Analyzing the Effect of Dimensionality Reduction in Document Categorization for Basque

Zelaia, A. and **Alegria, I.** and **Arregi, O.** and **Sierra, B.**
University of the Basque Country, UPV-EHU
Computer Science Faculty, 649 postakutxa
20.080 Donostia, Gipuzkoa, Euskal-Herria, Spain
{ccpzejaa, acpalloi, acparuro, ccpsiarb}@si.ehu.es

Abstract

This paper analizes the incidence that dimensionality reduction techniques have in the process of text categorization of documents written in Basque. Classification techniques such as Naïve Bayes, Winnow, SVMs and k-NN have been selected. The SVD (Singular Value Decomposition) dimensionality reduction technique together with lemmatization and noun selection have been used in our experiments. The results obtained show that the approach which combines SVD and k-NN combined with the cosine similarity measure for a lemmatized corpus gives the best categorization of all with a remarkable difference.

Introduction

Since the early 90s, automated categorization of texts into predefined categories has increased interest because the amount of available documents in digital form are rising fast. Most researchers propose approaches based on machine learning techniques (Sebastiani 2002), where automatically built classifiers learn from a set of previously classified documents.

The work we are presenting here analyzes the categorization of documents written in Basque. Several experiments have been made to classify documents written in extended languages such as English. But, the reality of lesser-used languages, as is the case of Basque, is different. In practice, one of the main problems we encounter is that only a short amount of manually classified documents is available. This fact restricts the capacity of the classifiers and may, consequently, produce poorer results.

In addition to that problem, we must take into account that Basque is an agglutinative and highly inflected language whose declension system has numerous cases (Alegria *et al.* 1996). This morphosintactical feature makes the categorization task more difficult, because semantic information is not really contained in word-forms but in their corresponding lemma. Therefore, it seems interesting to preprocess the corpus lemmatizing it and so, at the same time the dimension of the information to treat is reduced and an improvement in the efficiency of the system can be produced.

In this paper we analyze the effect that dimensionality reduction techniques such as lemmatization, noun selection and in particular SVD (Singular Value Decomposition) have in the process of text categorization of Basque documents. Latent Semantic Indexing (LSI) (Deerwester *et al.* 1990) implementation has been used to calculate the SVD of the matrix constructed for the training corpus. We have selected some of the most popular classification algorithms and two different experiments have been performed. We use three different corpora in both experiments: words, lemmas and nouns. In the first experiment, the classification techniques are used without applying SVD. In the second one, the same classification techniques are used but previously, the SVD technique has been applied to reduce the dimension. Obtained results show that the SVD dimensionality reduction technique combined with the k-NN classification algorithm gives the best results. Moreover, we find that they are obtained for the lemmatized corpus.

This paper is structured as follows. First, we survey the previous work on algorithms used for document categorization, and examine the foundations of LSI. Afterwards, the experimental setup is introduced, where both training and test corpora are described and lemmatization, noun selection and document frequency based feature selection processes are introduced. In the next section, experimental results are shown, compared and discussed. Finally, conclusions and future work are presented.

Related Work

Text categorization consists in assigning predefined categories to text documents (Sebastiani 2002). Simple but effective, the bag-of-words text document representation is one of the most frequently used. In this kind of text representation, the number of attributes of the corpus is usually considerable, and this can be problematic in inductive classification. Therefore, it is usually convenient to apply techniques that reduce the dimension of the representation. This reduction can be carried out in different ways: eliminating irrelevant features (words), substituting some words by others that represent them (lemmas, synonyms, hyperonyms, etc.), applying SVD technique etc.

In our two experiments we use classification algorithms which have reported good results for text categorization in other languages, such as Naïve Bayes (Minsky 1961), Winnow (Dagan, Karov, & Roth 1997), SVMs (Joachims 1999) and k-NN (Dasarathy 1991). Next, we briefly describe the foundations of LSI, which uses SVD for dimensionality reduction.

SVD using Latent Semantic Indexing (LSI)

LSI[1] was first introduced in 1988 originally developed in the context of Information Retrieval (Dumais 2004). It takes as input a collection of texts composed of n documents and m terms and represents it as an $m \times n$ term-document matrix. The elements m_{ij} of the term-document matrix are the occurrences of each term i in a particular document j. This way, we obtain matrix M, where each document is represented by a vector in an m-dimensional space (Berry & Browne 1999).

