WordNet Enrichment with Classification Systems

Andrés Montoyo and Manuel Palomar
Research Group of Language Processing and Information Systems
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
{montoyo, mpalomar}@dlsi.ua.es

German Rigau
TALP Research Center
Universitat Politècnica de Catalunya
Pau Gargalló 5
08028 Barcelona, Spain
g.rigau@lsi.upc.es

Abstract

This paper presents a new method to enrich semantically WordNet with categories from general domain classification systems. The method is performed in two consecutive steps. First, a lexical knowledge word sense disambiguation process. Second, a set of rules to select the main concepts as representatives for each category. The method has been applied to label automatically WordNet synsets with Subject Codes from a standard news agencies classification system. Experimental results show than the proposed method achieves more than 95% accuracy selecting the main concepts for each category.

1 Introduction and Motivation

Lexical knowledge resources are essential components of Natural Language Processing (NLP) system, such as Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Question Answering (QA) or Text Summarization (TS).

Many researchers have proposed several techniques for taking advantage of more than one lexical resource, that is, integrating several structured lexical resources from pre-existing sources. For instance, related to our work, (Byrd 1989) presented the integration of several structured lexical knowledge resources derived from monolingual and bilingual Machine Read Dictionaries (MRD) and Thesaurus. (Rigau 1994) presented an automatic method to enrich semantically the Spanish MRD, using a very simple conceptual distance formulas in conjunction with a Spanish/English bilingual dictionary and WordNet. (Yarowsky 1992) proposed a statistical method to enrich dictionary definitions with categories from the Roget's Thesaurus. This method was applied by (Rigau et al. 1998) to label a monolingual Spanish MRD with WordNet Subject Fields. (Chen and Chang 1998) used the Roget's Thesaurus to label LDOCE (Procter 1987). (Magnini and Cavaglia 2000) presented an augmented version of nominal part of WordNet, whose synsets have been semi-automatically annotated with one or more subject field codes.

Information brought by these Domain Labels would be complementary to what is already in WordNet. As these categories are grouping together words relevant for a specific domain, first, a category could include synsets of different syntactic categories: for example ECONOMY may groups together synsets from nouns like <money> and verbs like <sell>. Second, a domain label could also contain senses from different WordNet sub hierarchies: for example, HEALTH can be present in <doctor> and <hospital>. These classification systems are used by libraries to organise their books by subject, or by information providers to organize the World Wide Web.

This paper presents a new method to enrich WordNet with domain labels using a knowledge based Word Sense Disambiguation (WSD) system and a set of knowledge rules to select the main concepts of the sub hierarchies to be labelled. The WSD system used is the Specification Marks method (Montoyo and Palomar 2000). In this work, WordNet has been augmented with the IPTC[1] Subject Codes. The IPTC Subject Reference System has been developed to allow news agencies access to a universal language independent coding system for indicating the subject content of news items. This classification is also the standard classification system in NAMIC[2] project.

Although, this method can also be applied to other classification systems such as Library of Congress Classification (LC[3]), Dewey Decimal Classification (DDC[4]) or Roget's Thesaurus.

These classification systems provide a means of arranging information so that it can be easily located within a library, World Wide Web, newspapers, etc. Materials are usually classified by their category or class. Therefore, the field of human knowledge is divided into major categories, these are divided into subsections, and so on. The classification scheme is structured according to the state of current human knowledge.

On the other hand, several works in the literature (Ide and Véronis 1998) have shown that for many NLP tasks the fine-grained sense distinctions provided by WordNet are not necessary. We propose a way to deal with this problem, describing an automatic method to enrich semantically WordNet with categories or classes from the classification systems using the Specification Marks Method. Categories, such as AGRICULTURE, HEATH, etc, provide a natural way to establish semantic relations among word senses.

The organisation of this paper is as follows: After this introduction, Section 2 describes the WSD technique based on the Specification Marks method. Section 3 is devoted to the rules used to label for each domain their main concepts of WordNet. Section 4, presents the results of the experiments performed and finally, section 5 presents some conclusions and outlines further lines of research.

