Mining Linguistically Interpreted Texts

Cassiana Fagundes da Silva, Renata Vieira,
Fernando Santos Osório
PIPCA - Unisinos
Av. Unisinos, 950 - São Leopoldo, RS
Brasil – 93.022-000
{cassiana, renata, osorio}@exatas.unisinos.br

Paulo Quaresma
Departamento de Informática,
Universidade de Évora, 7000
Évora - Portugal
{pq}@di.uevora.pt

Abstract

This paper proposes and evaluates the use of linguistic information in the pre-processing phase of text mining tasks. We present several experiments comparing our proposal for selection of terms based on linguistic knowledge with usual techniques applied in the field. The results show that part of speech information is useful for the pre-processing phase of text categorization and clustering, as an alternative for stop words and stemming.

1 Introduction

Natural language texts can be viewed as resources containing uniform data in such a way that methods similar to those used in Data Base Knowledge Extraction can be applied to them. The adaptation of these methods to texts is known as Text Mining (Tan, 1999). Machine learning techniques are applied to document collections aiming at extracting patterns that may be useful to organize or recover information from the collections. Tasks related to this area are text categorization, clustering, summarization, and information extraction. One of the first steps in text mining tasks is the pre-processing of the documents, as they need to be represented in a more structured way.

Our work proposes a new technique to the pre-processing phase of documents and we compare it with usual pre-processing methods. We focus on two text mining tasks, namely text categorization and clustering. In the categorization task we associate each document to a class from a pre-defined set, in the clustering task the challenge is to identify groups of similar documents without being aware of pre-defined classes. Usually, the pre-processing phase in these tasks are based on the approach called bag-of-words, in which just simple techniques are used to eliminate uninteresting words and to reduce various semantically related terms to the same root (stop-words and stemming, respectively). As an alternative, we propose the use of linguistic information in the pre-processing phase, by selecting words according to their category (nouns, adjectives, proper names, verbs) and using its canonical form. We ran a series of experiments to evaluate this proposal over Brazilian Portuguese texts.

This paper is organized as follows. Section 2 presents an overview of text mining. Section 3 presents the methods used for collecting the linguistic knowledge used in the experiments. The experiments themselves are described in Section 4. Section 5 presents an analysis of the results and the paper is concluded in Section 6.

2 Text Mining

Text mining processes are usually divided in five major phases: A) *Document collection*: consists of the definition of the set of the documents from which knowledge must be extracted. B) *Pre-processing*: consists of a set of actions that transform the set of documents in natural language into a list of useful terms. C) *Preparation and selection of the data*: consists in the identification and selection of relevant terms form the pre-processed ones. D) *Knowledge Extraction*: consists of the application of machine learning techniques to identify patterns that can classify or cluster the documents in the collection. E) *Evaluation and interpretation of the results*: consists of the analysis of the results.

The pre-processing phase in text mining is essential and usually very expensive and time consuming. As texts are originally non-structured a series of steps are required to represent them in a format compatible with knowledge extraction methods and tools. The usual techniques employed in phase B are the use of a list of stop-words, which are discarded from the original documents and the use of stemming which reduces the words to their root.

Having the proper tools to process Portuguese texts, we investigate whether linguistic information can have an impact on the results of the whole process. In the next section we describe the tools we used for acquiring the linguistic knowledge in which we base our experiments.

3 Tools for acquiring linguistic knowledge

The linguistic knowledge we use in the experiments is based on the syntactic analysis performed by the PALAVRAS parser (Bick, 2000). This Portuguese parser is robust enough to always give an output even for incomplete or incorrect sentences (which might be the case for the type of documents used in text mining tasks). It has a comparatively low percentage of errors (less than 1% for word class and 3-4% for surface syntax) (Bick, 2003). We also used another tool that makes easier the extraction of features from the analyzed texts: the Palavras Xtractor (Gasperin et. al. 2003). This tool converts the parser output into three XML files, containing: a) the list of all words from the text and their identifier; b) morpho-syntactic information for each word; c) the sentence´s syntactic structures. Using XSL (*eXtensible Stylesheet Language*)[1] we can extract specified terms from the texts, according to their linguistic value. The resulting lists of terms according to each combination are then passed to phases C, D and E. The experiments are described in detail in the next section.

4 Experiments

4.1 Corpus

The corpus used in the experiments is composed by a subset of the NILC corpus (Núcleo Interdisciplinar de Lingüística Computacional[2]) containing 855 documents corresponding to newspaper articles of *Folha de São Paulo* from 1994. These documents are related to five newspaper sections: informatics, property, sports, politics and tourism.

4.2 Pre-processing techniques

We prepared three different versions of the corpus (V1, V2 and V3) for 3-fold cross validation. Each version is partitioned in different training and testing parts, containing 2/3 and 1/3 of the documents respectively.

