Keywords, k-NN and Neural Networks: A Support for Hierarchical Categorization of Texts in Brazilian Portuguese

Susana de Azeredo, Silvia Maria W. Moraes, Vera Lúcia Strube de Lima
PUCRS- Faculdade de Informática
Avenida Ipiranga, 6681. Prédio 32, Sala 629. CEP. 90619-900. Porto Alegre, RS, Brasil
E-mail: suzaz@terra.com.br, silvia.moraes@pucrs.br, vera.strube@pucrs.br

Abstract

A frequent problem in automatic categorization applications involving Portuguese language is the absence of large *corpora* of previously classified documents, which permit the validation of experiments carried out. Generally, the available *corpora* are not classified or, when they are, they contain a very reduced number of documents. The general goal of this study is to contribute to the development of applications which aim at text categorization for Brazilian Portuguese. Specifically, we point out that keywords selection associated with neural networks can improve results in the categorization of Brazilian Portuguese texts. The *corpus* is composed of 30 thousand texts from the Folha de São Paulo newspaper, organized in 29 sections. In the process of categorization, the *k-Nearest Neighbor* (k-NN) algorithm and the *Multilayer Perceptron* neural networks trained with the *backpropagation* algorithm are used. It is also part of our study to test the identification of keywords parting from the log-likelihood statistical measure and to use them as features in the categorization process. The results clearly show that the precision is better when using neural networks than when using the k-NN.

1. Introduction

Text categorization (or classification[1]) is the process of automatically attributing one or more predefined categories to a textual document (Sun and Lin, 2001; Sebastiani, 2006). The diversity of texts, however, can lead to the definition of a large number of categories, which makes the search for information difficult, demanding more computational resources. A commonly adopted solution is the organization of categories in a hierarchy. In the hierarchical categorization process, classification begins in the most generic category of the hierarchy (the root) and is repeated for the subcategories, until reaching the more specific categories (the leaves). This process reduces the computational costs, since only the classifiers of the branches of the hierarchy, which were activated during the process, are executed (Sun and Lin, 2001).

The emergence of large digital libraries and the Internet itself has motivated researches in automatic categorization, aiming, primarily, to improve the information retrieval process. The existence of various *benchmarks* for the English language with an expressive volume of labeled texts is one of the reasons that lead researchers to carry out experiments for the English language, since these *corpora* facilitate the analysis of the obtained results. Such a resource is practically inexistent when dealing with applications aimed at the Portuguese language, and it is precisely on this point that our contribution is concentrated.

The general goal of this study[2] is to contribute to the development of applications which aim at text categorization for Brazilian Portuguese. Specifically, we point out that keywords selection associated with neural networks can improve results in text categorization.

Several recent studies about text categorization present proposals which aim at improving the classification precision through new means for feature selection and new classification algorithms. Our study, however, is not exactly focused on improving these methods, but on contributing in a very slow and onerous task of manual classification of the *corpus*. Our support can deliver classification suggestions which can be confirmed and evaluated by a human being.

Our work proposes the hierarchical categorization of text from a large *corpus* of the Portuguese language, PLN-BR CATEG.[3] In our experiments, we use the hierarchy of categories defined by Langie in (Langie, 2004); techniques of document frequency[4] and *log-likelihood*[5] to select the most significant words, as well as the *k-Nearest Neighbor* algorithm and the *MultiLayer Perceptron* neural network classifiers.

This article is organized in 7 sections, including the bibliographic references. Section 2 presents work related to ours. Section 3 describes the methodology used in our experiments. Section 4 details the experiments carried out and the obtained results and, finally, Section 5 presents the conclusions.

2. Related Work

Several recent works in the field of text categorization present proposals which aim to improve the precision of

[1] In this document the terms "category" and "class" are used as synonyms. The same goes for categorization and classification.

[2] This work was partially funded by Brazilian Agency CNPq (PLN-BR project, CNPq #550388/2005-2)

[3] This *corpus* was obtained through the project "Resources and Tools for Information Retrieval in Textual Bases in Brazilian Portuguese" (PLN-BR). It is composed of 30 thousand texts from the Folha de São Paulo newspaper, organized in 29 sections.

