EXPLORING BASQUE DOCUMENT CATEGORIZATION FOR EDUCATIONAL PURPOSES USING LSI

A. Zelaia, I. Alegria, O. Arregi, A. Arruarte, A. Díaz de Ilarraza, J.A. Elorriaga and B. Sierra
University of the Basque Country, UPV-EHU, Spain

Keywords: Document Categorization, Latent Semantic Indexing (LSI), Computer Supported Learning Systems (CSLSs), Domain Module.

Abstract: In the process of preparing learning material for Computer Supported Learning Systems (CSLSs), one of the first steps involves finding documents relevant to the topics and to the students. This requires documents to be categorized according to some criteria. In this paper we analyze the behaviour of classification techniques such as Naïve Bayes, Winnow, SVMs and *k*-NN, together with lemmatization and noun selection, in the categorization of documents written in Basque. In a second experiment, we study the effect of applying the Singular Value Decomposition (SVD) dimensionality reduction technique before using the mentioned classification techniques. The results obtained show that the approach which combines SVD and *k*-NN for a lemmatized corpus gives the best categorization of all with a remarkable difference. The final aim pursued in this project is to facilitate the semiautomatic construction of the domain module of a CSLS.

1 INTRODUCTION

In the Information Age, learning occurs in contexts where information and knowledge are constantly changing. Finding documents relevant to topics and to the users involves one of the first steps in the process of preparing learning material (Vereoustre and McLean, 2003). Learning is opportunistic. It occurs in dynamic environments where new information, processes and people are appearing and disappearing. Current electronic document search engines do not provide a reasonable answer to most people's opportunistic learning needs. Following the guidelines set in (Aleven et al., 2003), the paper here presented tries to establish synergies between research occurring in the fields of Artificial Intelligence in Education and Electronic Document Technologies.

In Computer Supported Learning Systems (CSLSs) one of the main components is the *domain module*, where the subject to be learnt is modelled. The final aim of our project is to facilitate the construction of the domain module in a semi-automatic way. The process of creating it implies first the identification of learning material, i.e. the selection of the appropriate documents. This requires documents to be categorized according to some educational criteria. Most researchers propose approaches based on machine learning techniques, where automatically built classifiers learn from a set of previously classified documents. In our experiments, we use four classification techniques which have reported good results for categorizing documents: Naïve Bayes, Winnow, SVMs and *k*-NN.

Several experiments have been made to classify documents written in extended languages such as English. But, the reality of lesser-used languages, as it 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 classifier and may, consequently, produce poorer results. In addition, we find that for educational use, even for extended languages such as English, there is no educational collection of documents (Nakayama and Shimizu, 2003). Taking this fact into account we have decided to separate the experimental design into two steps. In the first step, presented in this paper, we analyze the behaviour of the classification algorithms using documents which correspond to a Basque newspaper and which are categorized according to a recognized standard classification. In a future second phase, we will analyze the behaviour of the classification techniques using an educational corpus for Basque, which will have to be previously constructed.

Zelaia A., Alegria I., Arregi O., Arruarte A., Díaz de Ilarraza A., Elorriaga J. and Sierra B. (2009).
EXPLORING BASQUE DOCUMENT CATEGORIZATION FOR EDUCATIONAL PURPOSES USING LSI.
In *Proceedings of the First International Conference on Computer Supported Education*, pages 5-9
DOI: 10.5220/0001834300050009

There is also another reason because of which we have to make a special effort in our classification task; the morphosintactical features of Basque. In fact, we must take into account that Basque is an agglutinative language whose declension system has numerous cases (Alegria et al., 1996). This makes the categorization task even more difficult, because semantic information is not really contained in word-forms but in their corresponding lemma. Thus, the categorization of documents written in Basque turns out to be challenging. In our experiments we analyze the effect of preprocessing the corpus in order to reduce the dimension of the information to treat. In this way, we analyze dimensionality reduction techniques such as lemmatization, noun selection and Singular Value Decomposition (SVD).

In this work we perform two experiments. In the first one, we apply the classification techniques to three different corpora. In the second experiment, we apply the SVD dimensionality reduction technique by means of Latent Semantic Indexing[1] (LSI) implementation, before applying the classification techniques to the same corpora.

