A Low Dimensionality Representation for Language Variety Identification

Abstract

Language variety identification aims at labelling texts in a native language (e.g. Spanish, Portuguese, English) with its specific variation (e.g. Argentina, Chile, Mexico, Peru, Spain; Brazil, Portugal; UK, US). In this work we propose a low dimensionality representation (LDR) to address this task with five different varieties of Spanish: Argentina, Chile, Mexico, Peru and Spain. We compare our LDR method with common state-of-the-art representations and show an increase in accuracy of ~35%. Furthermore, we compare LDR with two reference distributed representation models. Experimental results show competitive performance while dramatically reducing the dimensionality --and increasing the big data suitability-- to only 6 features per variety. Additionally, we analyse the behaviour of the employed machine learning algorithms and the most discriminating features. Finally, we employ an alternative dataset to test the robustness of our low dimensionality representation with another set of similar languages.

Introduction

Language variety identification aims at labelling texts in a native language (e.g. Spanish, Portuguese, English) with their specific variation (e.g. Argentina, Chile, Mexico, Peru, Spain; Brazil, Portugal; UK, US). Although at first sight language variety identification may seem a classical text classification problem, cultural idiosyncrasies may influence the way users construct their discourse, the kind of sentences they build, the expressions they use or their particular choice of words. Due to that, we can consider language variety identification as a double problem of text classification and author profiling, where information about how language is shared by people may help to discriminate among classes of authors depending on their language variety.

This task is specially important in social media. Despite the vastness and accessibility of the Internet destroyed frontiers among regions or traits, companies are still very interested in author profiling segmentation. For example, when a new product is launched to the market, knowing the geographical distribution of opinions may help to improve marketing campaigns. Or given a security threat, knowing the possible cultural idiosyncrasies of the author may help to better understand who could have written the message.

Language variety identification is a popular research topic of natural language processing. In the last years, several tasks and workshops have been organized: the Workshop on Language Technology for Closely Related Languages and Language Variants @ EMNLP 2014; the VarDial Workshop @ COLING 2014 - Applying NLP Tools to Similar Languages, Varieties and Dialects; and the LT4VarDial - Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialect @ RANLP BIBREF0 BIBREF1 . We can find also several works focused on the task. In BIBREF2 the authors addressed the problem of identifying Arabic varieties in blogs and social fora. They used character n -gram features to discriminate between six different varieties and obtained accuracies between 70%-80%. Similarly, BIBREF3 collected 1,000 news articles of two varieties of Portuguese. They applied different features such as word and character n -grams and reported accuracies over 90%. With respect to the Spanish language, BIBREF4 focused on varieties from Argentina, Chile, Colombia, Mexico and Spain in Twitter. They used meta-learning and combined four types of features: i) character n -gram frequency profiles, ii) character n -gram language models, iii) Lempel-Ziv-Welch compression and iv) syllable-based language models. They obtained an interesting 60%-70% accuracy of classification.

We are interested in discovering which kind of features capture higher differences among varieties. Our hypothesis is that language varieties differ mainly in lexicographic clues. We show an example in Table 1 .

In this work we focus on the Spanish language variety identification. We differentiate from the previous works as follows: i) instead of n -gram based representations, we propose a low dimensionality representation that is helpful when dealing with big data in social media; ii) in order to reduce the possible over-fitting, our training and test partitions do not share any author of instance between them; and iii) in contrast to the Twitter dataset of BIBREF4 , we will make available our dataset to the research community.

Low Dimensionality Representation

The key aspect of the low dimensionality representation (LDR) is the use of weights to represent the probability of each term to belong to each one of the different language varieties. We assume that the distribution of weights for a given document should be closer to the weights of its corresponding language variety. Formally, the LDR is estimated as follows:

Evaluation Framework

In this section, we describe the corpus and the alternative representations that we employ in this work.

HispaBlogs Corpus

We have created the HispaBlogs dataset by collecting posts from Spanish blogs from five different countries: Argentina, Chile, Mexico, Peru and Spain. For each country, there are 450 and 200 blogs respectively for training and test, ensuring that each author appears only in one set. Each blog contains at least 10 posts. The total number of blogs is 2,250 and 1,000 respectively. Statistics of the number of words are shown in Table 3 .