[4] Document frequency is a measure which determines in how many documents certain determined word appears (Lavelli, A.; Sebastiani, F.; Zanoli, R., 2004).

[5] *Log-likelihood* is a statistical measure used for the comparison of word frequencies (Rayson, P. and Garside, R., 2000).

the classification, through new means of feature selection, new classification algorithms or the combination of those already existing.

Srinivasan describes (Ruiz, M. and Srinivasan, P., 2002) a method for text categorization, based on a tree of backpropagation neural networks. The authors used in their experiments the UMLS Metathesaurus[6] hierarchy, the OHSUMED[7] *corpus* and the combination of three methods for feature selection: correlation coefficient, mutual information and odds ratio. In the preprocessing of the texts, the authors removed the stopwords and applied stemming, excluding stems which occurred in less than 5 documents from the training set. By category, the stems which appeared in less than 5% of the positive samples of the category were also removed. The training sets were selected using two different methods: centroid-based and *k-Nearest Neighbor*. The objective of the work was to find the set that better represents the domain of each category and is large enough to train the neural networks without overfitting. The texts were represented using the bag-of-words approach whose weights were calculated using *tf x idf* (Sebastiani, 2006) , where *tf* is the frequency of the term in the document, and *idf* is the inverse document frequency. In some cases this work presented results comparable to the optimized Rocchio algorithm, obtaining F1 values of 0.51.

Hao *et al.* (Hao, P.; Chiang, J. Tu, Y., 2007) describe a method for classifying documents in a hierarchical manner, using SVM classifiers. In their experiments, they utilize the Reuters[8] 21578 *corpus*, applying *stemming* to the words and considering only those features that occur at least 3 times in the training set and which are not stopwords. Moreover, they use the *information gain* measure in order to select the most relevant features for each classifier. The hierarchical structure used by the authors is automatically generated by applying iteratively a clustering method also based on SVM. In the process of hierarchical classification, the binary SVM classifiers used are capable of distinguishing with greater precision, according to the authors, one class in relation to the others. The proposed method was compared to other classifiers which used Naïve Bayes, Rocchio, Decision Trees, k–NN and Rule Induction. In almost all cases the performance of these classifiers was surpassed. The proposed method could reach a precision of 0.95 and a recall of 0.90 on average.

Ceci and Malerba in (Ceci *et al*, 2007) present a general hierarchical text categorization framework which includes a method for automatically determining thresholds for the classifiers. The classifiers combined the centroid-based, Naïve Bayes and SVM methods and were tested on Yahoo, DMOZ and RCV1 *corpora*. All the documents of the *corpora* were tokenized. HTML tags, punctuation marks, numbers and tokens of less than three characters were removed. Moreover, stopwords were removed and the Porter's algorithm for English was applied to the remaining words. The authors utilize a global approach in which a single set of features is chosen to be used in all the classifiers. For feature selection they use $max\ TF \times DF^2 \times ICF$, where *TF* is the maximum frequency of the term considering all training documents of a category *c*, DF is the percentage of documents of a category *c* in which the term occurs and ICF is the inverse of the category frequency, that is the number of subcategories of a category *c* in which the term occurs. This measure returns high scores for features that appear in many relevant documents and in documents with few category alternatives. The results obtained were satisfactory, since the classifiers were more efficient and effective, reducing the classification errors.

Langie (Langie, 2004) established a hierarchy structured in a tree with height 2 and with 28 categories (Figure 1) when he studied text categorization with k-NN. The experiments described by Langie served as a basis for our study, even though the author used a smaller volume of texts – 2,896 texts (from the year 1994) that are also part of the PLN-BR CATEG *corpus*. Langie developed and tested a hierarchical categorizer formed by various multi-label classifiers that implement the k-NN *algorithm* (Yang and Liu, 1999). Besides this, in an experimental way, Langie also analyzed the influence of some parameters in the text classification process: 1) number of nearest neighbors considered by the algorithm (*k*) and its relation with the amount of documents used in the training phase of the classification process; 2) number of features chosen during the feature selection, and 3) categorization strategy itself.

3. Methodology

In this section, we present the methodology used, describing the PLN-BR CATEG *corpus*, the hierarchy of categories and the document preprocessing.

