N-gram Language Models and POS Distribution for the Identification of Spanish Varieties

Marcos Zampieri[1], Binyam Gebrekidan Gebre[2], Sascha Diwersy[1]
[1]University of Cologne, Germany
[2]Max Planck Institute for Psycholinguistics, Nijmegen, Holland
mzampier@uni-koeln.de, bingeb@mpi.nl, sascha.diwersy@uni-koeln.de

RÉSUMÉ

Ngrammes et Traits Morphosyntaxiques pour la Identification de Variétés de l'Espagnol

Notre article présente expérimentations portant sur la classification supervisée de variétés nationales de l'espagnol. Outre les approches classiques, basées sur l'utilisation de ngrammes de caractères ou de mots, nous avons testé des modèles calculés selon des traits morphosyntaxiques, l'objectif étant de vérifier dans quelle mesure il est possible de parvenir à une classification automatique des variétés d'une langue en s'appuyant uniquement sur des descripteurs grammaticaux. Les calculs ont été effectués sur la base d'un corpus de textes journalistiques de quatre pays hispanophones (Espagne, Argentine, Mexique et Pérou).

ABSTRACT

This article presents supervised computational methods for the identification of Spanish varieties. The features used for this task were the classical character and word n-gram language models as well as POS and morphological information. The use of these features is to our knowledge new and we aim to explore the extent to which it is possible to identify language varieties solely based on grammatical differences. Four journalistic corpora from different countries were used in these experiments : Spain, Argentina, Mexico and Peru.

MOTS-CLÉS : classification automatique, ngrammes, espagnol, variétés nationales.

KEYWORDS: automatic classification, n-grams, Spanish, language varieties.

1 Introduction

Spanish is a world language with official status in 21 countries. It is regarded to be a Pluricentric language with a number of interacting centres and language varieties (Thompson, 1992). Each of these national varieties has their own characteristics in terms of phonetics, lexicon and syntax.

Computational applications can benefit from identifying the correct variety of Spanish texts when undertaking tasks such as Machine Translation or Information Extraction, as they are able to handle lexical, orthographic and syntactic variation more accurately. The task is modelled as a classification problem with very similar methods to those applied to general purpose language identification (Dunning, 1994).

To the best of our knowledge, very few attempts have been made to address the problem of identifying language varieties as evidenced in 2.1. In this work we try to classify texts retrieved from newspapers published in 2008 from four different Spanish speaking countries : Spain, Argentina, Mexico and Peru. Moreover, we propose the use of new features, not limited to the classical word and character n-grams. We experimented features based on POS distribution and morphosyntactic information. The use of knowledge-rich features is not an attempt to outperform word and character n-gram-based methods, but an attempt to examine the extent to which these varieties differ in terms of grammar.

2 Related Work

Language identification is the task of automatically identifying the language contained in a given document. State-of-the-art methods apply n-gram language models at the character and sometimes word-level with results usually above 95% accuracy. This level of success is very common when dealing with languages which are typologically not closely related. This is however not the case of language varieties in which the distinction is based on very subtle differences that algorithms can be trained to recognize.

One of the first general purpose language identification approaches was the work of Ingle (1980). Ingle applied Zipf's law distribution to order the frequency of stop words in a text and used this information for language identification. Dunning (1994) introduced the use of character n-grams and statistics for language identification. In this study, the likelihood of n-grams was calculated using Markov models and this was used as the most informative feature for identification. Other studies applying n-gram language models for language identification include Cavnar and Trenkle (1994) implemented as TextCat [1], Grefenstette (1995), and Vojtek and Belikova (2007).

In the recent years, a number of language identification methods were developed for internet data such as Martins and Silva (2005) and Rehurek and Kolkus (2009). The most recent general purpose language identification method to our knowledge is the one published by Lui and Baldwin (2012). Their software, called *langid.py*, has language models for 97 languages, using various data sources. The method achieved results of up to 94.7% accuracy, thus outperforming similar tools. All models described in this section neglect language varieties. Pluricentric languages, such as the case of Spanish, are represented by a unique class.

2.1 Models for Similar Languages, Varieties and Dialects

The identification of closely related languages is one of the bottlenecks of most n-gram-based models and there are only a few studies published about it. Ljubešić et al. (2007) propose a computational model for the identification of Croatian texts in comparison to other South Slavic languages reporting 99% recall and precision in three processing stages. One of these processing stages, includes a so-called *black list*, a list of forbidden words that appear only in

[1] http ://odur.let.rug.nl/vannoord/TextCat/

Croatian texts, making the algorithm perform better.

