# Does Semantic Information Help in the Text Categorization Task?

Edgardo Ferretti,[1] Marcelo Errecalde,[1] and Paolo Rosso[2]

[1]*Laboratorio de Investigación y Desarrollo en Inteligencia Computacional Universidad Nacional de San Luis. San Luis, Argentina;*
[2]*Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, España*

**ABSTRACT**

In this paper, we investigate how effective the use of semantic information could be in text categorization tasks. To this end, we consider distinct representations of documents differing in the kind of information incorporated: (a) information about terms only, (b) semantic information, and (c) a combination of both types of information. Moreover, we study how the vocabulary size reduction affects this task. The *k* Nearest Neighbours method was used to perform the categorization, and the vocabulary size was reduced by means of the Information Gain technique. A number of different document codifications were tested. The experimental results showed that the inclusion of semantic information in syntactically and semantically richer corpora could improve the text categorization task, if vocabularies with a sufficient number of features are considered. In our view, however, it is not yet possible to affirm that the introduction of semantic information ensures an improvement on the text categorization task. Our results suggest that the performance depends heavily on the particular characteristics of the corpus used in each case.

**KEYWORDS**

semantic text categorization, word sense disambiguation, *k* nearest neighbours method

---

Reprint requests to: Edgardo Ferretti; Laboratorio de Investigación y Desarrollo en Inteligencia Computacional Universidad Nacional de San Luis. San Luis, Argentina;

email: {ferretti, merreca}@unsl.edu.ar; prosso@dsic.upv.es

## 1. INTRODUCTION

*Text categorization* is the task of labeling natural language documents with thematic categories from a predefined set. This activity is usually carried out using a *classifier*, automatically derived from inductive learning processes that learn the correspondence between documents and categories, based on the evidence provided by a set of labeled documents (training set).

Several methods have been proposed for automated text categorization, such as nearest neighbours classification (Yang, 1994), decision trees (Lewis and Ringuette, 1994), neural networks (Wiener et al., 1995), rule learning (Slattery and Craven, 1998), inductive learning algorithms (Dumais et al., 1998), Bayesian classifiers (Lewis, 1998), support vector machines (Joachims, 1998), maximum entropy models (Nigam et al., 1999), boosting (Schapire and Singer, 2000), among others. One of the most used techniques is the *k* Nearest Neighbours (*k*-NN) method (Manning & Schütze, 1999). Many researchers have found that the *k*-NN algorithm achieves a very good performance in their experiments on different data sets (Yang & Liu, 1999). Given a set of labeled prototypes (i.e., text categories) and a test document to be classified, 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 with the highest score is chosen.

In the general classification task, a document could belong to more than one category, and different strategies could be used to consider this case. In our experiments, however, we worked with the classical paradigm assuming that each text is assigned to only one category. The goodness of the semantic *k*-NN was measured determining the error percentage obtained when categorizing texts for each data set.

An important problem for text categorization and for information retrieval tasks in general is to relate different words with the “same” information in order to perform a conceptual or semantic search (i.e. based on the *meaning* of the words). Therefore, it is necessary to consider synonyms and words that refer to the same concept (Rosso et al., 2004).

This work investigates how effective the use of semantic information during the categorization task could be when texts are indexed with word

senses (Gonzalo et al., 1998), and with the original terms plus their senses. The results obtained with these approaches are contrasted with those achieved when only terms are considered for the 20 Newsgroups, WebKB, and Reuters-21578 corpora. Thus, this paper pursues a previous work (Ferretti et al., 2003) where the categorization task was performed considering terms and senses separately, for the first two above-mentioned corpora.

The paper is organized as follows: Section 2 presents the text codifications used in this work and the method used to perform the selection of terms. Section 3 describes the data sets and pre-processing carried out. Section 4 shows the experimental design and the results obtained. Finally, Section 5 puts forward the conclusions and future work.

## 2. TEXT CODIFICATION AND SELECTION OF TERMS

In the present study, Salton's vector space model (Salton & Buckley, 1988) was used for the codification of texts. Each text was represented by an *n*-term vector, where *n* was the number of terms which belong to the training set. The Term Frequency * Inverse Document Frequency weighting scheme, commonly abbreviated as *TF*IDF*, was used for weighting the vector components.

