LexiRes: A Tool for Exploring and Restructuring EuroWordNet for Information Retrieval

Ernesto William De Luca and **Andreas Nürnberger** [1]

Abstract. The problem of word sense disambiguation in lexical resources is one of the most important tasks in order to recognize and disambiguate the most significant word senses of a term. Lexicographers have to decide how to structure information in order to describe the world in an objective way. However, the introduced distinctions between word meanings are very often too fine grained for specific applications. If we want to use or even combine lexical resources within information retrieval systems, for example, we might want to apply the lexical resources in order to disambiguate documents (retrieved from the web within an information retrieval system) given the different meanings (retrieved from lexical resources) of a search term having unambiguous description. Therefore, we are usually interested in a small list of meanings with very distinctive features. Since many lexical resources, especially WordNet, provide frequently too fine grained word sense distinctions, we implemented the tool LexiRes that gives the possibility to navigate lexical information, helping authors of already available lexical resources in deleting or restructuring concepts using automatic merging methods.

1 Introduction

Standard keyword based search engines retrieve documents without considering the importance of user oriented information presentation. It means that the user has to analyze every document and decide himself which are the documents that are relevant with respect to the context of his search. For example, users have to navigate every document in order to recognize to which meaning of their query words the documents belong to. Thus, it would strongly support a user if the context - which is defining the meaning of a word - could be recognized automatically and the documents could be labelled or grouped with respect to the meaning of the respective search terms. One way to obtain a context description of different word senses is to explore lexical resources using the word we are looking for in order to select concepts based on the linguistic relations of the lexical resource that define the different word senses. Such disambiguating relations are intuitively used by humans. However, if we want to automate this process, we have to use resources - such as probabilistic language models or ontologies - that define appropriate relations. One of these most important resources available to researchers for this purpose is WordNet [4] and its variations like MultiWordNet [3] and EuroWordNet [15] as discussed in the following.

However, since many lexical resources or ontologies, especially WordNet, provide frequently too fine grained word sense distinctions, we implemented the tool LexiRes that gives the possibility to navigate lexical information, helping authors of already available lexical resources in deleting or restructuring concepts using automatic merging methods. The restructured information can be navigated and explored. Authors can decide if word senses are unambiguous and important enough to let them in the hierarchy at the same place or if they express similar concepts and can be merged under the same (now, more general) meaning.

In the following, we first briefly introduce the structure of WordNet and EuroWordNet. Then we discuss the problem of word sense disambiguation in information retrieval and problems related to WordNet in order to motivate the LexiRes system, which is then presented in Sect. 4.

2 WordNet

WordNet [4] was designed by use of psycholinguistic and computational theories of human lexical memory. It provides a list of word senses for each word, organized into synonym sets (SynSets), each representing one constitutional lexicalized concept. Every element of a SynSet is uniquely identified by an identifier (SynSetID). It is unambiguous and carrier of exactly one meaning. Furthermore, different relations link these elements of synonym sets to semantically related terms (e.g. hypernyms, hyponyms, etc.). All related terms are also represented as SynSet entries. These SynSets also contains descriptions of nouns, verbs, adjectives, and adverbs. With this information we can describe the word context. Fig. 1 represents an example of the ontology hierarchy defined by WordNet [4]. This resource can be used for text analysis, computational linguistics and many related areas.

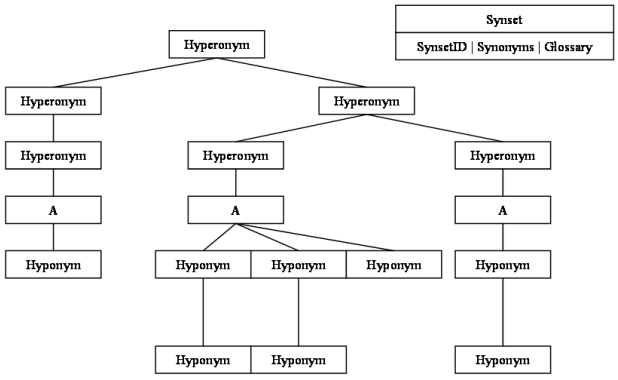

Figure 1. Example of an ontology hierarchy for a given term A.

