Towards Basque Oral Poetry Analysis: A Machine Learning Approach

Mikel Osinalde, Aitzol Astigarraga, Igor Rodriguez and **Manex Agirrezabal**
Computer Science and Artificial Intelligence Department,
University of the Basque Country (UPV/EHU), 20018 Donostia
teagenes@hotmail.com
aitzol.astigarraga@ehu.es
igor.rodriguez@ehu.es
manex.agirrezabal@ehu.es

Abstract

This work aims to study the narrative structure of Basque greeting verses from a text classification approach. We propose a set of thematic categories for the correct classification of verses, and then, use those categories to analyse the verses based on Machine Learning techniques. Classification methods such as Naive Bayes, k-NN, Support Vector Machines and Decision Tree Learner have been selected. Dimensionality reduction techniques have been applied in order to reduce the term space. The results shown by the experiments give an indication of the suitability of the proposed approach for the task at hands.

1 Introduction

Automated text categorization, the assignment of text documents to one or more predefined categories according to their content, is an important application and research topic due to the amount of text documents that we have to deal with every day. The predominant approach to this problem is based on Machine Learning (ML) methods, where classifiers learn automatically the characteristics of the categories from a set of previously classified texts (Sebastiani, 2002).

The task of constructing a document classifier does not differ so much from other ML tasks, and a number of approaches have been proposed in the literature. According to Cardoso-Cachopo and Oliveira (2003) , they mainly differ on how documents are represented and how each document is assigned to the correct categories. Thus, both steps, document representation and selection of the classification method are crucial for the overall success. A particular approach can be more suitable for a particular task, with a specific data, while another one can be better in a different scenario (Zelaia et al., 2005; Kim et al., 2002; Joachims, 1998).

In this paper we analyse the categorization of traditional Basque impromptu greeting verses. The goal of our research is twofold: on the one hand, we want to extract the narrative structure of an improvised Basque verse; and, on the other hand, we want to study to what extent such an analysis can be addressed through learning algorithms.

The work presented in this article is organized as follows: first we introduce Basque language and *Bertsolaritza*, Basque improvised context poetry, for a better insight of the task at hand. Next, we give a general review of computational pragmatics and text classification domains, examining discourse pattern, document representation, feature reduction and classification algorithms. Afterwards, the experimental set-up is introduced in detail; and, in the next section, experimental results are shown and discussed. Finally, we present some conclusions and guidelines for future work.

2 Some Words about Basque Language and *Bertsolaritza*

Basque, *euskara*, is the language of the inhabitants of the Basque Country. It has a speech community of about 700,000 people, around 25% of the total population. Seven provinces compose the territory, four of them inside the Spanish state and three inside the French state.

Bertsolaritza, Basque improvised contest poetry, is one of the manifestations of traditional Basque culture that is still very much alive. Events and competitions in which improvised verses, *bertso*-s, are composed are very common. In such performances, one or more verse-makers, named *bertsolaris*, produce impromptu compositions about topics or prompts which are given to them by a theme-prompter. Then, the verse-

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 119–125,
Hissar, Bulgaria, 9-11 September 2013.

maker takes a few seconds, usually less than a minute, to compose a poem along the pattern of a prescribed verse-form that also involves a rhyme scheme. Melodies are chosen from among hundreds of tunes.

Figure 1: *Bertsolari Txapelketa Nagusia*, the national championship of the Basque improvised contest poetry, held in 2009

When constructing an improvised verse strict constraints of meter and rhyme must be followed. For example, in the case of a metric structure of verses known as *Zortziko Txikia* (small of eight), the poem must have eight lines. The union of each odd line with the next even line, form a strophe. And each strophe, in turn, must rhyme with the others. But the true quality of the *bertso* does not only depend on those demanding technical requirements. The real value of the *bertso* resides on its dialectical, rhetorical and poetical value. Thus, a *bertsolari* must be able to express a variety of ideas and thoughts in an original way while dealing with the mentioned technical constraints.

The most demanding performance of Basque oral poetry, is the *Bertsolari Txapelketa*, the national championship of *bertsolaritza*, celebrated every four years (see Fig.1). The championship is composed by several tasks or contests of different nature that need to be fulfilled by the participants. It always begins with extemporaneous improvisations of greetings, a first verse called *Agurra*. This verse is the only one in which the poet can express directly what she/he wants. For the rest of the contest, the theme-prompter will prescribe a topic which serves as a prompt for the *bertso*, and also the verse metric and the number of iterations. For that reason, we thought the *Agurra* was of particular interest to analyse ways verse-makers use to structure their narration.