3.1 PLN-BR CATEG *Corpus*

This *corpus* gathers a total of 30 thousand texts published between the years 1994 and 2005 in the Folha de São Paulo Journal, produced in São Paulo/ Brazil.

These texts are organized in 29 sections: *Agrofolha* (193 texts), *Brazil* (5,606 texts), *Special Notebook* (509 texts), *Special Notebook 2* (50 texts), *Science* (182 texts), *Construction* (7 texts), *Daily* (6,458 texts), *Money* (4,153 texts), *Jobs* (238 texts), *Monday Interview* (4 texts), *Equilibrium* (28 texts), *Sports* (4,632 texts), *Invest Folha* (165 texts), *Business Folha* (36 texts), *Synapse Folha* (11 texts), *Teen Folha* (260 texts), *Little Folha* (78 texts), *Fovest* (82 texts), *Illustrated* (2,935 texts), *Real State* (120 texts), *Computers* (408 texts), *More!* (252 texts), *World* (2,410 texts), *Firs Page* (170 texts), *Folha Magazine* (3 texts), *Everything* (95 texts), *Tourism* (464 texts), *TVFolha* (236 texts) *and Vehicles* (215 texts).

The PLN-BR CATEG is a 9,900,234 tokens *corpus*.

[6] The Metathesaurus is a database of information on concepts used in the field of biomedicine. It preserves the meanings, hierarchical connections, and other relationships between terms of this area. Additional information and documentation may be found at http://www.nlm.nih.gov/research/umls/.

[7] OHSUMED is available at ftp://medir.ohsu.edu/pub/ohsumed. It includes medical abstracts from the year 1991.

[8] Available at: http://www.research.att.com/lewis/reuters21578.html

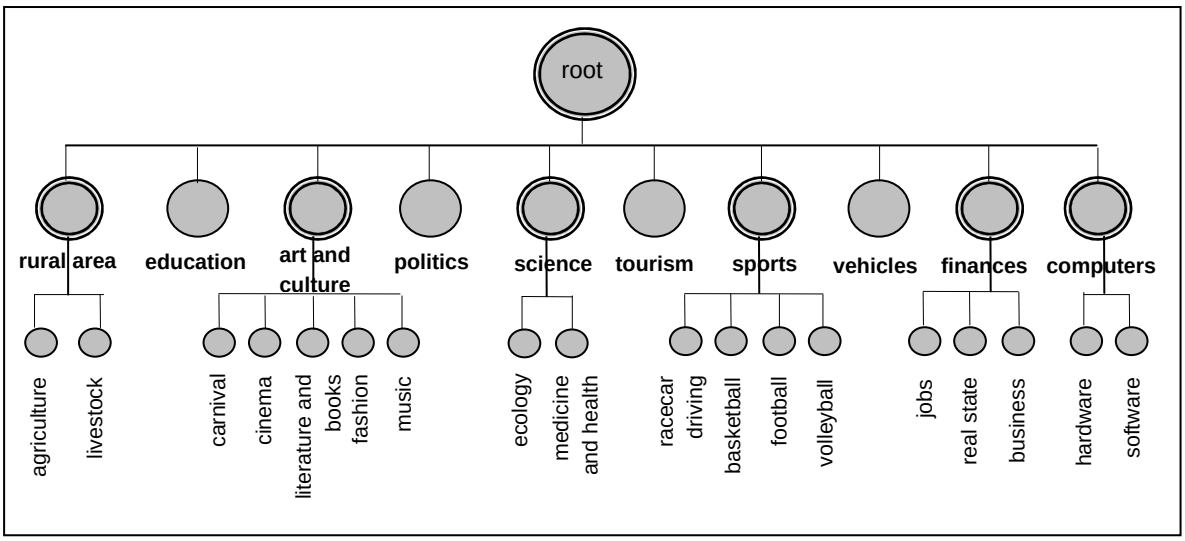

Figure 1: Langie's categories

3.2 Category Hierarchy

We have used in our work the hierarchy of categories defined by Langie (Langie, 2004). In our experiment it was necessary to establish some equivalences between Langie's categories and the PLN-BR CATEG *corpus* sections. The established relations are: **rural area:** *agrofolha;* **art and culture:** *illustrated, TVFolha, more! teen folha; little folha;* **science:** *science;* **sports:** *sports;* **finances:** *money;* **finances/jobs:** *jobs;* **finances/business:** *business folha;* **finances/real state:** *real state;* **computers:** *computers;* **tourism:** *tourism;* **vehicles:** *vehicles.*

