Classifying Pluricentric Languages: Extending the Monolingual Model

Marcos Zampieri[1], Binyam Gebrekidan Gebre[2], Sascha Diwersy[3]

University of Cologne[1,3], Max Planck Institute for Psycholinguistics[1]
Albertus-Magnus-Platz 1, 50931, Cologne, Germany[1,3]
Wundtlaan 1, 6525 XD, Nijmegen, Holland[2]
mzampier@uni-koeln.de, bingeb@mpi.nl, sascha.diwersy@uni-koeln.de

Abstract

This study presents a new language identification model for pluricentric languages that uses n-gram language models at the character and word level. The model is evaluated in two steps. The first step consists of the identification of two varieties of Spanish (Argentina and Spain) and two varieties of French (Quebec and France) evaluated independently in binary classification schemes. The second step integrates these language models in a six-class classification with two Portuguese varieties.

1. Introduction

Automatic Language Identification is a well-known research topic in NLP. State-of-the-art methods consist of the application of n-gram language models to distinguish a set of languages automatically. One of the first studies to apply n-gram models to language identification was Dunning (1994) and more recent experiments include Martins and Silva (2005) and Lui and Baldwin (2012).

Martins and Silva use internet data to identify a set of 12 languages and they report results ranging from 80% to 99% accuracy depending on the language. Lui and Baldwin report 91.3% for a set of 67 languages using Wikipedia data. Their software, called *langid.py*, claims to have language models for 97 languages by using various data sources.

These two examples do not take language varieties into account. Pluricentric languages are always modeled as a unique class and this is one reason for the good results these n-gram methods report. The level of success is usually high when classification includes languages which are typologically not closely related (e.g. Finnish and Spanish) and languages with unique character sets (e.g. Hebrew).

1.1 Classifying Varieties

Only recently the automatic identification of language varieties has received more attention. A few studies have been published such as Ljubesic et al. (2007) for the former Serbo-Croatian varieties, Huang and Lee (2008) for Mainland and Taiwanese Chinese and Trieschnigg et al. (2012) for Dutch dialects.

These methods aim to distinguish varieties and to our knowledge none of them has yet been integrated into a broader language identification setting. Here we try to replicate the experiments carried out by Zampieri and Gebre (2012) for Brazilian and European Portuguese on two varieties of French and two of Spanish. Subsequently, we integrate these four new language models into a six-class classification scheme.

2. Methods

We compiled four journalistic corpora containing texts from each of the four varieties. To create comparable samples, we retrieved texts published in the same year (2001 for French and 2008 for Spanish) and all metainformation and tags were removed. The length of texts in the corpora varies and as language identification benefits from longer texts, we only used texts consisting of up to 300 tokens.

Location	Newspaper	Year
Argentina	La Nación	2008
Spain	El Mundo	2008
France	Le Monde	2001
Quebec	Le Devoir	2001

Table 1: Corpora

The identification method works on three different aspects of language: orthography, lexicon and lexico/syntax. For the orthographical differences, we used character n-grams ranging from 2 to 6-grams. At the lexical level we used word uni-grams and finally, to explore lexico-syntactical differences, word bi-grams were used. The language models were calculated with Laplace probability distribution:

$$P_{lap}(w_1...w_n) = \frac{C(w_1...w_n)+1}{N+B} \quad (1)$$

In equation number 1: C is the count of the frequency of w_1 to w_2 in the training data, N is the total number of n-grams and B is the number of distinct n-grams in the training data. For probability estimation, we used the log-likelihood function:

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \quad (2)$$

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.

3. Results

Evaluation was done using each of the feature groups in a set of 1,000 documents sampled randomly. The sample contains 50% of the texts from each variety and it is divided into 500 documents for training and 500 for testing.

3.1 Binary Classification

We report results in terms of accuracy for binary classification as seen in table 2.

Feature	AR x ES	FR x QU
Word 1-grams	0.948	0.968
Word 2-grams	0.894	0.956
Character 2-grams	0.898	0.956
Character 3-grams	0.948	0.990
Character 4-grams	0.944	0.968
Character 5-grams	0.962	0.960
Character 6-grams	0.960	0.934

Table 2: Binary Classification

The results suggest that, on average, the French corpora have a stronger variation than the two varieties of Spanish. French scores were higher in most groups of features except character 5 and 6-grams. The results for French and Spanish are, however, lower than those obtained by Portuguese, which reached 0.998 accuracy for character 4-grams.

