WordNet Exploitation through a Distributed Network of Servers

I. D. Koutsoubos[1,2], Vassilis Andrikopoulos[1], and Dimitris Christodoulakis[1,2]

[1] Computer Engineering and Informatics Department, Patras University,
26500 Patras, Greece
Email: andrikop@ceid.upatras.gr
[2] Research Academic Computer Technology Institute,
61 Riga Feraiou, 26221, Patras, Greece
Email: koutsoub@cti.gr, dxri@cti.gr

Abstract. The architecture of a lexical database in which multilingual semantic networks would be stored requires the incorporation of flexible mechanisms and services, which would enable the efficient navigation within and across lexical data. We report on WordNet Management System (WMS), a system that functions as the interconnection and communication link between a user and a number of interlinked WordNets. Semantic information is being accessed through a distributed network of servers, forming a large-scale multilingual semantic network.

1 Introduction

WordNet has been identified as an important resource in the human language technology and knowledge processing communities. Its applicability has been cited in many papers and systems have been implemented using WordNet. Almost every NLP application nowadays requires a certain level of semantic analysis. The most important part of this process is semantic tagging: the annotation of each content word with a semantic category. WordNet serves as a useful resources with respect to this task and has so far been used in various applications including Information Retrieval, Word Sense Disambiguation, Machine Translation, Conceptual Indexing, Text and Document Classification and many others.

There is an increasing amount of wordnet resources being made available for NLP researchers. These resources constitute the basic raw materials for building applications such as the abovementioned. Semantic networks standardization is of prime importance in the case of WordNets incorporation in real life applications. Towards a vision of next-generation tools and services that will enable the widespread development and use of wordnet resources we present a distributed WordNet server architecture in which WordNet servers, analogous to database servers, provide facilities for storing and accessing wordnet data via a common network API. Apart from distributing wordnets over multiple servers the system is capable of distributing wordnet-related services over multiple servers.

2 Advantages of Distributed Systems

We can summarize the motivations for adopting a distributed architecture for WordNet management-exploitation:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 265–269.

Distributed Information Sources: WordNet resources may be scattered across multiple physical locations. Access to multiple resources may be mediated and rendered in a uniform way.

Sharing: Applications need to access several services or resources in an asynchronous manner in support of a variety of tasks. It would be wasteful to replicate problem-solving capabilities for each application. Instead it is desirable that the architecture supports shared access to agent capabilities and retrieved information.

Complexity Hiding: A distributed architecture allows specifying different independent problem-solving layers in which coordination details are hidden to more abstract layers.

Modularity and Reusability: A key issue in the development of robust analysis application is related to the enhancement and integration of existing stand-alone applications. Agent may encapsulate pre-existing linguistic applications, which may serve as components for the design of more complex systems. Inter-agent communication languages improve interoperability among heterogeneous services providers.

Flexibility: Software agents can interact in new configurations "on-demand", depending on the information flow or on the changing requirements of a particular decision making task.

Robustness: When information and control is distributed, the system is able to degrade gracefully even when some of the agents are not able to provide their services. This feature is of particular interest and has significant practical implications in natural language processing because of the inherent unpredictability of language phenomena.

Quality of Information: The existence of similar analysis modules able to provide multiple interpretation of the same input offers both 1) the possibility of ensuring the correctness of data through cross-validation and 2) a mean of negotiating the best interpretation(s).