For the experiments with the usual methods, irrelevant terms (stop-words) were eliminated from the documents, on the basis of a list of stop-words, containing 476 terms (mainly articles, prepositions, auxiliary verbs, pronouns, etc). The remaining terms were stemmed according to Martin Porter's algorithm (Porter, 1980). Based on these techniques we generated a collection of pre-processed documents called **PD1.**

To test our proposal we then pre-processed the 855 documents in a different way: we parsed all texts of our corpus, generating the corresponding XML files and extracted terms according to their grammatical categories, using XSL. Based on these techniques we generated a collection of pre-processed documents called **PD2.**

4.2.1 Other mining phases

All other text mining phases were equally applied to both PD1 and PD2. We used relative frequency for the selection of relevant terms. The representation of the documents was according to the vector space model. For the categorization task, vectors corresponding to each class were built, where the more frequent terms were selected. After that, a global vector was composed. We also tested with different numbers of terms in the global vector (30, 60, 90, 120, 150). For the clustering task we measured the similarity of the documents using cosine. After calculating similarity of the documents, the data was codified according to format required by the machine learning tool Weka (Witten, 2000). Weka is a collection of machine learning algorithms for data mining tasks that contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

In this work the adopted machine learning techniques are Decision Tree for the categorization process and K-means for text clustering.

Decision Tree is a supervised learning algorithm based on the recursive division of the training examples in representative subsets, using the metric of information gain. After the induction of a classifying tree, it can be applied to new examples, described with the same attributes of the training examples.

K-means divides a group of objects in k groups in a way that the resulting intracluster similarity is high, but the intercluster similarity is low. The similarity of groups is measured in respect to the medium value of the objects in a group, which can be seen as the center of gravity (centroid) of the group. The parameters used to run k-means are the default ones as suggested by the tool, seed 10 and 5 groups.

The evaluation of the results for the categorization task is based on the classification error, which was used to compare the results for PD1 and PD2. For the clustering task the evaluation of the results is given by recall and precision, based on the generated confusion matrices.

[1] Available in http://www.w3.org/Style/XSL/

[2] Available in http://nilc.icmc.sc.usp.br/nilc/

5 Results

5.1 Text Categorization

Table 1 shows the results for text categorization of PD1, given by the average error rates considering the three versions the corpus (V1, V2 and V3). We had around 20% of error for the categorization task. We can see minor variations in the results according to the size of the vectors. Best results were obtained for 150 terms.

Terms	30	60	90	120	150
Errors	21,64	21,99	20,47	20,35	**19,77**

Table 1: Average Classification Error for PD1%

Table 2 shows the results for different grammatical combinations in PD2, while Figure 1 summarizes the lowest error rates found for PD1 and all groups of PD2. The group nouns and adjectives presents the lower error rates of all experiments (18,01). However, due to the small size of the corpus, the improvement reported between usual methods (18,01) and nouns-adjectives (20,47), when considering the same number of terms (90), are at 75-80% confidence level only (t-test).

In general, the results show that the presence of nouns is crucial, the worst classification errors are based on groups that do not contain the category nouns, and here the confidence level for the differences reported reaches 95%. The groups containing nouns present results comparable to those found in the experiments based on usual methods of pre-processing. The use of verbs, either alone or with other grammatical groups is not an interesting option.

Terms	30	60	90	120	150
Nouns	24,91	21,75	23,98	23,51	22,69
Nouns-adjec.	23,15	20,35	**18,01**	19,18	18,71
Nouns-adjec.-proper names	20,82	22,92	20,94	21,05	21,17
Nouns-proper names	24,09	24,56	22,80	22,45	22,80
Adjec.-proper names	47,01	46,34	32,51	33,21	32,86
Verbs	63,73	62,33	57,75	58,45	55,64
Nouns-verbs	40	27,72	25,61	24,21	26,32
Nouns-verbs-adjectives	35,09	27,02	27,72	24,21	23,51

Table 2: Average Classification Error for PD2

It can be observed that usually the best results are obtained when the documents are represented by a larger number of terms (90, 120 and 150), for the group nouns, however, the best results were obtained for vectors containing just 60 terms.

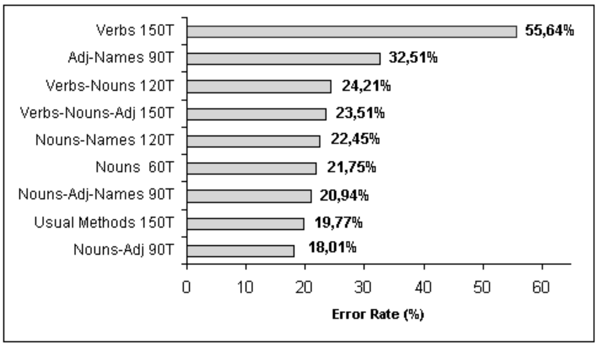

Figure 1: Lower error rates for PD1 and PD2

We looked at the terms resulting from different selection methods and categories to check the overlap among the groups. From PD1 to PD2 based on nouns and adjectives (the one with the best results) we could see that we had around 50% of different terms. That means that 50% of terms in PD1 are terms included in the categories nouns and adjectives and 50% of the terms selected on the basis of stop-words and stemming are from other grammatical categories. As adjectives added to nouns improved the results, we checked adjectives to figure out their significance. We found terms such as Brazilian, electoral, multimedia, political. Intuitively, these terms seem to be relevant for the classes we had. Analysing the groups containing verbs, we observed that the verbs are usually very common or auxiliary verbs (such as to be, to have, to say), therefore not relevant for classification.

