Rebuilding Lexical Resources for Information Retrieval using Sense Folder Detection and Merging Methods

Ernesto William De Luca and Andreas Nürnberger

Otto-von-Guericke University of Magdeburg, Department of Computer Science
Universitätplatz 2, 39106 Magdeburg
{deluca, nuernb}@iws.cs.uni-magdeburg.de

Abstract

In this paper we discuss the problem of sense disambiguation using lexical resources like ontologies or thesauri with a focus on the application of sense detection and merging methods in information retrieval systems. For an information retrieval task it is important to detect the meaning of a query word for retrieving the related relevant documents. In order to recognize the meaning of a search word, lexical resources, like WordNet, can be used for word sense disambiguation. But, analyzing the WordNet structure, we see that this ontology is fraught with different problems. The too fine grained distinction between word senses, for example, is unfavorable for a usage in information retrieval. We describe related problems and present four implemented online methods to merge SynSets based on relations like hypernyms and hyponyms, and further context information like glosses and domain. Afterwards we show a first evaluation of our approach, compare the different merging methods and discuss briefly future work.

1. Introduction

Polysemy (from Greek: *many meanings*) is the capacity for a word to have multiple meanings. If we consider the classical example given from the word bank, we have different meanings (bank of a river, bank to draw money, etc.) that can be recognized only in their related context. Humans often use polysemous words for searching for documents. However, an automatic determination of the related word senses is difficult (Mihalcea & Moldovan, 2001). Even though, several approaches have already been proposed, the disambiguation performance is still insufficient, especially for use in information retrieval. In the following, we discuss some fundamental problems of disambiguation methods based on lexical resources. Furthermore, we present an approach that uses automatically restructured linguistic knowledge from lexical resources in order to improve the disambiguation performance.

2. Lexical Resources

Lexical resources provide linguistic information about words of natural languages. This information can be represented in very diverse data structures, from simple lists to complex resources with many types of linguistic information and relations associated with the entries stored in the resource. These resources are used for preparing, processing and managing linguistic information and knowledge needed for the computational processing of natural language (Peters, 2001). An example of such large scale lexical resources is given by linguistic ontologies that cover many words of a language and have a hierarchical structure based on the relations between concepts. These ontologies can cover specific or general domains.

2.1. WordNet

Fellbaum (1998) discussed the design of the electronic lexical database WordNet – one of the most important general ontologies – and the theoretical motivations behind it. WordNet can be used for different applications, like word sense identification, information retrieval, and particularly for a variety of content-based tasks, such as semantic query expansion or conceptual indexing in order to improve the retrieval performance (Vintar et al. 2003).

2.2. Restructuring Lexical Resources

However, analyzing the WordNet structure, we see that this resource is fraught with different problems. Some of the semantic limitations of WordNet have been discussed in more detail in (Mihalcea & Moldovan 2001) and (Oltramari et al. 2002), where revisions were proposed. One major problem of WordNet is that usually too fine grained meanings are provided by the designers of this resource as distinct concepts, which are in almost all cases not relevant for a search process. The user is usually interested only in a few unique classes that are carrier of meaning and have clear distinctive features such as a category or a domain. Recently, this problem was also studied by Cooper (2005) who analyzed the number of senses listed in different monolingual dictionaries. Thus, especially for information retrieval purposes a redesign of the WordNet structure is recommended. One possible method of redesign is given by merging senses to more general concepts. Furthermore, very often meanings are distinguished that are semantically very close. This makes on the one hand the use of such resources for automatic categorization very difficult and on the other hand burdens the users with a much too detailed specialization. Therefore, we propose different pruning strategies in order to obtain a reduced set of (more expressive) concepts for a word sense categorization approach (see section 3.2).

Further problems have to be solved in order to obtain a better expressiveness of WordNet. Oltramari et al. (2002) revised 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 more detail in (Oltramari et al., 2002). Compared to this top level approach, we do not only revise the top level of WordNet but we analyze the expressiveness of every single SynSet related to the user query words.

Another critical point is given by the confusion between concepts and instances resulting in an "expressivity lack" (Gangemi et al. 2001). 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.

Another approach to solve some of the problems discussed above is to use the WordNet domains. These semantic domains define areas of human discussion, such as politics, economy, sport, which exhibit their own terminology and lexical coherence. They can be used in order to describe texts according to general subjects (topics) characterized by domain specific lexicon. In WordNet different SynSets belong to the same domain, so that we can merge them using the domain, in order to better summarize a context. We can use domains in order to better categorize the context for clustering purposes.