1 http://www.iptc.org

2 http://namic.itaca.it

3 http://lcweb.loc.gov/catalog

4 http://www.noblenet.org/wakefield/rdewey.htm

2 Specification Marks Method

WSD with Specification Marks is a method for the automatic resolution of lexical ambiguity of groups of words, whose different possible senses are related. The disambiguation is resolved with the use of the WordNet lexical knowledge base (1.6). The method requires the knowledge of how many of the words are grouped around a specification mark, which is similar to a semantic class in the WordNet taxonomy. The word-sense in the sub-hierarchy that contains the greatest number of words for the corresponding specification mark will be chosen for the sense-disambiguating of a noun in a given group of words.

2.1 Algorithm Description

The algorithm with Specification Marks consists basically of the automatic sense disambiguating of nouns that appear within the context of a sentence and whose different possible senses are related. Its context is the group of words that co-occur with it in the sentence and their relationship to the noun to be disambiguated. The input for the WSD algorithm will be the group of words $w=\{w_1, w_2, ..., w_n\}$. Each word w_i is sought in WordNet, each one has an associated set $s_i=\{s_{i1}, s_{i2}, ..., s_{in}\}$ of possible senses. Furthermore, each sense has a set of concepts in the IS-A taxonomy (hypernym/hyponym relations). First, the concept that is common to all the senses of all the words that form the context is sought. We call this concept the Initial Specification Mark (ISM), and if it does not immediately resolve the ambiguity of the word, we descend from one level to another through WordNet´s hierarchy, assigning new Specification Marks. The number of concepts that contain the sub-hierarchy will then be counted for each Specification Mark. The sense that corresponds to the Specification Mark with highest number of words will then be chosen as the sense disambiguation of the noun in question, within its given context.

We should like to point out that after having evaluated the method, we subsequently discovered that it could be improved with a set of heuristics, providing even better results in disambiguation. The set of heuristics are Heuristic of Hypernym, Heuristic of Definition, Heuristic of Common Specification Mark, Heuristic of Gloss Hypernym, Heuristic of

Hyponym and Heuristic of Gloss Hyponym. Detailed explanation of the method can be found in (Montoyo and Palomar 2000), while its application to NLP tasks are addressed in (Palomar *et al*. 2001).

3 Proposal for WordNet Enrichment

3.1 Method

In this section we describe, in detail, the method employed for enriching WordNet 1.6. This process is illustrated in Figure 1.

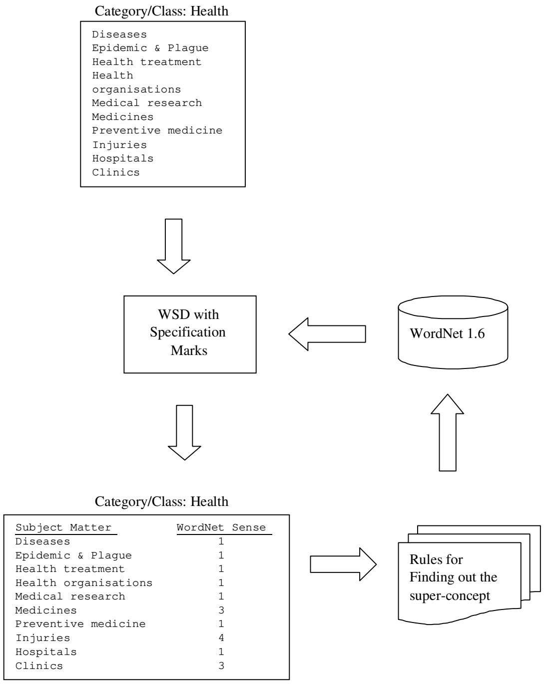
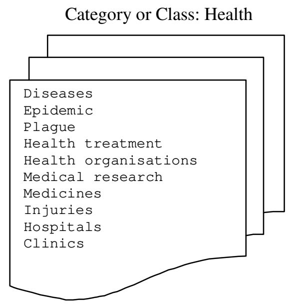