[1] University of Magdeburg, Germany, email: deluca@iws.cs.uni-magdeburg.de

2.1 EuroWordNet

WordNet was first developed only for the English language. Then different versions were developed for several other languages as for example EuroWordNet [15] for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Given that we want to retrieve from the web different documents in different languages analysing different contexts, we decided to use the EuroWordNet multilingual lexical database. Its structure is the same as the Princeton WordNet [4] in terms of SynSets with different semantic relations between them. Each individual wordnet represents a unique language-internal system of lexicalizations. The Inter-Lingual-Index (ILI) was introduced in order to connect the WordNets of the different languages. Thus, it is possible to access the concepts (SynSets) of a word sense in different languages.

In addition to the Inter-Lingual-index, there is also a Domain-Ontology and a Top-Concept-Ontology related to this lexical database. The shared Top-Ontology is a superordinate hierarchy of 63 semantic distinctions for the most important language independent concepts (e.g. Artifact, Natural, Cause, Building) and is interconnected with the ILI through the WordNet-Offsets. Hereby a common semantic framework for all the languages is given, while language specific properties are maintained individually. The Domain-Ontology was created for use in information retrieval settings in order to obtain specific concepts (only implemented exemplary for the computer terminology). Figure 2 gives an overview over the architecture of the EuroWordNet whereby the single components and its relations are represented among one another.

3 Word Sense Disambiguation in Information Retrieval

User studies have shown that categorized information can improve the retrieval performance for a user. Thus, interfaces providing category information are more effective than pure list interfaces for presenting and browsing information [2]. The authors of [2] evaluated the effectiveness of different interfaces for organizing search results. Users strongly preferred interfaces that provide categorized information and were 50% faster in finding information organized into categories. Similar results based on categories used by Yahoo were presented in [7].

The tool which we present in this paper, was developed as part of our work research towards a (multilingual) retrieval system that classifies documents with respect to the search terms in unambiguous classes, so-called Sense Folders. The main idea of our approach is to provide additional disambiguating information to the documents of a result set retrieved from a search engine in order to enable to restructure or filter the retrieved document result set. The use of web documents implies an on line categorization approach of the documents given the query terms provided from the user. Thus, we can support the user in choosing the relevant information by categorizing the documents using different classification techniques. In the system presented in [8, 10], we use user and query specific information in order to annotate - and thus categorize - search results from other search engines or text archives connected to the meta search engine by web services. The system currently supports methods to group documents based on semantic disambiguation of query terms using an ontology that can be selected by the user. The system analyzes every search term and extracts the belonging SynSets, that are, the sets defining the different meanings of a term and the linguistic relations from the used ontology. Based on these terms, prototypical word vectors of the disambiguating classes (”Sense Folders” [8]) are constructed. Every document is assigned to its nearest prototype (computed by using the cosine similarity) and afterward this classification is revised by a clustering process.

Agreeing with [16] we see one document having one sense per collocation and discourse. But differentiating us from [16], we do not want to learn and disambiguate word senses from untagged corpora.

The idea of this approach is to use ontologies in order to disambiguate query terms used in the retrieved documents [9]. Thus it is possible to categorize documents with respect to the meaning of a search term, i.e. each document is assigned to the best matching meaning (”Sense Folder”) of the search terms used in it. Obviously, only one sense per document can be distinguished in this setting, which is, however, appropriate for many typical retrieval problems where only short documents are considered as, for example, in Web Search.

For this annotation process we currently use WordNet (resp. EuroWordNet). However, if we analyze it, different problems have to be resolved. Very often meanings are distinguished that are semantically very close. For example, searching for the term ”bank” in an information retrieval environment, the user usually wants to know if the retrieved documents belong to the meaning ”bank” in the sense of ”furniture” or in the sense of ”banking”. The fine grained linguistic differentiation between the ”depository bank” meaning and the ”building bank” one is very often not so significant in order to select a relevant document.

