Automatic Identification of Language Varieties: The Case of Portuguese

Marcos Zampieri
University of Cologne
Albertus-Magnus-Platz 1, 50931
Cologne, Germany
mzampier@uni-koeln.de

Binyam Gebrekidan Gebre
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD
Nijmegen, Holland
bingeb@mpi.nl

Abstract

Automatic Language Identification of written texts is a well-established area of research in Computational Linguistics. State-of-the-art algorithms often rely on n-gram character models to identify the correct language of texts, with good results seen for European languages. In this paper we propose the use of a character n-gram model and a word n-gram language model for the automatic classification of two written varieties of Portuguese: European and Brazilian. Results reached 0.998 for accuracy using character 4-grams.

1 Introduction

One of the first steps in almost every NLP task is to distinguish which language(s) a given document contains. The internet is an example of a large text repository that contains languages that are often unidentified. Computational methods can be applied to determine a document's language before undertaking further processing. State-of-the-art methods of language identification for most European languages present satisfactory results above 95% accuracy (Martins and Silva, 2005).

This level of success is common when dealing with languages which are typologically not closely related (e.g. Finnish and Spanish or French and Danish). For these language pairs, distinction based on character n-gram models tends to perform well. Another aspect that may help language identification is the contrast between languages with unique character sets such as Greek or Hebrew. These languages are easier to identify if compared to language pairs with similar character sets: Arabic and Persian or Russian and Ukrainian (Palmer, 2010).

Martins and Silva (2005) present results on the identification of 12 languages by classifying 500 documents. Results varied according to language ranging from 99% accuracy for English to 80% for Italian. The case of Italian is particularly representative of what we propose here: among 500 texts classified, 20 were tagged as Portuguese and 42 as Spanish. Given that Italian, Portuguese and Spanish are closely related Romance languages, it is evident why algorithms have difficulty classifying Italian documents.

This example shows that a seemingly simple distinction task gains complexity when used to differentiate languages from the same family. In this study, we aim to go one step further and apply computational methods to identify two varieties of the same language: European and Brazilian Portuguese.

2 Related Work

The problem of automatic language identification is not new and early approaches to it can be traced back to 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. Ingle's experiments are different from those used in state-of-the-art language identification, which relies heavily on n-gram models and statistics applied to large corpora.

Dunning (1994) was one of the first to use character n-grams and statistics for language identi-

fication. In this study, the likelihood of n-grams was calculated using Markov models and this was used as the key factor for identification. After Dunning, other studies using n-gram models were published such as (Cavnar and Trenkle, 1994), which developed a language identification tool called TextCat[1] (Grafenstette, 1995), and more recently (Vojtek and Belikova, 2007).

Given its vast amount of multilingual material, the Internet became an important application of language identification. Documents are often unidentified regarding source language and the same document may contain more than one language. Examples of language identification applied to Internet data are (Martins and Silva, 2005) and later (Rehurek and Kolkus, 2009).

2.1 Identifying Similar Languages

If general purpose methods for automatic language identification were substantially explored, the same is not true for methods designed to deal specifically with similar languages or varieties. The identification of closely related languages seems to be the weakness of most n-gram based models and there are few recent studies published about it.

Recently this aspect of language identification received more attention, including a study by Ljubesic et al. (2007). Ljubesic et al. propose a computational model for identification of Croatian texts in comparison to other Slavic languages, namely: Slovenian, Bosnian and Serbian. The study reports 99% recall and precision in three processing stages.

The distinction between languages, dialects or varieties can be political. Serbian, Bosnian and Croatian were all variants of the Serbo-Croatian language spoken in the former Yugoslavia. After their independence, each of these countries adopted their variety as a national language. In the case of Portuguese, although there are substantial differences between Brazilian and European Portuguese, it is widely accepted that these two are varieties of the same language.

Another study worth mentioning is Piersman et al. (2010) on lexical variation. Piersman et. al. applied distributional lexical semantics to synonymy retrieval and it was also used for the identification of distinctive features (which the authors call lectal markers) in Dutch and Flemish. Experiments measuring lexical variation, focusing on convergences and divergences in lexicons, were recently carried out for Brazilian and European Portuguese by (Soares da Silva, 2010).

