ARTÍCULO

Evaluating the use of linguistic information in the pre-processing phase of Text Mining

Cassiana Fagundes da Silva, Fernando Santos Osório, Renata Vieira

PIPCA - UNISINOS
Av. Unisinos, 950
São Leopoldo, RS - 93.022-000 - Brasil
{cassiana,osorio,renata}@exatas.unisinos.br

Abstract

This work proposes and evaluates the use of linguistic information in the pre-processing phase for text mining tasks applied to Portuguese texts. We present several experiments comparing our proposal to the usual techniques applied in the field. The results show that the use of linguistic information in the pre-processing phase brings some improvements for both text categorization and clustering.

Keywords: Text Mining, Linguistic Information, Pre-processing, Text Categorization and Text Clustering

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, 99]. 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 term extraction. One of the first steps for a text mining task is the pre-processing of the documents, since texts 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 [Yang and Pederson, 97]. In the clustering task the challenge is to identify groups of similar documents without being aware of pre-defined classes [Theodoridas and Koutroumbas, 98].

Usually, the pre-processing phase in these tasks is based on the approach called bag-of-words, in which simple techniques are used to eliminate unimportant 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.

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.

2. Overview of Text Mining

Text mining processes are usually divided in five major phases:

– **Document collection:** consists of the

definition of the set of the documents from which knowledge must be extracted.

- **Pre-processing:** consists of a set of actions that transform the set of documents in natural language into a list of useful terms.
- **Preparation and selection of the data:** consists in the identification and selection of relevant terms from the pre-processed ones.
- **Knowledge Extraction:** consists of the application of machine learning techniques to identify patterns that can classify or cluster the documents in the collection.

Evaluation and interpretation of the results: consists of the analysis of the results.

The pre-processing for text mining is an essential and usually very expensive phase. As texts are originally non-structured a series of steps are required to represent them in a format compatible for knowledge extraction. However, much of the research in this area is concentrated in phases preparation and selection of the data and knowledge extraction. Usually, few methodological changes can be observed in phase pre-processing where the usual techniques employed are the use of a list of stopwords, which are discarded from the original documents, and the use of stemming, which reduces the words to their root. Sometimes, if a dictionary is available, instead of using the stemming, one can substitute each word by its canonical form (singular, masculine for nouns and the infinitive for verbs) [Gonçalves and Quaresma, 03]. In this work we focus on the analysis of phase pre-processing in relation to the Portuguese language. 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 on which we base our experiments.

3. Tools for linguistic knowledge

The linguistic knowledge we use in the experiments is based on the results of the syntactic analysis performed by the PALAVRAS parser [Bick, 00]. 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). Once the texts were parsed, we were able to select terms based on their grammatical categories. The canonical form of the words was also available. Figure 1 shows the parser output for the sentence *"Janeiro começa com grandes liquidações"* (January beggins with great sales).

```
STA:fcl
=SUBJ:n('janeiro' M S)  Janeiro
=P:v-fin('começar' PR 3S IND) começa
=ADVL:pp
==H:prp('com')   com
==P<:np
===>N:adj('grande' F P) grandes
===H:n('liquidação' F P)liquidações
=.
```

Figure 1. PALAVRAS output

We also used another tool that makes easier the extraction of features from the analyzed texts: the Palavras Xtractor [Gasperin et. al., 03]. This tool converts the parser output into three XML files containing i) (Figure 2), ii) morpho-syntactic information for each word listed in (Figure 3) and iii) the sentence structures (Figure 4).

```
<words>
<word id="word_1">Janeiro</word>
<word id="word_2">começa</word>
<word id="word_3">com</word>
<word id="word_4">grandes</word>
<word id="word_5">liquidações</word>
<word id="word_6">.</word>
</words>
```

Figure 2. Words

Using XSL[1] (*eXtensible Stylesheet Language*) we can extract specified terms from the texts, according to their linguistic value. In our work we extract the following combination of terms (each combination corresponding to one experiment): nouns; nouns and adjectives; nouns and proper names; nouns, adjectives and proper names; adjectives and proper names; verbs; verbs and nouns. The resulting lists of terms according to each combination are then passed to phases preparation and selection of the data, knowledge extraction and evaluation and interpretation of the results. The experiments are described in detail in the next section.