Figure 1 : Process of WordNet enrichment

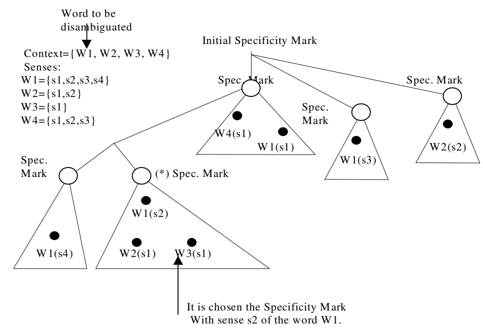
The group of words pertaining to a category, that is, to be disambiguated come from different files of the classification systems. These groups of nouns are the input for the WSD module. This module will consult the WordNet knowledge base for all words that appear in the semantic category, returning all of their possible senses. The disambiguation algorithm will then be applied and a new file will be returned, in which the words have the correct sense as assigned by WordNet. After a new file has been obtained, it will be the input for the rules module. This module will apply a set of rules for finding out a main-concept in WordNet for each category. This main-concept in WordNet is labelled with its corresponding category of the classification system.

The method performs the following steps to enrich and label WordNet.

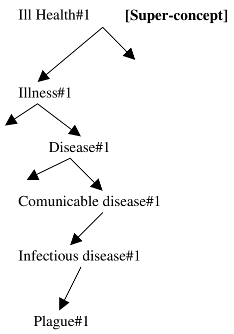
Step 1. Starting with the categories of the classification systems. We would like to clear up any ambiguities at this stage. There are words in the categories that form two words or more. These word combinations of two or more words are not in WordNet, therefore it would be impossible to disambiguate. To resolve this problem we use the utility of WordNet "Find Keywords by Substring" (grep). This substring is a synset in WordNet and relates to the words of the category. (i.e., the substring "Health organization" isn´t in WordNet but finding it with this utility we obtain the substring "Health maintenance organization").



Step 2. To locate the synset or number sense associated with each one of the words of the category, using the Specification Marks Method. For example:



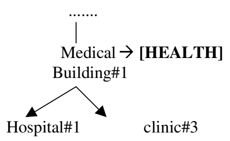
Step 3. To obtain the main-concept from each category, using the hyper/hyponym relationships in the taxonomy of WordNet. For example, the main-concept for disease is *ill_health*:



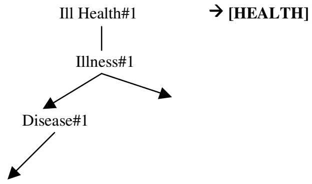
Step 4. To label the main-concept, obtained in WordNet, with the category belonging to the group of words in the classification systems. For example, the main-concept obtained in the step 3 is labeled with *Health*.
Step 5. To propagate down the category following the hyponym relations of the hierarchy of WordNet.

3.2 Main-Concept Rules

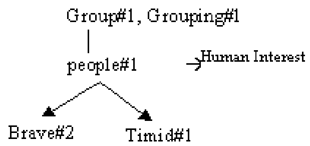
The way to combine the semantic categories of classification systems and WordNet would be to obtain the main-concept of WordNet for each group of words that belong to a semantic category. For obtaining these main-concepts we apply the following set of rules.
Rule 1. If a synset contains only hyponym words belonging to the category for disambiguating, it is chosen as the main-concept. The category is assigned to that main-concept as to full hyponyms and meronyms. For example, the category *Health* is made up of a group of words including *clinic* and *hospital* as it is shown in:



Rule 2. If the synset selected has a hypernym that is made up of the same word as the chosen entry, it is selected as the main-concept. The category is assigned to that main-concept as to full hyponyms and meronyms. For example, the synset *ill_health* is made up of *ill* and *health* and therefore it is a hypernym of *disease#1*, as it is shown in:



Rule 3. This rule resolves the problem of those words that are neither directly related in WordNet nor are in some composed synset of a hyper/hyponym relationships. We search for categories with the hyponym gloss. The hypernym of the word disambiguated is obtained in the taxonomy of WordNet. Then, all of the other words included in the category in some gloss of an immediate hyponym synset of WordNet are checked, and the label of the category is assigned to it. Also, this category label is assigned to all the hyponym and meronym relationships.
Rule 4. When the word to be disambiguated is next to the root level (one or two levels), that is, near the top of the taxonomy, this rule assigns the category only to this synset and at all its hyponyms and meronyms. For example, the category *Human Interest* is assigned to *people#1*, as it is shown in:



4 Experiments & Results

The goal of the experiments is to assess the effectiveness of the proposed method to enrich semantically WordNet 1.6. with categories from IPTC. A brief description of the resources used is included in this section to introduce the reader in the test environment.

4.1 Experiment 1

The first goal was to assess the effectiveness of the disambiguation of the Specification Marks method. It was carried out on random sentences taken from the Semantic Concordance Corpus Semcor (Miller et al., 1993) and Microsoft Encarta Encyclopaedia Deluxe (Encarta), and the IPTC Subject Reference System.
In the first approach we wanted to verify that, the Specification Mark Method can obtain

successful results, and therefore this method can be applied successfully on any corpus. The percentages of correct resolutions achieved with these two corpora were Semcor 67,4% and Encarta 65,9% respectively. We should like to make a clear distinction, however, it does not require any sort of training, no hand-coding of lexical entries, or the hand-tagging of texts.

In the second approach we tested the Specification Mark Method on word clusters related by categories over IPTC Subject Reference System. The percentage of correct resolution was 96.1%. This successful percentage was because the method uses the knowledge of how many of the words in the context are grouped around a semantic class in the WordNet taxonomy.

The experimental results, using general corpus and classification systems, are those shown in the following table 1.

Texts	Corpus Names	Ratio	Values
Unrestricted Text	SEMCOR	Precision	67.4 %
		Recall	66.5 %
		Coverage	98.5 %
	ENCARTA	Precision	65.9 %
		Recall	65.1 %
		Coverage	98.8 %
Classification System	IPTC	Precision	96.1 %
		Recall	92.5 %
		Coverage	96.8 %

Table 1: Main results.

4.2 Experiment 2

Once it has been shown that the WSD Specification Marks Method works well with classification systems, we tested the method of combining the semantic categories of IPTC and WordNet.

Table 2 presents some IPTC categories with the different test sets, computed as the amount of synsets of WordNet correctly labelled, synsets incorrectly labelled and words unlabelled (synsets are not in WordNet).

To evaluate the precision, coverage and recall of the method, we applied the rules of the section 2.2. and we hand checked the results for each word belonging to an IPTC category.

Categories IPTC	Total Number Words IPTC	Correctly Labelled Synsets	Incorrectly Labelled Synsets	Words Unlabelled
Arts, culture & entertainment	23	21	2	0
Crime, law & Justice	8	8	0	0
Disasters & accidents	10	7	2	1
……..	….	….	….	….
……..	….	….	….	….
Weather	6	6	0	0
TOTAL	399	358	16	25

Table 2: Total results of the IPTC categories

Precision is given by the ratio between correctly synsets labelled and total number of answered (correct and incorrect) synsets labelled. Coverage is given by the ratio between total number of answered synsets labelled and total number of words. Recall is given by the ratio between correctly labelled synsets and total number of words. The experimental results are those shown in the table 3.

%	Coverag e	Precision	Recall
WordNet Enrichment	93.7 %	95.7 %	89.8 %

Table 3: Results for WordNet Enrichment

We saw that if the Specification Mark Method disambiguates correctly and the rules of the section 2.2. are applied, the method works successfully. However, if the Specification Mark Method disambiguates incorrectly, the labelling of WordNet with categories of IPTC is also done incorrectly.

