Exploring Balkanet Shared Ontology for Multilingual Conceptual Indexing

Sofia Stamou* Goran Nenadic+ Dimitris Christodoulakis*

*Computer Engineering and Informatics Department
Patras University, 26500, and
Research Academic Computer Technology Institute
61 Riga Feraiou, 26221, Patras, Greece
{stamou, dxri}@cti.gr
+Department of Computation
UMIST, Manchester, UK
G.Nenadic@umist.ac.uk

Abstract

As the size of the Web grows, it becomes an imperative to equip search engines with sophisticated indexing modules in order to enable a meaningful organization of the stored data. In this paper we present a structured multilingual conceptual repository that has been employed as the backbone of a conceptual indexing and retrieval system. Our conceptual warehouse originates from a multilingual semantic network (Balkanet) and its Inter-Lingual-Index, which was enriched with domain ontology information inherited from the SUMO ontology. We report on the ontology's design principles and provide a description of its structure. We argue that an important attribute of the Balkanet's ILI is its flexibility in incorporating new concepts and/or languages by allowing the percolation of shared semantic attributes to all concepts represented within taxonomies. We further present our approach to conceptual indexing, and introduce an indexing algorithm that utilizes Balkanet's classified conceptual taxonomies. Finally, we discuss how conceptual taxonomies can help retrieval algorithms in making links between terms used in search requests and semantically related terms that might be found in the indexed documents.

Introduction

The advent of the World Wide Web has made available a great wealth of digital information, which continuously proliferates as users of Information Technology Systems (ITS) increase. Web search engines are among the most widely used ITS and as such their competence poses many challenges to the Information Retrieval (IR) community. While access, index coverage and speed of Web search engines are being improved, end users are more and more faced with the problem of how to deal with the massive amount of information, and where to find what one needs in the huge network of data sources. As traditional keyword-matching retrieval approaches seem not to meet adequately users' information needs, the IR community is more and more challenged by the *paraphrase problem* (Woods, 1997), i.e. the problem of retrieving relevant documents that are indexed with terms which are different from (but conceptually related to) query terms. To cope with the paraphrase problem many approaches have been addressed in the literature, the most promising of which imply the utilization of conceptual taxonomies towards conceptual and contextual indexing of Web documents (Stairmand and Black, 1996; Gilarranz et al., 1997).

In this paper we report on the design of a multilingual conceptual ontology, and its contribution in tackling issues pertaining to the paraphrase problem. Specifically, we explore how a language-independent conceptual ontology that exhibits a tree-like structure of its concepts, can be employed to conceptually index Web documents, and how it can assist information seekers to navigate within the Web's conceptual space. We also argue that a conceptual ontology can help retrieval algorithms to locate qualitative data sources by making connections between terms used in a search request and semantically related terms that might be found in the indexed documents.

The semantic resource out of which our sense inventory is obtained is the Balkanet lexical database (Oflazer et al., 2001). Balkanet is a multilingual semantic network, comprising monolingual Wordnets for Central and Eastern European languages. Each individual Wordnet stores concepts organized into semantic taxonomies, which are further mapped against their English semantic equivalents via an Inter-Lingual-Index (ILI). Mapping is achieved through lexico-semantic relations so that all monolingual nodes are conceptually aligned across languages. To allow the efficient manipulation of Balkanet against conceptual indexing, we classified ILI's conceptual taxonomies under broad conceptual domains, that have been adopted from the Suggested Upper Merged Ontology (SUMO, cf. (Niles and Pease, 2001)). To demonstrate the potential of a structured sense inventory, we further suggest a general infrastructure that employs Balkanet's ontology as a baseline for a more meaningful organization of the data sources that are indexed by Web search engines.

The remainder of the paper is organized as follows: in Section 2 we describe the methodology we adopted for building a language-independent conceptual ontology and we demonstrate how the ontology evolved within a multilingual semantic network. We also sketch the way in which conceptual hierarchies are classified under broad conceptual domains. In Section 3 we present our approach towards conceptual indexing and we propose an indexing algorithm that utilizes Balkanet's classified conceptual taxonomies. We conclude the paper with a discussion on the challenges associated with using the shared ontology as the basis for conceptually driven IR, and we point to future research directions.