3 Related Work

3.1 Computational Pragmatics

As stated in the introduction, the aim of this paper is to notice if there is any discourse pattern in greeting verses. In other words, we are searching certain defined ways verse-improvisers in general use to structure their discourse.

If the study of the meaning is made taking into account the context, we will have more options for getting information of the factors surrounding improvisation (references, inferences, what improvisers are saying, thinking, self-state, context). The field that studies the ways in which context contributes to meaning is called pragmatics. From a general perspective, Pragmatics refers to the speaker and the environment (Searle, 1969; Austin, 1975; Vidal, 2004).

The study of extra-linguistic information searched by pragmatics is essential for a complete understanding of an improvised verse. In fact, the understanding of the text of each paragraph does not give us the key for the overall meaning of the verse. There is also a particular world's vision and a frame of reference shared with the public; and, indeed, we have been looking for those keys. We believe that the verse texts are not linear sequences of sentences, they are placed regarding a criterion and the research presented here aims to detect this intent.

Therefore, searching for the discourse facts in greeting verses led us to study their references.

3.2 Text Categorization

The goal of text categorization methods is to associate one or more of a predefined set of categories to a given document. An excellent review of text classification domain can be found in (Sebastiani, 2002).

It is widely accepted that how documents are represented influences the overall quality of the classification results (Leopold and Kindermann, 2002). Usually, each document is represented by an array of words. The set of all words of the training documents is called vocabulary, or dictionary. Thus, each document can be represented as a vector with one component corresponding to each term in the vocabulary, along with the number that represents how many times the word appears in the document (zero value if the term does not occur). This document representation is called the bag-of-words model. The major drawback of this text representation model is that the number of features in the corpus can be considerable, and thus, intractable for some learning algorithms.

Therefore, methods for dimension reduction are required. There exists two different ways to carry out this reduction: data can be pre-processed, i.e., some filters can be applied to control the size of the system's vocabulary. And, on the other hand, dimensionality reduction techniques can be applied.

3.2.1 Pre-processing the Data

We represented the documents based on the aforementioned bag-of-word model. But not all the words that appear in a document are significant for text classification task. Normally, a pre-processing step is required to reduce the dimensionality of the corpus and, also, to unify the data in a way it improves performance.

In this work, we applied the following pre-processing filters:

- **Stemming**: remove words with the same stem, keeping the most common among them. Due to its inflectional morphology, in Basque language a given word lemma makes many different word forms. A brief morphological description of Basque can be found in (Alegria et al., 1996). For example, the lemma *etxe* (house) forms the inflections *etxea* (the house), *etxeak* (houses or the houses), *etxeari* (to the house), etc. This means that if we use the exact given word to calculate term weighting, we will loose the similarities between all the inflections of that word. Therefore, we use a stemmer, which is based on the morphological description of Basque to find and use the lemmas of the given words in the term dictionary (Ezeiza et al., 1998).

- **Stopwords**: eliminate non-relevant words, such as articles, conjunctions and auxiliary verbs. A list containing the most frecuent words used in Basque poetry has been used to create the stopword list.

3.2.2 Dimensionality Reduction

Dimensionality reduction is a usual step in many text classification problems, that involves transforming the actual set of attributes into a shorter, and hopefully, more predictive one. There exists two ways to reduce dimensionality:

- **Feature selection** is used to reduce the dimensionality of the corpus removing features that are considered non-relevant for the classification task (Forman, 2003). The most well-known methods include: Information Gain, Chi-square and Gain Ratio (Zipitria et al., 2012).

- **Feature transformation** maps the original list of attributes onto a new, more compact one. Two well-known methods for feature transformation are: Principal Component Analysis (PCA) (Wold et al., 1987) and Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Hofmann, 2001).

The major difference between both approaches is that feature selection selects a subset from the original set of attributes, and feature transformation transforms them into new ones. The latter can affect our ability to understand the results, as transformed attributes can show good performance but little meaningful information.

