Using WordNet Relations and Semantic Classes in Information Retrieval Tasks

Javi Fernández, Rubén Izquierdo, and José M. Gómez

University of Alicante, Department of Software and Computing Systems
San Vicente del Raspeig Road, 03690 Alicante, Spain
javifm@ua.es, ruben@dlsi.ua.es, jmgomez@ua.es
http://www.dlsi.ua.es

Abstract. In this paper we explore the use of semantic classes in an existing information retrieval system in order to improve its results. Thus, we use two different ontologies of semantic classes (WordNet domain and Basic Level Concepts) in order to re-rank the retrieved documents and obtain better recall and precision. Finally, we implement a new method for weighting the expanded terms taking into account the weights of the original query terms and their relations in WordNet with respect to the new ones (which have demonstrated to improve the results). The evaluation of these approaches was carried out in the CLEF Robust-WSD Task, obtaining an improvement of 1.8% in GMAP for the semantic classes approach and 10% in MAP employing the WordNet term weighting approach.

1 Introduction

The two main goals of the Robust-WSD task are to measure the robustness of the retrieval systems (good stable performance over all queries) and test the benefits of the use of Word Sense Disambiguation (WSD) on this kind of systems. We decided to use an already implemented and evaluated system in last year's edition of CLEF as starting point for our approach. From all the available systems, we have chosen the Universidad Complutense de Madrid system [11], because of its good results, availability and the possibility of adapting the code easily to our objectives. Our main goal consists on experimenting the benefits of the use of *semantic classes* in Information Retrieval (IR) systems. However, we propose, also, a new and flexible way of weighting terms for the query expansion based on *WordNet relations.*

WSD, the task of assigning the correct sense to words depending on the context in which they appear, is a hard task and still a long way from being useful in other natural language processing applications, as shown in recent international evaluations [16,13]. The word senses these systems use are taken from a particular lexical semantic resource (most commonly WordNet [4].) WordNet has been widely criticized because of its too fine–grained sense distinctions, which are not useful for higher level applications like machine translation or question answering, and they are too subtle to be captured by automatic systems with the

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part I, LNCS 6241, pp. 158–165, 2010.

current small volumes of word–sense annotated examples. This can be a reason for the poor results of current WSD systems.

A possible solution is the use of *semantic classes* instead of word senses, because they group together senses of different words. This has several advantages: the average polysemy of texts is decreased, they provide richer and more useful information than word senses, and the amount of training data for each classifier is increased. Izquierdo et al. empirically explored the performnace of different levels of abstraction on the supervised WSD task [7]. These levels were provided by *WordNet Domains* (WND) [9], *SUMO labels* [10], *Lexicographer Files of WordNet* [4] and *Basic Level Concepts* (BLC20) [6]. Izquierdo et al. [7] referred to this approach as class–based WSD since the classifiers were created at a class level instead of at a sense level. As we have said, class–based WSD clusters senses of different words into the same explicit and comprehensive grouping. Only thoses cases belonging to the same semantic class are grouped to train the classifier. For example, the coarser word grouping obtained by Snow et al. [15] only has one remaining sense for "church". Using a set of Base Level Concepts [6], the three senses of "church" are still represented by *faith.n#3*, *building.n#1* and *religious_ceremony.n#1*.

We think that IR could take advantage with the use of word sense disambiguation, but from a semantic class point of view instead of the traditional word sense point of view. As to the data of the robust adhoc IR task has been processed automatically by two WSD systems, and the information of word senses is available, we did not run any class–based WSD system over the data. The next section describes the architecture of our system. In section 3 we discuss the results of this system at the CLEF 2009 Robust-WSD Task. Finally, in section 4 we draw conclusions and outline future works.

2 Description of the System

The system architecture is shown in Figure 1.

The user query is pre-parsed to obtain a set of terms without stopwords and any special symbol. Next, a ranked list of relevant documents are retrieved using the Lucene search engine[1]. With the retrieved documents, the initial query, the relations of the external resource WordNet and state-of-art query expansions methods an expanded query is obtained. The terms of this new query are weighted taking into account the weights of the original query terms, their relations in WordNet with respect to the new ones, the weight assigned by the WSD system to each sense and the weight returned by the expansion method. Once we have a new list of weighted terms, we perform another search but using the expanded query instead of the original one in order to retrieved a new ranked list of documents. Finally, we use the semantic class information from two different semantic resources (WordNet Domains and Base Level Concepts) in order to obtain a re-ranked document list as result.

In the following section we explain each of these processes in more detail.