The SVD technique compresses vectors representing documents into vectors of a lower-dimensional space. It consists in factoring matrix $M \in \mathbb{R}^{m \times n}$ into the product of three matrices, $M = U\Sigma V^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, where $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$ being $k = \min\{m, n\}$, and U and V are orthogonal matrices of singular vectors.

Once matrix M has been factored, it can be approximated by a lower rank M_p which is calculated using the p largest singular triplets of M. This operation is called dimensionality reduction, and the p-dimensional space to which document vectors are projected is called the reduced space. When using the reduced space generated by M_p instead of the one generated by M, most of the important underlying structure that associates terms with documents is captured and consequently, noise is reduced. This results in a representation where similar documents have similar vectors.

For text categorization purposes, LSI represents each of the documents to be categorized by a p-dimensional vector. Afterwards, the similarity among it and all the documents in the training set (reduced space) is calculated using the cosine similarity measure. LSI has been successfully used in the categorization of written documents (Pierre 2001).

Experimental Setup

The aim of this section is to describe the document collection used in our experiments and to give an account of the lemmatization, noun selection and document frequency based feature selection technique we have applied.

Document Collection

We are interested in the categorization of documents written in Basque. Among all the electronic documents available in Basque, we have selected newspaper texts, because there are standardized categories for this domain, and we have access to a sufficient amount of documents manually categorized. The documents used in this experiment correspond to the Basque newspaper *Euskaldunon Egunkaria* corresponding to the articles published during two months of 1999. They are a total of 6.064 documents categorized to the 17 standard first level IPTC categories[2]. Each of the documents has a unique category associated to it. It must be noted that all categories do not have the same number of documents, as we can see in Table 1.

[1] http://lsi.research.telcordia.com, http://www.cs.utk.edu/∼lsi

[2] http://www.iptc.org

Category	Training	Test
1. Culture	600	202
2. Law & Justice	129	42
3. Disasters	75	26
4. Economy	234	78
5. Education	82	27
6. Environmental Issues	69	22
7. Health	35	12
8. Human interests	36	11
9. Labour	132	43
10. Lifestyle	40	13
11. Politics	1.184	393
12. Religion	25	8
13. Science	35	12
14. Social Issues	464	156
15. Sport	1.283	429
16. Conflicts	100	33
17. Weather	25	9
TOTAL	4.548	1.516

Table 1: Number of documents distributed by categories.

Document categorization is achieved in two steps: during the *training* step an inductive generalization of the set of documents is obtained, and during the *test* step the effectivity of the system is measured. Therefore, the 6,064 documents have been split into two different sets of documents: 4,548 documents for training (75 %) and 1,516 documents for testing (25 %). This proportion stands in each one of the 17 categories, as can be observed in Table 1.

Feature selection

As we have mentioned in the introduction, Basque is an agglutinative and highly inflected language. In order to face the difficulties derived from these morphosintactical features, we have applied two types of feature selection. On the one hand, stopword lists have been used to eliminate non-relevant words, i.e. the most frequent words and words that appear less than a threshold in the training corpus ($>$1 doc, $>$2 doc, $>$3 doc).

On the other hand, we use linguistic methods such as lemmatization and noun selection to reduce the number of features. The studies of the effects that stemming algorithms produce in text categorization are controversial for languages with a low level of inflection such as English (Spitters 2000) (Riloff 1995), but recent experiments show that lemmatization helps in the process of categorization of documents written in an inflected language using LSI (Nakov, Valchanova, & Angelova 2003). Because of the morphosintactical features of Basque, we expect that lemmatization should allow us to maintain the same semantic information, reducing the number of attributes to be processed. We have used the lemmatizer designed by the IXA [3] group (Ezeiza *et al.* 1998), which obtains for each word in the document, its corresponding lemma, as well as its part-of-speech tag. This system reduces the different number of features from each

[3] http://ixa.si.ehu.es

category by more than 50%. While the number of different word-forms in the whole document collection is 92,373, there are 38,654 different lemmas, among which 14,213 are nouns. So, we have used three different corpora: bag-of-words (W), bag-of-lemmas (L) and bag-of-nouns (N).

Experimental Results

In order to evaluate the results obtained in our experiments, we have concentrated in effectiveness issues, rather than on efficiency ones, and calculate the accuracy rate for each categorization.