3.2.3 Learning Algorithms

Once the text is properly represented, ML algorithms can be applied. Many text classifiers have been proposed and tested in literature using ML techniques (Sebastiani, 2002), but text categorization is still an active area of research, mainly because there is not a general faultless approach.

For the work presented here, we used the following algorithms: Nearest Neighbour Classifier (IBk) (Dasarathy, 1991), Nave Bayes Classifier (NB) (Minsky, 1961), J48 Decision Tree Learner (Hall et al., 2009) and SMO Support Vector Machine (Joachims, 1998).

All the experiments were performed using the Weka open-source implementation (Hall et al., 2009). Weka is written in Java and is freely available from its website [1].

In Fig.2, the graphical representation of the overall Text Classification process is shown.

4 Experimental Setup

The aim of this section is to describe the document collection used in our experiments and to give an account of the stemming, stopword deletion and dimensionality reduction techniques we have applied.

[1] http://www.cs.waikato.ac.nz/ml/weka/

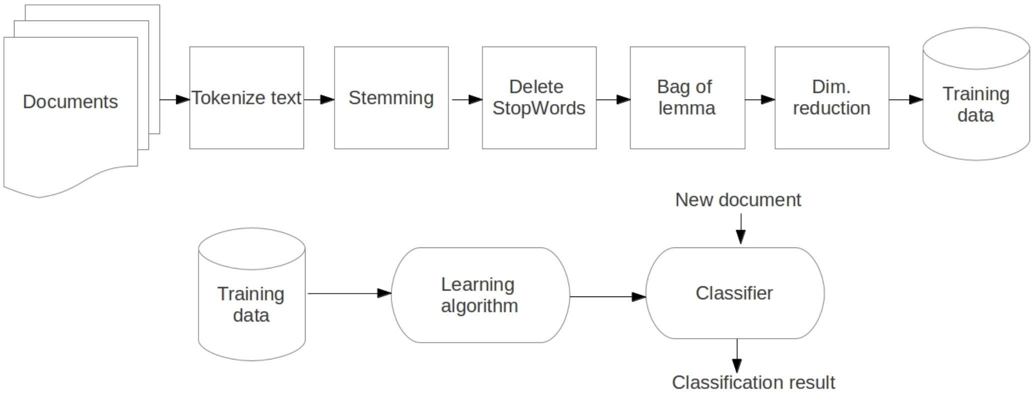

Figure 2: The overall process of text categorization

4.1 Categorization

To make a correct categorization of the verses, before anything else the unit to be studied needs to be decided. We could take as a unit of study the word, the strophes or the entire verses. Considering that we want to extract the structure that would provide information about the decisions made by the improviser and the discourse organization, we decided that the strophe[2] was the most appropriate unit to observe those ideas. Therefore, the first job was to divide the verses in strophes. After that, we began to identify the contents and features in them. The goal was to make the widest possible characterization and, at the same time, select the most accurate list of attributes that would make the strophes as much distinguishable as possible.

We sampled some strophes from the verse corpus described in section 4.2 and analysed them one by one. We had two options when categorizing the strophes: first, analyse and group all the perceived topics, allowing us to propose a realistic classification of the strophes from any verse. And second, make a hypothesis and adjust the obtained data to the hypothesis. We decided to take both paths.

After analysing each of the strophes and extracting their topics, we made the final list, sorted by the relevance of the categories. We obtained a very large list of contents and we arranged it by the importance and by the number of appearance. But that thick list did not help us in our mission as we wanted. So we agreed to try to define and limit the collection of attributes. And we decided to use the second option. Therefore, we studied the foundations of discourse analysis (Roberts and Ross, 2010; Gumperz, 1982), and the classifications proposed by critics of the improvisation field (Egaña et al., 2004; Diaz Pimienta, 2001); and then, we compared them with our predicted one. Merging both approaches we tried to build a strong set of categories.

Combining inductive and deductive paths we formed a list of six categories. So the initial big list that we gathered was filtered to a more selective classification. Therewith, we found possible to label the majority of the strophes in the analysed verses, and also get a significant level of accuracy.

Thus, these are the categories to be considered in the verse classification step:

1. Message: the main idea
2. Location: references to the event site
3. Public: messages and references relating to the audience
4. Event: messages and references relating to the performance itself
5. Oneself aim or Oneself state
6. Miscellaneous: padding, junk. Sentences with no specific meaning or intend.