This problem of too fine grained description of meanings in WordNet makes on the one hand the automatic categorization very difficult and on the other hand burdens the users with a much too detailed specialization. Therefore, we propose a simple pruning strategy in order to obtain a reduced set of (more expressive) concepts for the categorization approach (see Sect. 3.2). Furthermore, we describe in the following some further problems that should be tackled for a better expressiveness of WordNet.

3.1 Problems of the EuroWordNet Hierarchy

In the following we briefly examine the main semantic limitations of WordNet and describe some problems that have to be solved for its better expressiveness (see also [6, 5, 13]).

Some lexical links of WordNet should be interpreted using formal semantics in order to express ”things in the world”. The authors of [13] revise the Top Level of WordNet (upper or general level) where the criteria of identity and unity are very general, in order to recognize the constraint violations occurring in it. The concepts of identity and unity are described in [13].

However, we analyze the expressiveness of every SynSet in order to better categorize the context for clustering purposes. It means that we merge categories that are in the same domain and that are not much different from another. This decision is based on our need of few unique classes that are carrier of an expressive meaning for a user as well as for an improved clustering performance.

An example is given in [10]. If we retrieve a word from WordNet, several meanings are assigned to the domain ”Factotum” that could be described as the class ”other domain, generic”. The reason for this assignment is simply the problem that the WordNet authors have to assign a domain to each SynSet. If a term can not be categorized (by the author) to a more specific domain, the generic domain ”Factotum” is used. Therefore, if we want to categorize documents with WordNet senses, we have to choose which senses are relevant and which are not, in order to obtain appropriate disambiguation results.

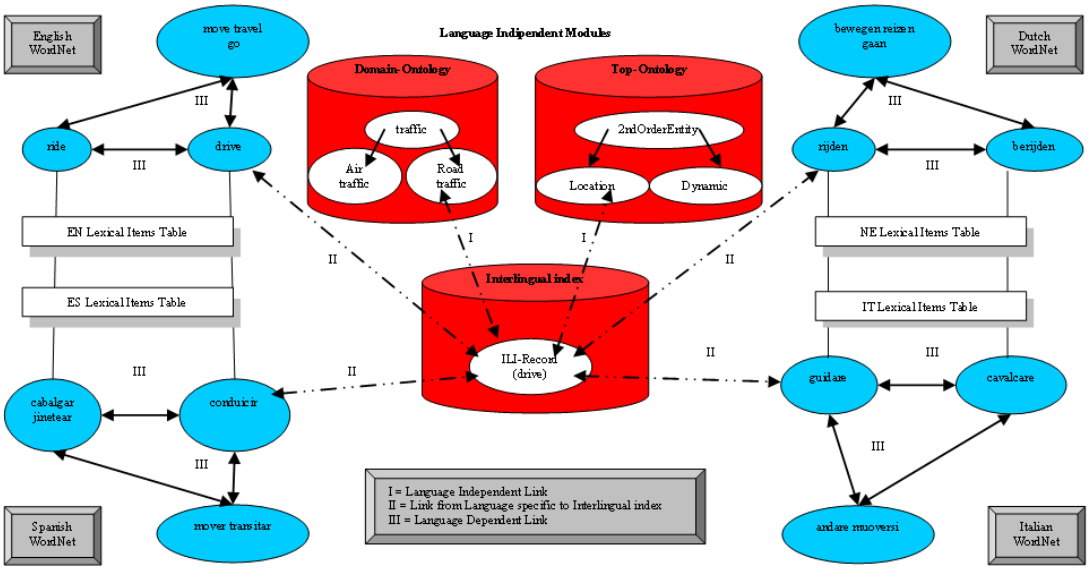

Figure 2. EuroWordNet Architecture (see [15]).

However, if we maintain all senses that are labelled with "Factotum", we have in many cases to distinguish between only slightly different contexts defined by different SynSets. One possibility to derive terms that have a very similar meaning is to analyze their hyponyms or hypernyms. If there are two senses described in WordNet belonging to the same domain, they often have the same hyponyms or hypernym. This frequently causes disambiguation problems that can not be solved if we keep all classes. For this reason, we decided to exclude some irrelevant (for the context disambiguation process) "Factotum" SynSets.