This paper is structured as follows. In Section 2 the research context is presented. In Section 3, we reference the classification algorithms used in our experiments, and examine the applications of LSI in text categorization problems and for educational purposes. In Section 4 the experimental setup is introduced, where both training and test corpora are described and lemmatization and noun selection processes are introduced. In Section 5, experimental results are shown, compared and discussed. Finally, Section 6 contains some conclusions and comments on future work.

2 RESEARCH CONTEXT

This work is part of a project that aims to acquire semiautomatically the domain for CSLSs. Concretely, the system that is being developed takes an electronic document as the base for building the Domain Module (Larrañaga et al., 2003). This module is enriched with additional documents and other didactic material. The process is divided into three different phases: Domain Module structure acquisition, generation of didactic material and domain enrichment and maintenance. In the Domain Module structure acquisition phase, first the document table of contents is analyzed obtaining the main topics of the domain and the relations among them. These topics and their relations constitute the first version of the domain ontology. Once the initial process has been finished, the whole document is analyzed in order to look for new topics and relations. The generation of didactic material is an ontology-driven analysis which splits out the whole document into Learning Objects (homepage, 2001) categorizing them according to some pedagogical purpose.

Finally, in order to enrich the Domain Module with more didactic material and to maintain it up to date, new documents are analyzed and incorporated to the domain module. The work presented in this paper will help in this last phase. Document classification will allow to connect the new documents to the concepts of the domain.

3 CLASSIFICATION TECHNIQUES AND LATENT SEMANTIC INDEXING (LSI)

Text categorization consists in assigning predefined categories to text documents (Sebastiani, 2005). When the bag-of-words text document representation is used, the number of attributes in 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 (terms), substituting some words by others that represent them (lemmas, etc.), applying SVD technique, etc.

The SVD technique compresses vectors representing documents into vectors of a lower-dimensional space (Berry and Browne, 1999). This operation is called dimensionality reduction, and the space to which document vectors are projected is called the reduced space. When using the reduced space, most of the important underlying structure that associates terms with documents is captured and consequently, noise is reduced.

In our experiments we use LSI (Deerwester et al., 1990) (Dumais, 2004) to calculate the SVD and the cosine similarity measure among the document to be categorized and all the documents in the reduced space (training set). LSI has been successfully used in the categorization of documents written in english (Dolin et al., 1999) (Dumais, 1995). It has also been used for a variety of educational applications, such as the representation of knowledge in CSLSs (Zampa and Lemaire, 2002), tutoring dialog (Graesser et al.,

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

2001) and automatic essay grading (Miller, 2003).

We use classification algorithms which have reported good results for text categorization in other languages; in this way, we use Naïve Bayes (Minsky, 1961), Winnow (Dagan et al., 1997), SVMs (Joachims, 1999) and *k*-NN (Dasarathy, 1991).

4 EXPERIMENTAL SETUP

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

4.1 Document Collection

As we have pointed out in the introduction, we are interested in the categorization of documents written in Basque with educational purposes. The ideal would be to have available an educational collection of documents categorized according to some standard labelling, but there is neither such educational corpus nor a standard educational classification. 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. This will allow us to analyze the behaviour of the selected classification techniques when applied to Basque documents.

The documents used in this experiment correspond to the *Euskaldunon Egunkaria* newspaper, 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 can be seen in Table 1.

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 effectiveness 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.

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

Table 1: Number of documents distributed by categories.

Category	Training	Test
1. Culture	600	202
2. 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

4.2 Feature Selection. Lemmatization

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 techniques. On the one hand, stopword lists have been used to eliminate non-relevant words, i.e. the most and least frequent words in the training corpus. On the other hand, we use linguistic methods such as lemmatization and noun selection to reduce the number of features. Indeed, recent experiments show that lemmatization helps in the process of categorizing documents written in an inflected language using LSI (Nakov et al., 2003). Therefore, we expect that lemmatization, and noun selection in particular, should allow us to maintain the same semantic information, reducing the number of attributes to be processed.

We have used the Basque lemmatizer designed by the IXA natural language processing 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 category by more than 50%.