5 Conclusion and Further Work

This paper has shown the WSD Specification Marks Method to assign a category of a classification system to a WordNet synset and their descendants and meronyms. We enrich the WordNet taxonomy with categories of the classification system.

The experimental results, when the method is applied to IPTC Subject Reference System, indicate that this may be an accurate and effective method to enrich the WordNet taxonomy.

We have seen in these experiments a number of suggestive indicators. The WSD Specification Marks Method works successfully with classification systems, that is, categories subdivided into groups of words that are semantically related. Although, this method has been tested on IPTC Subject Reference Systems, but can also be applied to other systems that group words about a single category. These systems are Library of Congress Classification (LC), Roget's Thesaurus or Dewey Decimal Classification (DDC).

A relevant consequence of the application of the Method to enrich WordNet is the reduction of the word polysemy (i.e., the number of categories for a word is generally lower than the number of senses for the word). That is, category labels (i.e., HEALTH, SPORTS, etc), provide a way to establish semantic relations among word senses, grouping then into clusters. Therefore, this method intends to resolve the problem of the fined-grainedness of WordNet´s sense distinctions.

Furthermore, now we able to perform variants of WSD systems using domain labels rather than synset labels (Magnini and Strapparava 2000).

Acknowledgements

This research has been partially funded by the UE Commission (NAMIC IST-1999-12302) and the Spanish Research Department (TIC2000-0335-C03-02 and TIC2000-0664-C02-02).

References

Byrd R. (1989). *Discovering Relationship among Word Senses*. In proceedings of the 5th Annual Conference of the UW Centre for the New OED, pages 67-79. Oxford, (England).

Chen J. and Chang J. (1998). *Topical Clustering of MRD Senses Based on Information Retrieval Techniques*. Computational Linguistic **24**(1): 61-95.

Ide N. and Véronis J. (1998). *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art*. Computational Linguistics **24** (1): 1-40.

Magnini B. and Cavaglia G. (2000) *Integrating subject field codes into WordNet*. In Proceedings of the LREC-2000, Athens, Greece.

Magnini B. and Strapparava C. (2000) *Experiments in Word Domain Disambiguation for Parallel Texts*. In Proceedings of the ACL Workshop on Word Senses and Multilinguality, Hong Kong, China.

Miller G., Leacock C., Randee T. and Bunker R. (1993). *A Semantic Concordance*. Proc. 3rd DARPA Workshop on Human Language Tecnology, pages 303-308, Plainsboro, (New Jersey).

Montoyo, A. and Palomar M. (2000). *WSD Algorithm applied to a NLP System*. In Proceedings 5th International Conference on Application of Natural Language to Information Systems (NLDB´2000). Versailles, (France).

Montoyo, A. and Palomar M. (2000). *Word Sense Disambiguation with Specification Marks in Unrestricted Texts*. In Proceedings 11th International Workshop on Database and Expert Systems Applications (DEXA 2000), pages 103-108. Greenwich, London, UK.

Palomar M., Saiz-Noeda M., Muñoz, R., Suárez, A., Martínez-Barco, P., and Montoyo, A. (2000). *PHORA: NLP System for Spanish*. In Proceedings 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). México DF. (México).

Procter P. (1987). *Longman Dictionary of common English*. Longman Group. UK.

Rigau G. (1994). *An Experiment on Automatic Semantic Tagging of Dictionary Senses*. In International Workshop the Future of the Dictionary. Grenoble, (France).

Rigau G., Rodríguez H. and Agirre E. (1998) *Building Accurate Semantic Taxonomies from Monolingual MRDs*. COLING-ACL'98. Montreal, Canada.

Yarowsky D. (1992). *Word Sense disambiguation using statistical models of Roget´s categories trainined on large corpora*. Proc. 14th COLING, 454-460, Nantes, France.