3 Our Approach

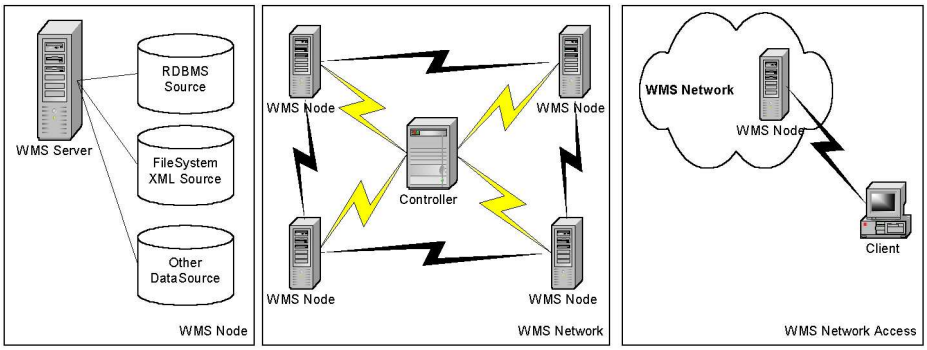
For the implementation of a flexible multilingual lexicographic database where navigation in the linguistics information would be facilitated there is an apparent need that flexible mechanisms and services are provided by a main technical infrastructure of the multilingual network. The WordNet Management System (WMS) is the system that acts as the interconnection and communication link between a user and any of the involved monolingual systems. As part of this communication someone should have the ability to submit requests for wordnet data contained in the WMS network. Moreover, 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.

From its definition, WMS falls into the Data Integration framework, being able to manage a distributed, dynamic network of homogeneous data. Previous systems built for this purpose [8,11,12] are often characterized by a centralized system that controls and manages interactions among distributed information sources in order to serve requests for data. As a consequence, in a distributed environment where no a priori knowledge of the location of specific data is possible, the traditional mediator and federated databases approaches are not appropriate. Furthermore, approaches such as [7,9,10] that provide a source- and query-independent mediator do not deal with decentralized systems with participants and

information sources location upredictability. As mentioned in [6] a P2P approach would be a more appropriate solution, since it lacks a centralized structure and promotes the equality among peers and their collaboration only for the time necessary to fulfill a request.

On the other hand, a feature that was considered very important during the design of the system, was the ability of the system to provide data to the wider possible set of data consumers, ranging from simple users to industrial solution-based applications. This requirement called for a variety of rendering mechanisms and interfaces with variable complexity for the communication of the system with its users. The ideal solution to this problem would be an API for wordnet access, as described in [4] or an extension of it, covering more recent achievements in the interface technologies like the Web Services technologies [`http://java.sun.com/webservices/`].

Taking both requirements into account, WMS was designed following a mixed approached, borrowing elements from both architectures to solve specific problems. Specifically, it was decided that the WordNet providers, i.e. the sources of WordNet data, should form a network of servers, using P2P techniques and thus creating a unified semantic data resource which could be accessed from data consumers, linked as clients to the servers of the system, without taking into account resource-specific details which are hidden to them. The architecture of WMS is summarized in the following figures.



3.1 Network of Servers

Each WMS server hosts one (or more) wordnet data sources which are interconnected via the ILI structure [3]. WordNet data sources are identified by language and version (creating a unique pair). A WMS server is considered a node in the P2P network and is treated equally by its peers. For each peer to acquire knowledge of the data available in the network (and additionally their location and how to access them), a super-node was added to the system. It serves as a yellow pages provider, or a directory service, registering WordNet hosts and distributing this information to the other nodes of the network. The super-node maintains all information about the servers of the network and the data hosted in each one. By communicating with the super-node, each node registers itself on the network and acquires

information concerning all the distributed WordNet data sources, which validates on the grounds of accessibility and availability.

Furthermore, the server operates on two modes. In the first mode, it provides the data exclusively for its hosted data sources and links to remote ones to the clients, with the client responsible for acquiring the remote data. In its second mode, the server is responsible for both local and remote data sources, providing remote data by executing remotely the requested operations and forwarding the results to the clients.

3.2 Clients

For the purpose of architecture, we consider any kind of semantic data consumer, either simple solutions as a site or more sophisticated ones as information brokering systems, possible clients to the system. In order to accommodate the multiple needs defined by such a variety of systems, each WordNet provider was decided to also act as a server for these clients. Using the standard client-server schema, the data consumer has to submit its requests to a WMS server in order to retrieve the necessary results. Additionally, a uniform API is provided to the interested parties in the form of a number of services provided by a Web Services mechanism, which add a level of abstraction between the clients and the data resources, facilitating the usage of the system for the implementation of different in their nature applications which use semantic data in very different ways.