3.3 Preprocessing

All the documents of PLN-BR CATEG *corpus* used in our experiments were lemmatized by a FORMA tool whose precision, according to the author Marco Gonzalez, is around 97% (Gonzalez *et. al*, 2007). Possibly due to some formatting problem 1,105 texts from de *corpus* could not be processed and were removed. We choosed lemmatisation because according to Vilares *et al* in (Vilares, 2002), some languages with a more complex flexional morphology as German, Spanish, French and Portuguese, stemming simple algorithms are not used to be enough and besides they have a high computacional cost. Another reason is taken from Gonçalves' study in (Gonçalves, 2007) who carried out categorization experiments with lemmatisation in a subset of the PLN-BR and got better results with this normalization technique.

The set of documents was divided two parts: training set and test set. As training set 2,896 documents from the year 1994 were used. These documents correspond to the Folha Hierarq[9], which contains 30,398 words. They were chosen as a training set because they had already been manually classified and, so, we already knew their class.

The remaining 26,606 texts were used to test the classifiers. It is worth highlighting that the stopwords[10] were removed from all the documents.

After this, we identified the keywords with Wordsmith Tools (WSTolls)[11](Scott, 1996). A keyword analysis normally involves at least two Wordlist files. For example, in order to get the keywords from *agrofolha* section, the WSTools makes a Wordlist of the most frequent words from this section. After this, it is made a Wordlist of the most frequent words from all other sections. These two lists are compared and, using log-likelihood statistics WSTools will display a table containing the words whose frequency is either unusually high or unusually low in comparison to the remaining words. With this tool we could observe that, for example, in *agrofolha* section, the word *animals* appears 138 times (which represents 0.026% of the text) and, in other sections, this word appears 554 times (which represents 0.001% of the texts or less). This means that the word *animals* appears much more in texts from *agrofolha* than in texts from other sections, making it a statistically relevant word in the PLN-BR CATEG *corpus*. The total number of keywords displayed by WSTools is 500. Since there is not a consensus about what a relevant sample from these 500 keywords is, we used the totality. The texts were represented using the bag-of-words approach and word weights were calculated using *tf-idf* .

[9] Folha Hierarq is a subset of PLN-BR CATEG, which contains around 50% of the documents from the year 1994 previously classified in the categories of Langie's hierarchy.

[10] Stopwords are words like articles, prepositions and conjunctions, which use to be removed before some kinds of text analysis. Langie's stoplist was used in our work (Langie, 2004).

[11] Wordsmith Tools is a suite of programs for those interested in computing and studying words frequencies and word patterns. (http://www.lexically.net/wordsmith/)

4. Experiments

The first experiment of categorization is based on Langie's work (Moraes and Lima, 2007). It was developed with use of a hierarchical categorizer with 7 multi-label k-NN classifiers, denominated local classifiers. The position of the local classifiers was marked in the Figure 1 with double circles. The local *root* classifier was responsible for categorizing the text in one of the level 1 categories of the hierarchy. The remaining 6 classifiers were responsible for categorizing in the subcategories of the categories: *rural area*, *art and culture*, *science*, *sports*, *finances* and *computers*. The similarity among the documents was defined by cosine, according to the threshold classification strategy based on rank. In this approach, the category of the neighboring document most similar to the text is assumed as its category. Following previous works (Langie, 2004; Moraes and Lima, 2007), the categorizer uses 13, 17 and 23 as the *k* neighbors to be considered. The value of *k* is chosen according to the quantity (*q*) of documents from the category of the training set. If *q* is less than 250, the value 13 is used, if it is between 251 and 500, 17 is used and for a *q* greater than 500, 23 is used.

The results were evaluated according to the Precision (Pr), Recall (Re) and F1 measures, discussed in Sebastiani (2006).

