Approaching Wordnets through a Structural Point of View

Dimitris Avramidis, Maria Kyriakopoulou, Manolis Tzagarakis, Sofia Stamou, and Dimitris Christodoulakis

Computer Engineering and Informatics Department
University of Patras
GR-265 00, Rion Patras, Greece
Research Academic Computer Technology Institute
Riga Feraiou 61,
GR-262 21, Patras, Greece
{avramidi, kyriakop, tzagara, stamou, dxri}@cti.gr

Abstract. The analogy of a semantic network to hypertext has long been recognized, and semantic networks have been considered as a logical model of hypertext – especially for those hypertexts with typed nodes and links. Moreover, wordnets form the most representative type of semantic networks in the field of Natural Language Processing and semantics in particular. It is obvious that hypertext and wordnets share many common points regarding their fundamental principles and the objectives towards which they both aim. This paper expresses our initial thoughts towards incorporating the Balkan WordNet in Callimachus CB-OHS, as such systems can conveniently support structure. We strongly believe that such tasks can be addressed by using already implemented domain abstractions along with a new set of behaviors.

1 Introduction

Hypertext has always been closely related to the idea of freedom to associate, making it to be considered as an alternative means of structuring information [2]. This new promising field provides its users (namely, authors and readers) with effective ways of presenting and exploring information. For authors, hypertext systems offer a high degree of flexibility for connecting pieces of information and presenting it as an assembled collection in an information network. For readers, hypertext provides tools for navigating in these information networks and for exploring them freely. Therefore, hypertext can be a valuable dialogic means, facilitating the organization of information according to the user needs.

On the other hand, semantic networks form a highly structured linguistic resource enabling a flexible navigation through the lexical items of a language. Wordnet forms a kind of conventional dictionary where semantic information of the terms it contains is represented. The main structural entities of wordnets are language internal relations through which words are linked based on their semantic properties. The main contribution of wordnets in lexicography is the systematic patterns and relations that exist among the meanings expressed via lexical units. In this respect, wordnet as a particular type of semantic networks resembles much hypertext as far as the structural organization of information is concerned.

The need for linguistic support in various user tasks, such as search engines [20], document processing [6], etc., raises new requirements for developers of linguistic applications. Such developers are in constant search for suitable tools. Given that linguistics has a strongly structural nature, the deployment of structure oriented platforms for such tasks can be beneficial. In the attempt to provide structure oriented tools, we found the deployment of Component Based Open Hypermedia Systems (CB-OHS) to create a platform for linguistic processing very challenging.

This paper is a first attempt to model wordnets using Callimachus CB-OHS [21], which provides a baseline for modeling different structural domains. Wordnets, being heavily structured language resources, could take significant advantage when modeled on the ground of a CB-OHS. In such cases, computations over structure, needed by many wordnet applications, can be efficiently provided. Therefore, applications built on top of this framework, would exploit at large the potential offered by wordnets.

Our motivation was BalkaNet[1] (**Balka**n Word**Net**), an IST project undertaken by our laboratory, that aims at combining effectively Balkan lexicography and modern computation. The most ambitious feature of BalkaNet is its attempt to represent semantic relations and organize lexical information of Balkan languages in terms of word meanings. One envisaged application of BalkaNet concerns its incorporation in Information Retrieval (IR) systems in order to support *conceptual* text retrieval as opposed to exact keyword matching. Trying to realize conceptual indexing, computations over wordnet structure must be applied, a task that can be better facilitated by CB-OHSs.

2 Structure in Semantic Networks

Wordnets form the most representative type of semantic networks in the field of Natural Language Processing and semantics in particular. Motivated by theories of human knowledge organization, wordnet emerged as a highly structured language repository, where words are defined relative to each other. Unlike machine-readable dictionaries and lexica in book format, wordnet makes the commonly accepted distinction (between conceptual-semantic relations) which link concepts and lexical relations [4]. Thus, despite their resemblance to typical thesauri, wordnets in general clearly separate the conceptual and the lexical levels of language. Such a distinction is reflected via semantic-conceptual and lexical relations that hold among synsets and words respectively. Wordnets form semantic dictionaries that are designed as networks, partly because representing words and concepts as an interrelated system seems to be consistent with evidence for the way speakers organize their mental lexicons [9, 15].

[1] For more information please visit `http://www.ceid.upatras.gr/Balkanet`.