Design Principles and Implementation of the Conceptual Ontology

The lexical knowledge resource out of which our conceptual repository emerged is the Balkanet lexical database. Balkanet is a multilingual semantic network that comprises monolingual Wordnets for six Balkan languages, namely Turkish, Greek, Bulgarian, Serbian, Romanian and Czech. Lexicalized concepts are represented in terms of synonym sets (synsets), which form the core structural elements of each Wordnet (Fellbaum, 1998). Each synset encodes all terms sharing the same sense. Synsets are organized into semantic taxonomies through hypernymic and hyponymic links. Balkanet is currently under development and by the time of this contribution each monolingual Wordnet stores approximately ~15K synsets. Table 1 summarizes some quantitative data of each monolingual Wordnet.

Wordnet	*Synsets*	*Av.sysnet length*	*Av.senses per literal*	*Common synsets*
Bulgarian	15007	1,79	1,31	8516
Czech	26525	1,49	1,37	8195
Greek	15781	1,33	1,33	5933
Romanian	14707	1,92	1,72	8484
Serbian	4772	1,73	1,27	4772
Turkish	10280	1,52	1,35	8516

Table 1: Statistics on the monolingual Balkan Wordnets[1]

All concepts within each monolingual semantic taxonomy are mapped against their English semantic equivalents via an Inter-Lingual-Index (ILI). Balkanet's ILI originates from the latest version of Princeton WordNet (ftp.cogsci.princeton.edu), which is a structured hierarchy, comprising approximately 153K unique literals organized into 115K synsets. Mapping of monolingual taxonomical elements onto the ILI's nodes is achieved through lexico-semantic relations so that all monolingual nodes are conceptually aligned across languages. Language-specific concepts, which are not lexicalized in English (e.g., concepts describing specific professions, food, etc.) are manually embedded into ILI by complex-equivalence inter-ILI relations.

To ensure that terminological overlap across Wordnets is not hampered by inconsistent projections of the monolingual concepts onto the ILI nodes, two validation tasks have been performed. The first one is the *Hierarchy Preservation Principle* (Tufis and Cristea 2002), which allows importing hierarchies across monolingual aligned synsets and checks the validity of their taxonomic structures through a soft ILI clustering approach. ILI clustering concerns the grouping of similar senses of different occurrences of the same word. In addition to this validation control policy, a semantic validation task was carried over the Orwell's *1984* multilingual corpus (http://nl.ijs.si/ME/CD/docs/1984.html) delivered by the Multext-East project (Erjavec et al., 2001). Semantic validation aims at checking inter-lingual mappings across Wordnets by examining terms' translations in the parallel corpus. Both tasks aim at verifying the consistent mapping of monolingual concepts across the ILI's taxonomies.

[1] Figures represent statistics as of January 2004. The Czech Wordnet contains more synsets as its development started earlier (within the EWN project), while the Serbian Wordnet contains fewer synsets because its development started in a later phase of the Balkanet project.

Structuring Conceptual Hierarchies

Balkanet is organized similarly to EuroWordNet (EWN) (Vossen, 1998). However, while EWN implements its ILI via simple, unstructured equivalence links, we designed Balkanet's ILI as a shared, more complex conceptual warehouse by providing a tree-like structure of its concepts, building this way a conceptual taxonomy. The main rationale for structuring the ILI is that a language independent conceptual taxonomy employed as the backbone of a conceptual indexing infrastructure would result in a semantically meaningful organization of the indexed data. In order to utilize the conceptual taxonomy to efficiently locate where in the taxonomy a concept belongs to, it is necessary to first organize the concepts of the taxonomy in such a way so that every concept has explicit pointers to its most specific concepts (hyponyms) and from its most general concepts (hypernyms).

