Text Categorization and Information Retrieval Using WordNet Senses

Paolo Rosso[1], Edgardo Ferretti[2], Daniel Jiménez[1], and Vicente Vidal[1]

[1] Dept. of Computer Systems and Computation,
Polytechnic University of Valencia, Spain.
Email: prosso@dsic.upv.es, djimenez@dsic.upv.es, vvidal@dsic.upv.es
[2] LIDIC-Dept. of Computer Science,
National University of San Luis, Argentina.
Email: ferretti@unsl.edu.ar

Abstract. In this paper we study the influence of semantics in the Text Categorization (TC) and Information Retrieval (IR) tasks. The K Nearest Neighbours (K-NN) method was used to perform the text categorization. The experimental results were obtained taking into account for a relevant term of a document its corresponding WordNet synset. For the IR task, three techniques were investigated: the direct use of a weighted matrix, the Singular Value Decomposition (SVD) technique in the Latent Semantic Indexing (LSI) model, and the bisecting spherical k-means clustering technique. The experimental results we obtained taking into account the semantics of the documents, allowed for an improvement of the performance for the text categorization whereas they were not so promising for the IR task.

1 Introduction

Nowadays, nearly all kind of information is stored in electronic format: digital libraries, newspapers collections, etc. Internet itself can be considered as a great world database which everybody can access to from everywhere in the world. In order to provide inexperienced users with a flexible access to information, it is crucial to take into account the meaning expressed by the documents, that is, to relate different words but with the "same" information. The classical vector space model introduced by Salton [10] for IR was shown by Gonzalo et al. [4] to give better results if WordNet synsets are chosen as the indexing space instead of terms: up to 29% improvement in the experimental results was obtained for a manually disambiguated test collection derived from the SemCor corpus.

2 Document Codification: Vector of Terms and Vector of Synsets

In the present study, we used the vector space model for the codification of a document with a vector of terms. The vector space model was also used when WordNet synsets were chosen as the indexing space instead of word forms, in order to relate different terms with the same information. Due to the phenomenon of polysemy, it was important to identify the exact meaning of each term. The disambiguation of the meaning of the term was obtained through its context (i.e., the portion of the text in which it is embedded), the WordNet ontology [7]

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 299–304.

and a collection of sense-tagged samples, to train the supervised method for the Word Sense Disambiguation (WSD) [8]. In order to perform the WSD, each term of a document needed first to be tagged (as noun, verb, adjective or adverb) according to its morphological category. This Part-Of-Speech (POS) task was performed by the TnT POS-tagger [1]. The POS-tagged vector of each document was used as input data for the supervised sense-tagger. The final output was a *sense-tagged vector*, that is, a vector tagged with the disambiguated sense for each term of the document of the data sets. In the final vector of each document (and query of the IR task), those terms that were not sense-tagged were removed.

3 The Semantic K Nearest Neighbours Technique

The K Nearest Neighbours is one of the most used techniques for the text categorization task due to its good performance. Given a set of labelled prototypes (i.e., categories) and a test document, the K-NN method finds its k nearest neighbours among the training documents. The categories of the K neighbours are used to select the nearest category for the test document: each category gets the sum of votes of all the neighbours belonging to it and that one with the highest score is chosen. Other strategies calculate these scores taking into account the distances between the K neighbours and the test document or, alternatively, using a similarity measure like the scalar product. In this last strategy, which is the one that we used in our work, each document is represented through a vector of terms and each category gets a score equal to the sum of the similarities between the K neighbours and the test document.

The number of terms of any given collection of documents of medium size may be approximately ten of thousands. Therefore, it was very important to optimise the list of terms that identified the collection. This optimisation was focused to reduce the number of terms eliminating those with poor information. A list of stopwords was used to reduce the number of terms that identify the collection. It included terms which did not provide any relevant information: typically, words as prepositions, articles, etc. Some of these techniques help to improve the results of categorization in determined data sets, once noisy vocabulary is eliminated. There are several methods for selecting the terms to remove. In our work, we employed the *Information Gain* (IG) method [13]. IG measured the amount of information which contributed a term for the prediction of a category, as a function of its presence or absence in a given document. Once calculated the IG_i value for each term i, those terms with the highest value were selected being the most relevant.

4 The Techniques for Information Retrieval

The IR models used in this work are classified within the vector space model and are based in the well-known matrix of terms by documents. With the weighted matrix we modelled the IR system induced by the document collection. We also investigated the LSI model, which is based on the SVD technique, and a clustering model which uses the bisecting spherical k-means algorithm.

4.1 The LSI Technique

There are several techniques in the LSI model. Our approach is based on the SVD technique, in which a part of the spectrum of the singular values of the matrix is calculated [2]. Given a partial SVD of an arbitrary matrix M, we must find p numbers $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ and p vectors $u_i \in \Re^m$ and $v_i \in \Re^n$ such that:

$$M \approx M_p = U_p \Sigma_p V_p^T = \sum_{i=1}^{p} u_i \sigma_i v_i^T \tag{1}$$

The evaluation of queries within the SVD technique is based on the calculation of te angle between the query vector with all the document vectors of the collection.