Besides, the SMART system conventional code scheme was used (Salton, 1971). Each codification is composed of three letters: the first two letters refer to *TF* and *IDF* components, respectively, whereas the third one indicates whether normalization is employed or not. The cosine normalization is equivalent to converting the similarity function of *k*-NN classifier into the calculation of cosine between the two vectors, which is invariant with respect to the size of the two documents.

**SMART nomenclature for texts codification**

- $d_i$ : It is the i-th component of vector $\overline{d}$ of size *n*.
- *N*: Number of training documents.
- $TF_{d,i}$ : Term frequency of the *i*-th term in document *d*.

- $DF_i$ : Document Frequency of *i*-th term over the collection.

Definition: $d_i = TF'_{d,i} IDF'_{d,i} NORM$

Where:

$TF'_{d,i}$ =

If $TF_{d,i} = 0$ then 0

If $TF_{d,i} \neq 0$ then:

*n*: *none* = $TF_{d,i}$

*b*: *binary* = 1

*m*: $\max - norm = \frac{TF_{d,i}}{\max_i (TF_{d,i})}$

*a*: $avg - norm = 0.5 + 0.5 \frac{TF_{d,i}}{\max_i (TF_{d,i})}$

*l*: log = 1+ log ($iTF_{d,i}$)

$IDF'_{d,i}$ =

*n*: *none* = 1

*t*: *tfidf* = log *t*: $\left( \frac{N}{DF_i} \right)$

*NORM* =

*n*: *none* = 1

*c*: *cosine* = $\frac{1}{\sqrt{\sum_i \left( TF'_{d,i} IDF'_{d,i} \right)^2}}$

The vector space model that was employed for the codification of each text as a vector of terms was also used when senses and terms plus senses were chosen as the indexing space instead of word forms, so as to relate different terms with the same information during the text categorization. As external lexical resource, we used the WordNet ontology (Miller, 1995), which is based on the concept of synset (set of synonyms). Therefore, in WordNet, a polysemic term belongs to more than one synset. WordNet ontology is partitioned into hierarchies, each one associated to a morpho-syntactical category. In order to perform the Word Sense Disambiguation (WSD), each term of a document needs first to be morphologically tagged (as noun, verb, adjective, or adverb) according to its POS (Part-Of-Speech) category. This POS task was performed by the TnT POS-tagger (Brants,

1998). The POS-tagged document was used as input data for a supervised Hidden Markov Model sense-tagger (Molina et al., 2002). The final output was a sense-tagged document, that is, a document tagged with the dis-ambiguated sense for each term of the text of the corpus. Finally, those terms that were not sense-tagged were removed.

Figure 1 shows the steps performed to convert an original document of the collection in its corresponding vectors of terms, synsets, and terms plus synsets. In the first place, the document is pre-processed according to the collection characteristics (See Section 3). Secondly, the pre-processed document is used as input by the TnT POS-tagger. The output of the POS-tagger is used for the sense-tagger (WSD) which interacts with WordNet to obtain the sense-tagged document. Then, this information is finally used to generate the vectors of terms, synsets, and terms plus synsets.