In addition, we introduce the notion of *conceptual domains*, which are treated as conceptual ontologies and which serve to the transfer of the respective semantic attributes within monolingual Wordnets and across the ILI network. The Balkanet ILI is organized as a set of conceptual taxonomies for certain conceptual domains, which are inherited from the SUMO ontology (http://ontology.teknowledge.com/). SUMO is an upper ontology that contains concepts general enough to address a broad range of domain areas. Concepts specific to particular domains are included within ILI's taxonomies, whereas SUMO provides a structure upon which ontologies need to be constructed for particular domains. The architecture of the conceptual taxonomies linked to the SUMO ontology domains is illustrated in Figure 1. We chose SUMO as a base ILI ontology for three reasons. First and foremost, it was already mapped to Princeton WordNet's synsets, which are contained in the Balkanet ILI. Secondly, it combines resources from many fields, and, most importantly, it is freely available and extensible.

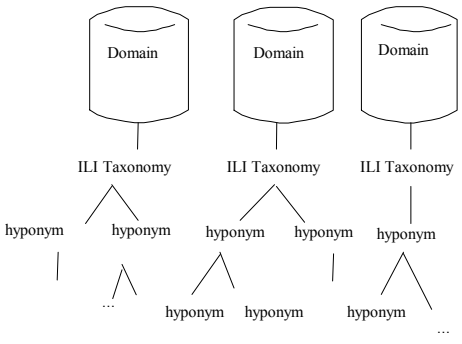

Figure 1: Balkanet ILI classified taxonomies

Each element of a conceptual domain is built into a taxonomic structure and each taxonomy links concepts that belong to that particular domain. All ILI hierarchies that belong to the SUMO ontology domains are marked-up with explicit domain information, which is automatically transferred to the equivalent monolingual Wordnet taxonomies through inter-ILI equivalence links. This way, conceptual domains are assigned automatically to monolingual Wordnet synsets.

Conceptual Indexing using Domain Taxonomies

To demonstrate the potential that conceptual taxonomies have in Web indexing, we employ the Balkanet shared ontology as a baseline for a more meaningful organization of the data records that are to be indexed by Web search engines. The main component of our conceptual indexing approach is a conceptual classification formula, which clusters the contents of the engine's index on the basis of their topical relations and semantic similarity. To perform conceptual clustering, we treat ILI's conceptual domains as topics under which Web documents are classified. Conceptual clustering takes place via an internal mapping between documents' representative terms and ILI's concepts, and by calculating their semantic similarity. Based on corresponding index terms, each document is assigned to a specific domain(s).

The first step towards classification concerns the morphological pre-processing of documents in order to extract a core set of lexicalized concepts, represented in each document. To address multilingual conceptual indexing, the clustering module employs the language denoting tags accompanying each document as a guide towards morphological processing and towards the use of the information encoded within the respective monolingual Wordnet. Morphological processing involves document tokenization, part-of-speech tagging and lemmatization. Henceforth, term weighting schemes (for example the normalized *tf*idf* formula (Salton and Buckley, 1988)) are employed against all documents' content terms[2]. Terms with high frequency weights are those that lexicalize the most representative concepts of a given document, and are the ones on which indexing and clustering are based. These terms are then located in the corresponding monolingual Wordnet, and their ILI's conceptual equivalents are retrieved simply by following the semantic links. Document clustering then takes place by traversing the conceptual taxonomies of the retrieved ILI nodes. The closer the matching nodes are to a topmost node (the shortest path), the more likely that a given document belongs to that cluster. However, relying exclusively on the idea of the shortest path for measuring conceptual distance is not sufficient per se for ensuring the successful conceptual clustering of documents. This is essentially the case where a document's terms are mapped against several ILI concepts, each of which belongs to a different taxonomy and whose distances from each taxonomy's root node are equal (or comparable). To account for such conflicting cases we allow for a document to be clustered under multiple conceptual domains.[3] For calculating conceptual distances we follow Resnik's (1995) approach that captures semantic similarity by means of the information content of the concepts in a hierarchical network. Conceptual distance is not only used to reflect semantic similarities between terms, but also to tackle sense ambiguities issues in cases a term is distributed over several ILI nodes.