Rule - #Word Sense, SynSets (Domain)
#0 rule,ruler (Metrology)
#1 convention, normal pattern, rule, formula (Sociology)
#2 rule, regulation (Factotum – behavior)
#3 rule, formula (Mathematics)
#4 principle, rule (Factotum – rule, law)
#5 principle, rule (Factotum – generalization)
#6 rule (Factotum – religion)
#7 rule, prescript (Factotum – guide)
#8 rule (Factotum – game, sport)
#9 rule, linguistic_rule (Linguistics)
#10 dominion, rule (Factotum – legal authority)
#11 rule (History Time_Period)

Table 1. Rule example for disambiguation.

One of the main problems related to the WordNet domains is represented by the "Factotum" domain. If we retrieve the word "rule" from WordNet (Table 1), we get 12 different meanings of this term. 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.

A typical example of this problem is given for the mentioned word "rule", where the SynSets #6, #7, #10 are labeled with the "Factotum" entry and are not very expressive, but too general to be joined. This frequently causes disambiguation problems that can not be solved if we keep all classes. If we stay with the 'rule' example, we can also notice that there are two categories that are labeled with the same words "principle rule" in the domain "Factotum". There are also two other categories that are not really relevant for disambiguating the senses, as for example "rule linguistic rule" and "rule regulation" that should be considered as instances of "rule", related to the problem of expressivity lack.

Another example is given in Table 2, where the word senses of "bank" are listed. Here we can see, differently to the "rule" example, that the SynSets #0 and #3 (Banking), #2 and #5 (Money), #4 and #6 (Factotum) belong pair wise to the same domain and could be merged in an easier way. A problem of using only domain information for merging is that also SynSets with different meanings (for example for the "Factotum" domain) in the same domain are merged.

Bank - Word Sense, SynSets (Domain)
#0, bank, depository, financial institution (Banking)
#1, bank (Geography, Geology)
#2, bank (Money)
#3, bank, bank building (Banking)
#4, bank (Factotum - arrangement of similar objects)
#5, savings bank, coin bank, money box (Money)
#6, bank (Factotum - a long ridge or pile)
#7, bank (Economy - games)
#8, bank, cant, camber (Town_Planning, Transport)
#9, bank (Aeronautic)

Table 2. Bank example for disambiguation.

3. Lexical Resources in Information Retrieval

After having discussed the problems related to WordNet, we describe in the following the use of lexical resources in an information retrieval environment. As we said before, we need to recognize the linguistic context for disambiguating the search word. This context can be modelled in two ways (Ide & Véronis, 1998):

1. Bag of words (as in some window surrounding the searched word, as in a bag).
2. Relational information (including information about distance from searched word, syntactic relations, semantic categories, etc.).

For our purpose, we apply a combination of both approaches, retrieving information from lexical resources, like machine readable dictionaries (MRDs), thesauri, ontologies or computational lexicons for the identification of meaning of polysemous word in a word sense disambiguation (WSD) task. The use of linguistic knowledge for this task is described as knowledge-driven WSD approach (Ide & Véronis, 1998). In order to obtain a linguistic context description of different word senses we have to explore these lexical resources using the word we are looking for, selecting the concepts based on the linguistic relations that define the different word senses and their linguistic context.

3.1. Disambiguating Word Senses

In order to disambiguate the word senses of words, we implemented an approach to classify documents in Sense Folders. Currently we use the WordNet SynSets and their linguistic relations in order to create prototype vectors that define the Sense Folders for the different meanings of the query terms. These context descriptions are used then in order to categorize and annotate retrieved documents with their best matching Sense Folder. Every document is compared with all Sense Folders and then first assigned to its most similar Sense Folder. Afterwards this classification is revised by the clustering process.

In the retrieval interface we have implemented (De Luca and Nürnberger, 2005) labels defining the disambiguating classes are then added to each document of the result set. The visualization of such additional information (see Figure 1) should enable a simple navigation through the huge number of documents and if possible should restrict information only to the relevant query-related results. With this approach, we help the user in recognizing the meaning of a given search term within

the retrieved documents for a quicker access to the relevant results. More details about this approach can be found in (De Luca and Nürnberger 2004; 2005).