5 EXPERIMENTAL RESULTS

In this section we show the results obtained in the two experiments. In both of them we use the general-purpose classifier named SNoW (Carlson et al., 1999) for Naïve Bayes and Winnow algorithms and Weka

Table 2: Accuracy rates before applying SVD.

		all	>1	>2	>3
Naïve Bayes	Words	**80.09**	78.89	78.10	77.77
	Lemmas	**81.53**	81.07	80.74	80.28
	Nouns	79.49	**79.62**	79.35	79.62
Winnow	Words	80.09	**81.13**	80.47	79.49
	Lemmas	80.15	**80.47**	78.10	77.77
	Nouns	**79.35**	78.83	76.78	76.45
SVMs	Words	81.53	82.72	83.18	**83.71**
	Lemmas	84.10	**84.56**	83.58	83.11
	Nouns	81.40	**82.58**	81.60	81.99
k-NN	Words	37.80	**54.75**	38.32	40.96
	Lemmas	50.66	40.11	58.91	**59.17**
	Nouns	61.08	69.53	70.84	**72.16**

(Witten and Frank, 2005) for SVMs. We apply the classification algorithms to three different corpora: a corpus of text documents (words), a second one of lemmatized documents and a third one in which only nouns appearing in documents have been kept.

5.1 Experiment before Applying SVD

In this experiment, elimination of irrelevant words, lemmas and nouns has been performed based on the word frequency in documents; terms that appear in more than 1, 2 or 3 documents (>1, >2, etc.) are kept. The accuracy rates using the test-corpus for each classification technique are shown in Table 2. The best results obtained for each technique and corpus appear printed in boldface.

As shown in Table 2, the best result has been obtained by using SVMs after removing words that appear in only 1 document (>1) and using the lemmatized corpus (84.56 %). We want to emphasize that, taking into account the morphosintactical features of Basque and the reduced corpora used, the accuracy rates obtained with this method are high for all the three corpora. In fact, they are as good as some results reported for other similar corpora and language features (Nakov et al., 2003).

Results obtained using Naïve Bayes and Winnow are also very good. Both have been obtained using SNoW, and we argue that the processing it performs 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 for Winnow.

However, results show 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). The accuracy rates in the table have been obtained for different *k* values (*k*=1,..., 10), and using the Euclidean distance.

Table 3: Accuracy rates for SVMs and *k*-NN after SVD.

		LSI dim.	Accuracy
SVD+SVMs	Words	1000	75.00%
	Lemmas	500	**81.46**%
	Nouns	500	80.34%
SVD+*k*-NN	Words	300	84.89%
	Lemmas	400	**87.33**%
	Nouns	200	85.36%

5.2 Experiment after Applying SVD

In this second experiment, LSI has been used to create the three reduced spaces for the training document collections. Different number of dimensions have been experimented $(100, 200, 300, 400, 500, 1000)$. The weighting scheme used has been 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.

The best results in this experiment have been obtained by using *k*-NN. In Table 3 the best result for each corpus is shown, and it can be observed that, the highest accuracy rate has been obtained for the lemmatized corpus, which significantly improves and increases 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. This is because SVMs are good enough when the number of features is high, and consequently, the dimensionality reduction does not benefit to them.

We have also used Naïve Bayes and Winnow to categorize the documents after applying SVD, but we do not include the results in Table 3 because they are fairly worse that the ones obtained before applying SVD. The reason may be that the way SNoW treats data makes it adequate to work with raw texts instead of with the reduced dimensional vectors obtained after the SVD.

Finally, given that the best results have been obtained by combining SVD and *k*-NN, we consider interesting to show all the accuracy rates obtained for different dimensions and number of neighbours. In Table 4 the results for the best *k* are shown: *k*=10 (Words) and *k*=3 (Lemmas and Nouns).

Table 4: SVD + *k*-NN accuracy rates for Words, Lemmas and Nouns.

	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

6 CONCLUSIONS AND FUTURE WORK

Along this paper, we have analyzed the categorization of documents written in Basque with the purpose of facilitating the construction of the domain module in a CSLS. This work constitutes an important step in the process of semi-automatically acquiring the domain module of CSLSs. The two experiments performed in this study show that advances in the field of Electronic Document Technologies can find interesting applications in the field of Artificial Intelligence in Education. Results demonstrate that the *k*-NN classification algorithm combined with the SVD dimensionality reduction technique gives very good results even for a lesser-used and highly inflected language such as Basque. We would like to emphasize that when lemmatization is used, results increase up to 87.33%.