Wordnets' hierarchical structure – as shown in Figure 1 – allows a searcher to access information stored in lexical chains along more than one path, semantics being among them. Conceptual structures are modeled as a hierarchical network enabling a graphical representation of the lexicalized concepts when the latter are denominated by words [19]. The theoretical analysis shows dependencies among semantic relations, such as inheritance of relations from sub-concepts to super-concepts. Therefore, related senses grouped together under the same lexical chain form preliminary conceptual clusters. Words belonging to the same lexical chain are connected via language internal relations, each one denoting the type of relation that holds among the underlying word meanings. Some of the language relations are reflective in the sense that if a link holds between terms A and B then a link also holds between term B and term A. However, bi-directionality of the relations strongly depends on the language particularities and semantic properties of the underlying word meanings.

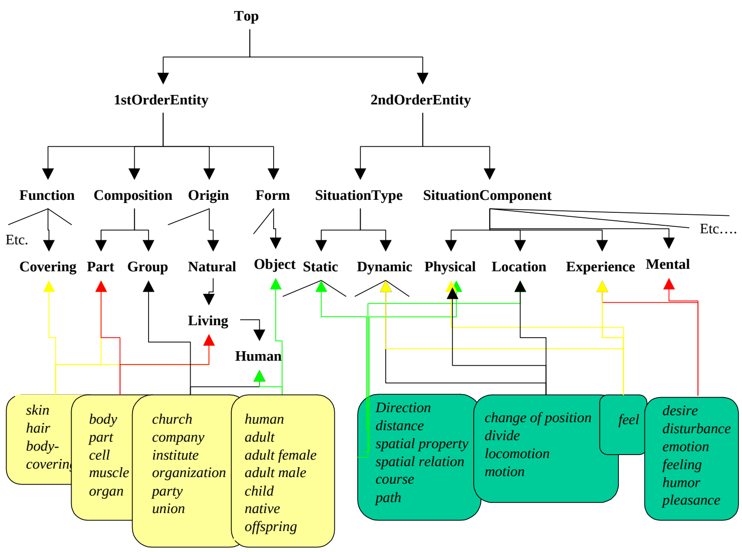

Fig. 1. This is an example of a WordNet depicting only the basic relations of *hyponymy* and *hyperonymy*, each one forming a tree. However, there are many other typed relations, such as antonymy, meronymy, etc., that have been recognized [5].

In order to account for particularities in lexicalized concepts, tags are assigned to each lexical relation denoting specialized semantic characteristics of a word's meaning. Tags can be viewed as a means of semantic constraint imposed upon semantic relations that link word meanings rather than word forms. Moreover, tags provide information about which of the semantic properties represented in a lexical chain are inherited to its components. In this respect, words represent an atomic and unbiased level of individuality that becomes meaningful via anchoring of semantic relations. As Hasan [7] pointed out, any word in a chain can be related to multiple other words in that chain. All lexical relations form a graph in which cycles are disallowed, all they contribute very little, if of any, new information.

Summarizing, the structure of lexical data within wordnets is what differentiates the latter from traditional lexicographic aids (both dictionaries and thesauri). The motivation behind constructing semantic networks in the form of a graph relies on the fact that lexical data becomes meaningful only via predefined linguistics structures. Navigation through the content of wordnets becomes feasible via language internal relations, which form the main notion around which structure is defined.

3 Approaching Wordnet via Hypertext

Adopting the "primacy of structure over data" [17], hypertext can be seen as a technology well suited to exploring different kinds of representational structures [13]. Viewing different parts of information as objects, users, often referred to as readers, can navigate through them in a more effective and convenient fashion. Additionally, authors can manipulate information according to their needs [10]. Therefore, hypertext can be regarded as an informal mechanism that describes the attributes of these objects and captures relationships that possibly exist between them. Such a characteristic allowed hypertext become known as an alternative way of structuring information.

The analogy of a semantic network to hypertext has long been recognized [3]. A semantic network is a knowledge representation scheme consisting of a directed graph in which conceptual units are represented as nodes, and relations between the units are represented as links. The graph becomes semantic when each node and link is assigned a particular type, making them meaningful [22, 23]. The essential idea of semantic networks is that the graph-theoretic structure of relations can be used for inference as well as understanding [12]. We claim that wordnets, the most representative type of semantic networks, can be supported by a CB-OHS [16, 24].