Ranaivo-Malancon (2006) presents a semi-supervised character-based model to distinguish between Indonesian and Malay, two closely related languages from the Austronesian family and Huang and Lee (2008) proposes a bag-of-words approach to distinguish Chinese texts from Mainland and Taiwan reporting results of up to 92% accuracy. More recently, Trieschnigg et al. Trieschnigg et al. (2012) described classification experiments for a set of sixteen Dutch dialects using the Dutch Folktale Database.

For romance languages, the DEFT2010[2] shared task aimed to classify French journalistic texts not only with respect to their geographical location but also incorporating a temporal dimension. For Portuguese, Zampieri and Gebre (2012) proposed a log-likelihood estimation method to distinguish between European and Brazilian Portuguese texts with results above 99.5% for character n-grams. The model was later applied to a multilingual setting with French and Spanish texts (Zampieri et al., 2012).

3 Methods

We collected four comparable corpora to use in our experiments, one for each language variety. To collect comparable samples, we retrieved texts published in the same year from local newspapers regarded to have similar register, as follows :

Country	Newspaper	Year
Argentina	La Nación	2008
Mexico	El Universal	2008
Peru	El Comércio	2008
Spain	El Mundo	2008

TABLE 1 – Corpora

Each sub-corpus contains a set of 1,000 documents randomly sampled to avoid bias towards a given topic or genre. These sub-corpora were divided in training and test settings of 500 documents each. Following the compilation of the corpora, four groups of features were selected. The list of features used and the aspect of language that these features aim to analyse are presented next :

- **Character n-grams (2 to 5)** : orthography and lexicon
- **Word uni-grams** : lexicon
- **Word bi-grams** : lexicon and syntax
- **POS and morphological features** : morphology and syntax

The first three groups of features (knowledge-poor features) are standard in language identification and they were widely used in previous approaches. The fourth group of features (knowledge-rich features) is to our knowledge new and it consists of the use of POS and morphological feature annotation. The POS tags and morphological information were used as one unit in form of a compound tags (e.g. *N-msc-sg* or *V-inf*).

A snapshot of the tagset with nouns, adjectives and verbs is presented in table 2.

[2] http ://www.groupes.polymtl.ca/taln2010/deft.php

POS	Morph. Inf.	Example
N	msc sg	coche
N	msc pl	coches
N	fem sg	silla
N	fem pl	sillas
A	msc sg	bonito
A	msc pl	bonitos
A	fem sg	bonita
A	fem pl	bonitas
V	ind pres sg p1	hago
V	inf	hacer

TABLE 2 – Tagset

Although research in language identification and text classification shows that character and word n-gram-based methods outperform knowledge-rich features, we believe that these features are still worth experimenting with. Firstly, from an NLP perspective, these new features model a different aspect of language that cannot be addressed by neither character nor word n-grams. Secondly, because the average results obtained and the corresponding most informative features might be an important resource for contrastive linguistics providing an indication of how varieties converge and diverge.

The classification method is based on n-gram language models and document log-likelihood estimation (Dunning, 1993) as described in Zampieri and Gebre (2012). Its performance is comparable to state-of-the-art methods in language identification which focus on similar languages. It was tested on Bosnian, Croatian and Serbian documents[3] achieving 91.0% accuracy. Models described in Ljubešić et al. (2007) achieved 90.3% and 95.7% accuracy using the same dataset.

The method calculates language models using Laplace probability distribution for smoothing and after this calculation computes the probability of each document to belong to a certain class using a log-likelihood function as shown in equation 1.

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \tag{1}$$

N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability for each of the language models. The language model with higher probability determines the identified language of the text.

4 Results

The first experiments used knowledge-poor features to classify the four Spanish varieties evaluated using precision (P), recall (R) and f-measure (F). Results ranged from 0.813 f-measure for character 4-grams to 0.876 f-measure for word bi-grams. The results for each class remained constant for all features and this can be seen in table 3.

[3] http ://www.nljubesic.net/resources/tools/bs-hr-sr-language-identifier/

Feature	P	R	F
C 2-grams	0.835	0.804	0.819
C 3-grams	0.848	0.806	0.826
C 4-grams	0.842	0.787	0.813
C 5-grams	0.854	0.811	0.832
W 1-grams	0.879	0.848	0.848
W 2-grams	0.880	0.870	0.876

TABLE 3 – 4-Class Classification

The Peninsular Spanish class seemed to be the most difficult for the algorithm to identify in this setting. As an example, table 4 presents a confusion matrix for the character 4-grams feature in which the algorithm obtained its worst performance.