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

```
<words>
<word id="word_1">
<n canon="janeiro" gender="M" number="S"/>
</word>
<word id="word_2">
<v canon="começar">
<fin tense="PR" person="3S" mode="IND"/>
….
```

Figure 3. Part-of-Speech

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1"
span="word_1..word_6">
<chunk id="chunk_1" ext="sta" form="fcl"
span="word_1..word_5">
…
```

Figure 4. Chunks

4. Experiments

Several experiments were undertaken to evaluate different pre-processing approaches for the categorization and clustering tasks of text mining. We compared the usual techniques (based on stop-words and stemming) with our proposal of selecting terms according to certain grammatical categories. In our experiments we used a corpus composed by a subset of the NILC corpus (Núcleo Interdisciplinar de Lingüística Computacional[2]) containing 855 documents corresponding to newspaper articles from Folha de São Paulo of 1994. These documents are related to five newspaper sections: informatics, property, sports, politics and tourism. We prepared three versions of the same corpus (V1, V2 e V3), each version is partitioned in different training and testing parts, containing 114 (2/3) and 57 (1/3) of the documents by class respectively. All results presented in the paper are related to the average error rates considering these three versions (3-fold cross validation).

The total of 855 documents was pre-processed for testing the two different approaches. Irrelevant terms were eliminated from the documents, on the basis of a list of stop-words from European Portuguese[3], which was adapted to Brazilian Portuguese, containing 476 terms (mainly articles, prepositions, auxiliary verbs, pronouns, etc). The remaining terms were stemmed according to Martin Porter's[4] algorithm, which removes the final letters from the words according to a set of rules, avoiding the presence of words that are the same but vary in gender, number or inflection. We will refer to this first set of pre-processed documents as PD1.

To test our proposal we then pre-processed the documents again but in a different way: we parsed the texts, generated the corresponding XML files and extracted terms according to their grammatical categories, using XSL. The resulting pre-processed documents following this last method will be called PD2. The

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 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 by the union of the local vectors. We also tested with different numbers of terms in the global vectors (30, 60, 90, 120, 150). We used Weka's [Witten, 00] implementation of the following machine learning algorithms: decision trees, artificial network neural for categorization and k-means for clustering.

For the clustering task we measured the similarity of the documents using cosine. After calculating similarity of the documents, the entry was generated according to the ARFF format required by Weka. The parameters used to run k-means are: random number seed equal 10 and set number of cluster equal 5.

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 based on the confusion matrix (as we can see the examples pertaining to each group identified through k-means) and their corresponding Recall and Precision tables.

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

[3] Provided by Paulo Quaresma from the University of Évora.

[4] Available in http://snowball.sourceforge.net

5. Results

Here we present the results of our experiments for PD1 and PD2. We first present the results for categorization and after that the results for clustering.

We applied multiple categorization to our corpus, where a class from a set of classes is defined for each document. We used Decision Trees (DT) and Artificial Network Neural (ANN) - Multi-layer Perceptron using Backpropagation - as the machine-learning algorithms, both implemented in Weka. We tested several variations on the number of selected terms, considering the most frequent 30, 60, 90, 120 and 150 to constitute the document vectors.

5.1 Text Categorization

5.1.1 Pre-processing documents according to usual methods (PD1)

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.

Table 2 presents the experiments using ANN MLP-BP with momentum 0.9, learning rate 0.1 and number of neurons 8 and 16. For each topology, 10 simulations, with different seeds were made.

Table 1. Average Classification Error (%) for DT in PD1

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

The lowest error rate (14,64%) was obtained on the basis of 8 neurons in the intermediate layer and 90 terms. There is, however, a large variation in the error according to the number of entries.