3.1 Hypertext and Wordnet: Similarities

Hypertext and wordnets share many common points regarding their fundamental principles and the objectives towards which they both aim. In particular, they are both targeted towards capturing relations that possibly exist between objects and thus providing information of the underlying objects via various types of links used for describing the relations. Therefore, the main characteristic of wordnets and hypertext systems is the ability to create associations between semantically related information items. On the one hand, these associations imply purposeful and important relationships between associated materials, whereas on the other hand the emphasis upon creating associations stimulates and encourages habits of relational thinking of the user [11].

Relations form the notion around which both semantic networks and hypertext are organized. In the case of semantic networks, relations are denoted explicitly between the lexical units they contain via predefined lexical links, and capture information on the semantic properties of words. In the case of hypertext, although the notion of association can be met in all hypertext domains, the navigational domain with the use of *links* is most closely related to it. Consequently, lexical relations form the fundamental entity of semantic networks in the same way as associations in hypertext form the basic structural element around which domains are modeled.

In both cases, information objects (either lexical or not) are heavily structured in order to enable users of wordnets or hypertext to navigate successfully through the information they contain. Structure is achieved via internal links, which form the basis on which information is stored and expressed. However, links in semantic networks and hypertext, until recently, have been viewed as two distinct elements and no attempt has been made towards comparing the two. We report on the similarities that exist between hypertext relations and semantic links in an attempt to model the latter in hypertext systems.

In order to create associations in an effective way, hypertext researchers have created a flexible link structure incorporating different levels of functionality. More specifically, in hypertext one can create single or bi-directional links, binary or n-ary links, links to links, automatically activated links, etc. Similarly, links in wordnet are reflective and generally have no restriction on the number and types of links they could include, as the relatedness between the information items is properly and adequately expressed.

However, reflective links does not always apply to all wordnet links posing, thus, the need for semantic tags to be attached on single-direction relations. Namely, tags are used on semantic network relations to indicate that a lexical item is related to another via a particular type of link but not vice versa. However, in the case of hypertext, tags are used implicitly and are defined within structural elements during the modeling phase of a structural domain (e.g. an attribute may exist denoting that a link is bi-directional or not).

Furthermore, besides creating associations among semantically related information items, another characteristic shared between hypertext and semantic networks is inheritance. This feature implies that objects are inherited by their descendants. More specifically, the notion of generalization and specialization forms the principle on which relations are expressed. Specialization and generalization define a containment relationship between a higher-level entity set and one or more lower-level entity sets. Specialization is the result of taking a subset of a higher-level entity set to form a lower-level entity set, whereas generalization is the result of taking the union of two or more disjoint (lower-level) entity sets to produce a higher-level entity set.

Inheritance in wordnets is described via the *H/H tree* (Figure 1) – that is the complementary hypernymy/hyponymy relations. This type of relationship between objects results in viewing wordnets like tree-structured sources of information, and thus circular loops are disallowed. As far as hypertext is concerned, these organizational structures exist in the taxonomic domain under the respective terminology of *supertaxon* and *subtaxon*. The subtaxon is associated with the supertaxon via an "is-a" relationship, inheriting all the characteristics that the latter might have. In particular, the user can classify objects (known as specimens) into sets according to their features, search within the members of a set to find relationships or discreet subsets, and create new sets from the already existing ones [18].

3.2 Using Callimachus to Model Wordnet

The Callimachus CB-OHS attempts to provide the framework in which different hypertext domains co-exist and structure servers - providing new abstractions and services – for new domains can be developed. Special attention has been paid on the provision of suitable tools, which are part of the methodology, to facilitate structure servers' development. One such tool is comprised of the structure templates that aim at maintaining the specifications of the structure model of a particular hypertext domain. Structure servers are guided by these structure templates to provide domain specific abstractions and constraints.

In Callimachus, the methodology for defining the structure model of a domain (e.g. navigational, taxonomic) consists of the specification and interrelation of structural types [21]. A structural type is either an instantiation of a basic abstract class, the Abstract Structural Element (ASE), or a specialization of another existing structural type. Depending on the structural domain to which a structural type belongs, an arbitrary number of *properties* and *endsets* can be specified. Such structural type definitions constitute a template that describes the structural domain to be modeled. In this view, structure servers based on templates provide the ground for delivering structural domain specific abstractions to clients. In particular, they operate on structural objects[2] transforming them to useful abstractions, thus enabling clients to use them according to the structural domain specific restrictions.