Another critical point is given by the confusion between concepts and instances resulting in an "expressivity lack" [5]. For example, if we look for the hyponyms of "mountain" in WordNet, we will find the "Olympus mount" as a subsumed concept of the word treated as "volcano" and not as instance of it. Thus, we do not have a clear differentiation between what we use to describe (concepts) and their instantiation (instances). We also have the problem that we can not use only concepts or only instances because there is no intended separation between them in WordNet.

The authors of [12] treat also the important difference between endurance and perdurance of the entities that should be included in WordNet. Enduring and perduring entities are related to their behaviour in time. Endurants are always wholly present at any time they are present. Perdurants are only partially present, in the sense that some of their proper parts (e.g., their previous phases) may be not present. However, these aspects of instances are not discussed in this paper since they seem to be of less importance for the considered disambiguation problem.

When we deal with EuroWordNet, these problems persist, and other problems come along. The problem of automatically finding multilingual translation of word senses over languages can be solved using such a resource. The use of the Inter-Lingual-Index helps for this purpose, but the coverage of language-dependent word senses varies from language to language. The number of Synsets varies from an amount of 20.000 (german) to 150.000 (english) Synsets. Using this lexical resource, we have to take into account the missing (or incomplete) translations contained in the lexical resource, apart from the lexical gaps (word senses that exist in a language and not in another).

3.2 Merging the EuroWordNet SynSets

One possible way to tackle some of the problems described above is to merge SynSets manually, when the author means that they belong together. Another possibility is to use methods that restructure EuroWordNet by merging SynSets that have a very similar meaning. Therefore, we studied methods in order to automatically merge SynSets based on the analysis of the linguistic relations defined in EuroWordNet.

We implemented four online methods to merge SynSets based on relations like hypernyms and hyponyms, and further context information like glosses and domain. The first merging approach is based on context information extracted from the hypernymy relation (superordinate words) in order to define the Sense Folders. It means that we first build word vectors for every word sense (Sense Folder), containing the whole hypernymy hierarchy related to the query word. Then we compare all Sense Folders with one another and merge them when the similarity exceeds a given threshold (i.e., when their word vectors are sufficiently close to each other). A similar approach is applied for the hyponyms (subordinate words). In the third approach we merge the Sense Folders if their linguistic relations and context information (glosses) are similar. The fourth approach exploits the domain concept of MultiWordNet [3]. Here we merge the Sense Folders only if they belong to the same domain (having exactly the same domain description).

An evaluation of this methods was done on a small corpus of 252 documents retrieved from web searches that had been manually annotated. Hereby, we compared the manual annotated classes with the Sense Folders assigned using the approach described in [8] together with the merging functions implemented. Based on this first evaluation, the hypernym approach seemed to nicely merge Sense Folders that had similar hypernyms which even might be labeled with different domain descriptions. However, a better classification was

obtained for words that had fewer meanings (SynSets) before merging starts. The second approach based on hyponyms almost never merged SynSets due to the usually very different hyponyms assigned to each sense. Using the third approach, a lattice was built between the merged Sense Folders. This approach merges SynSets not having the same hypernyms, but similar words given from the descriptions of all relations and words together. With the fourth approach we are sure to merge Sense Folders that belong to the same context, describing it in a different way. The classification was always the best, but the Factotum problem as discussed in Sect. 3.1 persisted. If this merged class contains very different meanings and is used for classification, this classification is worse than before. The possibility to exclude such classes (labeled with the "Factotum" domain) will be studied in future work, e.g. by analyzing approaches that exploits combined information from the first three merging methods. For details of the evaluation see [11].

4 The lexical restructuring tool (LexiRes)

The main idea of this tool is to give authors the possibility to navigate the ontology hierarchy in order to restructure it, by manual merging or using the merging functions described in Section 3.2.