Alternative representations

We are interested in investigating the impact of the proposed representation and compare its performance with state-of-the-art representations based on n -grams and with two approaches based on the recent and popular distributed representations of words by means of the continuous Skip-gram model BIBREF6 .

State-of-the-art representations are mainly based on n -grams models, hence we tested character and word based ones, besides word with tf-idf weights. For each of them, we iterated n from 1 to 10 and selected 1,000, 5,000 and 10,000 most frequent grams. The best results were obtained with the 10,000 most frequent BOW, character 4-grams and tf-idf 2-grams. Therefore, we will use them in the evaluation.

Due to the increasing popularity of the distributed representations BIBREF7 , we used the continuous Skip-gram model to generate distributed representations of words (e.g. n -dimensional vectors), with further refinements in order to use them with documents. The continuous Skip-gram model BIBREF8 , BIBREF9 is an iterative algorithm which attempts to maximize the classification of the context surrounding a word. Formally, given a word $w(t)$, and its surrounding words $w(t-c),~w(t-c+1),...,~w(t+c)$ inside a window of size $2c+1$, the training objective is to maximize the average of the log probability shown in Equation 23 :

$$\frac{1}{T} \displaystyle \sum _{t=1}^T \displaystyle \sum _{-c \le j \le c,j \ne 0} \log p(w_{t+j}|w_t)$$ (Eq. 23)

To estimate $p(w_{t+j}|w_t)$ we used negative sampling BIBREF9 that is a simplified version of the Noise Contrastive Estimation (NCE) BIBREF10 , BIBREF11 which is only concerned with preserving vector quality in the context of Skip-gram learning. The basic idea is to use logistic regression to distinguish the

target word W_O from draws from a noise distribution $P_n(w)$, having k negative samples for each word. Formally, the negative sampling estimates $p(w_O|w_I)$ following Equation 24 :

$$\log \sigma (v^{\prime }_{w_O}{}^T v_{w_I}) + \displaystyle \sum _{i=1}^k \mathbb {E}_{w_i}\sim P_n(w) \bigg [\log \sigma (-v^{\prime }_{w_i}{}^T v_{w_I}) \bigg]$$ (Eq. 24)

where $\sigma (x)=1/(1+\exp (-x))$. The experimental results in BIBREF9 show that this function obtains better results at the semantic level than hierarchical softmax BIBREF12 and NCE.

In order to combine the word vectors to represent a complete sentence we used two approaches. First, given a list of word vectors $(w_1,w_2,...,w_n)$ belonging to a document, we generated a vector representation v of its content by estimating the average of their dimensions: $v=n^{-1}\sum _{i=1}^n w_i$. We call this representation Skip-gram in the evaluation. In addition, we used Sentence vectors (SenVec) BIBREF13 , a variant that follows Skip-gram architecture to train a special vector sv representing the sentence. Basically, before each context window movement, SenVec uses a special vector sv in place of $w(t)$ with the objective of maximizing the classification of the surrounding words. In consequence, sv will be a distributed vector of the complete sentence.

Following state-of-the-art approach BIBREF13 , in the evaluation we used a logistic classifier for both SenVec and Skip-gram approaches.

Experimental Results

In this section we show experimental results obtained with the machine learning algorithms that best solve the problem with the proposed representation, the impact of the preprocessing on the performance, the obtained results in comparison with the ones obtained with state-of-the-art and distributed

representations, the error analysis that provides useful insights to better understand differences among languages, a depth analysis on the contribution of the different features and a cost analysis that highlights the suitability of LDR for a big data scenario.

Machine learning algorithms comparison

We tested several machine learning algorithms with the aim at selecting the one that best solves the task. As can be seen in Table 4 , Multiclass Classifier obtains the best result (results in the rest of the paper refer to Multiclass Classifier). We carried out a statistical test of significance with respect to the next two systems with the highest performance: SVM ($z_{0.05} 0, 880 < 1, 960$) and LogitBoost ($z_{0.05} = 1, 983 > 1, 960$).