5.2 Text Clustering

We tested our hypothesis through clustering experiments for PD1 and variations of PD2. For the experiments on clustering we used vectors containing 150 features from V2 and we set k to 5 groups. The resulting confusion matrix for PD1 is presented in Table 3.

	Cl.0	Cl.1	Cl.2	Cl.3	Cl.4
Sp.	1	**31**	2	0	23
Prop.	**2**	0	4	0	51
Inf.	0	0	1	0	**55**
Pol.	0	0	2	**39**	16
Tour.	5	0	**17**	0	33

Table 3: Confusion Matrix PD1 (150 terms)

Considering the larger group in each row and column (highlighted in the table) as the intended cluster for each class, the corresponding precision is of 50,52%.

We repeated the same set of experiments for PD2. We tested several grammatical groups, the best result was related to nouns and proper names. The results are shown in Tables 4. The corresponding precision is 63,15%.

	Cl.0	Cl.1	Cl.2	Cl.3	Cl.4
Sp.	0	**38**	19	0	0
Prop.	**11**	0	44	1	1
Inf.	0	0	19	0	**38**
Pol.	0	1	20	**36**	0
Tour.	0	0	**57**	0	0

Table 4: Confusion Matrix PD2 (nouns + proper names, 150 terms)

6 Conclusions

This paper presented a series of experiments aiming at comparing our proposal of pre-processing techniques based on linguistic information with usual methods adopted for pre-processing in text mining.

We find in the literature other alternative proposals for the pre-processing phase of text mining. (Gonçalves and Quaresma, 2003) use the canonical form of the word instead stemming, for European Portuguese. (Feldman et al, 1998) proposes the use of compound terms as opposed to single terms for text mining. Similarly, (Aizawa, 2001) uses morphological analysis to aid the extraction of compound terms. Our approach differs from those since we propose single terms selection based on different part of speech information.

The results show that a selection made solely on the basis of category information produces results at least as good as those produced by usual methods (when the selection considers nouns and adjectives or nouns and proper nouns) both in categorization and clustering tasks. In the categorization experiments we obtained the lowest error rate for PD2 when the pre-processing phase was based on the selection of nouns and adjectives, 18,01%. However, the second best score in the case of categorization was achieved by the traditional methods, 19,77%. Due to the small corpus, further experiments are needed to verify the statistical significance of the reported gains. The results of the clustering experiments show a difference in precision from 50,52% to 63,15%. As we are planning to test our techniques with a larger number of documents and consequently a larger number of terms, we are considering applying other machine-learning techniques such as Support Vector Machines that are robust enough to deal with a large number of terms. We are also planning to apply more sophisticated linguistic knowledge than just grammatical categories, as, for instance, the use of noun phrases for terms selection, since this information is provided by the parser PALAVRAS. Other front for future work is further tests for other languages.

References

Aizawa A., 2001. Linguistic Techniques to Improve the Performance of Automatic Text Categorization. *Proc. of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 307-314.

Bick, E. 2000. *The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework*. Århus University. Århus: Århus University Press.

Bick, E. 2003. A Constraint Grammar Based Question Answering System for Portuguese. *Proceedings of the 11º Portuguese Conference on Artificial Intelligence*, pages 414-418. LNAI Springer Verlag.

Feldman R., et al. 1998. Text Mining at the Term Level. *Proc. of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 65-73. LNCS Springer.

Gasperin, C.; Vieira, R.; Goulart, R. and Quaresma, P. 2003. Extracting XML Syntactic Chunks from Portuguese Corpora. *Proc. of the TALN Workshop on Natural Language Processing of Minority Languages and Small Languages*, pages 223-232. Batz-sur-Mer France.

Gonçalves, T. and Quaresma, P. 2003. A prelimary approach classification problem of Portuguese juridical documents. *Proceedings of the 11º Portuguese Conference on Artificial Intelligence*, pages 435-444. LNAI Springer Verlag.

Porter, M. F. 1980. An Algorithm for Suffix Stripping. *Program*, 14 no. 3, pages 130-137.

Tan, Ah-Hwee. 1999. Text mining: the state of the art and the challenges. *Proc. of the Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases*, pages 65-70, Beijing.

Witten, I. H. 2000. *Data mining: Pratical Machine Learning tools and techniques with Java implementations*. Academic Press.