Document	Predicted			
Language	**ARG**	**MEX**	**PER**	**SPA**
ARG	(496)			4
MEX		(280)	120	
PER		20	(480)	
SPA	280		2	(218)

TABLE 4 – Confusion Matrix

From the 500 texts from Spain used for testing, only 218 were correctly classified, 280 were tagged as Argentinian and 2 as Peru. We subsequently classified the varieties in binary settings. Results are reported in terms of accuracy and can be seen in table 5.

Feature	ARGxMEX	ARGxPER	MEXxPER	SPAxARG	SPAxMEX	SPAxPER	Average
C 2-grams	0.999	0.996	0.860	0.852	0.957	0.940	0.934
C 3-grams	0.999	1.000	0.911	0.847	0.987	0.991	0.956
C 4-grams	1.000	0.999	0.922	0.827	0.992	0.996	0.965
C 5-grams	0.999	0.999	0.927	0.802	0.991	0.993	0.952
W 1-grams	0.999	0.999	0.945	0.851	0.994	0.992	0.963
W 2-grams	0.999	0.997	0.951	0.881	0.998	0.989	0.969
Average	0.999	0.998	0.919	0.843	0.986	0.983	0.955

TABLE 5 – Binary Classification

The best results were obtained for the classification of texts from Argentina and Mexico reaching 0.999 average accuracy. As the confusion matrix in 4 indicated, the worst setting was again Spain x Argentina with an average result of 0.842 accuracy. All the results obtained were substantially higher than the 4-class classification setting. As classification algorithms tend to perform better in binary settings, this was an expected outcome.

4.1 POS and Morphology

Next we present the results obtained using POS distribution and morphological features, combined in sets of 2, 3 and 4 compound tags as explained in section 3. The classification between Mexican and Spanish texts obtained the best results reaching 0.831 using combinations of two tags. These two varieties also obtained satisfactory scores for character and word-based features, 0.986 on average. Accuracy results for all binary classification settings are presented in table 6.

Feature	ARGxMEX	ARGxPER	MEXxPER	SPAxARG	SPAxMEX	SPAxPER	Average
PoS 2-grams	0.766	0.650	0.742	0.637	0.831	0.702	0.721
PoS 3-grams	0.815	0.670	0.753	0.673	0.821	0.741	0.746
PoS 4-grams	0.823	0.732	0.737	0.690	0.806	0.667	0.743
Average	0.801	0.684	0.744	0.666	0.819	0.703	0.736

TABLE 6 – Classification with POS Tags

The poorest results were obtained once again in the classification of Spanish and Argentinian texts, which also obtained the worst performance using knowledge-poor features. Even though the results are lower than those obtained using knowledge-poor features, the algorithm scored better than the expected 0.50 baseline, indicating that it is able to identify patterns in the datasets using only sets of morphosyntactical information. Named entities which usually help algorithms to identify varieties at the lexical level are not present in the experiments using POS tags and therefore do not influence the performance of the classifier.

4.2 Relationship Between Features

To evaluate the relationship between the features explored here, we analysed results using hierarchical clustering. For each cluster, two p-values (between 0 and 1) are calculated via multiscale bootstrap resampling. These values indicate how strong the cluster is supported by data. The two p-values are : the AU (Approximately Unbiased), in red, computed by multiscale bootstrap resampling and BP (Bootstrap Probability) in green, computed by normal bootstrap resampling. The graphic shows the difference between the performance of knowledge-poor and knowledge-rich features, arranging each in a different cluster 1.

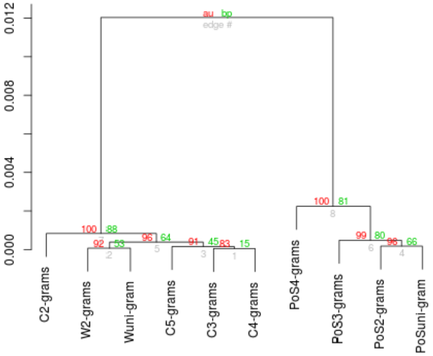

FIGURE 1 – Cluster Dendogram with AU/BP Values

The analysis grouped the two word-based feature groups in the same cluster, as they performed on average better than the character-based methods. Another interesting point of the analysis is that the results of character 4- and 5-grams are grouped in the same cluster due to an increase in performance when a larger amount of characters are taken into account. Character 4- and 5-grams features are closer to the lexical level taking whole words into account, which suggests that the model is more effective when using complete lexical items as features.