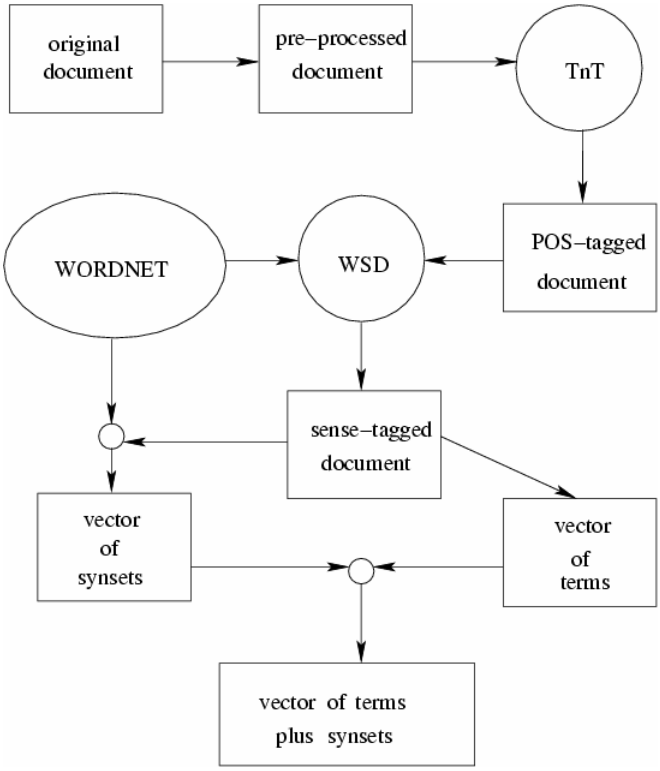


**Fig. 1:** Conversion process of the document into vectors of terms, synsets, and terms

plus synsets

For computational efficiency reasons in space and time, the study of methods for reducing the numbers of terms in the vocabulary is of great interest. Moreover, some of these techniques help to improve the results of categorization in certain data sets, once noisy vocabulary is eliminated. There are several methods for selecting the terms to be removed (Yang & Pedersen, 1997): Documents Frequency Thresholding, Information Gain, Mutual Information, Term Strength, etc.

In this work, the Information Gain (IG) method was used. The *IG* measures the amount of information a term contributes to the prediction of a category, as a function of its presence or absence in a given text. Once calculated the $IG_i$ value for all the terms, those terms with the highest values were selected because they were the most relevant for the category selection.

## 3. DATA SETS

In this section, a brief description of the 3 data sets used to perform our experiments is presented.

The 20 Newsgroups data set contains 19997 news messages of 20 Usenet discussion groups that were sent in 1993. The data are organized into 20 different news groups, each corresponding to a different topic. Some news groups are very closely related one another, while others are highly unrelated. Each category contains 1000 documents except for the *soc.religion.christian* category that contains 997. The pre-processing of this data set consisted in removing the headers of the news groups, but “From:” and “Subject:” fields were maintained. The training set was composed of 16000 texts (the first 800 texts of each category) whereas the other 3997 texts were used as test set.

The WebKB data set is a set of web pages gathered from the departments of computer science of different universities: Cornell, Texas, Washington, Wisconsin, and miscellaneous. The pages are divided into seven categories. In our experiments, we worked with the following four categories: *student*, *faculty*, *course*, and *project*. These four categories together contain 4199 pages. The error rate was calculated using a “one university out for testing, three in for training” approach. On average, the

training set contains 3939 pages and the testing set, 1040. The pre-processing consisted in removing the headers and HTML labels. The aim was reproducing similar experimental conditions to those of a previous work (Ferretti et al., 2003).

Distribution 1.0 of the Reuters-21578 collection is distributed in 22 files in SGML format. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 1000 documents, while the last (reut2-021.sgm) contains 578 documents. For the Reuters-21578 collection, the documents are Reuters' newswire stories and the categories are five different sets of content-related categories. For each document, a human indexer decided to which set that document belonged. The TOPICS categories are economic subject categories. This set of categories has been used in almost all previous research with the Reuters data, and is used in our work as well. In order to perform our experiments, we used the Modified Apte (ModApte) split of training and test documents (Section VIII.B. of the README file in Reuters collection). However, to maintain consistency with the experiments carried out with the WebKB and 20 Newsgroups data sets as well as to be able to pursue the work previously done (Ferretti et al., 2003), those documents belonging to more than one category were assigned to only one category. The pre-processing of this data set was carried out as follows.

Firstly, the 21578 documents in SGML format contained in the 22 files provided by the collection were separated one per file, obtaining, thus, 21578 files. Secondly, from the 135 original topic categories, 90 were selected, taking into account those that were present in at least one training and one test document. Thirdly, the documents were assigned to a category. Those documents having only one category assigned in the collection by the human indexer were directly categorized, while those belonging to more than one category were assigned to one category, so as to maintain the number of documents belonging to each category as uniform as possible.

Moreover, in the categorization of the documents, the issue of having at least one training document and one test document per category was considered. Only two categories (*lin-oil* and *sun-meal*) could not satisfy this issue so, only training documents were assigned to them. Finally, those documents that did not have any topic category or that had a category not included in the 90 selected were removed.

Then, the 7770 training documents and the 3019 test documents were converted to plain text considering the content of the documents delimited by the following SGML labels:

1. ⟨DATE⟩, ⟨\DATE⟩
2. ⟨TEXT⟩, ⟨\TEXT⟩
    a. ⟨TITLE⟩, ⟨\TITLE⟩
    b. ⟨AUTHOR⟩, ⟨\AUTHOR⟩
    c. ⟨BODY⟩, ⟨\BODY⟩

After the pre-processing of all the documents belonging to the data sets above mentioned, the documents were POS and sense-tagged.