In our experiments we have confirmed that categorization results are also good when documents are written in Basque. This will permit us to face the Basque document categorization problem for an educational environment in a more established way. It will be a great advance in the process of constructing the domain module for CSLSs in a semi-automatic way. However, the lack of a Basque educational collection of documents makes this first step of acquisition of learning material be harder. Our future work will be conducted to construct such a corpus (Ghani et al., 2001) and repeat the experiments in order to confirm the good results.

Regarding the domain acquisition task, we are currently working in the automatic extraction of the main topics and the pedagogical relations among them represented, explicitly or implicitly, in the table of contents of a document. A set of heuristics that infer such relations and the part-of-speech information have been already defined (Larrañaga et al., 2004) (Larrañaga et al., 2008).

ACKNOWLEDGEMENTS

This work is supported by the MEC (TIN2006-14968-C02-01) and by the University of the Basque Country (UE06/19).

REFERENCES

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

Aleven, V., Hoppe, U., Kay, J., Mizoguchi, R., Pain, H., Verdejo, F., and Yacef, K., editors (2003). *Technologies for Electronic Documents for Supporting Learning*.

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

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*, pages 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.

Dolin, R., Pierre, J., Butler, M., and Avedon, R. (1999). Practical evaluation of ir within automated classification systems. *Proceedings of the International Conference on Information and Knowledge Management CIKM*, pages 322–329.

Dumais, S. (1995). Using lsi for information filtering: Trec-3 experiments. In Harman, D., editor, *Third Text REtrieval Conference (TREC3)*, pages 219–230.

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*.

Ghani, R., Jones, R., and Mladenic, D. (2001). Using the web to create minority language corpora. In *International Conference on Infroamtion and Knowledge Management (CIKM 2001)*.

Graesser, A., Person, N., Harter, D., and Group, T. T. R. (2001). Teaching tactics and dialog in autotutor. *International Journal of Artificial Intelligence in Education*, 12(3):257–279.

homepage, L. L. O. M. W. G. (2001). IEEE P1484.12. http://ltsc.ieee.org/wg12/.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of*

ICML-99, 16th International Conference on Machine Learning, pages 200–209.

Larrañaga, M., Elorriaga, J., and Arruarte, A. (2008). A heuristic nlp based approach for getting didactic resources from electronic documents. In *Proceedings of the 3th Europeanl Conference on Technology-Enhanced Learning, Springer, LNCS 5192*, pages 197–202.

Larrañaga, M., Rueda, U., Elorriaga, J., and Arruarte, A. (2003). Index analysis: A means to acquire the domain module structure. In *X CAEPIA - V TTIA*, volume II, pages 339–342.

Larrañaga, M., Rueda, U., Elorriaga, J., and Arruarte, A. (2004). Acquisition of the domain structure from document indexes using heuristic reasoning. In Lester, J., Vicari, R., and Paraguacu, F., editors, *Intelligent Tutoring Systems, LNCS 3220*, pages 175–186.

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, 28.

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

Nakayama, M. and Shimizu, Y. (2003). Subject categorization for web educational resources using mlp. In *European Symposium on Artificial Neural Networks, ESANN'2003*, pages 9–14.

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"*, pages 311–318, Bulgaria.

Sebastiani, F. (2005). Text categorization. *Text Mining and its Applications*, pages 109–129.

Vereoustre, A. and McLean, A. (2003). Reusing educational material for teaching and learning: Current approaches and directions. In Aleven, V., Hoppe, U., Kay, J., Mizoguchi, R., Pain, H., Verdejo, F., and Yacef, K., editors, *Supplementary Proceedings of AIED2003*, pages 621–630.

Witten, I. and Frank, E. (2005). Data mining. practical machine learning tools and techniques. *Morgan Kaufmann Publishers*.

Zampa, V. and Lemaire, B. (2002). Latent semantic analysis for user modeling. *Journal of Intelligent Information Systems. Special Issue on Education Applications.*, 18(1):15–30,.