4.2 The Clustering Technique

When searching for a document, it is often useful (for speed, efficiency, or undestandability) to provide it with a structure. In an electronic document collection, such structure should be provided automatically, and may be based on several similarity criteria: by contained terms, by document structure, by document category, by meaning of content. A popular structure is provided by grouping, or *clustering*. The clustering technique used in this work to evaluate semantic lemmatisation (i.e, the expansion to synonyms) was the Bisecting-Spherical K-Means [5]. This algorithm tries to join the advantages of the Bisecting K-Means algorithm with the advantages of a modified version of the Spherical K-Means. The Bisecting-Spherical K-Means clustering algorithm tries to find k disjoint clusters $\{\pi_j\}_{j=1}^{k}$, from the document collection expressed by matrix M such that it maximizes the following objective function:

$$f\left(\{\pi_j\}_{j=1}^{k}\right) = \sum_{j=1}^{k} \sum_{m \in \pi_j} m^t c_j \tag{2}$$

where $c_j$is the normalised *centroid or concept vector of the cluster* π_j, which it is calculated given the following expression:

$$t_j = \frac{1}{n_j} \sum_{m \in \pi_j} m \; ; c_j = \frac{t_j}{\|t_j\|} \tag{3}$$

where n_j is the number of documents in the cluster π_j.

5 Experimental Results: The Influence of Semantics

5.1 The Text Categorization Task

Different experiments were carried out over the modified 20Newsgroups corpus [9] which was pre-processed taking into account for each relevant term its WordNet synset. For each document, its vectors of terms and WordNet synsets were obtained using the *Rainbow system* [6]. The text categorization task was performed employing the K-NN method, where K was set equal to 30. The 30-KNN classifier carried out the text categorization taking into account the semantics of each document. For this experiment, the vector of synsets of each document was used, instead of its vector of terms.

The goodness of the semantic K-NN classifier was measured determining the error percentage obtained classifying a set of test documents. Figure 1 shows the comparison of the error percentage obtained with (WordNet synsets) and without (terms) the introduction of the semantics with respect to the size of the vocabulary.

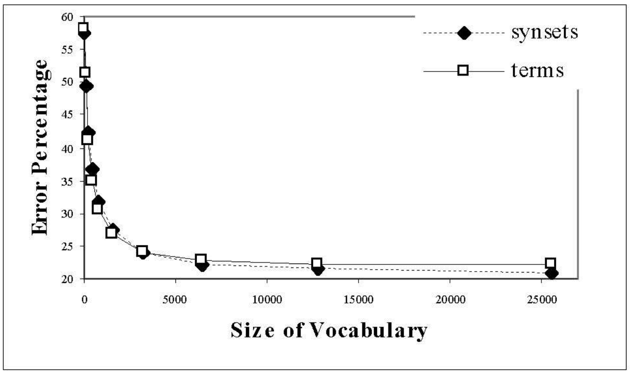

Fig. 1. Text categorization (20Newsgroups corpus): terms vs. WordNet synsets

5.2 The Information Retrieval Task

The criteria used to evaluate the IR experiments, was the average precision-recall ratio:

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

where $\bar{P}(r)$ is the average precision at the recall level r, N_q the number of queries used, and $P_i(r)$ the precision at recall level r for the i-th query. To get each $P_i(r)$, first we evaluated the i-th query obtaining a sorted document set ordered by relevance. Then we calculated the precision each time a relevant document appeared in the answering set. In this data set we interpolated 11 standard recall levels as follows: let $r_j \in\{0,...,10\}$, be a reference to the j-th standard recall level, then, $P(r_j) = \max_{r_j \leq r \leq r_j+1} P(r)$.

The collection used for the experiments contains articles from the 1963 Times Magazine [12]. Query statistics were also obtained for the query collection, formed by a total of 83 queries with an average of 15 words and one line per query. In Figure 2 the most representative results of the study are presented: concretely, the SVD and clustering comparisons between semantic lemmatisation and stemming, which associates words by the root. In fact, words usually have different morphological variants with similar semantic interpretations which would be considered as the same term in IR systems. Stemming algorithms (or stemmers) attempt to reduce a word to its stem or root form. The joining of words with the same information to a single term, also reduces the number of terms that identify the document collection. The experiments were carried out employing the Paice stemming algorithm [3]. In all the studied cases, the semantic lemmatisation had a worse performance than the stemmer. We can observe that the performance of the semantic lemmatisation with the SVD is slightly better than the semantic lemmatisation with the rest of the methods.