In Figure 2 we present the detailed indexing algorithm that utilizes ILI's classified conceptual taxonomies to organize indexed documents. The algorithm takes as input the list of the most representative lexicalized concepts of a given document (determined as discussed earlier), as well as the monolingual Wordnet taxonomy against which indexing will take place.

```
search_terms against ILI
  if terms found
    traverse_taxonomies_up_to_the_domains
      if all belong to the same domain
        index_doc_under_that_domain
      else
        count_matching_nodes_of_each_domain
          if matching nodes are equal
            count_conceptual_distance
              if equal
                index_doc_in_all_matching_domains
              else
                index_doc_in_domain_of_shortest_pat
                h
          else
            index_doc_in_domain_of_the_more_matchin
            g_nodes
  else
index_doc_in_the_plain_index
```

Figure 2: The Indexing Algorithm

The indexing algorithm maps extracted terms against the respective Wordnet taxonomy, and attempts to classify each document under one or more conceptual domains. When matching Wordnet nodes are located, the algorithm computes semantic similarities between document's terms and taxonomies' nodes in order to determine the conceptual domain under which the given document will be stored. The algorithm proceeds until all documents whose terms correspond to the hierarchies' nodes are assigned to one or more conceptual domains. If the algorithm fails to map the extracted terms to the taxonomies' nodes, it stores the document under the engine's plain index. Note that failure mappings might be either due to lemmatization errors, or due to Wordnets' incompleteness.

At the engine's repository, multiple indices are kept, each one corresponding to an ILI conceptual domain. The engine's indexing modules are responsible for directing crawled Web pages to the engine's clustering modules and for storing conceptually related pages under the same index. Organizing the engine's repository into conceptual clusters facilitates the performance of the engine's retrieval modules. Also, note that Balkanet ontology can be employed to conceptually index Web documents irrespective of their natural language through the maintenance of cross-lingual topical-focused indices.

Challenges and Conclusions

Developing a language-independent, consistent and comprehensive conceptual ontology that can be used for semantic indexing is not an easy task. The major difficulty we encountered while structuring our sense inventory concerned inter-lingual alignment issues. In particular, we were challenged to incorporate (into the ILI) language-specific concepts that are common across the Balkan languages, but for which there were no lexicalized English counterparts. We tackled such cases by allowing complex ILI relations, an approach that reassures that ILI remains a language neutral conceptual knowledge base. Inter-ILI links also guarantee a level of consistency across Wordnet mappings. Moreover, the adoption of the SUMO ontology

[2] As content terms we consider nouns, verbs, adjectives and adverbs.

[3] This way a document about *tuition fees* would be clustered under both *education* and *economy* domains.

domains helped us structure the ILI taxonomy in a meaningful way and gave us the flexibility to enrich the ILI with new concepts without imposing any need for structural changes. This flexibility is due to the percolation of the shared semantic attributes to all the concepts represented in each ILI taxonomy.

Further, the Balkanet shared ontology can serve as a baseline for multilingual conceptual indexing. We have presented an approach that clusters documents according to the conceptual domains to which their representative terms belong. Documents can be classified under multiple domains, while the problem of ambiguous terms is addressed on the grounds of conceptual distances within the taxonomy. So far in our experiments we have used Resnik's approach to calculate semantic similarities, but we are also considering other approaches, like the conceptual density approach (Agirre and Rigau, 1996).