Table 2. Average Classification Error (%) for ANN in PD1

	# Terms				
Neurons	30	60	90	120	150
8	20,26	16,56	14,64	32,32	55,37
16	20,68	16,74	16,19	24,28	54,23

The two techniques are compared in Table 3. We can see that the lowest error is obtained on the basis of ANN MLP-BP. However the general results are very instable, whereas decision trees present a more stable result overall. If we take the average for all experiments we have 20,84% e 27,12%, for decisions trees and ANN respectively.

Table 3. Comparing DT and ANN in PD1

Technique	Lowest error	Overall average error
DT	19,18%	20,84%
ANN MLP-BP	14,64%	27,12%

5.1.2 Pre-processing documents based on linguistic information (PD2)

Table 4 shows the results (%) for different grammatical groups in PD2 using DT. We considered the following grammatical categories:

- Nouns (*nn*)
- Nouns and adjectives (*nn* + *adj*)
- Nouns, adjectives and proper nouns (*nn* + *adj* +*prpr*)
- Nouns and proper nouns (*nn* + *prp*)
- Nouns and verbs (*nn* + *vrb*)
- Nouns, adjectives and verbs (*nn* + *adj* + *vrb*)
- Verbs (*vrb*)
- Adjectives and proper nouns (*adj* + *prp*)

The group *nn* + *adj* 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 (20,47%) and *nn* + *adj* (18,01%), when considering the same number of terms (90), are at 75-80% confidence level only (t-test).

Table 4. Avarage Classification Error (%) for grammatical groups in PD2 using DT

# Terms	30	60	90	120	150
nn	24,91	21,75	23,98	23,51	22,69
nn + adj	23,15	20,35	18,01	19,18	18,71
nn + adj + prp	20,82	22,92	20,94	21,05	21,17
nn + prp	24,09	24,56	22,80	22,45	22,80
adj + prp	47,01	46,34	32,51	33,21	32,86
vrb	63,73	62,33	57,75	58,45	55,64
nn + vrb	40	27,72	25,61	24,21	26,32
nn + adj + vrb	35,09	27,02	27,72	24,21	23,51

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.

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.

Table 5 shows the lowest error rates for PD1 and for all groups of 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 nn + adj (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 *nn + adj* and then other 50% 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.

Table 5. Lower Error Rates for PD1 and PD2

Groups	# Terms	Error Rate (%)
nn + adj	90	18,01
nn + adj +prp	90	20,94
nn	60	21,75
nn + prp	120	22,45
nn + adj + vrb	150	23,41
nn + vrb	120	24,21
adj + prp	90	32,51
PD1	150	19,77

Results obtained with ANN are presented in Table 6. Here the grammatical group with the lowest error rate is *nn + adj +prp.*

Table 6. Average Classification Error (%) for ANN in PD2

#Terms	30	60	90	120	150
nn-8	22,85	19,49	21,54	48,01	71,13
nn-16	23,09	19,22	20,27	54,38	67,54
nn+adj-8	21,86	19,66	21,59	34,82	56,38
nn+adj-16	22,56	19,18	19,05	32,35	56,52
nn+prp-8	22,27	19,18	21,62	49,15	65,78
nn+prp-16	22,94	19,77	19,01	38,82	66,58
nn+adj+prp-8	21,39	18,00	17,55	33,29	56,46
nn+adj+prp-16	22,44	16,96	18,27	37,98	54,21
adj+prp-8	49,45	38,25	35,57	45,17	72,77
adj+prp-16	48,53	38,62	37,75	47,63	74,64

Table 7 shows a summarized comparison between different grammatical groups using ANN, taking into account the lowest error obtained rates.

Table 7. Lowest Error for ANN in PD1 and PD2

Groups	# Terms	# Neurons	Error Rate (%)
adj + prp	90	8	35,57
nn + adj	90	16	19,05
nn + prp	90	16	19,01
nn + adj +prp	60	16	16,96
nn	60	16	29,22

Table 8 presents the overall average error by grammatical category. The best grammatical combination is (*nn + adj + prp*), however, the usual methods present the lowest error overall.

Table 8. Overall average error for ANN in PD1 and PD2

Groups	Average Error Rate (%)
nn + adj	19,88
nn + adj +prp	21,38
nn	23,37
nn + prp	23,34
nn + adj + vrb	24,96
nn + vrb	26,02
vrb	56,83
adj + prp	38,39
PD1	20,84

Finally, Table 9 presents the comparison between DT and ANN for classification in PD2. The results obtained with decision tree have shown more stability and consequently better overall average results in the whole set of experiments, varying from 19,88% for *nn + adj*, to 21,38% for *nn + adj + prp*, whereas the ANN present an overall average of 29,97% for the same group.

Table 9. Comparing ANN and DT in PD2

Technique	Category	Lowest error	Overall average error
DT	*nn+adj*	18,01%	19,88%
	nn+adj+prp	20,82%	21,38%
ANN MLP-BP	*nn+adj+prp*	16,96%	29,97%

5.2 Text Clustering

Although clustering is a technique to be applied to a set of documents with no previous classification, we wanted to verify if a clustering algorithm such as k-means could identify groups according to the classes we knew in advance, and how would it differ according to different pre-processing methods. We, therefore, tested our hypothesis through clustering experiments for PD1 and variations of PD2. As we can see below, k-means was able to identify some of the groups according to our previous classification for PD1 and some more for PD2.