[1] `http://lucene.apache.org`

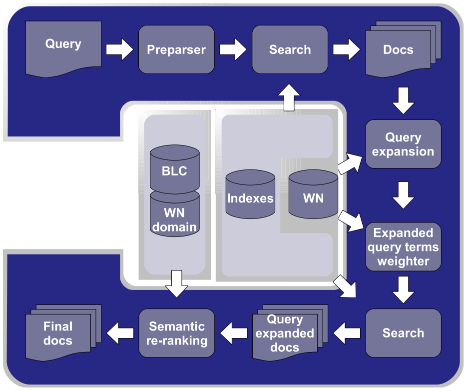

Fig. 1. Architecture of the system

2.1 Search Engine and Query Expansion

The search engine, which we are using, is the one provided by the Universidad Complutense de Madrid [11]. Their implementation is a modified version of Lucene which uses the BM25 probabilistic model [14] for document retrieval. They have also implemented two state-of-art query expansion methods: Kullback-Liebler Divergence [3] (an information-theoretic approach) and the Bo1 model [8,12] (based on Divergence From Randomness [2]). For our system we have chosen the Bo1 model because it is the approach with the best results in their evaluations. We also have decided to use the same constant values than they used in the last CLEF Robust-WSD edition in order to compare the effectiveness of our methods of semantic classes.

As we can see in Figure 1, we make two search processes. For the first retrieval process, query terms are lemmatized and stemmed in order to increase the system recall. The first *Search* module gets these terms as input and returns a list of relevant documents using the BM25 probabilistic model. Next, in the *Query expansion module*, we expand the original query obtaining new terms by means of the Bo1 model.

Although [11] proposed a method for weighting the expanded query terms based on WordNet, we have prefered to use our own method due to the fact that they do not use all senses of each term but the one with the highest weight. In our system we have decided to use all senses retrieved by the WSD system in order to improve the recall. In this way, the system searches all expanded terms in the relations of WordNet with respect to the synonyms, hyperonyms and hyponyms until a certain *distance*. For example, if the distance was 2, we search any expanded term among the hyperonym and hyponym synsets of the original terms but, also, the hyperonyms of the hyperonyms and the hyponyms of the hyponyms. The distance value sets the number of jumps to make in the WordNet relations from the synsets of the query terms. As we have mentioned above, we use all senses supplied for the WSD system for each query term but we take into account the score given by these systems to each sense in order to calculate the weight of the expanded terms. Thefore, this distance factor is calculated by the following equation:

$$weight(synset_{i,d}) = weight(synset_{i,d-1}) * \alpha^d \tag{1}$$

We defined $synset_{i,1}$ as a given WordNet synset and $synset_{i,d}$ as another WordNet synset which is related to the $synset_{i,1}$ of a distance of d jumps (taken into acount only hyperonym and hyponym relations). Thus, $weight(synset_{i,d})$ is the weight of the synset i, d and $weight(synset_{i,1})$ is the score given by the WSD system to the synset i, d. α is a constant whose value is between 0 and 1 and d the distance of $synset_{i,d}$ to the $synset_{i,1}$.

Once we have calculated the previous synset weight, we combine this weight with the weight assigned by the expanded method bo1 in order to calculate the final term weight using the following equation:

$$weight(term_t) = \frac{weight(synset_{i,d}) + weight_0(term_t)}{2} \tag{2}$$

Where $weight(term_t)$ is the weight of the expanded term t which is grouped in the WordNet synset i, d, and $weight_0(term_t)$ is the weight assigned by bo1 to the term t.

With these equations, we give importance to those expanded terms which are related to more likely the original query terms and, in addition, we include the score given by the WSD system for each query term in the final term weight. Thus, we include all senses of a term in the search but we give more importance to those terms which are related to more likely senses and closer to the original query terms.

2.2 Semantic Classes

Our approach consists in mapping the assigned word senses to semantic classes, specifically to WordNet Domains labels and Basic Level Concepts.

WordNet Domains [9] is a hierarchy of 165 Domain Labels which have been used to label all the WordNet synsets. Information brought by Domain Labels is complementary to what is already in WordNet. First of all, a Domain Label can include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as doctor or hospital, and from verbs, such as to operate. Second, a Domain Label may also contains senses from different WordNet subhierarchies. For example, SPORT contains senses such as athlete, deriving from life form, game equipment from physical object, sport from act and playing field from location.

Basic Level Concepts [6] are a set of concepts that result from the compromise between two conflicting principles of characterization: represent as many concepts as possible and represent as many features as possible. As a result of this, Basic Level Concepts typically occur in the middle of hierarchies and less frequently than the maximum number of relations.