As well as the five categories closely linked to the communication situation, there is another that we called Miscellaneous (padding, filling). Due to

[2] a pair of stanzas of alternating form on which the structure of a given poem is based

the demanding nature of the improvisation performances, they usually are sentences not very full of content and intent.

We have decided to consider each one of them as a separate goal, and hence six classifiers were to be obtained, one for each category. Thus, each categorization task was addressed as a binary classification problem, in which each document must be classified as being part of $category_i$ or not (for example, Location vs. no Location).

4.2 Document Collection

For the task in hands, we decided to limit our essay to greeting verses from tournaments. We selected 40 verses of a corpus of 2002 verses and divided them into strophes (212 in total). But when we began assigning categories (1-6) to each strophe, we realized we were in blurred fields. It was pretty difficult to perform that task accurately and we thought it was necessary to ask some expert for help. Mikel Aizpurua[3] and Karlos Aizpurua[4] (a well-known judge the former and verse improviser and Basque poetry researcher the latter) agreed to participate in our research, and they manually labelled one by one the 212 strophes.

In that study, we considered each binary class decision as a distinct classification task, where each document was tested as belonging or not to each category. Thus, the same sentence could effectively belong to more than one categories (1 to 6 category labels could be assigned to the same sentence).

As an example, let us have a look to an initial greeting verse composed by Anjel Larrañaga, a famous verse-maker (see Fig.3).

There we can see that each strophe (composed of two lines), was labelled in one, two or even tree different categories.

- (1) (3): Message, Public
- (5): Oneself aim
- (4) (5): Event, Oneself state
- (1) (5) (3): Message, Oneself aim, Public

The document categorization process was accomplished in two steps: during the training step, a general inductive process automatically built a classifier by learning from a set of labelled documents. And during the test step, the performance of the classifier was measured. Due to the small size of our manually categorized corpus, we used the k-fold cross-validation method, with a fold value of k=10.

Agur ta erdi bertsozaleak
*lehendabiziko **sarreran**,*
behin da berriro jarri gerade
*kantatutzeko **aukeran**,*
ordu ilunak izanagaitik
*txapelketan gora-**beheran**,*
saia nahi degu ta ia zuen
*gogoko izaten **geran**.*

As a first introduction,
greetings to all improvisation fans. (1) (3)
Many times we were ready
to sing like now! (5)
Even though there are hard times
in our championship contest, (4) (5)
We will try to make our best
and we hope you find it to your liking! (1) (5) (3)

Figure 3: A welcome verse composed by Anjel Larrañaga

4.3 Pre-processing the Data

In order to reduce the dimensionality of the corpus, two pre-processing filters were applied. On the one hand, a stopword list was used to eliminate non-relevant words. On the other hand, a stemmer was used to reduce the number of attributes.

The number of different features in the unprocessed set of documents was 851, from which were extracted 614 different stems and 582 terms after eliminating the stopwords. So finally, we obtained a bag-of-lemmas with 582 different terms.

5 Experimental Results

In this section we show the results obtained in the experiments. There are various methods to determine algorithms' effectiveness, but precision and recall are the most frequently used ones.

It must be said that a number of studies on feature selection focused on performance. But in many cases, as happened to us, the are few in-

[3] http://bdb.bertsozale.com/en/web/haitzondo/view/-Mikel-Aizpurua

[4] http://bdb.bertsozale.com/en/web/haitzondo/view/-Karlos-Aizpurua

Category	ML method	Attribute selection	Performance	F-measure
Message	1-nn	None	64.62%	0.62
Location	SMO	InfoGain	89.62%	0.86
Public	SMO	ChiSquare	83.01%	0.81
Event	5-nn	None	78.30%	0.76
Oneself	SMO	InfoGain	62.26%	0.60
Miscellaneous	1-nn	GainRatio	87.74%	0.83

Table 1: Best results for each category

stances of positive classes in the testing database. This can mask the classifiers performance evaluation. For instance, in our testing database only 22 out of 212 instances correspond to class 2 (”Location”), giving an performance of 90.045 % to the algorithm that always classifies instances as 0, and thereby compressing the range of interesting values to the remaining 9.954 %. Therefore, in text categorization tasks is preferred the F-measure, the harmonic average between precision and recall.