Trying to model the Balkan wordnet using Callimachus, its structure must be explicitly defined by introducing the notion of templates. Based on the common points that hypertext and wordnets share, we realized that the latter borrow many characteristics from different hypertext domains, thus raising issues that have been met in cross-domain interoperability topics [1, 14]. Taking this observation into account, strong evidence exists in building the BalkaNet structure server on top of already existing structure servers (e.g. navigational, taxonomic) requesting the appropriate services (see Figure 2). The latter could be efficiently achieved by creating the BalkaNet template, which will emerge from the utilization of the navigational and the taxonomic ones.

More specifically, in the BalkaNet structure server both navigational and taxonomic structural characteristics coexist. Depending on the requests it is called to serve, the corresponding structure server is activated. If, for example, an "is-a" request is submitted to the server, the respective H/H tree has to

[2] A structural object is an instantiation of a structural type.

be traversed. Serving such requests presupposes that they are directed to the taxonomic structure server, due to the correspondence H/H trees have with the underlying domain. Conversely, if a request doesn't fall in the "is-a" relation, the navigational structure server is called to serve it. This is due to the fact that non "is-a" relations (e.g. synonymy, antonymy) link nodes in a non hierarchical way. Summarizing, other requests can be erved in a similar manner.

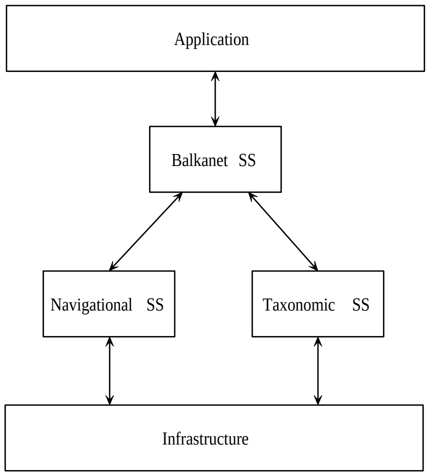

Fig. 2. The BalkaNet structure server and its incorporation in Callimachus system.

Consequently, specifying an adequate set of behaviors would result in better exploitation of the wordnet structure. In this way, applications built on top of Callimachus, such as conceptual indexing, can perform structural computations, increasing the possibility for better results in information retrieval. For example, speaking of query expansion[3] in search engines, given a query by the user, the structure server could traverse the synonymy relation for each appearing query term and replace it with its synonym. Similarly, in case the user wishes to broaden the search conceptually, the hyperonymy relation could be traversed adding the hyperonyms to the query, whereas for narrowing the query, the hyponymy relation should be traversed. Supporting this task within the CB-OHS framework provides the ground for delivering the computational tools to exploit wordnet structures.

4 Discussion

As it has been already mentioned, hypertext is not mainly used for the organization of information but can be considered as a significant means of structuring

[3] Query expansion is a method for performing sense-based retrieval by linking words based on their semantics [8].

information. Viewing wordnets under the notion of hypertext, the power of the latter is enforced even more, making us infer that any kind of information can be structured under its fundamental characteristics. Taking advantage of the structural characteristics of hypertext, while developing wordnets, can prove quite beneficial for both the lexicographic and linguistic communities.

Open Hypermedia Systems (OHS) in their move from domain-specific frameworks to cross-domain Component Based Open Hypermedia Systems, provide frameworks for supporting combinations among different domains. Based on the similarities that wordnets and already existing hypertext domains share, we reached the conclusion that even if wordnet is a new application domain it cannot be seen as a new structural domain. In particular, the wordnet structure server relies on the basic structural abstractions of the navigational and the taxonomic domain. However, it supports a new set of behaviors providing the wordnet functionality.

The motivation of this work emerged from the BalkaNet project whose application focuses on information retrieval applications. However, we strongly believe that by adopting structures implied by the hypertext community in other applications, such as lexicography, the potential and performance of the latter can be significantly improved. When it comes to the storage of lexicographic data, the need for efficient structures becomes apparent due to the large amount of information that has to be handled and especially due to the dynamic nature of the underlying information.