Without applying SVD

In this experiment, elimination of irrelevant words, lemmas and nouns has been performed based on the word frequency in documents. Different low thresholds (all, >1 doc, etc.) and a constant high threshold (in order to discard functional words/lemmas) are applied. The resulting number of attributes in the training corpora are shown in the top part of Table 2. The average accuracy using the test-corpus for each classification technique is shown in the rest part of the table. The best results obtained for each technique and corpus appear printed in boldface.

		all	>1 doc	>2 doc	>3 doc
Numb. of Attrib.	W	73728	33294	22821	17776
	L	34729	15175	10750	8542
	N	10381	7301	5913	5050
NB	W	**80.09**%	78.89%	78.10%	77.77%
	L	**81.53**%	81.07%	80.74%	80.28%
	N	79.49%	**79.62**%	79.35%	79.62%
Ww	W	80.09%	**81.13**%	80.47%	79.49%
	L	80.15%	**80.47**%	78.10%	77.77%
	N	**79.35**%	78.83%	76.78%	76.45%
SVMs	W	81.53%	82.72%	83.18%	**83.71**%
	L	84.10%	**84.56**%	83.58%	83.11%
	N	81.40%	**82.58**%	81.60%	81.99%
k-NN	W	37.80%	**54.75**%	38.32%	40.96%
	L	50.66%	40.11%	58.91%	**59.17**%
	N	61.08%	69.53%	70.84%	**72.16**%

Table 2: Accuracy without applying SVD

As we can see in Table 2, the best categorization has been obtained by using SVMs (Weka software (Witten & Frank 1999)) after removing words that appear in only 1 document (>1 doc) and using the lemmatized corpus. We want to enphasize that the accuracy rates obtained are high (83.71 %(W), 84.56%(L), 82.58%(N), taking into account the morphosintactical features of Basque and the reduced corpora used.

Results obtained using Naïve Bayes and Winnow are also very good. Both have been obtained using the general-purpose classifier named SNoW (Carlson *et al.* 1999), and we argue that the preprocessing they perform is very adequate for text categorization tasks. Both work better with more attributes, in general. Moreover, we can see that lemmatization and noun selection help Naïve Bayes in general, but this is not the case with Winnow, making results poorer.

However, the results obtained tell us that k-NN algorithm is not suitable for text categorization using raw data, even though noun selection gives acceptable accuracy rates (72.16 % the best). Results in the table have been obtained for different k values ranging between 1 and 10, and using the euclidean distance.

Finally, we want to state that most of the best accuracy rates have been obtained by eliminating words that only appear in a document (>1 doc case).

Applying SVD

In this second experiment, LSI has been used to create the three reduced spaces for the training document collections. The sizes of the training matrices created are 34288×4548 (W), 14648×4548 (L) and 7209×4548 (N). Different number of dimensions have been experimented ($p = 100, 200, 300, 400, 500, 1000$). The weighting scheme used was logarithm for local weighting and entropy for global one.

When using k-NN, different experiments for different number of neighbours ($k = 1, \cdots, 10$) have been made and the following criteria has been followed: regarding the categories of the k closest (with the highest cosine), the most frequent one was selected. In case the result is a tie, the category with the highest mean is chosen.

		LSI dim.	Accuracy
SVD+SVMs	W	1000	75.00%
	L	500	**81.46**%
	N	500	80.34%
SVD+k-NN	W	300	84.89%
	L	400	**87.33**%
	N	200	85.36%

Table 3: Accuracy for different methods using SVD

The best results in this experiment have been obtained by using k-NN to categorize the documents in the reduced space. In Table 3 the best results for each corpus are shown, and it can be observed that, using k-NN they are all superior to the best results obtained in the previous experiment for each of the corpus. The highest accuracy has been obtained for the lemmatized corpus, which significantly improves and raises up to 87.33 %. This confirms our hypothesis that lemmatization helps improving results in agglutinative languages such as Basque. Selecting nouns also gives better results than word-forms, but they do not give the best ones.

However, when SVMs are used after applying SVD, results become poorer than without applying it. This may be because SVMs are good when the number of features is high.