A couple of recent studies for spoken Portuguese were published by scholars like (Rouas et al., 2008) and later (Koller et al., 2010). These studies model the substantial phonetic differences among Brazilian, African and European Portuguese and discussed how to distinguish them automatically. For written Portuguese, however, to our knowledge there have been no studies published.

The experiments presented here can open new research perspectives in two areas: contrastive linguistics and NLP. For contrastive linguistics, our experiments provide a quantitative estimation on the differences between the two varieties, hence the three groups of features used: lexico-syntactical, lexical and orthographical. Secondly, in real-world NLP applications. Brazilian and European Portuguese do not share a common orthography and identifying the variety of a Portuguese text will help NLP tasks such as spell checking.

3 Linguistic Motivation

Although they are considered to be the same language, there are substantial differences between European and Brazilian Portuguese in terms of phonetics, syntax, lexicon and orthography. For the analysis of written texts, differences in syntax, lexicon and orthography were considered.

Orthography in these language varieties differs in two main aspects: graphical signs and mute consonants. Due to phonetic differences, Brazilian and European Portuguese use different orthographical signs for the same word, such as:

- *econômico* (BP); *económico* (EP): *economic* (EN)

Mute consonants are still used in the Portuguese orthography and are no longer used in Brazil:

- *ator* (BP); *actor* (EP):*actor* (EN)

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

Differences also appear at the syntactic level. Some contractions are only uses in one of the varieties; for instance: *mo* (pronoun *me* + definite masculine article *o*) is exclusive to Portugal. Past verb tenses (perfect and imperfect) are used in different contexts. The use of pronouns also differs, the Brazilian variety tends to prefer the pronoun before the verb whereas the European variety uses it primarily afterwards:

- *eu te amo* (BP) and *eu amo-te* (EP): *I love you* (EN)

Lexical variation is also a distinctive characteristic of these varieties. Some words are frequent in one of the varieties and rare in the other: *nomeadamente* (EP), *namely* (EN) is widely used in Portugal and rare in Brazil. Additionally, there are cases in which each variety may heavily favor a different word in a set of synonyms, such as: *coima*(EP), *multa* (BP), *fine , penalty* (EN).

4 Methods

In order to create a reliable classification method to distinguish Brazilian and European Portuguese, we compiled two journalistic corpora containing texts from each of the two varieties. The Brazilian corpus contains texts published in 2004 by the newspaper *Folha de São Paulo* and the Portuguese corpus contains texts from *Diário de Notícias*, published in 2007. Texts were pre-processed using Python scripts: all meta-information and tags were removed.

The length of texts in the corpora varies and we therefore classified them and grouped them together according to their length in tokens. Language identification and classification tasks tend to be favoured when using large documents and we explore this variable in section 5.2.

4.1 Experiments

The features used take into account differences in lexicon, syntax and orthography. For the orthographical differences, we used character n-grams ranging from 2 to 6-grams. At the lexical level identification was performed using word unigrams and finally, to explore lexico-syntactical differences we used word bi-grams. The language models were calculated using the Laplace probability distributions using a function available in NLTK (Bird et al., 2009) as shown in equation 1:

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

In the 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 (Dunning, 1993) represented in equations 2 and 3:

$$P(L|text) = argmax_L P(text|L)P(L) \quad (2)$$

$$P(L|text) = argmax_L \prod_{i=1}^{N} P(n_i|L)P(L) \quad (3)$$

Equation 3 is a detailed version of equation number 2 and it shows how the classification is made. 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 (log-likelihood) for each of the language models. The language model with higher probability determines the identified language of the text.

5 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. We report results in terms of accuracy calculated as follows:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (4)$$

The formula can be interpreted as the number of instances correctly classified ($tp+tn$) divided by all instances classified.

5.1 Word Uni-Grams

The word uni-gram features were used to perform classification taking into account lexical differences. Accuracy results are reported using texts of maximum 300 tokens each.

Max. Len.	Accuracy
300 words	0.996

Table 1: Word Uni-Gram Results

Proper nouns play an important role when using word uni-grams. It is very likely that texts from Portugal will contain named entities that are almost exclusively used in Portugal and vice-versa (e.g. names of important or famous people from Brazil/Portugal or names of companies).