The authors developed a method for the automatic selection of BLC from WordNet. They use a very simple method for deriving a small set of appropriate meanings using basic structural properties of WordNet. The approach considers:

- The total number of relations of every synset or just the hyponymy relations.
- Discard those BLCs that do not represent at least a number of synsets.
- Optionally, the frequency of the synsets (summing up the frequency of the senses provided by WordNet).

The process of automatic selection of BLC follows a bottom-up approach using the chain of hypernym relations. For each synset in WN, the process selects as its Base Level Concept the first local maximum according to the relative number of relations. For synsets having multiple hypernyms, the path having the local maximum with higher number of relations is selected. Usually, this process finishes having a number of false Base Level Concepts. That is, synsets having no descendants (or with a very small number) but being the first local maximum according to the number of relations considered. Thus, the process finishes by checking if the number of concepts subsumed by the preliminary list of BLC is higher than a certain threshold. For those BLC not representing enough concepts according to a certain threshold, the process selects the next local maximum following the hypernym hierarchy.

Thus, depending on the type of relations considered to be counted and the threshold established, different sets of BLC can be easily obtained for each WN version. For our work, we have selected the set of BLC built using all kind of relations and a threshold of 20 as the minimum number of synsets that each BLC must subsume.

We explain now the **representation of documents or queries with semantic classes** of the words contained on them. In the task data, each ambiguous word is annotated with its possible senses, each one with a certain probability. Starting from this information, we create a domain vector, for a query or for a document, containing all the semantic classes information of the query or document. The Domain vector consists of a vector where each element represents a WordNet Domain or a Basic Level Concept and its associated weight. Note that there are 165 Domain Labels in WordNet Domains and 558 Basic Level Concepts for nouns. The way to build this vector is: each word has annotated several senses, with the associated probability; each word sense is mapped to its proper semantic class, and the element of the vector corresponding to this Domain is increased with the probability associated with the word sense. When all the words are processed, we obtain a domain vector representing the semantic information of the document or query. Finally to compare two documents, or a document and a query, and obtain their similarity in terms of their semantic content, we use the value of the cosine defined by the two domain vectors.

2.3 Integration of Semantic Classes in Robust Ad Hoc

Once the final list of documents from the expanded query is retrieved, the *Semantic re-ranking* module rearranges this list taking into account both the similarity returned by the BM25 probabilistic model and the similarity calculated by the semantic class system. In order to do this, we have studied several equations described in [5]. In this paper, the best results were the ones obtained with the following equation:

$$semsim(i,j) = \begin{cases} simmax_i + sim_{ij} * sem_{ij} & \text{if } sem_{ij} > h \\ \\ sim_{ij} & \text{otherwise} \end{cases} \quad (3)$$

Where $semsim(i,j)$ is the final similarity between the query i and the document j, sim_{ij} is the similarity of the query i with respect to the document j returned by the search engine, sem_{ij} is the same similarity but returned by the semantic class system, $simmax_i$ is the greatest value of similarity returned by the search engine for the query i and h is a constant which determines a semantic similarity threshold defined empirically. This equation gives more relevance those documents with high semantic similarity but takes into account the semantic class score in the final similarity value.

3 Evaluation

In this section we report the results of each one of our proposals separately.

For the evaluation of the *Expanded query terms weighter*, we have to set the value for two variables: α and d (distance). In order to get the best values for these variables, we have experimented with several different values for them. In table 1 we present two of the best results of those experiments. With $\alpha = 0.8$ and $d = 1$ we improve the baseline GMAP in a 9.97%. With $\alpha = 0.92$ and $d = 6$ we improve both the baseline MAP in a 0.02% and the baseline GMAP in a 8.19%.

For the evaluation of the *Semantic re-ranking*, the only variable is the threshold h for the reranker. We have experimented with different values for this variable in order to obtain the best results. In Table 2 we present the best results

Table 1. Evaluation of the *Expanded query terms weighting* module

	MAP	GMAP	R-Prec	P@5	P@10
BM25 + Bo1 (Baseline)	.3737	.1294	.0.3585	.4475	**.3825**
BM25 + Bo1 + WD ($\alpha = 0.8$, $d = 1$)	.3706	**.1423**	.3624	.4500	.3750
BM25 + Bo1 + WD ($\alpha = 0.92$, $d = 6$)	**.3738**	.1400	**.3655**	**.4513**	.3775

Table 2. Evaluation of the *Semantic re-ranking* module

	MAP	GMAP	R-Prec	P@5	P@10
BM25 + Bo1 (Baseline)	.3737	.1294	.3585	**.4475**	.3825
BM25 + Bo1 + WND + RR ($h = 0.5$)	.3752	.1298	**.3638**	.4462	**.3862**
BM25 + Bo1 + BLC20 + RR ($h = 0.8$)	**.3776**	**.1317**	.3609	.4437	.3806

of those experiments for each semantic classes model. The integration of the semantic classes to the search engine improves the baseline results. With WND we improve both the baseline MAP in a 0.4% and the baseline GMAP in a 0.31%. With BLC20 we improve both the baseline MAP in a 0.64% and the baseline GMAP in a 1.77%.

