Method for WordNet Enrichment using WSD

Andrés Montoyo[1], Manuel Palomar[1] and German Rigau[2]

[1]Department of Software and Computing Systems, University of Alicante, Alicante, Spain
{montoyo, mpalomar}@dlsi.ua.es
[2]Departament de Llenguatges i Sistemes Informàtics, Universitat Politécnica de Catalunya
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

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.

Byrd in [3], proposes the integration of several structured lexical knowledge resources derived from monolingual and bilingual Machine Read Dictionaries (MRD) and Thesauri. The work reported in [19] used a mapping process between two thesauri and two sides of a bilingual dictionary. Knight in [7], provides a definition match and hierarchical match algorithms for linking WordNet [9] synsets and LDOCE [15] definitions. Knight and Luk in [8], describe the algorithms for merging complementary structured lexical resources from WordNet, LDOCE and a Spanish/English bilingual dictionary. A semiautomatic environment for linking DGILE [2] and LDOCE taxonomies using a bilingual dictionary are described in [1]. A semiautomatic method for associating Japanese entries to an English ontology using a Japanese/English bilingual dictionary is described in [13]. An automatic method to enrich semantically the monolingual Spanish dictionary DGILE, using a Spanish/English bilingual dictionary and WordNet is described in [16]. Several methods for linking Spanish and French words from bilingual dictionaries to WordNet synsets are described in [17]. A mechanism for linking LDOCE and DGILE taxonomies using a Spanish/English bilingual dictionary and the notion of Conceptual Distance between concepts are described in [18]. The work reported in [4] used LDOCE and

Roget's Thesaurus to label LDOCE. A robust approach for linking already existing lexical/semantic hierarchies, in particular WordNet 1.5 onto WordNet 1.6, is described in [5].

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 [11].

The organisation of this paper is as follows: After this introduction, in Section 2 we describe the technique used (Word Sense Disambiguation (WSD) using Specification Marks Method) and its application. In Section 3 we describe the rules used in the method for labelling the noun taxonomy of the WordNet. In section 4, some experiments related to the proposal method are presented, and finally, conclusions and an outline of further lines of research are shown.

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. 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. Detailed explanation of the method can be found in [12], while its application to NLP tasks are addressed in [14].

3 Proposal for WordNet Enrichment

The 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, WordNet presents word senses that are too fine-grained for NLP tasks. We define a way to deal with this problem, describing an automatic method to enrich semantically WordNet 1.6. with categories or classes from the classification systems using the Specification Marks Method. Categories, such as Agriculture, Health, etc, provide a natural way to establish semantic relations among word senses.

3.1 Method

In this section we describe, in detail, the method employed for enriching WordNet 1.6. 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 the super-concept in WordNet. This super-concept in WordNet is labelled with its corresponding category of the classification system. This process is illustrated in Figure 1.

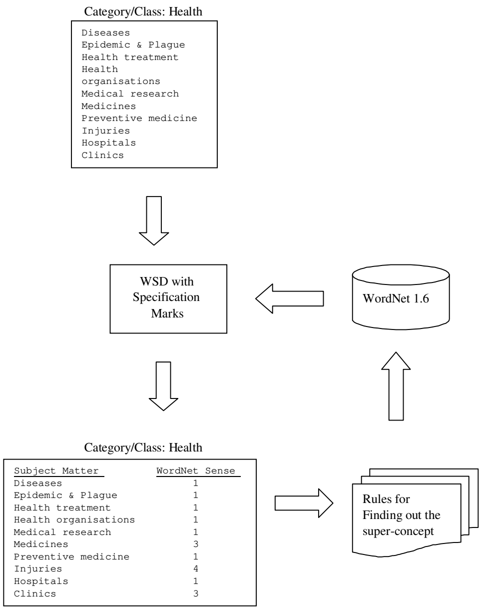

Figure 1 : Process of WordNet enrichment

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.
Step 3. To obtain the super-concept from each category, using the hyper/hyponym relationships in the taxonomy of WordNet. For example, the super-concept for disease is *ill_health*.
Step 4. To label the super-concept, obtained in WordNet, with the category belonging to the group of words in the classification systems. For example, the super-concept obtained in the step 3 is labeled with *Health*.