Table1 shows the configurations that have achieved the best results for each category.

Based on the results of the table, we can state that they were good in three out of six categories (Location, Public and Miscellaneous); quite acceptable in one of them (Event); and finally, in the remaining two categories (Message and Oneself) the results were not very satisfactory.

Regarding to the learning algorithms, it should be pointed out that SMO and k-nn have shown the best results. We can state also that in most cases best accuracy rates have been obtained using dimensionality reduction techniques. Which in other words means that the selection of attributes is preferable to the raw data.

6 Conclusions and Future Work

In this paper we shown the foundations of the automated analysis of Basque impromptu greeting verses. The study proposes novel features of greeting-verses and analyses the suitability of those features in the task of automated feature classification. It is important to note that our primary goals were to establish the characteristics for the correct classification of the verses, and so to analyse their narrative structure. And, secondly, to validate different methods for categorizing Basque greeting verses.

Towards this end, we introduced different features related to improvised greeting verses and categorized them into six groups of Message, Location, Public, Event, Oneself and Miscellaneous. Then, we implemented six different approaches combining dimensionality reduction techniques and ML algorithms. One for each considered categories.

In our opinion, the most relevant conclusion is that k-nn and SMO have shown to be the most suitable algorithms for our classification task, and also, that in most cases attribute selection techniques help to improve their performance.

As a future work, we would like to assess the problem as a multi-labelling task (Zelaia et al., 2011), and see if that improves the results.

Finally, we must say that there is still much work to do in order to properly extract discourse-patterns from Basque greeting verses. To this end, we intend to use our classifiers to label larger corpora and find regular discourse patterns in them.

7 Acknowledgements

The authors gratefully acknowledge *Bertsozale Elkartea*[5] (Association of the Friends of Bertsolaritza), whose verse corpora has been used to test and develop the *Bertsobot* project.

This work has been partially conducted under the SAIOTEK projects XIMUR and POETAUTO, the Basque Government Research Team grant and the University of the Basque Country UPV/EHU, under grant UFI11/45 (BAILab).

References

Iñaki Alegria, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Harvard University Press.

Ana Cardoso-Cachopo and Arlindo Oliveira. 2003. An empirical comparison of text categorization methods. In

[5] http://www.bertsozale.com/en

String Processing and Information Retrieval, pages 183–196. Springer.

Belur V Dasarathy. 1991. Nearest neighbor ({NN}) norms:{NN} pattern classification techniques.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Alexis Diaz Pimienta. 2001. *Teoría de la improvisación: primeras páginas para el estudio del repentismo*. Ediciones Unión.

Andoni Egaña, Alfonso Sastre, Arantza Mariskal, Alexis Diaz Pimienta, and Guillermo Velazquez. 2004. *Ahozko inprobisazioa munduan topaketak: Encuentro sobre la improvisación oral en el mundo : (Donostia, 2003-11-3/8)*. Euskal Herriko Bertsozale Elkartea.

Nerea Ezeiza, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

John J Gumperz. 1982. Discourse strategies: Studies in interactional sociolinguistics. *Cambridge University, Cambridge*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142.

Sang-Bum Kim, Hae-Chang Rim, Dongsuk Yook, and Heui-Seok Lim. 2002. Effective methods for improving naive bayes text classifiers. *PRICAI 2002: Trends in Artificial Intelligence*, pages 479–484.

Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1):423–444.

Marvin Minsky. 1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

W Rhys Roberts and WD Ross. 2010. *Rhetoric*. Cosimo Classics.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March.

María Victoria Escandell Vidal. 2004. Aportaciones de la pragmática. *Vademécum para la formación de profesores. Enseñar español como segunda lengua (12) 1 lengua extranjera (LE)*, pages 179–197.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52.

Ana Zelaia, Iñaki Alegria, Olatz Arregi, and Basilio Sierra. 2005. Analyzing the effect of dimensionality reduction in document categorization for basque. *Archives of Control Sciences*, 600:202.

Ana Zelaia, Iñaki Alegria, Olatz Arregi, and Basilio Sierra. 2011. A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing*, 11(8):4981–4990.

Iraide Zipitria, Basilio Sierra, Ana Arruarte, and Jon A Elorriaga. 2012. Cohesion grading decisions in a summary evaluation environment: A machine learning approach.