4.1 Related Work

Different work has been already done using the variants of WordNet. The authors of [1] developed VisDic for browsing and editing multilingual information taken from EuroWordNet. Here users can browse static information on text blocks.

Another web interface for multilingual information browsing is presented in [14]. Here a parallel corpus annotated with MultiWordNet [3] can be browsed as well as the words with their related annotated word senses, but the corpus is very restricted. All accessible information is static. This interface is used only for a bilingual search in a closed domain.

Other work dealing with the lexicography has shown that researchers in this area mostly deal with multilingual lexical resources or corpora only, without the possibility of merging similar word senses.

Given that the EuroWordNet format is defined by the EuroWordNet Database Editor Polaris that uses a proprietary specification, we first converted the EuroWordNet Database in an XML format, in order to access it with standard XML query tools. In order to retrieve information from this resource, we use the Exist Open Source native XML database.

4.2 The tool

In order to use the LexiRes tool, we have to load an ontology into its scratch framework. The tool currently supports the EuroWordNet structure, but can easily be extended for different ontologies. Considering that we use a multilingual lexical resource, we give the possibility to define the language we want to work with and the linguistic relations we want to show for recognizing the query word in the context menu. After having set it the hierarchy will be displayed.

Figure 3 shows a screenshot of the LexiRes editor. On the left side, we can enter the query words. On the right side, we can choose which collection we want to retrieve and which language we want to use as a source language. Looking for "bank", in the english language, the ontology engine retrieves 19 meanings. These meanings are describing the different word senses. Every word sense is represented as a SynSet. We can apply different actions for these SynSets. Some meanings that belong to the same domain, as the two "bank" - SynSets under the superordinate "incline" SynSet could be merged. If authors decide that the description of these SynSets is too fine grained, they can choose to merge the "source" SynSets to a "target". The goal is to obtain only word senses describing contexts as unambiguous as possible. Based on the merging a new SynSet is created to which all relations of the original SynSets are assigned. Authors can also decide that a SynSet should not be a carrier of meaning for the intended application of the ontology; this SynSet can be removed just clicking on it and choosing to remove it.

The linguistic relations as also the properties of every SynSet can be shown just picking the corresponding fields. These can be first set within the check boxes under the "show relations" area. If the author activates the check boxes, the linguistic relations related to the selected SynSet will be shown. The author can choose to "show properties" or "hide properties" with a right mouse click on a SynSet. Here all SynSet-related information is shown. The original XML code part of the SynSet can also be chosen clicking on the right mouse button and choosing the "show XML" option. The properties and the XML code are shown on the right side down of the interface under "Details".

The SynSets can be also automatically retrieved and translated in the different languages available in the ontology (see Figure 4). These can be set within the menu button language and can be shown, always SynSet-dependent within a click. We can notice that not all SynSet have a translation, due to the missing entries in the lexical resource.

As we said before, the tool gives the possibility to manually merge SynSets, when the authors decide that two SynSets belong to the same meaning and/or describe the same concept. The author working with LexiRes can also use an automatically created list of candidate SynSets that can be merged. This list can be created with the approaches discussed in 3.2. The system proposes the list of changes and the user can select to accept all or check each proposal for merging manually. At the moment these merging methods are implemented outside the tool. The resulting list of possible merging SynSets is first examinated from the authors and then done manually. After having restructured the ontology hierarchy, a new set of SynSets is created. This set is supposed to contain only word senses that are carrier of a distinctive meaning in the context of the considered application. This is a very important step for a use of lexical resources in information retrieval. The possibility to merge SynSets in advance gives the advantage to categorize the retrieved documents disambiguating them with structured word senses that facilitate an automatic classification process [8]. A detailed description of the evaluation of the automatic merging methods applied to the WordNet SynSets in given in [11].