As one can observe in Figure 2, the best results were obtained for the *sports* section, using document frequency (*>=4*) for feature selection. A precision of 0.84, a recall of 0.95 and an F1 of 0.90 were obtained. The worst results were obtained in the science section, whose precision did not reach 0.1. *Sports* presented good results, since it has a smaller and more constant vocabulary. Throughout the 11 years of newspaper 1,274 keywords in common were found. Science, however, besides having few documents, has a much broader vocabulary which has certainly evolved in this

With the purpose of improving categorization precision, we decided to use the *Multilayer Perceptron* networks which were trained with the backpropagation algorithm. Figure 3 shows 28 Neural Networks (NN) that were built: 10 for level 1 of the hierarchy and 18 for level 2. There is 1 network for each category at level 1: *rural area*, *art and culture*, *science*, *education*, *sports*, *finances*, *computers*, *politics*, *tourism and vehicles.*

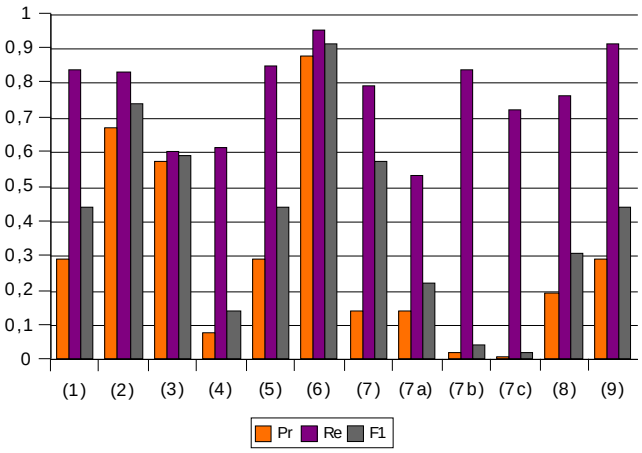

Figure 2: Performance of the k-NN Categorizer for: (1) *rural area*, (2) *art and culture*, (3) *politics*, (4) *science*, (5) *vehicles*, (6) *sports*, (7) *finances*, (7a) *jobs*, (7b) *real estate*, (7c) *business*, (8) *tourism* and (9) *computers*.

At level 2, we find: 1 network for each subcategory in the *rural area*: *agriculture* and *livestock;* 5 networks for the subcategories in *art and culture: carnival, cinema, literature and books, fashion and musica;* 2 networks for the subcategories in *science: ecology, medicine and health;* 4 networks for the subcategory in *sports: racecar driving, basketball, football and volleyball;* 3 networks for the subcategories in *finances: jobs, real state and business;* and 2 networks for the subcategories in *computers: hardware and software.*