5.2.1 Pre-processing documents according to usual methods (PD1)

For the experiments on clustering we used vectors containing 30, 90 and 150 features, and three versions of the corpus as before (V1, V2, V3). We had a total of 9 experiments, our random seed was 10 and we tested k-means for 5 groups. From these 9 runs, k-means was able to identify clearly only two groups (Politics and Sports) with 150 terms, is presented in Table 10.

Table 10. Confusion matrix PD1 (150 terms)

	Cl. 0	Cl. 1	Cl. 2	Cl. 3	Cl. 4
Sport	1	31	2	0	23
Property	2	0	4	0	51
Informatics	0	0	1	0	55
Politic	0	0	2	39	16
Tourism	5	0	17	0	33

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

5.2.2 Pre-processing documents based on linguistic information (PD2)

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 show in Table 11. The corresponding overall precision is 63,15%.

Table 11. Confusion Matrix PD2 (group nn + prp, 90 terms, V1)

	Cl. 0	Cl. 1	Cl. 2	Cl. 3	Cl. 4
Sport	0	38	19	0	0
Property	11	0	44	1	1
Informatics	0	0	19	0	38
Politic	0	1	20	36	0
Tourism	0	0	57	0	0

K-means identified also 3 groups for V1 with 150 terms for other two grammatical groups: nouns and adjectives, and nouns, proper names and adjectives, all for the classes sports, politics and informatics. We tested a grammatical combination without nouns (only proper names and adjectives) but only one group was identified. We tried also to cluster the documents for k larger than 5 (6, 7 and 8) but again only sports, informatics and politics were clustered into the same group.

Table 12 presents precision and recall of the clusters in PD1 and PD2, respectively.

Table 12. Precision and Recall in PD1 and PD2

	PD1			PD2		
	Precision	Recall	F	Precison	Recall	F
Cl.0	0,28	0,03	0,05	1	0,19	0,32
Cl.1	1	0,54	0,70	0,97	0,67	0,79
Cl.2	0,65	0,30	0,41	0,36	1	0,53
Cl.3	1	0,68	0,81	0,97	0,63	0,77
Cl.4	0,31	0,96	0,47	0,97	0,67	0,79

We can see that PD2 presented in general better precision and recall when compared to PD1.

6. Conclusion and future works

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 pre-processed a corpus according to the two approaches, resulting in different documents representations that we call PD1 (usual methods) and PD2 (using linguistic information).

We find in the literature other alternative proposals for the pre-processing phase of text mining. Once in the use of canonical form of the word instead stemming, for European Portuguese [Gonçalves and Quaresma, 03]. [Feldman et. al., 98] proposes the use of compound terms as opposed to single terms for text mining. Similarly, [Aizawa, 01] 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% for PD1 to 63,15% for PD2.

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.

7. Acknowledgments

This work was supported by CNPq-Brazil.

References

[Aizawa, 01] Aizawa A.: Linguistic Techniques to Improve the Performance of Automatic Text Categorization. In: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (2001), 307-314.

[Bick, 00] Bick, E.: The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework. Århus University. Århus: Århus University Press (2000).

[Feldman et al, 98] Feldman R., et al.:Text Mining at the Term Level. In: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, LNCS Springer (1998), 65-73.

[Gasperin et. al., 03] Gasperin, C. et al: Extracting XML Syntactic Chunks from Portuguese Corpora. In: Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages - Batz-sur-Mer France June 11 – 14, (2003).

[Gonçalves and Quaresma, 03] Gonçalves, T.; Quaresma, P.: A prelimary approach to the multilabel classification problem of Portuguese juridical documents. In: 11º Portuguese Conference on Artificial Intelligence. Lectures Notes in Artificial Intelligence. Berlim: Springer Verlag, (2003).

[Tan, 99] Tan, Ah-Hwee: Text mining: the state of the art and the challenges. In: Pacific-Asia Workshop on Knowledge Discovery from

Advanced Databases-PAKDD'99, Beijing, April (1999), 65-70.

[Theodoridas and Koutroumbas, 98] Theodoridas, S. and Koutroumbas, K.: Pattern Recognition. Academic Press (1998), 351-459.

[Witten, 00] Witten, I. H.: Data mining: Pratical Machine Learning tools and techniques with Java implementations. Academic Press (2000).

[Yang and Pederson, 97] Yang, Y; Pederson, J.: A Comparative Study on Feature Selection in Text Categorization. In Proceedings of 14th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US (1997), 412-420.