## 4. EXPERIMENTAL STUDY

This section reports several experiments carried out over the modified 20 Newsgroups, WebKB, and Reuters-21578 corpora, pre-processed as explained above.

In the first place, the three best codifications and *k* values that reported the "lowest error percentages" with vectors of terms only were selected to compare their results with those obtained with the vectors of synsets and terms plus synsets, respectively. The results are presented for the 20 Newsgroups, WebKB, and Reuters corpora. Then, an analysis of how the vocabulary size reduction affects the text categorization task is done.

It is important to note that, for each document of the 20 Newsgroups and WebKB corpora, we built only their corresponding vectors of terms plus synsets to pursue the work previously done (Ferretti et al., 2003), while for the Reuters-21578 corpus, its vectors of terms, synsets, and terms plus synsets were obtained. The complete vocabularies obtained were:

- 20 Newsgroups: 31786 terms and 27652 synsets.
- WebKB: 17251 terms and 13475 synsets.
- Reuters: 28969 terms and 14580 synsets.

Text categorization task was performed employing the *k* Nearest Neighbours method provided by the Rainbow system (McCallum, 1996).

Several trials with different *k* values belonging to the set {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50} were performed. Furthermore, the different vocabulary sizes we worked on were {50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600} features.

### 4.1. Results and Analysis

**4.1.1** *Experiments with complete vocabulary*. In Figure 2, the lowest error percentages corresponding to the best *k* values used in our experiments (in some cases, they differ from each other in the codifications) are shown for the 20 Newsgroups, WebKB, and Reuters corpora with complete vocabularies. As can be observed, for the 20 Newsgroups corpus, the best results were those obtained with the *atc*, *btc*, and *ltc* codifications. This fact confirms the importance of *IDF* information which sub-estimates those terms that occur in many texts and are not relevant, and of the cosine normalization

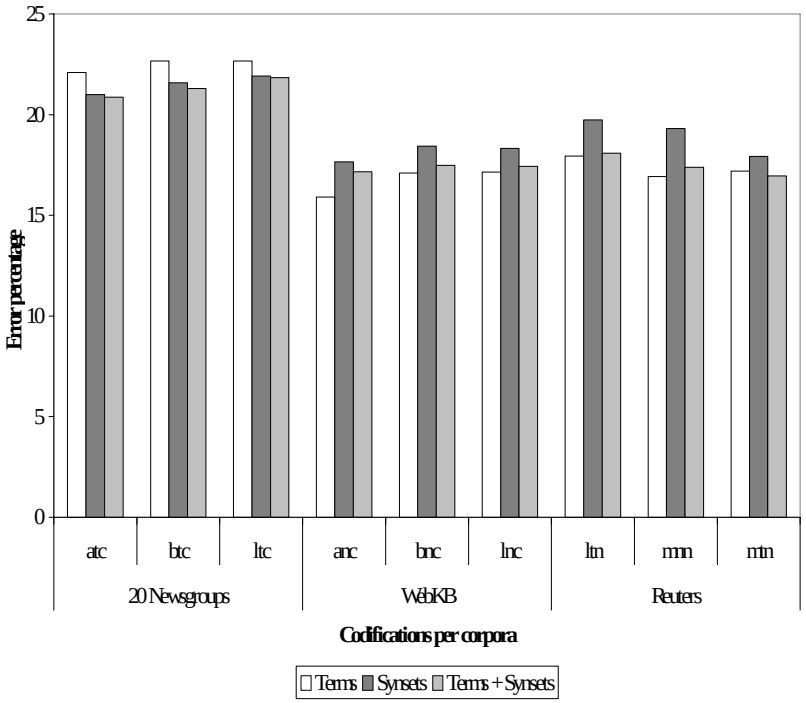


**Fig. 2:** Error percentages for the best k-NN classifiers

that weights a term in a given text with respect to the other terms of the same text and does not weight the term in its absolute frequency of occurrence. For these codifications, it is evident that the use of semantic information improves the results obtained when only terms are considered. In particular, the lowest error percentages were achieved with the vocabulary of terms plus synsets.

For the Reuters corpus, the best results were obtained with the *ltn*, *mnn*, and *mtn* codifications. It can be noted that, when isolated semantic information is considered (synsets only), the highest error percentages are obtained. On the other hand, the results obtained with terms plus synsets are comparable to those achieved with terms only. However, it must be pointed out that these results are not representative of those obtained with the rest of the codifications, and due to space constraints they are not shown in the figure. In the majority of these cases, the use of terms plus synsets slightly outperformed the use of terms only.