5 Conclusions

In this paper we motivated and presented LexiRes, a tool to help lexicographers in exploring available lexical resources for navigating and restructuring them, especially for use in information retrieval frameworks. Furthermore, we have discussed how lexical resources, here EuroWordNet, can be used in order to disambiguate documents (retrieved from the web within an information retrieval system) given different meanings (retrieved from lexical resources). After having discussed the problems related to the EuroWordNet structure, we presented the functionality of our tool. Using LexiRes we obtain a hier-

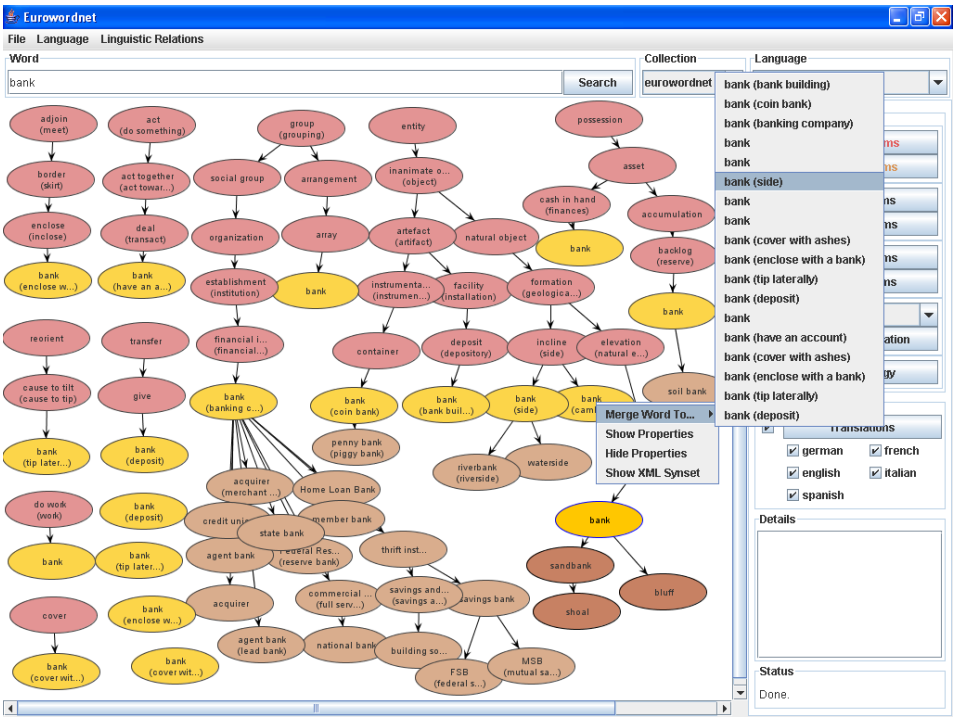

Figure 3. Example of the word ”bank” - manual merging functions - in the LexiRes Editor.