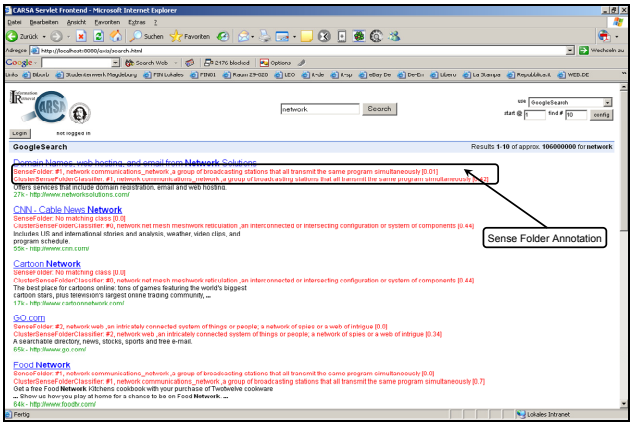

Figure 1 Sense Folder Annotation

3.2. Merging Word Senses

First of all, 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. The too fine grained distinction of word senses complicates, as described in Section 2.2 categorization in this case. Therefore, a revision of the WordNet structure is required. One way to obtain a higher granularity is to merge SynSets if they describe a very similar meaning of the same word (see also De Luca and Nürnberger, 2006). For web search, such methods could be used for creating a reduced structure of the ontology hierarchy, having fewer word senses that are carrier of a more distinctive meaning in order to categorize the documents retrieved (De Luca and Nürnberger, 2004).

However, if we maintain all senses that are labeled with "Factotum", we have in many cases to distinguish between only slightly different contexts defined by different SynSets. For this reason, we have to find methods to exclude (or merge) irrelevant (for the context disambiguation process) "Factotum" 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 hypernyms. Another possibility is to consider information about the linguistic context of the word sense, thus considering its correlated words.

3.2.1. Merging methods

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). The threshold value for this first evaluation was set to 0.5. 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 (Pianta et al. 2002). Here we merge the Sense Folders only if they belong to the same domain (having exactly the same domain description). With these merging methods we build automatically, for every user query, a new slim subset of the WordNet ontology that is then used for the classification process. Doing the subset creation automatically, the WordNet SynSets are not modified in their original structure. This revised subset contains only the relevant meanings related to the user query terms, according to the merging functions used. The user can disambiguate at this point the retrieved documents selecting the annotated documents he was looking for. Furthermore, the automatic online disambiguation allows to adapt the restructuring methods dynamically to the current user and result set by simply modifying the used threshold values and thus modifying the granularity of sense distinction. For example, if only results of a merged subset (Sense Folder) are included in a result set due to very specific user query terms, we can dynamically decide to use a more fine grained structure in order to provide the user with more specific meanings.

3.2.2. Evaluation

Table 3 shows a first evaluation of the merging methods where the documents are classified with the "pure" Sense Folder classification approach (SF) and with the clustering process (CL) presented in more detail in (De Luca and Nürnberger, 2004). The evaluation of our approach is based on a small corpus of 252 documents retrieved from web searches that had been manually annotated. We compared the manual annotated classes with the Sense Folders assigned using the approach described in (De Luca & Nürnberger 2004) together with the merging functions described above. Moreover, we show the improvements of the categorization process in an information retrieval application (De Luca & Nürnberger 2005) compared to the experiments described in (De Luca & Nürnberger 2004). The results show that the performance of the disambiguation process almost always improves when the merging methods are applied. The most significant improvement is obtained by the domain merging method.

The top level problem described in (Oltramari et al., 2002) is solved in our case using the domain merging method. Referring to our discussion above, we can notice that merging SynSets belonging to the same domain shows mostly a better categorization. The "bank" example shows that a great improvement of the categorization performance is done using clustering and merging methods together.

However, the SynSets labeled with the "Factotum" domain affect strongly the classification performance. The "rule" example demonstrate that already the categorization performance without any merging methods is poor, because of the too fine grained word sense distinction that cannot be used for an information retrieval purpose. The SynSets labeled with the 'Factotum' entry are not very expressive, but too general to be joined.

word (classification)	contexts	domains	Hyperonym	hyponym	Without merging
argument (SF)	0.20	**0.47**	0.20	0.20	0.20
argument (CL)	0.16	0.18	**0.20**	**0.20**	**0.20**
bank (SF)	**0.44**	0.38	0.42	0.42	0.42
bank (CL)	0.31	**0.55**	0.29	0.31	0.31
chair (SF)	**0.61**	**0.61**	0.37	**0.61**	**0.61**
chair (CL)	**0.77**	**0.77**	0.71	**0.77**	**0.77**
network (SF)	0.42	0.42	**0.57**	0.42	0.42
network (CL)	0.42	0.42	**0.57**	0.42	0.42
rule (SF)	0.10	0.09	0.09	**0.19**	**0.19**
rule (CL)	0.12	**0.26**	0.09	0.19	0.19

Table 3. Evaluation of the combination of classification and merging methods.