If we consider now the results obtained for WebKB, the three best codifications were *anc*, *bnc*, and *lnc*. In contrast to what happened with the other corpora, better results were not obtained using semantic information over text categorization than using vector of terms.

**4.1.2** *Experiments with reduced vocabulary.* In order to reproduce the same experimental conditions of (Ferretti et al., 2003), we used a 30-NN classifier.

Figure 3 depicts the results of how the reduction in the vocabulary size affected the categorization task of the classifier in the 20 Newsgroups corpus, for the *atc* codification. This figure is a representative case of what also happened with the other codifications. As can be observed, in all the cases the error percentage decreases when the vocabulary size increases. Furthermore, for small vocabulary sizes (50 and 100), the lowest error values were obtained when using the vectors of synsets only. This could be due to the fact that 50 or 100 synsets gather more information than the same quantity of terms, or terms plus synsets, because a synset could be associated with more than a single term. In addition, for vocabulary sizes ranging from 200 to 3200, the classifier performed better using the vectors of terms only, while for a vocabulary size of 12800 or higher, the 30-NN behaved better using vectors with semantic information instead of terms only.

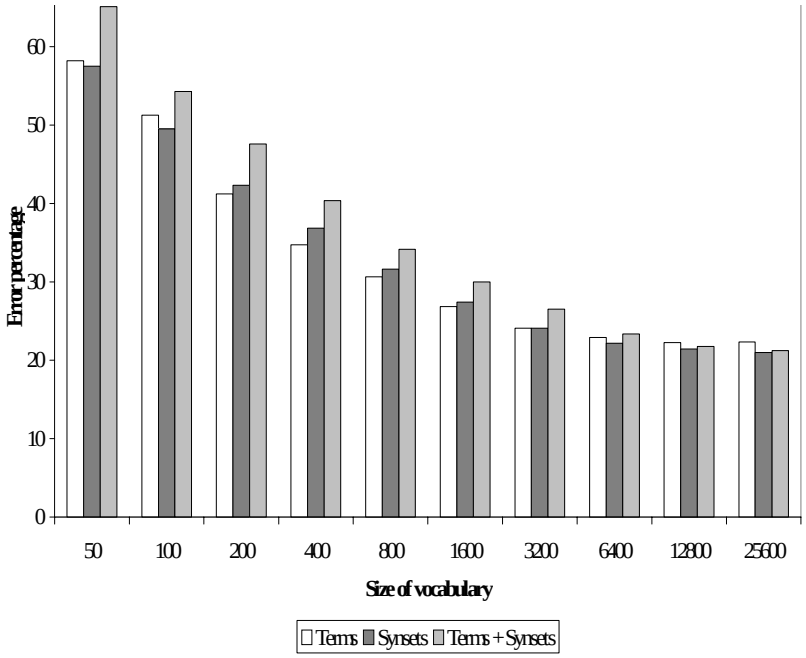


**Fig. 3**: 20 Newsgroups corpus: error percentage of 30-NN classifier as function of the vocabulary size for the *atc* codification of text documents

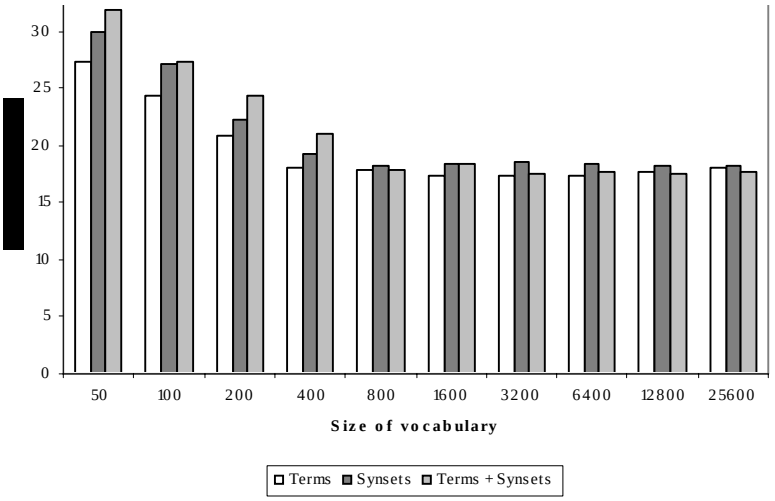


**Fig. 4:** Reuters corpus: error percentage of 30-NN classifier as function of the vocabulary size for the *mtn* codification of text documents

For the Reuters corpus, Figure 4 shows the results with reduced vocabu-

laries for the *mtn* codification (as a representative case). Contrary to what happens in 20 Newsgroups, for this corpus, the reduction of vocabularies does not affect negatively text categorization until the size of the vocabulary is lower than 400 features. Besides, until the number of features selected does not reach a size of approximately 50% of the original vocabulary, the use of vectors of terms plus synsets is not justified.