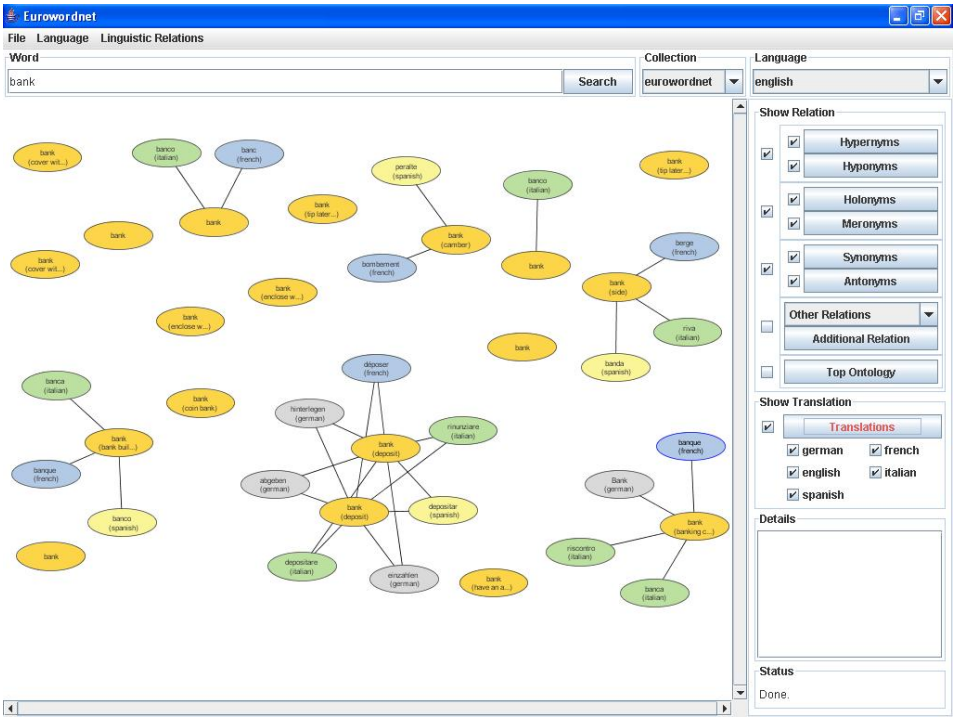

Figure 4. Example of the word ”bank” - SynSet translations - in the LexiRes Editor.

archical word specific overview that gives the possibility to restructure concepts using automatic or manual merging methods. These methods are important to obtain a lexical resource that is more appropriate in order to disambiguate user query words in documents retrieved from an information retrieval system.

REFERENCES

[1] Hork A. and Smr P., 'Visdic - wordnet browsing and editing tool.', in *Proceedings of the Second International WordNet Conference (GWC2004)*, (2004).

[2] Susan T. Dumais, Edward Cutrell, and Hao Chen, 'Optimizing search by showing results in context', in *CHI*, pp. 277–284, (2001).

[3] L. Bentivogli E. Pianta and C. Girardi., 'Multiwordnet: developing an aligned multilingual database.', in *First International Conference on Global WordNet*, Mysore, India, (2002).

[4] C. Fellbaum D. Gross G. Miller, R. Beckwith and K. Miller., 'Five papers on wordnet.', *International Journal of Lexicology*, **3(4)**, (1990).

[5] Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari, 'Conceptual analysis of lexical taxonomies: the case of wordnet top-level', in *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pp. 285–296, New York, NY, USA, (2001). ACM Press.

[6] N. Guarino and C. A. Welty., *An overview of OntoClean.*, 151–172, Handbook on Ontologies, Springer, 2004.

[7] Yannis Labrou and Timothy W. Finin, 'Yahoo! as an ontology: Using yahoo! categories to describe documents', in *CIKM*, pp. 180–187, (1999).

[8] Ernesto William De Luca and Andreas Nürnberger, 'Improving ontology-based sense folder classification of document collections with clustering methods', in *Proc. of 2nd Int. Workshop on Adaptive Multimedia Retrieval (AMR 2004), part of ECAI 2004*, eds., Philippe Joly, Marcin Detyniecki, and Andreas Nürnberger, (2004).

[9] Ernesto William De Luca and Andreas Nürnberger, 'Ontology-based semantic online classification of documents: Supporting users in searching the web', in *Proc. of the European Symposium on Intelligent Technologies (EUNITE 2004)*, (2004).

[10] Ernesto William De Luca and Andreas Nürnberger, 'Supporting mobile web search by ontology-based categorization', in *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen, Proc. of GLDV 2005*, eds., Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner, pp. 28–41, (2005).

[11] Ernesto William De Luca and Andreas Nürnberger, 'The use of lexical resources for sense folder disambiguation.', in *Workshop Lexical Semantic Resources (DGfS-06)*, Bielefeld, Germany., (2006).

[12] E. Motta, S. Buckingham, and J. Domingue. Ontology-driven document enrichment: Principles and case studies, 1999.

[13] A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo. Restructuring wordnet's top-level: The ontoclean approach.

[14] Pianta E. Ranieri M. and Bentivogli L., 'Browsing multilingual information with the multisemcor web interface', in *Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora*, pp. 38–41, Portugal, (2004).

[15] P. Vossen. Eurowordnet general document.

[16] David Yarowsky, 'Unsupervised word sense disambiguation rivaling supervised methods', in *Meeting of the Association for Computational Linguistics*, pp. 189–196, (1995).