3.2 Super-Concepts Rules

The way to combine the semantic categories of classification systems and WordNet would be to obtain the super-concept of WordNet for each group of words that belong to a semantic category. For obtaining these super-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 super-concept. The category is assigned to that super-concept as to full hyponyms and meronyms. For example, the category *Health* is made up of a group of words including *clinic* and *hospital*.

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 super-concept. The category is assigned to that super-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*.

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 use the gloss of each synset of the hyponym relationship. 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, that is, in the top of the taxonomy, this rule assigns the category to the synset and at all its hyponyms and meronyms. For example, the category *Health* is assigned to *injury#4*.

4 Discussion

The goal of the experiments is to assess the effectiveness of the proposed method to enrich semantically WordNet 1.6. with categories from IPTC. Table 1 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).

Categories IPTC	Total Number Words IPTC	Correctly Labelled Synsets	Incorrectly Labelled Synsets	Words Unlabelled
Arts, culture & entertainment	23	21	2	0
Disasters & accidents	10	7	2	1
Agriculture	6	5	0	1
Chemical	9	8	0	1
Computing & Technology	10	9	1	0
Construction & property	5	3	1	1
Energy & resource	14	10	0	4
Financial & business services	13	12	1	0
Consumer goods	10	10	0	0
Media	12	12	0	0
Tourism & leisure	7	7	0	0
……….	…..	…..	…..	…..
……….	…..	…..	…..	…..
……….	…..	…..	…..	…..
Health	12	8	3	1
TOTAL	399	358	16	25

Table 1: IPTC categories with the different test sets

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.

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 following table.

%	Coverage	Precision	Recall
WordNet Enrich-ment	93.7 %	95.7 %	89.8 %

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

Several works in the literature [6] 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.

This paper applies the WSD Specification Marks Method to assign a category of a classification system to a WordNet synset as to full hyponyms 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 strongly 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.

Furthermore, now we able to perform variants of WSD systems using domain labels rather than synset labels [10].

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

1. Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., and Samiotou A. 1994. TGE: Tlink Generation Environment. *In proceedings of the 15th International Conference On Computational Linguistic (COLING´94)*. Kyoto, (Japan).
2. Alvar M. 1987. Diccionario General Ilustrado de la Lengua Española VOX. *Bibliograf S.A.*. Barcelona, (Spain).

3. 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).
4. Chen J. and Chang J. 1998. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. *Computational Linguistic* **24**(1): 61-95.
5. Daudé J., Padró L. And Rigau G. 2000. Mapping WordNets Using Structural Information. *In Proceedings 38th Annual Meeting of the Association for Computational Linguistics(ACL00)*. Hong Kong. (Japan).
6. 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.
7. Knight K. 1993. Building a Large Ontology for Machine Translation. *In proceedings of the ARPA Workshop on Human Language Technology*, pages 185-190. Princenton.
8. Knight K. and Luk S. 1994. Building a Large-Scale Knowledge Base for Machine Translation. *In proceedings of the American Association for Artificial Inteligence*.
9. Miller G. A., Beckwith R., Fellbaum C., Gross D., and Miller K. J. 1990. WordNet: An online lexical database. *International Journal of Lexicography* **3**(4): 235-244.
10. 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.
11. 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).
12. 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).
13. Okumura A. and Hovy E. 1994. Building japanese-english dictionary based on ontology for machine translation. *In proceedings of ARPA Workshop on Human Language Technology*, pages 236-241.
14. 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 D.F. (México).
15. Procter P. 1987. Longman Dictionary of common English. *Longman Group*. England.
16. Rigau G. 1994. An Experiment on Automatic Semantic Tagging of Dictionary Senses. *In International Workshop the Future of the Dictionary*. Grenoble, (France).
17. Rigau G. and Agirre E.1995. Disambiguating bilingual nominal entries against WordNet. *Seventh European Summer School in Logic, Language and Information (ESSLLI´95).* Barcelona, (Spain).
18. Rigau G., Rodriguez H., and Turmo J. 1995. Automatically extracting Translation Links using a wide coverage semantic taxonomy. *In proceedings fifteenth International Conference AI´95, Language Engineering´95*. Montpellier, (France).
19. Risk O. 1989. Sense Disambiguation of Word Translations in Bilingual Dictionaries: Trying to Solve The Mapping Problem Automatically. *RC 14666, IBM T.J. Watson Research Center*. Yorktown Heights, (United State of America).