be adopted, meaning that each WordNet will be built separately from independently developed resources and then linked to the most equivalent concepts in the ILI. We aim at a total set of 15.000 comparable synsets in each language, corresponding with more or less 30,000 literals, covering generic vocabulary of the involved languages. The Part-Of-Speech (POS) distribution will be 65% nouns, 25% verbs, 5% adjectives and 5% adverbs. In addition, monolingual WordNets developed from scratch within the framework of the project will comprise approximately 8,000

synsets whereas the number of synsets that will be added in already existing ones will be determined at a later stage. Finally, in order to keep compatibility with EWN the Language Independent Module, namely the Top-Concept Ontology, will be maintained along with the ILI records. One differentiation from EWN concerns the structuring of the ILI. The motivation behind structuring the ILI originates from various problems related to the mapping of senses in EWN. More specifically, because of high level of sense differentiation in the ILI there is a danger that conceptual equivalencies across WordNets are not linked to exactly the same sense of the English equivalent but instead connected to distinct ILI concepts reflecting different senses of the same word. In order to account for these diverging mappings from local WordNets onto ILI concepts, domain labels are going to be included in the ILI and the latter will be structured on the basis of the top ontology so that terms linked to the ILI correspond to the same conceptual domain even if they are not exact translational equivalents. In addition, a structured ILI would mean a grouping of the ILI concepts that belong to the same conceptual domain enabling thus a preliminary clustering of terms and as a consequence a preliminary clustering of documents indexed on the basis of the ILI. The project started a few months ago, thus no new synsets have been developed so far. Members of the consortium are in the process of setting the requirements, specifying the methodology to be followed for the development of the WordNet Management System, the structuring of the ILI and developing tools for processing the monolingual lexical resources. In addition, the Base Concepts and the Top Ontology of EWN along with their internal relations are being carefully examined and checked against Balkan lexical resources in order to be enriched with Local Base Concepts and conclude on their applicability to BalkaNet.

2 Implementing the BalkaNet through a WordNet Management System

The main differentiation between BalkaNet and EWN relies on the reusability and openness of the tools and software. More specifically, EWN has been constructed by using Polaris tool (Bloksma, 1996), which we feel has a few drawbacks. First and foremost it is a commercial stand-alone tool designed solely for WordNet maintenance that cannot be easily adapted to a new application and runs only in Microsoft Windows. For BalkaNet a (inter) networked tool will be developed to help partners coordinate their work online. Although it would be technically possible, we do not want to create a fully Internet-based Web Polaris tool, since the Web cannot yet deliver a full-blown graphical user interface, and this would unnecessarily restrict local editing of WordNets. However, keeping all the benefits of the Web, such as distributed work environment, concurrent access to the data and multiple views of the data will be achieved through the WMS. Thus, we intend to develop a WMS that will allow the local tools to retrieve the required information. However, since the Internet is not always reliable the offline operation of local tools will be the primary mode of the WMS whereas the online one will be an extra facility. This way, a WMS that supports both online and offline integration with local tools, plus a good dedicated online interface is going to be a powerful federated platform for coordinated development of the monolingual WordNets while at the same time the construction of a multilingual database will be feasible. So far, EWN shares the same concept of a multilingual synset in the ILI. New records can only be added at the tail of the file and are maintained by a central authority that issues periodical releases of a new ILI record replacing the previous one. The WMS will provide a more flexible reference scheme that enables local WordNets to keep references to the ILI even while the latter is significantly restructured. The benefit behind using the WMS is that project

developers will be tightly linked with other WordNets and valuable suggestions for new terminology fields will be facilitated. The WMS will be as open to the user as possible since it will be fairly easy for the users to develop and add their own components to the system. This can be accomplished by either encapsulating in the system capabilities for "plugging in" other applications, or by deploying the system under a free source license. This way, users will be able to use the same platform for their work and keep at the same time the data compatible. A new browser (editor) developed for the BalkaNet project will be able to work with WordNet files written in XML and it will also employ client-server architecture. The above tools will be developed in Linux platform and the results will be widely available. The central infrastructure of the WMS is going to be a federated database along with necessary communication protocols and Linux-based tools, which will run locally and provide central services. Summarizing, the (inter) networked WMS is going to be a platform-independent tool that will enable development of monolingual WordNets and their linking into a central database.

Conclusions

A central multilingual database with WordNets for a set of Central and Eastern European languages along with a WordNet Management System will be developed. Furthermore, an adjustment of BalkaNet to EWN semantic network will be attempted so as to extend the latter and make cross-language information retrieval efficient for the less-studied Balkan languages. Finally, with the implementation of BalkaNet it will be possible to trace and explore relationships among Romance and Balkan languages.

Acknowledgements

This research was supported by the EC in the framework of the BalkaNet project, Ref.No. IST-2000-29388. We wish to thank all partners of the project for their valuable contribution and Dr. Piek Vossen and Prof. Christiane Fellbaum for their support.

References

Bloksma L., Diez-Orzas, Vossen P (1996) *The User Requirements and Functional Specification of the EuroWordNet project* EWN-deliverable D.001, Le-4003

Vossen P. (1998) *A Multilingual Database with Lexical Networks,* Kluwer Academic Publishers, Dordrecht

Affiliations of the authors

S. Stamou *{stamou@cti.gr}* and D. Christodoulakis *{dxri@cti.gr}*, can be reached at Databases Laboratory, of Computer Engineering & Informatics Department, Patras University, Greece. K. Oflazer {*oflazer@sabanciuniv.edu}* can be reached at Human Language Technology Laboratory, Sabanci University, Orhanli, Tuzla, Istanbul, Turkey. K. Pala {*pala@fi.muni.cz}* can be reached at Faculty of Informatics, Masaryk University, Brno, Czech Republic. D. Cristea *{dcristea@infoiasi.ro}* can be reached at Faculty of Informatics, University Alexandru Ioan Cuza, Iasi, Romania. D. Tufis *{tufis@racai.ro}* can be reached at Centre for Advanced Research in Machine Learning, NLP and Conceptual Modeling, Academia Romana, Bucharest, Romania. S. Koeva *{svetla@ibl.bas.bg}* can be reached at Bulgarian Academy of Science, Institute of Bulgarian Language, Sofia, Bulgaria. G. Totkov *{totkov@ulcc.uni-plovdiv.bg}* can be reached at Computer Science Department, University of Plovdiv, Plovdiv, Bulgaria. D. Dutoit *{memodata@wanadoo.fr}* can be reached at Memodata Natural Language Department, Caen, France and M. Grigoriadou *{gregor@di.uoa.gr}* Department of Informatics & Telecommunications, Athens University, Greece.