As stated before, the morphological features were not expected to outperform the knowledge-poor models, but to be used to investigate differences in grammar. An interesting outcome of these experiments is the direct relationship between the algorithm's performance using knowledge-poor and knowledge-rich features. One clear example is the classification of Argentina and Spain which obtained the worst results with characters and words as well as when using POS and morphology : 0.843 and 0.666 accuracy respectively. Another example is Argentina and Mexico which achieved the best results using characters and words, 0.999 accuracy and the second best results with POS tags, 0.801 accuracy.

For these reasons, the results presented here are an encouraging perspective for further studies. It is possible to use the outcome of the classification as a source of information for contrastive linguistics to provide quantitative overview on how these varieties converge and diverge in terms of grammar and lexicon. Linguistic analysis may be carried out using the most informative features in classification.

5 Conclusion and Future Perspectives

We presented a first attempt to identify a set of four Spanish varieties in written texts with f-measure results ranging from 0.813 to 0.876. As expected, the binary classification settings have achieved significantly better results in comparison to the 4-class classification setting. The algorithm was able to distinguish between texts from Argentina and Mexico with an average accuracy of 0.999. As previously discussed, the integration of these language models in real-world NLP applications, should improve results in a number of NLP tasks.

The experiments used not only the classical character and word n-gram models but also morphosyntactic information combined with POS. This is to our knowledge a new contribution of our work to this kind of experiments. The classification with knowledge-rich features achieved up to 0.831 accuracy for Mexican and Peninsular Spanish. We observed a direct relationship between the performance of knowledge-poor and knowledge-rich features, binary settings which obtained good performance using characters and words also present good results using morphosyntactic information. This aspect should be better explored in future work through a careful linguistic analysis.

As future perspectives, first we wish to compare the performance of our method with general purpose language identification methods such as *langid.py* (Lui and Baldwin, 2012). Second, we are replicating our experiments to a set of French varieties. Finally, we would like to experiment the combination of POS and word n-grams to investigate if performance increases.

Acknowledgements

We would like to thank the anonymous reviewers for their careful feedback.

Références

Cavnar, W. and Trenkle, J. (1994). N-gram-based text catogorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics - Special Issue on Using Large Corpora*, 19(1).

Dunning, T. (1994). Statistical identification of language. Technical report, Computing Research Lab - New Mexico State University.

Grefenstette, G. (1995). Comparing two language identification schemes. In *Proceedings of JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data*, Rome.

Huang, C. and Lee, L. (2008). Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC 2008*, pages 404–410.

Ingle, N. (1980). *A Language Identification Table*. Technical Translation International.

Ljubešić, N., Mikelic, N., and Boras, D. (2007). Language identification : How to distinguish similar languages ? In *Proceedings of the 29th International Conference on Information Technology Interfaces*.

Lui, M. and Baldwin, T. (2012). langid.py : An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL*.

Martins, B. and Silva, M. (2005). Language identification in web pages. *Proceedings of the 20th ACM Symposium on Applied Computing (SAC), Document Engineering Track. Santa Fe, EUA.*, pages 763–768.

Ranaivo-Malancon, B. (2006). Automatic identification of close languages - case study : Malay and indonesian. *ECTI Transactions on Computer and Information Technology*, 2 :126–134.

Rehurek, R. and Kolkus, M. (2009). Language identification on the web : Extending the dictionary method. In *Proceedings of CICLing. Lecture Notes in Computer Science*, pages 357–368. Springer.

Thompson, R. (1992). Spanish as a pluricentric language. In Clyne, M., editor, *Pluricentric Languages : Different Norms in Different Nations*, pages 45–70. CRC Press.

Trieschnigg, D., Hiemstra, D., Theune, M., de Jong, F., and Meder, T. (2012). An exploration of language identification techniques for the dutch folktale database. In *Proceedings of LREC2012*.

Vojtek, P. and Belikova, M. (2007). Comparing language identification methods based on markov processess. In *Slovko, International Seminar on Computer Treatment of Slavic and East European Languages*.

Zampieri, M. and Gebre, B. G. (2012). Automatic identification of language varieties : The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237, Vienna, Austria.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2012). Classifying pluricentric languages : Extending the monolingual model. In *Proceedings of the Fourth Swedish Language Technlogy Conference (SLTC2012)*, pages 79–80, Lund, Sweden.