5.2 Word Bi-Grams

For the word bi-gram model evaluation, we explored how the maximum length of texts affects the performance of the algorithm. The results were classified according to the maximum text size, ranging from 100 words to 700 words. The best results were reached with a maximum length of 500 words, after that, the model seems to indicate saturation, as can be seen in table 2:

Max. Len.	Accuracy
100 words	0.851
200 words	0.886
300 words	0.889
400 words	0.904
500 words	0.912
600 words	0.912
700 words	0.905

Table 2: Text Size and Word N-Grams

Results from table 2 are presented in figure number 1:

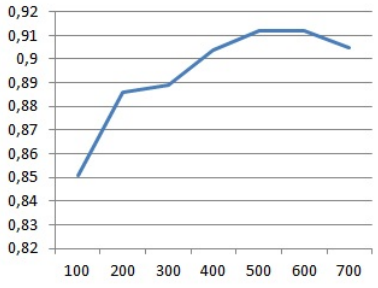

Figure 1: Text Size and Word Bi-Grams

One explanation for the results is that only a few journalistic texts in both corpora are larger than 500 words. Adding these few texts to the classification brings no improvement in the algorithm's performance.

5.3 Character N-Grams

The best results obtained in our experiments relied on a character n-gram model and were designed using texts of maximum 300 tokens. Results reached 0.998 accuracy for 4-grams, and they are presented in table number 3:

N-Grams	Accuracy
2-Grams	0.994
3-Grams	0.996
4-Grams	0.998
5-Grams	0.988
6-Grams	0.990

Table 3: Character N-Gram Results

The good results obtained by character n-grams in comparison to the word n-gram models presented in the previous sections, indicate that the orthographical differences between Brazilian and European Portuguese are still a strong factor for distinguishing these varieties.

6 Conclusions and Future Work

This paper explored computational techniques for the automatic identification of two varieties of Portuguese. From the three groups of features tested, the character-based model using 4-grams performed best. The results for word bi-grams were not very good reaching an accuracy result of 0.912. These outcomes suggest that for this task lexico-syntactic differences are not as important as differences in orthography and lexicon.

The small number of classes contributes to the encouraging results presented here. When performing binary classification, the baseline result is already 0.50 and when provided with meaningful features, algorithms are quite likely to achieve good results.

Experiments are being carried out to integrate the character-based model described in this paper in a real-world language identification tool. Preliminary results for this model in a 6-fold classification reached above 90% accuracy and f-measure.

Acknowledgments

The authors express gratitude to *Folha de São Paulo* for the Brazilian corpus and Sascha Diwersy for the European corpus. Many thanks to the anonymous reviewers who provided important feedback to increase the quality of this paper.

References

Bird, S.; Klein, E.; Loper, E. 2009. *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit* O'Reilly Media

Cavnar, W.; Trenkle, J. 1994. N-gram-based Text Categorization. *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* Volume 19, Issue 1. MIT Press

Dunning, T. 1994. Statistical Identificaion of Language *Technical Report MCCS-94-273* Computing Research Lab, New Mexico State University

Grafenstette, G. 1995. Comparing two Language Identification Schemes. *JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data* Rome

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

Koller, O.; Abad, A.; Trancoso, I.; Viana, C. 2010. Exploiting Variety-dependent Phones in Portuguese Variety Identification Applied to Broadcast News Transcription. *Proceedings of Interspeech 2010*

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

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

Palmer, D. 2010. Text Processing. In: Indurkhya, N.; Damerau, F. (Ed.) *Handbook of Natural Language Processing - Second Edition.*. CRC Press. p. 9-30

Piersman, Y.; Geeraerts, D.; Spelman, D. 2010. The Automatic Identification of Lexical Variation Between Language Varieties. *Natural Language Engieeniring 16 (4).*. Cambridge University Press. p. 469-491

Rehurek, R.; Kolkus, M. 2009. Language Identification on the Web: Extending the Dictionary Method. *Proceedings of CICLing. Lecture Notes in Computer Sciences*. Springer. p. 357-368

Rouas, J.; Trancoso, I.; Ribeiro, M.; Abreu, M. 2008. Language and Variety Verification on Broadcast News for Portuguese. *Speech Communication vol. 50 n.11*. Elsevier

Soares da Silva, A. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese: endo/exogeneousness and foreign and normative influence. *Advances in Cognitive Sociolinguistics*. Berlin. De Gruyter

Vojtek, P;. Belikova, M. 2007. Comparing Language Identification Methods Based on Markov Processess. *Slovko, International Seminar on Computer Treatment of Slavic and East European Languages*