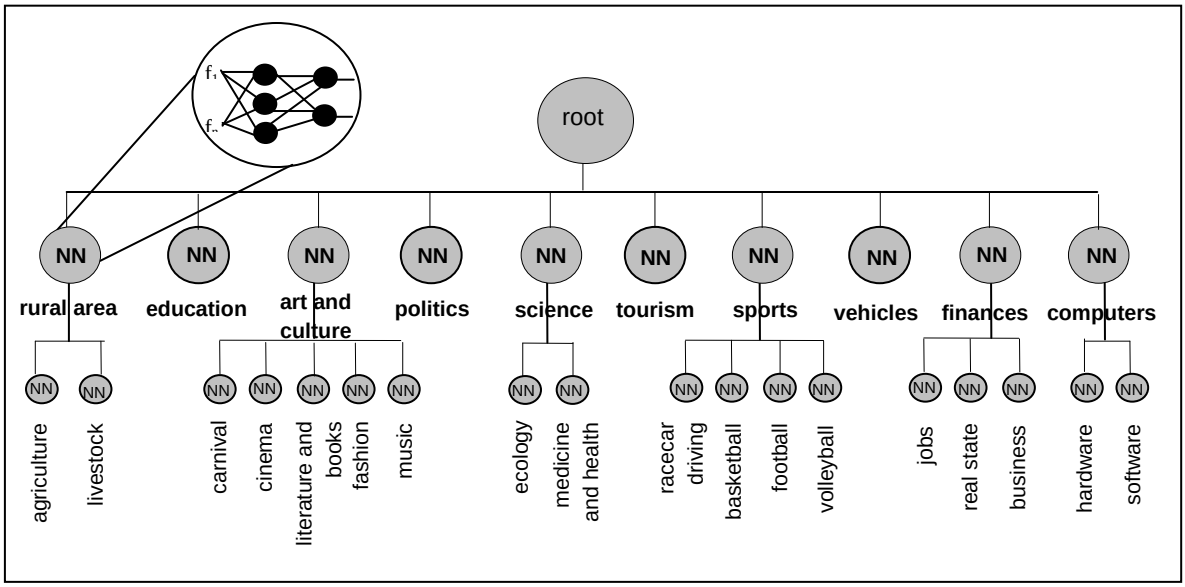

Figure 3: Neural classifiers

All the neural networks were trained with the same training set (2,896 texts from 1994) and two types of training were carried out: the first one with the 500 keywords taken from WStools on the training set and the second one using the 500 keywords from WStools on the whole *corpus*. In training, *k-fold cross-validation* was used to find the best network configuration. In this way, using *k = 10*, all the 2,896 texts from the training set were separated in categories and divided in to 10 subsets. These 10 subsets were alternately combined in groups of 9 and formed the estimation set while the remaining subset formed the validation set. We tested several neural network topologies varying the number of hidden neurons from 29 to 89 and changing learning rate to 0.1, 0.3, 0.5 and 0.7. From the tested topologies the best one was *features x 37 x 2*. The number of *features* used depends on the words selected in the document set from a specific category. The two neurons of the output layer respectively represent *belongs to the category* and *does not belong to the category*. Thus, the text will be categorized in a category as *rural area*, for example, when the neuron *belongs to the category* has the greatest exit value. A learning rate of 0.1 and a sigmoid transfer function were used in all the networks. They obtained a mean square error of less than 9 x 10^{-2} and converged up to 150 epochs. The networks used the multi-label classification as a categorization strategy, allowing for the same text to be classified in more than one category.

The results have shown that the precision significantly increases with neural networks. Precision reached 0.3 in *science* category and in *sports* it was close to 0.99. Figure 4 shows precision, recall and F1 when using neural networks and Figure 5 shows them when using the k-NN algorithm.

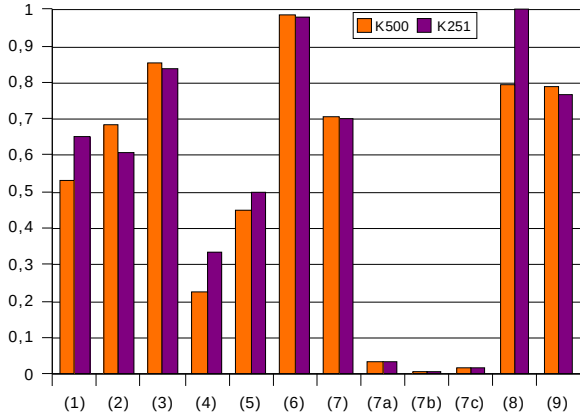

Figure 4. Performance of the Neural Categorizer, using 500 keywords (K500) and 251 (K251) on the whole *corpus*.

10% of the archives which had been classified by the k-NN as *rural area* were carefully read. This reading revealed that 45% of the texts were correctly classified. Though of a low percentage, not all the archives from *agrofolha* were, in fact, from the *rural area* category. Moreover, the large majority of the texts from the *money* section identified by the classifier are, actually, from the *rural area*. In the case of the subcategories *agriculture* and *livestock*, it was observed, through reading, that 57% of the texts classified as *agriculture* were, in fact, about this topic. In the *livestock* subcategory, on the other hand, we had a rate of 45% of the texts classified correctly as *livestock*. In order to compare these results, we read the archives classified by the Neural Network and we observed that the texts which had been correctly classified coincide, in the large majority, with those classified by the k-NN.

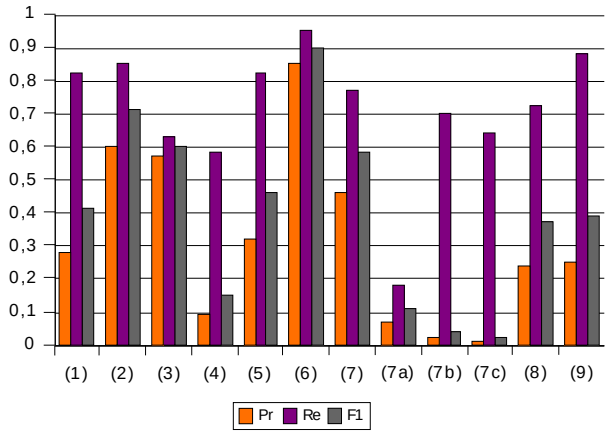

Figure 5. Performance of the k-NN Categorizer, using 500 keywords from the training set.