References

1. Dirk Bucka-Lassen, Claus Aagaard Pedersen, and Ólavur H. Reinert. Cooperative Authoring using Open Spatial Hypermedia. Master's thesis, Aarhus University, 1998.
2. Vanevar Bush. As We May Think. *Atlantic Monthly*, pages 101–108, July 1945.
3. Jeff Conklin. Hypertext: An Introduction and Survey. *IEEE Computer*, 20(9):17–41, 1987.
4. Martha W. Evens, editor. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks.* Cambridge University Press, Cambridge, England, 1988.
5. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
6. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11):964–971, 1987.
7. Ruqaiya Hasan. Coherence and Cohesive Harmony. In James Flood, editor, *Understanding Reading Comprehension*, pages 181–219. IRA, 1984.
8. Y. Jing and W. Bruce Croft. An Association Thesaurus for Information Retrieval. In *Proceedings of RIAO-94,* 4^{th} *International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.
9. Martin Kay. The Concrete Lexicon and the Abstract Dictionary. In *Proceedings of the* 5^{th} *Annual Conference of the UW Center for the New Oxford English Dictionary*, pages 35–41, Waterloo, Ontario, Canada, 1989.

10. Maria Kyriakopoulou, Dimitris Avramidis, Michalis Vaitis, Manolis Tzagarakis, and Dimitris Christodoulakis. Broadening Structural Computing Systems Towards Hypermedia Development. In *Proceedings of the* 3^{rd} *International Workshop on Structural Computing*, pages 131–140, Århus, Denmark, 2001. Springer-Verlag.
11. George P. Landow. Relationally Encoded Links and the Rhetoric of Hypertext. In *Proceedings of the ACM Conference on Hypertext*, pages 331–343. ACM Press, 1987.
12. Fritz W. Lehmann. Semantic Networks in Artificial Intelligence. In Fritz W. Lehmann, editor, *Semantic Networks*, pages 1–50. Pergamon Press Ltd., 1992.
13. Catherine C. Marshall. Exploring Representation Problems Using Hypertext. In *Proceedings of the ACM Conference on Hypertext*, pages 253–268, Chapel Hill, North Carolina, United States, 1987. ACM Press.
14. Dave E. Millard, Luc Moreau, Hugh C. Davis, and Siegfried Reich. FOHM: A Fundamental Open Hypertext Model for Investigating Interoperability between Hypertext Domains. In *Proceedings of the* 11^{th} *ACM Conference on Hypertext and Hypermedia*, pages 93–102, San Antonio, Texas, United States, 2000. ACM Press.
15. George A. Miller. Nouns in Wordnet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 23–46. MIT Press, 1998.
16. Peter J. Nürnberg, John Leggett, and Uffe K. Wiil. An Agenda for Open Hypermedia Research. In *Proceedings of the* 9^{th} *ACM Conference on Hypertext and Hypermedia: links, objects, time and spacestructure in hypermedia systems*, pages 198–206, Pittsburgh, Pennsylvania, United States, 1998. ACM Press.
17. Peter J. Nürnberg, John J. Leggett, and Erich R. Schneider. As We Should Have Thought. In *Proceedings of the* 8^{th} *ACM Conference on Hypertext*, pages 96–101, Southampton, United Kingdom, 1997. ACM Press.
18. H. Van Dyke Parunak. Dont't Link Me In: Set Based Hypermedia for Taxonomic Reasoning. In *Proceedings of the* 3^{rd} *annual ACM Conference on Hypertext*, pages 233–241, San Antonio, Texas, United States, 1991. ACM Press.
19. Uta Priss. The Formalization of Wordnet by Methods of Relational Concept Analysis. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 179–196. MIT Press, 1998.
20. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
21. Manolis Tzagarakis, Dimitris Avramidis, Maria Kyriakopoulou, monica c. schraefel, Michalis Vaitis, and Dimitris Christodoulakis. Structuring Primitives in the Callimachus Component-Based Open Hypermedia System. *Journal of Network and Computer Applications*, October 2002. To be published.
22. Weigang Wang and Roy Rada. Experiences with Semantic Net Based Hypermedia. *International Journal of Human Computer Studies*, 43(3):419–439, 1995.
23. Weigang Wang and Roy Rada. Structured Hypertext with Domain Semantics. *ACM Transactions on Information Systems (TOIS)*, 16(4):372–412, 1998.
24. Uffe K. Will and Peter J. Nürnberg. Evolving Hypermedia Middleware Services: Lessons and Observations. In *Proceedings of the 1999 ACM Symposium on Applied Computing*, pages 427–436. ACM Press, 1999.