We have also used Naïve Bayes and Winnow to categorize the documents after applying SVD, but we do not include the results obtained in Table 3 because they are poorer that the

ones obtained without applying SVD. The reason may be that the way Snow treats the data makes it adequate to work with raw texts instead of with the vectors obtained after the SVD.

	100	200	300	400	500
W	82.98%	84.30%	**84.89**%	84.76%	84.63%
L	**85.95**%	**86.61**%	**86.81**%	**87.33**%	**87.07**%
N	84.37%	**85.36**%	84.83%	85.03%	84.76%

Table 4: SVD + k-NN accuracy rates.

Finally, given that the best results have been obtained by combining SVD and k-NN, we consider interesting to show the results obtained for different dimensions and number of neighbous. In Table 4 the results for the best k are shown: k=10 (W) and k=3 (L)(N). We want to enphasize that when the lemmatized corpus is used, the results for every dimensionality experimented increase the best result met without applying SVD (84.56 % in Table 2).

Conclusions and Future Work

In this paper we have shown the foundations and results of an experiment conducted to validate different methods for categorizing documents written in Basque. In our opinion, the most important conclusion is that combining SVD for dimensionality reduction together with the cosine similarity measure and the k-NN algorithm yields to an important improvement in the categorization accuracy. We would like to emphasize that when lemmatization is used, results raise up to 87.33% accuracy.

For future work, we intend to test other combinations of methods constructing a multi-classifier (Wolpert 1992) and trying to perform a more sophisticated feature selection technique (Yang & Pedersen 1997) (Inza *et al.* 2000) over the features given by the SVD.

Finally, we intend to confirm the good results of combining LSI and k-NN algorithm for other languages and corpora (Reuters-21578).

Acknowledgements

This work is funded by the University of the Basque Country (UPV00141.226-T-14816/2002), the Basque Government (UE02/B11), and Gipuzkoa Council in a European Union Program.

References

Alegria, I.; Artola, X.; Sarasola, K.; and Urkia, M. 1996. Automatic morphological analysis of basque. *Literary & Linguistic Computing* 11.

Berry, M., and Browne, M. 1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Philadelphia: SIAM Society for Industrial and Applied Mathematics, ISBN: 0-89871-437-0.

Carlson, A.; Cumby, C.; Rosen, J.; and Roth, D. 1999. Snow. *UIUC Tech report UIUC-DCS-R-99-210*. University of Illinois.

Dagan, I.; Karov, Y.; and Roth, D. 1997. Mistake-driven learning in text categorization. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, 55–63.

Dasarathy, B. 1991. Nearest neighbor (nn) norms: Nn pattern recognition classification techniques. *IEEE Computer Society Press*.

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41:391–407.

Dumais, S. 2004. Latent semantic analysis. *ARIST (Annual Review of Information Science Technology)* 38:189–230.

Ezeiza, N.; Aduriz, I.; Alegria, I.; Arriola, J.; and Urizar, R. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *COLING-ACL'98*.

Inza, I.; Larrañaga, P.; Etxeberria, R.; and Sierra, B. 2000. Feature subset selection by bayesian network-based optimization. *Artificial Intelligence* 123:157–184.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. *Proceedings of ICML-99, 16th International Conference on Machine Learning* 200–209.

Minsky, M. 1961. Steps toward artificial intelligence. In *Proceedings of the Institute of Radio Engineers*, volume 49, 8–30.

Nakov, P.; Valchanova, E.; and Angelova, G. 2003. Towards deeper understanding of the lsa performance. In *Proc. of the Int. Conference RANLP-03 "Recent Advances in Natural Language Processing"*, 311–318.

Pierre, J. 2001. On the automated classification of web sites. *Linköping Electronic Articles in Computer and Information Science* 6.

Riloff, E. 1995. Little words can make a big difference for text classification. In *Proceedings of the 18th Annual International ACM SIGIR*, 130–136.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Spitters, M. 2000. Comparing feature sets for learning text categorization. *Recherche D'information Assistee Par Ordinateur Sur Internet, RIAO2000*.

Witten, I., and Frank, E. 1999. Data mining. practical machine learning tools and techniques with java implementations. *Morgan Kaufmann Publishers*.

Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5:241–259.

Yang, Y., and Pedersen, J. 1997. A comparative study on feature selection in text categorization. In Kaufmann, M., ed., *Proceedings of the Fourteenth International Conference on Machine Learning, ICML'97*, 412–420.