Information Retrieval by Means of Word Sense Disambiguation

L. Alfonso Ureña[1], J.M. Gómez[2] and M. de Buenaga[2]

[1] Dpto. de Informática, Universidad de Jaén
Avda. Madrid 35, 23071 Jaén, Spain,
laurena@ujaen.es,
[2] Dpto. de Inteligencia Artificial, Universidad Europea - CEES
Villaviciosa de Odón, 28670 Madrid, Spain,
{jmgomez,buenaga}@dinar.uem.es

Abstract. The increasing problem of information overload can be reduced by the improvement of information access tasks like Information Retrieval. Relevance Feedback plays a key role in this task, and is typically based only on the information extracted from documents judged by the user for a given query. We propose to make use of a thesaurus to complement this information to improve RF. This must be done by means of a Word Sense Disambiguation process that correctly identifies the suitable information from the thesaurus WordNet. The results of our experiments show that the utilization of a thesaurus requires Word Sense Disambiguation, and that with this process, Relevance Feedback is substantially improved.

1 Introduction

Information access methods must be improved to overcome the information overload that most professionals face nowadays. Information access tasks, like Information Retrieval (IR), help the users to access a great amount of text they find in the Internet and their organizations. For instance, about 90% of the information in corporations exists in the form of text [8].

In a typical IR environment, users perform a series of iterative searches refining the query after each search. Many IR systems can automatically refine the user query making use of a set of documents that have been shown to the user, and that he or she has marked as relevant or not relevant. This Relevance Feedback (RF) process usually improves the quality of the search results.

The RF process is usually based only on the use of documents retrieved by the system and judged by the user. This information can be complemented with other information extracted from additional resources like *thesauri*. In this work, we explore the utilization of the lexical database WordNet to complement documents' information for RF.

Our experiments demonstrate that the use of WordNet can improve RF results when a Word Sense Disambiguation (WSD) process is applied. The raw data obtained from WordNet must be refined by a sense identification process for them to be effective. For automatically disambiguating query terms, more information than the query itself is required, and provided in the form of the documents judged by the user in a RF setting. We have applied a WSD algorithm based on the integration of resources that enables us to use WordNet for RF.

The results of our experiments show that: first, using WordNet without WSD to complement documents' information decreases RF perfomance; and secondly, using WSD applied to WordNet information substantially improves RF based on the information of documents.

This paper is organized as follows. First we describe the RF approach used in this work. After that, the use of WordNet and WSD for RF is presented. Next, we describe our experiments and results. Finally, the conclusions of our work are shown.

2 Relevance Feedback by Query Expansion

In most retrieval situations, it is customary to conduct searches iteratively as a sequence of partial search operations, which allows the results of earlier searches to be used to improve the output of later searches. One possibility consists of asking the user the relevance of certain documents retrieved in earlier search operations, and using these assessments to construct improved query formulations. This process, known as Relevance Feedback, has been shown experimentally to improve retrieval effectiveness greatly [11].

The most popular model for Information Retrieval is the Vector Space Model (VSM). In this framework, it has been demonstrated that the *Ide dec-hi* RF method was the best performing one [10]. This method is based on expanding the original query with terms from retrieved documents, using all the ones considered relevant by the user, and the first non-relevant document.

When evaluating RF effectiveness, it is preferrable to employ the residual collection method. For a given query, the first n documents retrieved are examined to assess relevance. Then, the query formulation is improved using those documents, and a new search using the new query is performed over the document collection, excluding the first n documents retrieved for the previous query. This way, we can avoid the unfair and unrealistic situation in which already seen documents can be retrieved again, biasing the results to be much better than the original ones.

3 Using Word Sense Disambiguation in Relevance Feedback

The Query Expansion process described earlier is based only on the use of the documents retrieved. Nevertheless, other resources like *thesauri* can be used to complement the information contained in documents [11]. An example of advanced thesaurus is the lexical database WordNet [6], which has been successfully employed in other classifications tasks like Text Categorization [2]. In RF, and for every word sense in a query, a set of synonyms can be extracted from WORDNET and used to complement the documents' information obtained from the previous query.

Unfortunately, the incorporation of many heterogeneous words can impact negatively on the RF perfomance. This is caused by the great number of senses a word can have. We would not like to retrieve documents concerning "river" when we perform a query like "bank" to know about finantial interests. We require a WSD process to identify the correct sense of the words occurring in the query in order to effectively use a thesaurus. This is just an example of a Natural Language Processing task where WSD can improve the performance of the system, among others like Text Categorization [2], Cross-Language Information Retrieval [3], Information Extraction, or Machine Translation [5].