A similar result is given by the documents related to the word "argument". Here four of five word senses of this word were labeled (from the WordNet authors) with the "Factotum" domain. Using the domain merging with the "pure" Sense Folder approach, we achieve a better categorization, than using it with the clustering process. Here we reach a worse classification due to the too high similarity of the documents. We can notice that the success of these merging processes strongly depend from the distribution of the word senses for a given word.

The confusion between concepts and instances pointed out in (Gangemi et al., 2001) could be solved with the Sense Folders merging method (context merging method). Here Sense Folders are merged considering all words contained in the descriptions of the word senses. Thus, these descriptions could be words (instances) that are already contained in the word vectors of one Sense Folder and could describe a similar context that is already present in another one.

An interesting result is given by the combination of the classification approaches with the hypernym merging methods for the word "network". There are three word senses of "network". The first related to "Computer Science" domain, the second to the "Telecommunication" domain and the third to the "Factotum" domain, where network is considered in its general word sense (as web).

The first two word senses have similar hypernyms (their similarity exceeds the threshold) and therefore they are merged together. A clearer distinction between concepts is done and the best classification is reached.

After this first evaluation of our approaches, we have to consider other different aspects. Different combinations of all merging functions should be evaluated and hierarchical merging methods (different from the word vectors approach we use) should be assessed. Furthermore, the possibility to use such an approach for the development of a multilingual information retrieval system is already under development. All these factors could improve the performance of our retrieval approach improving once again the results.

4. Conclusions

In this paper we discussed the problem of sense disambiguation using lexical resources with a focus on sense detection and merging methods in information retrieval systems. We analyzed the WordNet structure and proposed different merging methods. Afterwards, we evaluated and compared the implemented merging methods. Even though, the first results are very promising, the disambiguation performance is still insufficient for use in an information retrieval system. Therefore, we are currently studying methods on how to further improve the disambiguation performance finding the best combination of merging methods for solving the related problems described above.

5. References

Cooper, Martin C. (2005). A Mathematical Model of Historical Semantics and the Grouping of Word-Meanings into Concepts. In *Computational Linguistics*, Volume 31, Part 2.

De Luca, E. W. and Nürnberger A. (2004). Improving Ontology-Based Sense Folder Classification of Document Collections with Clustering Methods. In *Proceedings of the 2nd Int. Workshop on Adaptive Multimedia Retrieval (AMR 2004)*, pp 72 - 86.

De Luca, E. W. and Nürnberger, A. (2005). Supporting Mobile Web Search by Ontology-based Categorization. In *Proceedings of GLDV Tagung 2005*, pp. 28 - 41.

De Luca, E. W. and Nürnberger, A. (2006). The Use of Lexical Resources for Sense Folder Disambiguation. In *Workshop Lexical Semantic Resources (DGfS-06)*, Bielefeld, Germany.

Fellbaum, C. (1998). WordNet, an electronical lexical database, *Cambridge, MIT Press*.

Gangemi A., Guarino N. and Oltramari A. (2001): Conceptual analysis of lexical taxonomies: the case of WordNet top-level. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, pp. 285 - 296.

Ide N. and Véronis J. (1998). Word Sense Disambiguation: The State of the Art. In *Computational Linguistics*, Volume 14, Part 1.

Mihalcea, R. and Moldovan D.I. (2001). Automatic generation of a coarse grained WordNet. In *Proceedings of the SIGLEX*, Pittsburgh, USA.

Miller. G. A. (2001). Ambiguous Words. In *Impacts Magazine*. Published on KurzweilAI.net.

Oltramari A., Gangemi A., Guarino N. and Masolo C. (2002). Restructuring WordNet's Top-Level: The OntoClean approach. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*.

Peters W. (2001). Lexical Resources. In *NLP group, Dept. of Computer Science, University of Sheffield*, http://phobos.cs.unibuc.ro/roric/lex_introduction.html.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of 1st International Conference on Global WordNet*.

Vintar, S., Buitelaar, P., Volk, M. (2003). Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. In *Proc. of the ECML/PKDD Work. on Adapt. Text Extraction and Mining*, Croatia.