4 Conclusions

The results of the experiments with our two proposals have shown improvements to the initial information retrieval system.

In the first one, the *Expanded query terms weighter* module, we have experimented with the weights of the terms in a probabilistic IR system. We have applied a smoothing function based on the WordNet distance to the weights given by the IR system. The experiments made have shown GMAP improvements of nearly 10% but not significant MAP improvements.

As future work we propose to continue with the experiments on this module. For the propagation function 2, the search of the best values for α and d can be more exhaustive, finding better values for this variables. Moreover, new relations can be explored in WordNet (not only hyponyms and hyperonyms), in order to improve recall. Even new weight propagation functions can be proposed to better exploit the concept of *distance* in WordNet.

In the second of our proposals, the *Semantic re-ranking* module, we have integrated the semantic classes to a IR system. We have done this integration recalculating the weight of the documents retrieved depending on the similarity between the semantic class of each document and the semantic class of the query. The results of the experiments made reveal that the semantic classes resources can be effectively integrated to the IR systems.

This module can also be applied at new levels. We only have used five simple integration functions for the search engine and the semantic classes weights. More functions can be studied to find the best way to integrate the available resources of semantic classes.

Acknowledgements

This paper has been partially supported by the Spanish government, project TIN-2006-15265-C06-01 and by the framework of the project QALL-ME, which is a 6th Framework Research Programme of the European Union (EU), contract number: FP6-IST-033860 and by the University of Comahue under the project 04/E062.

References

1. Aguirre, E., Di Nunzio, G.M., Mandl, T., Otegi, A.: Clef 2009 ad hoc track overview: Robust-wsd task. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I, Corfou, Greece. LNCS, vol. 6241,pp. 36–49. Springer, Heidelberg (2010)

2. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4), 357–389 (2002)
3. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York (1991)
4. Fellbaum, C. (ed.): WordNet. An Electronic Lexical Database. The MIT Press, Cambridge (1998)
5. Fernández, J., Izquierdo, R., Gómez, J.M.: Alicante at clef 2009 robust-wsd task. In: Working notes of Cross Language Evaluation Forum 2008, Corfou, Greece, CLEF (2009)
6. Izquierdo, R., Suárez, A., Rigau, G.: Exploring the automatic selection of basic level concepts. In: Angelova, G., et al. (eds.) International Conference Recent Advances in Natural Language Processing, Borovets, Bulgaria, pp. 298–302 (2007)
7. Izquierdo, R., Suárez, A., Rigau, G.: An empirical study on class-based word sense disambiguation. In: EACL, pp. 389–397. The Association for Computer Linguistics (2009)
8. Macdonald, C., He, B., Plachouras, V., Ounis, I.: University of glasgow at trec 2005: Experiments in terabyte and enterprise tracks with terrier. In: TREC (2005)
9. Magnini, B., Cavaglià, G.: Integrating subject field codes into wordnet. In: Proceedings of LREC, Athens, Greece (2000)
10. Niles, I., Pease, A.: Towards a standard upper ontology. In: Weltyand, C., Smith, B. (eds.) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001), pp. 17–19 (2001)
11. Peréz Agüera, J.R., Zaragoza, H.: Query clauses and term independence. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 138–145. Springer, Heidelberg (2009)
12. Plachouras, V., He, B., Ounis, I.: University of glasgow at trec 2004: Experiments in web, robust, and terabyte tracks with terrier. In: Voorhees, E.M., Buckland, L.P., Voorhees, E.M., Buckland, L.P. (eds.) TREC, NIST (2004)
13. Pradhan, S., Loper, E., Dligach, D., Palmer, M.: Semeval-2007 task-17: English lexical sample, srl and all words. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, June 2007, pp. 87–92. Association for Computational Linguistics (June 2007)
14. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR 1994: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232–241. Association for Computational Linguistics (1994)
15. Snow, R., Prakash, S., Jurafsky, D., Ng, A.: Learning to merge word senses. In: Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1005–1014 (2007)
16. Snyder, B., Palmer, M.: The english all-words task. In: Mihalcea, R., Edmonds, P. (eds.) Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, pp. 41–43. Association for Computational Linguistics (July 2004)