3.3 Data Management in the WMS

For the internal communication of the nodes of the WMS network, a custom XML messaging schema is used. Provision was taken during the design of the schema to keep it as flexible and extendable as possible to accommodate possible future enhancements of wordnet data. For the same purpose, the API that describes the functions provided by WMS is also designed with openness in mind, allowing the extension of the available operations and the flexible incorporation of new ones.

As far as the communication with the clients is concerned, a variety of access methods are provided, ranging from simple HTTP requests and SOAP to RMI. The actual messaging uses XML to describe the requests and the data, but lacking a standardized WordNet protocol, describing data and functions, the system provides templating mechanisms for defining the requests and the responses.

WMS provides the developers with the capability to use and maintain different types of storage mechanisms for their respective WordNets, from simple solutions as text files to more sophisticated ones like binary structures. The requirements for such an abstraction are set by the system in the form of an API, which a developer that wants to use a specific medium has to implement. Currently, WMS provides by default mechanisms for access to the majority of commercial Relational Database Management Systems and to XML files that use the VisDic formalism [5].

4 Discussion and Future Enhancements

We have presented the architecture of WordNet Management System, a distributed network of servers that provides facilities for WordNet exploitation. In the future, it is envisaged to

incorporate other types of lexical resources besides wordnets and to provide the mechanisms for interaction with other NLP modules, such as a module for Semantic Indexing of documents.

Acknowledgements

This research is supported under project BalkaNet: "Design and Development of a Multilingual Balkan WordNet", funded by the European Commission under the framework of the IST Programme (IST-2000-29388). We would like to especially thank professor Thanos Skodras for his support and encouragement.

References

1. Fellbaum Ch. (ed.) (1998) WordNet: An Electronic Lexical Database, Cambridge, M.A: MIT Press.
2. Stamou S., Oflazer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002) "BALKANET: A Multilingual Semantic Network for Balkan Languages". In Proceedings of the GWA 1st International WordNet Conference, Mysore, India, Jan. 21–25, 2002.
3. Vossen P. (ed.) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.
4. Miatidis M., Assimakopoulos D., Koutsoubos I.-D., Kourousias G., Tzagarakis M., Christodoulakis D. (2001) "Access to WordNets Through API".
5. Pavelek, T. and Pala K. (2002) "WordNet standardization from a practical point of view". In Proceedings of the LREC Workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation, Las Palmas, Gran Canaria. 30–34, May 28th, 2002.
6. Panti M., Penserini L., Spalazzi L. (2002) "A pure P2P approach to information integration". Tec. Report 2002-02-19, Istituto di Informatica, University of Ancona, 2002.
7. R. J. Bayardo, W. Bohrer, et al. (1997) "InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments". In Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 26, No. 2, June 1997.
8. H. Garcia-Molina, Y. Papakonstantinou, et al. (1997) "The TSIMMIS approach to mediation: data models and languages". In Journals of Intelligent Information Systems, 8:117–132, 1997.
9. A. Levy, A. Rajaraman, J. Ordille (1996) "Querying Heterogeneous Information Sources Using Source Descriptions". In Proceedings of the 22nd VLDB Confernece, September, 1996.
10. M. Nodine, W. Bohrer, A. H. Hiong Ngu (1999) "Semantic Brokering over Dynamic Heterogeneous Data Sources in InfoSleuth". In Proceedings of the 15þ International Conference on Data Engineering, March 1999.
11. A. Sheth and J. Larson (1990) "Federated Database Systems for Managing Distributed, Heterogeneous and Anonymous Databases". In ACM Transaction on Database Systems, 22(3), 1990.
12. G. Wiedehold (1992) "Mediators in the Architecture of Future Information Systems". In IEEE Computer Magazine, 25:38–49, March 1992.