Finally, the results corresponding to WebKB corpus with reduced vocabularies for the *anc* codification are depicted in Figure 5. As before, this figure is a representative case of what also happened with the other codifications. Contrary to what happened with 20 Newsgroups, the error percentage does not always decrease when the vocabulary size increases. When working with vocabulary sizes of 50, 100, and 200, the error percentage decreases when the vocabulary size increases, but from vocabulary sizes of 400 to 12800, the opposite occurs. This fact agrees with what happened in (Ferretti et al., 2003) when using synsets versus terms only.

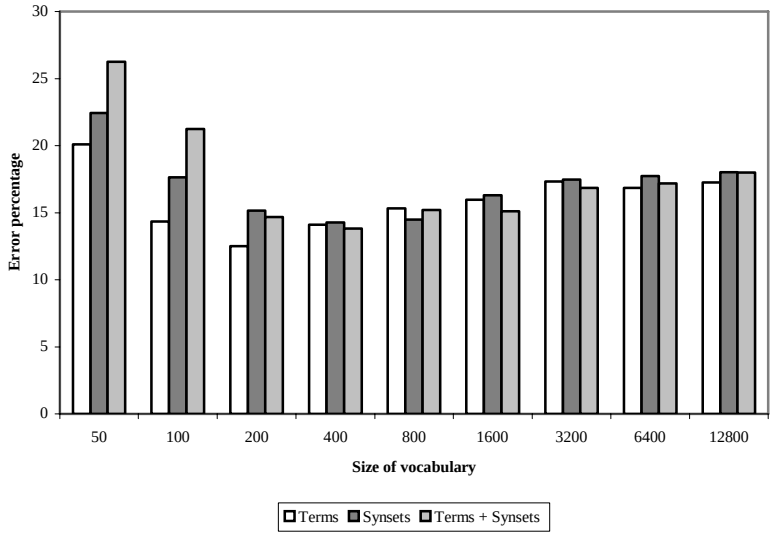


**Fig. 5:** WebKB corpus: error percentage of 30-NN classifier as function of the vocabulary size for the *anc* codification of text documents

Owing to the nature of this corpus, composed by web pages that are not syntactically and semantically as rich as the 20 Newsgroups corpus, the use of a vocabulary of 50 and 100 words is perhaps not enough to perform the categorization, but when increasing it too much, there are noisy terms that do not help in the categorization task.

## 5. CONCLUSIONS AND FURTHER WORK

In this paper, we investigated whether the introduction of semantic information helps to improve the text categorization task using the *k*-NN method. Two different approaches were considered: (a) synsets only and (b) terms plus synsets. We also analyzed how different alternatives of documents codification, as well as the vocabulary size reduction, could affect the task. In order to carry out this study, three corpora (20 Newsgroups, WebKB, and Reuters-21578) with very different characteristics were considered.

We concluded that the impact of the use of semantic information depends to a great extent on the particular characteristics of the corpora. For instance, corpora such as 20 Newsgroups and Reuters-21578 could justify the use of semantics with complete vocabularies, whereas WebKB could not because the former corpora are syntactically and semantically richer than WebKB.

A second observation was that the vocabulary size plays an important role in deciding whether it is convenient to incorporate semantic information in the text categorization task. For example in 20 Newsgroups, it is possible to obtain acceptable results with a reduction up to 50% of the vocabulary size, whereas for the Reuters corpus, good results are achieved with a vocabulary of 800 features. Therefore, we must note that the use of terms plus synsets usually outperforms the results obtained with synsets only. Nonetheless, the use of synsets only instead of terms plus synsets seems to be less affected if aggressive politics of vocabulary reduction are adopted (e.g. vocabulary size lower than 400).

Taking into account that there was not one codification of the documents reporting the best results for the three corpora, it can be stated that an optimal codification does not exist and that the best codification depends on the

characteristics of the texts that are used in the categorization.