The lexical database WORDNET has been used in IR before, but for it to be useful is essential to apply WSD [14]. In a general IR setting, the information of the query words is not enough to identify word senses, but in RF we can take advantage of the retrieved documents as an additional context to perform an effective WSD process.

In the latest years, many WSD approaches have been proposed, which can be classified according to the knowledge sources they make use of. Some approaches are based on the utilization of some kind of lexicon or lexical database [16]. Other people make use of unlabelled [9] or semantically labelled [1] text corpora as a training collection. Finally, recent works propose the combination of several knowledge sources like lexical databases, text corpora, heuristics, collocations, etc. for WSD [15]. We follow the latter approach, combining the utilization of the lexical database WORDNET, and the semantically-tagged text corpus SEMCOR [7] as described in [13].

The main idea of our WSD method is taking advantage of WORDNET information as an initial base for refining the representation of word senses through a learning algorithm. For instance, the Rocchio and the Widrow-Hoff learning algorithms admit the definition of an initial representation of classes to be learned. A set of WORDNET synonyms are extracted and used as initial representation of each word sense. Then, the learning algorithm is applied taking as input a collection of labeled data (SEMCOR) to refine the initial representation of word senses, obtaining better results in final tagging of new words than the use of any of the resources alone [13]. The senses of new words (the query words) can be identified using the context of documents where they occur, that happen to be the retrieved documents in the previous query.

As a concluding remark, we can observe that the utilization of WSD in other Natural Language Processing tasks like RF is a natural and interesting way of evaluating WSD algorithms. WSD can hardly be considered an independent task, and WSD techniques should be compared in the framework of the task where WSD is being applied. Then, a WSD approach is better than others when the task where it is applied in, is improved by that approach more than by others. This indirect evaluation must complement the direct, standalone evaluation of WSD approaches.

4 Evaluation and Discussion of Results

For the experiments developed in this paper, we have employed the IR system Smart [11]. This system is a well known standard that implements the RF option described erlier in this paper. The number n of documents considered in the RF process is 15. Additionaly, a set of 5.000 documents has been extracted from the Wall Street Journal corpus in the IR test collection TREC, widely used in IR system evaluation [4]. We have also randomly selected a set of 50 queries (topics) among the first hundred, each one with at least one relevant document in the subsed selected from the Wall Street Journal corpus.

The thesaurus employed in this work is the lexical database WordNet [6]. Among the several lexical and concept relations in WordNet, we have used only the synonymy relation.

The VSM employed promotes a recall/precision based evaluation. Recall is defined as the number of relevant documents retrieved by the system over the number of relevant documents in the collection. Precision is defined as the number of relevant documents retrieved by the system over the number of documents retrieved. These metrics are calculated by a special module of Smart. Precision is calculated at the eleven levels of recall 0.0, 0.1, 0.2, ..., 1.0, and averaged over them. Also the average over the three levels 0.2, 0.5 and 0.8 is calculated.

Table 1. Precision averaged over the 11 standard levels of recall, and over the three levels 0.2, 0.5 and 0.8, for the initial run and the different queries constructed by the RF process

	Original Query		Expanded Query	Expanded+WSD Query
Level	*initial run*	*feedback*	*feedback*	*feedback*
11-pt average:	0,1094	0,1853	0,1693	0,2088
% change:		69,4	54,75	84,73
3-pt average:	0,1106	0,1894	0,1719	0,2176
% change:		71,20	55,42	96,74

The results of our experiments are presented in Table 1 and Figure 1. The Table 1 shows the precision averages over the eleven and three recall levels for: the original query; the RF process based only on documents; the RF process based on documents and thesaurus without disambiguation; and the RF process based on documents and thesaurus with disambiguation. The Figure 1 shows the four approaches in a typical recall/precision graph obtained from the precision at the eleven recall levels.