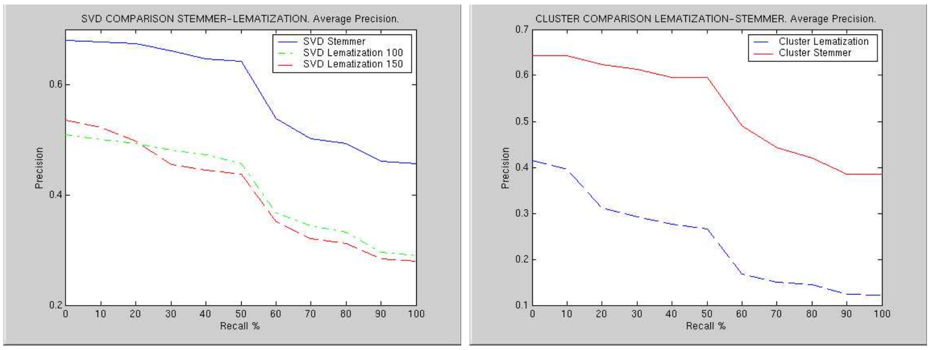

Fig. 2. Semantic lemmatization vs. stemming (Times Magazine corpus): SVD (left) and clustering (right) comparisons.

6 Conclusions and Further Work

In this paper, we investigated whether the introduction of semantic information could help to improve the tasks of TC and IR. With regard to the study of how the semantic 30-KNN performed, it can be remarked that when documents are indexed with WordNet synsets, the performance slightly improved. Therefore, the use of words which refer to the same concept is a research direction we plan to investigate further. As future work, it would be interesting to carry out some experiments using other data sets (e.g. the TREC document collection). In these experiments, the two vector representations should be also combined, in order to take into account with different weights, terms and WordNet synsets at the same time. With regard to the poor performance we obtained for the IR task, it could be due to mainly three reasons. First, the queries of 15 words were pretty long (normal queries are 1.5 words on average) and such long queries implicitly have a disambiguation effect. We should expect better effect of using WordNet for the normal 1 or 2 queries. Second, the semantic lemmatisation related synonyms when they are in the same morphologic group: it should be combined with standard morphological lemmatisation because they could complement each other. Moreover, also other relations could be exploited in the semantic lemmatisation, possibly including the contextual information of the glosses of all the hyponyms. Last, but not least, indexing by WordNet synsets can be very helpful for text retrieval tasks only if the error rate is below 30% [4] and, unfortunately, the state-of-the-art of WSD techniques perform with error rates ranging from 30% to 60% which cannot guarantee better results than standard word indexing.

Acknowledgements

The work of P. Rosso was partially supported by the TIC2000-0664-C02 and TIC2003-07158-C04-03 projects. The work of E. Ferretti was made possible by AECI. We are grateful to A. Molina and F. Pla for making their sense-tagger available.

References

1. Brants, T.: TnT – A Statistical Part-Of-Speech Tagger. In: `http://www.coli.uni-sd.de/~thorsten/tnt`.
2. Dumais, S., Furnas, G., and Landauer, T.: Using latent semantic analysis to improve access to textual information. In: Proc. of Computer Human Interaction (1988).
3. Fox, C., Fox, B.: Efficient Stemmer Generation Project. In: `http://www.cs.jmu.edu/common/projects/Stemming/`.
4. Gonzalo, J., Verdejo, F., Chugur, I., Cigarrán, J.: Indexing with WordNet Synsets can improve Text Retrieval. In: Proc. of the Workshop on Usage of WordNet for NLP (1998).
5. Jiménez, D., Ferretti, E., Vidal, V., Rosso, P., and Enguix, C. F.: The Influence of Semantics in IR using LSI and K-Means Clustering Techniques. In: Proc. of Workshop on Conceptual Information Retrieval and Clustering of Documents, ACM Int. Conf., Dublin, Ireland (2003): 286–291.
6. McCallum, A.: Bow: A Toolkit for Statistical Language Modelling, Text Retrieval, Classification and Clustering. In: `http://www.cs.cmu.edu/~mccallum/bow/`.
7. Miller, A.: WordNet: Lexical Database for English. In: Communications of the ACM, Vol. 38 (1995): 39–41.
8. Molina, A., Pla, F., Segarra. E.: A Hidden Markov Model Approach to Word Sense Disambiguation. In: Proc. of IBERAMIA2002, Seville, Spain, Lecture Notes in Computer Science (2002).
9. Rennie, J.: Original 20 Newsgroups Data Set. In: `http://www.ai.mit.edu/~jrennie`.
10. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. In: Information Processing and Management, Vol. 24 (1998): 513–523.
11. Text Retrieval Conference (TREC) document collection. In: `http://www.trec.nist.gov`.
12. Times Magazine corpus. In: `ftp://ftp.cs.cornell.edu/pub/smart/time/`.
13. Yang, Y., Pedersen, O.: A Comparative Study on Feature Selection in Text Categorization. In Proc. of the Int. Conf. on Machine Learning, (1997): 412–420.