Yang and Liu (1999) state that because texts often belong to more than one category, when assigning only one category to each document, a low performance of *k*-NN method may occur. Further experiments taking into account this aspect should be carried out to see whether it is possible to improve precision. A possibility could be to design a *k*-NN classifier for each category in order to decide whether the test document belongs to a category or not.

As further work, it would be also interesting to carry out experiments using other data sets like a collection of very short documents belonging to a narrow-domain (e.g. the abstracts of the CICLing conference on Computational Linguistics [Alexandrov et al., 2005]), other techniques for reducing vocabulary size, and other WSD methods (Buscaldi et al., 2004) for sense-tagging the terms. Moreover, it would be interesting to analyze the inclusion of semantic information in other text categorization methods.

## ACKNOWLEDGMENTS

The authors, Edgardo Ferretti and Marcelo Errecalde, thank the Universidad Nacional de San Luis and the ANPCYT for their unstinting support. The work of Paolo Rosso was supported by the MCyT TIN2006-15265-C06-04 research project. We are grateful to A. Molina, F. Pla, and E. Segarra for making their sense-tagger available.

## REFERENCES

Alexandrov, M., Gelbukh, A. and Rosso, P. 2005. An approach to clustering abstracts, Natural Language Processing and Information Systems, *Tenth International Conference on Applications of Natural Language to Information Systems*, 275-285.

Brants, T. 1998. *TnT: A statistical part-of-speech tagger*. http://www.coli.unisaarland.de/~thorsten/tnt/

Buscaldi, D., Rosso, P. and Masulli, F. 2004. Integrating conceptual density with wordnet domains and cald glosses for noun sense disambiguation,

*Advances in Natural Language Processing: 4th International Conference*, 183-194.

Dumais, S., Platt, J., Heckerman, D. and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization, *Proceedings of the 7th International Conference on Information and Knowledge Manage-ment*, 148-155.

Ferretti, E., Lafuente, J. and Rosso, P. 2003. Semantic text categorization using the k nearest neighbours method, *First Indian International Conference on Artificial Intelligence*, 434-442.

Gonzalo, J., Verdejo, F., Chugur, I. and Cigarrán, J. 1998. Indexing with wordnet synsets can improve text retrieval, *Workshop on Usage of WordNet for Natural Language Processing*, 38-44.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features, *Proceedings of the 10th European Conference on Machine Learning*, London, 137-142.

Lewis, D.D. and Ringuette, M. 1994. A comparison of two learning algorithms for text classification, *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, 81-93.

Lewis, D.D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval, *Proceedings of the 10th European Conference on Machine Learning*, 4-15.

Manning, C.D. and Schütze, H. 1999. *Foundations of statistical natural language processing*, MIT Press.

McCallum, A.K. 1996. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*.

Miller, G.A. 1995. Wordnet: a lexical database for English., *Communications of the ACM*, **38**, 39-41.

Molina, A., Pla, F. and Segarra, E. 2002. A hidden markov model approach to word sense disambiguation, *Proceedings of the 8th Ibero-American Conference on AI*, 655-663.

Nigam, K., Lafferty, J. and McCallum, A. 1999. Using maximum entropy for text classification, *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 61-67.

Rosso, P., Molina, A., Pla, F., Jimenez, D., Vidal, V. 2004. Information retrieval and text categorization with semantic indexing, Computational Linguistics *and Intelligent Text Processing: 5th International Conference, CICLing*, 596-600.

Salton, G. 1971. The smart retrieval system: experiments in automatic document *processing*, Prentice Hall.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, **24**, 513-523.

Schapire, R.E. and Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization, *Machine Learning*, **39**, 135-168.

Slattery, S. and Craven, M. 1998. Combining statistical and relational methods for learning in hypertext domains, *Proceedings of the 8th International Workshop on Inductive Logic Programming*, 38-52.

Wiener, E.D., Pedersen, J.O. and Weigend, A.S. 1995. A neural network approach to topic spotting, *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, 317-332.

Yang, Y. 1994. Expert network: effective and efficient learning from human decisions in text categorization and retrieval, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 13-22.

Yang, Y. and Pedersen, J.O. 1997. A comparative study on feature selection in text categorization, *Proceedings of the 14th International Conference on Machine Learning*, 412-420.

Yang, Y. and Liu, X. 1999. A re-examination of text categorization methods, *22nd Annual International SIGIR*, 42-49.