The experiments results presented in Table 1 and Figure 1 show that:

- Using a thesaurus to complement the information of documents in a RF process ussually performs worse than using only the documents. The average precision shows a drop of 15% and 25%. This is caused by the ambiguities

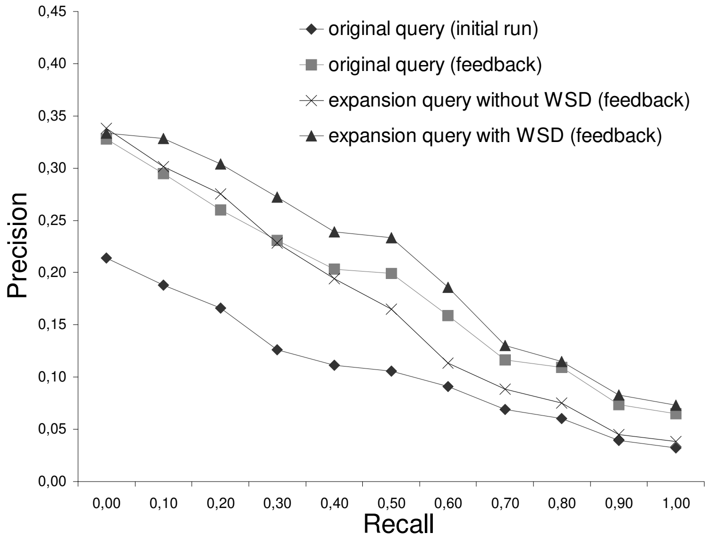

Fig. 1. Recall/precision graph for the initial run and the different queries expanded with documents, and synonyms with and without WSD

in the query words and the high number of senses, and the great number of synonyms extracted from WordNet.

– Using a thesaurus and a WSD algorithm to complement the documents information leads to a substantial improvement in the perfomance. The average precision exhibits an improvement of 15% and 25%. The reasons for this are the accuracy of the WSD algorithm and the restriction of words obtained from the thesaurus to those strictly needed.

The set of words extracted from the retrieved documents for the query expansion process is complemented by a set of precise and informative words obtained from the thesaurus, increasing the amount and quality of information available for constructing the new query.

These results also show that the disambiguation method employed in this work is as effective as to be of practical use. A direct evaluation was performed and reported elsewhere [12], showing that the WSD was very effective. These indirect results complement the direct ones.

5 Conclusions

In this paper, we have presented an approach that integrates the use of documents and thesaurus information for Relevance Feedback. The process is based on the utilization of a Word Sense Disambiguation algorithm that effectively takes advantage of the information extracted from the lexical database WORDNET for Relevance Feedback. The results of our experiments show that making use of WORDNET improves Relevance Feedback only when performing a Word Sense Disambiguation process.

As a conclusion, we present an advance in the RF task that helps to alleviate the information overload problem. Also, we outline an evaluation model for WSD approaches based on the evaluation of the task they are applied in.

References

1. R. Bruce and W. Janyce. Word sense disambiguation using decomposable models. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, 1994.
2. M. Buenaga, J.M. Gómez, and B. Díaz. Using WORDNET to complement training information in text categorization. In *Proceedings of Second International Conference on Recent Advances in Natural Language Processing (RANLP)*, 1997.
3. G. Grefenstette. *Cross-Language Information Retrieval.* Ed. by G. Grefenstette Kluwer Academic Publishers, 1998.
4. D. Harman. Overview of the fourth Text Retrieval Conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference*, 1996.
5. A. Kilgarriff. What is word sense disambiguation good for? In *Proceedings of Natural Language Processing Pacific Rim Symposium*, 1997.
6. G. Miller. WORDNET: A lexical database for English. *Communications of the ACM*, 38(11), 1995.
7. G. Miller, C. Leacock, T. Randee, and R. Bunker. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, 1993.
8. Oracle. Managing text with Oracle8(TM) ConText Cartridge. In *An Oracle Technical White Paper*, 1997.
9. P. Pedersen and R. Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 1997.
10. G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.
11. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
12. L.A. Ureña, M. Buenaga, M. García, and J.M. Gómez. Integrating and evaluating WSD in the adaptation of a lexical database in text categorization task. In *Proceedings of the First Workshop on Text, Speech, Dialogue –TSD'98–*, 1998.
13. L.A. Ureña, M. de Buenaga, and J.M. Gómez. Integrating linguistic resources in TC through WSD. *Computers and the Humanities*, 2000. In press.
14. E. Voorhees. Using WORDNET to disambiguate word senses for text retrieval. In *Proceedings of SIGIR'93*, 1993.
15. Y. Wilks and M. Stevenson. Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of COLING-ACL*, 1998.
16. L. Xiaobin and S. Spakowicz. WORDNET-based algorithm for word sense disambiguation. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995.