The proposed approach for clustering documents based on the classified ILI taxonomy exhibits several advantages. One benefit for clustering ILI's taxonomies under the SUMO domains is that each taxonomy can be viewed as a domain-specific Wordnet and, as such, it can be employed by applications that require specialized knowledge sources. Another advantage of our structured ILI is that it can be extended with other languages and/or concepts without requiring any modifications. Moreover, the conceptual indexing infrastructure we have designed maintains distinct multilingual indices for each conceptual domain, a feature that makes the engine's repository manageable upon updates and has a strong potential in supporting specialized cross-lingual Web searches. In addition to indexing, the suggested classified and structured sense inventory enables the efficient maintenance of the ILI's hierarchies, and contributes in dealing with the proliferation of ILI's concepts among individual Wordnets.

We believe that the Balkanet shared ontology can be further used to improve IR performance by using conceptual indexing, as conceptual taxonomies have a strong potential in helping information seekers satisfy their needs. We argue that a core component of a conceptual retrieval system is a conceptual indexing module that groups indexed documents under conceptual domains on the basis of their semantics, and organizes them on the basis of their conceptual closeness. The objective of the conceptual taxonomy is, therefore, to feed the engine's indexing modules with information on the documents' semantics so as to index them under conceptual domains. Thus, the main idea for employing Balkanet's shared ontology towards IR is that the ontology could be used as a deep conceptual map of the data sources stored by a Web search engine, allowing users to navigate within the Web's conceptual graph. In that respect, the conceptual ontology can help retrieval algorithms make connections between terms used in a search request and semantically related terms that might be found in the relevant indexed documents.

A significant amount of work remains to be accomplished prior the proposed indexing module is fully functional for retrieval purposes. To that end we are currently testing the performance of our algorithm in indexing a small set of Web documents collected from the Southeast European Times (http://www.balkantimes.com) Web site, which contains multilingual news articles. In the future, we plan to develop a searching mechanism that would explore the conceptual taxonomies while processing search queries, in order to retrieve high quality results. We also plan to embed advanced searching modes, which would allow the system's users specify the conceptual domain(s) out of which they wish to retrieve information. Also, we want to explore more gradual and fine-grained clustering of documents. Still, many challenging issues need to be addressed before we end up with an online, scalable conceptual IR system.

Acknowledgements

The research presented here has been carried out in the framework of the Balkanet project, supported by the European Commission (IST-2000-29388).

References

Agirre E. & Rigau G. (1996). Word Sense Disambiguation Using Conceptual Density. In Proceedings of the COLING '96, Copenhagen, Denmark, pp.16-22

Erjavec T., Ide N., Tufis D. (2001). Automatic Sense Tagging Using Parallel Corpora. In Proceedings of the 6th NLP Pacific Rim Symposium, Japan, pp.212-219

Fellbaum C., (Ed.) (1998). Wordnet: An Electronic Lexical Database. MIT Press

Gilarranz J., Gonzalo J., Verdejo F. (1997). An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA

Oflazer K., Stamou S., Christodoulakis D. (2001). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In the Elsnet Newsletter of the European Network in Human Language Technologies, vol. Autumn 2001

Niles I. & Pease A. (2001). Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS), Ogunquit, Maine, pp. 2-9

Princeton WordNet 2.0. ftp.cogsci.princeton.edu

Resnik P., (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. In Proceedings of the 3rd Workshop on Very Large Corpora, MIT, pp. 54-68

Salton G. & Buckley C. (1988). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, vol.24, no.5, pp.513-523

Southeast European Times. http://www.balkantimes.com

Stairmand M.A. & Black W.J. (1996). Conceptual and Contextual Indexing using WordNet-derived Lexical Chains. In Proceedings of the 18th BCS IRSG Annual Information Retrieval Research Colloquium, pp.47-65

SUMO. http://ontology.teknowledge.com

Tufis D. & Cristea D. (2002). Methodological Issues in Building the Romanian Wordnet and Consistency Checks in Balkanet. In Proceedings of the LREC Special Workshop on Wordnets, Las Palmas, Spain

Vossen P. (Ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. In Kluwer Academic Publishers, Dordrecht

Woods W. (1997). Conceptual Indexing: A Better Way to Organize Knowledge. Technical Report TR-9761, Sun Microsystems Laboratories, Mountain View, CA