5. Conclusions

The methodology of keywords selection, linguistically motivated with the help of WSTools bring better precision, when using neural networks than when using the k-NN. Despite of using a relatively small training set (only 2,896 texts) the neural network has provided satisfactory results when precision is calculated.

We believe that the worse results can be related to language evolution and language variation through time, since in a *corpus* of journalistic texts as the PLN-BR CATEG *corpus*, news vary and the vocabulary changes year by year.

We recognize that precision can be even better and we believe that domain ontology could be a very useful support. Domain ontology is a semantic resource that represents a set of concepts and relationships between these concepts in a specific domain. Using concepts instead of only words might improve hierarchical categorization of texts. Thus, among our future perspectives is the task of developing a conceptual map which could help us to better understand the text we are working with.

6. Acknowledgements

This project was supported by CNPq.

7. References

Ceci, M. and Malerba, Donato (2007). Classifying web documents in a hierarchy of categories: a comprehensive study. In Journal of Intelligent Information Systems Volume 28, Issue 1 (February 2007), p. 37 – 78.

Gonçalves, T.C.F. (2007) Utilização de Informação Lingüística na classificação de documentos em Língua Portuguesa. Tese de Doutorado. Departamento de Informática. Universidade de Évora.

Gonzalez, M., Lima, V.L.S. and Lima, J.V. (2006). Tools for Nominalization: an Alternative for Lexical Normalization. In Workshop on Comp. Proc. Of Portuguese Language – Written and Spoken, 7; PROPOR, Proceedings…, Springer-Verlag, p.100-109.

Hao, P.; Chiang, J. Tu, Y. (2007). Hierarchically SVM classification based on support vector clustering method and its application to document categorization. In Expert Systems and Applications Volume 33, Issue 3, October 2007, p. 627-635.

Haykin, S. (1998). Neural Networks: a Comprehensive Foundation. 2nd edition. Prentice Hall. PTR. 842 pgs.

Langie, L. C. (2004). Um Estudo sobre a Aplicação do algoritmo k-NN à Categorização Hierárquica de Textos. Dissertação de Mestrado. Faculdade de Informática, PUCRS, 126 p.

Lavelli, A. ; Sebastiani, F. e Zanoli, R. (2004) Distributional term representations: an experimental comparison. *Proceedings of CIKM-04, 13th ACM International Conference on Information and Knowledge Management*, Washington, US, pp. 615-624.

Moraes, S.M.W., Lima, V.L.S. (2007). Um estudo sobre categorização hierárquica de uma grande coleção de textos em língua portuguesa. In Anais XXVI Congresso da SBC. TIL V Workshop em Tecnologia da Informação e da Linguagem Humana. Rio de Janeiro.

Rayson, P. and Garside, R. (2000) Comparing corpora using frequency profiling. In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp.1–6.Availablein: http://ucrel.lancs.ac.uk/llwizard.html.

Ruiz, M. Srinivasan, P. (2002). Hierarchical Text Categorization Using Neural Networks. In Journal of Information Retrieval, Volume 5, Number 1, January, Springer.

Sebastiani, F. (2006) Classification of text, automatic, In Keith Brown (ed.), The Encyclopedia of Language and Linguistics, vol. 14, 2a edição, Elsevier Science Publishers, Amsterdam, NL, p. 457-462.

Scott, M. (1996). WordSmith Tools. Oxford: Oxford University Press.

Sun, A. e Lim, E. (2001) Hierarchical Text Classification and Evaluation, In: IEEE International Conference on Data Mining, Proceedings…, Califórnia, USA, p.521-528.

Vilares, J., Barcala, F.M., Alonso, M.A. (2002) Using Syntactic dependency-pairs conflation to improve retrieval performance in Spanish. Computational Linguistics and Intelligent Text Processing, Springer-Verlag, Lectures Notes in Computer Science, 2276, pp 381—390.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), p. 42-49.