3.2 Identifying Varieties or Identifying Newspapers?

Studies on identification of language varieties use standard corpora sampled from newspapers and magazines (Huang, 2008). They do not, however, address the question of textual genres and stylistics that underlie these samples. Newspapers and magazines could contain distinctive features that influence the performance of the classifiers. To explore this variable we carried out a controlled experiment using two newspapers from Spain. A corpus with texts from *El País*, published in 2008, was compiled and classified with the *El Mundo* corpus.

Feature	Mundo x País	Difference
W 1-grams	0.614	-33.4%
W 2-grams	0.498	-39.6%
C 2-grams	0.658	-24.0%
C 3-grams	0.654	-29.4%
C 4-grams	0.728	-21.6%
C 5-grams	0.688	-27.4%
C 6-grams	0.564	-39.6%

Table 3: El Mundo x El Pais

Results are 21.6% to 39.6% worse than the classification of Argentinian and Peninsular Spanish. In one case, word bi-grams, the classification result is lower than the 50% baseline expected for binary classification. The poor results obtained suggest that the language models applied here are actually distinguishing the varieties and that the text types and genres do not substantialy influence the algorithm's choice.

3.3 Multilingual Classification

To evaluate the classification model we integrated the four language models described so far: Spain, Argentina, France and Quebec, with two Portuguese varieties: Brazil and Portugal (Zampieri and Gebre, 2012). The results for this six-class classification model are presented in terms of Accuracy, Recall, Precision and F-Measure.

Feature	A	R	P	F
W 1-grams	0.917	0.917	0.905	0.911
W 2-grams	0.878	0.878	0.866	0.872
C 2-grams	0.898	0.898	0.880	0.889
C 3-grams	0.947	0.947	0.933	0.940
C 4-grams	0.910	0.910	0.890	0.899
C 5-grams	0.924	0.924	0.905	0.915
C 6-grams	0.935	0.935	0.932	0.933

Table 4: 6-Class Classification

4. Conclusion and Future Work

Studies on language indentification neglect pluricentric languages. We therefore presented a language identification model focusing on language varieties. The best results on the binary classification were obtained by using character n-grams. For Argentinian and Peninsular Spanish, 0.962 using 5-grams and for Quebec and Hexagonal French, 0.990 using 3-grams. The six-class model reached 0.947 accuracy and 0.940 f-measure using character 3-grams.

This work shed light on two areas. First, it shows that language identification models may include language varieties without substantial loss of performance. This should help to increase performance in NLP applications such as spell checking and MT systems. The second area is contrastive linguistics. Pluricentric languages are often the object of study due to their variation in grammar and syntax. Classification experiments such as this can provide a quantitative overview on how varieties diverge and converge.

Further experiments include the integration of these models into broader classification schemes (up to 20-fold) and the use of more knowledge-rich features such as POS bi-grams, to measure the extent to which these varieties differ in terms of grammar.

5. References

T. Dunning. 1994. *Statistical Identification of Language* Technical Report MCCS-94-273 New Mexico State University

C. Huang; L. Lee. 2008 Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity *Proceedings of PACLIC 2008* p. 404-410

N. Ljubesic; N. Mikelic; D. Boras. 2007. Language Identificaiton: How to Distinguish Similar Languages? *Proceedings of the 29th International Conference on Information Technology Interfaces*

M. Lui; T. Baldwin. 2012 langid.py: An Off-the-shelf Language Identification Tool *Proceedings of the 50th Meeting of the ACL* p. 25-30

B. Martins; M. Silva. 2005. Language Identification in Web Pages *Proceedings of the 20th ACM Symposium on Applied Computing (SAC)* Santa Fe, EUA. p. 763-768

D. Trieschnigg; D. Hiemstra; M. Theune; F. de Jong; T. Meder. 2012. An exploration of language identification techniques for the Dutch Folktale Database. *Proceedings of LREC2012*

M. Zampieri; B. G. Gebre. 2012. Automatic Identification of Language Varieties: The Case of Portuguese *Proceedings of KONVENS2012*. Vienna, Austria. p. 233-237