Preprocessing impact

The proposed representation aims at using the whole vocabulary to obtain the weights of its terms. Social media texts may have noise and inadequately written words. Moreover, some of these words may be used only by few authors. With the aim at investigating their effect in the classification, we carried out a preprocessing step to remove words that appear less than n times in the corpus, iterating n between 1 and 100. In Figure 1 the corresponding accuracies are shown. In the left part of the figure (a), results for n between 1 and 10 are shown in a continuous scale. In the right part (b), values from 10 to 100 are shown in a non-continuous scale. As can be seen, the best result was obtained with n equal to 5, with an accuracy of 71.1%. As it was expected, the proposed representation takes advantage from the whole vocabulary, although it is recommendable to remove words with very few occurrences that may alter the results. We show examples of those infrequent words in Table 5 .

In Figure 2 , when analysing the evolution of the number of remaining words in function of the value of

n , we can see a high number of words with very low frequency of occurrence. These words may introduce a high amount of noise in our LDR weight estimation. In addition, removing these words may be also beneficial in order to reduce the processing time needed to obtain the representation. This fact has special relevance for improving the performance in big data environments.

Language variety identification results

In Table 6 we show the results obtained by the described representations employing the Multiclass Classifier. As can be appreciated, the proposed low dimensionality representation improves more than 35% the results obtained with the state-of-the-art representations. BOW obtains slightly better results than character 4-grams, and both of them improve significantly the ones obtained with tf-idf 2-grams. Instead of selecting the most frequent n -grams, our approach takes advantage from the whole vocabulary and assigns higher weights to the most discriminative words for the different language varieties as shown in Equation 10 .

We highlight that our LDR obtains competitive results compared with the use of distributed representations. Concretely, there is no significant difference among them (Skip-gram $z_{0.05} = 0,5457 < 1,960$ and SenVec $z_{0.05} = 0,7095 < 1,960$). In addition, our proposal reduces considerably the dimensionality of one order of magnitude as shown in Table 6 .

Error analysis

We aim at analysing the error of LDR to better understand which varieties are the most difficult to discriminate. As can be seen in Table 7 , the Spanish variety is the easiest to discriminate. However, one of the highest confusions occurs from Argentinian to Spanish. Mexican and Spanish were considerably confused with Argentinian too. Finally, the highest confusion occurs from Peruvian to Chilean, although

the lowest average confusion occurs with Peruvian. In general, Latin American varieties are closer to each other and it is more difficult to differentiate among them. Language evolves over time. It is logical that language varieties of nearby countries — as the Latin American ones — evolved in a more similar manner that the Spanish variety. It is also logical that even more language variety similarities are shared across neighbour countries, e.g. Chilean compared with Peruvian and Argentinian.

In Figure 3 we show the precision and recall values for the identification of each variety. As can be seen, Spain and Chile have the highest recall so that texts written in these varieties may have less probability to be misclassified as other varieties. Nevertheless, the highest precisions are obtained for Mexico and Peru, implying that texts written in such varieties may be easier to discriminate.

Most discriminating features

In Table 8 we show the most discriminant features. The features are sorted by their information gain (IG). As can be seen, the highest gain is obtained by average, maximum and minimum, and standard deviation. On the other hand, probability and proportionality features has low information gain.

We experimented with different sets of features and show the results in Figure 4 . As may be expected, average-based features obtain high accuracies (67.0%). However, although features based on standard deviation have not the highest information gain, they obtained the highest results individually (69.2%), as well as their combination with average ones (70,8%). Features based on minimum and maximum obtain low results (48.3% and 54.7% respectively), but in combination they obtain a significant increase (61.1%). The combination of the previous features obtains almost the highest accuracy (71.0%), equivalent to the accuracy obtained with probability and proportionality features (71.1%).

Cost analysis

We analyse the cost from two perspectives: i) the complexity to the features; and ii) the number of features needed to represent a document. Defining l as the number of different language varieties, and n the number of terms of the document to be classified, the cost of obtaining the features of Table 2 (average, minimum, maximum, probability and proportionality) is $O(l\cdot {n})$. Defining m as the number of terms in the document that coincides with some term in the vocabulary, the cost of obtaining the standard deviation is $O(l\cdot {m})$. As the average is needed previously to the standard deviation calculation, the total cost is $O(l\cdot {n}) + O(l\cdot {m})$ that is equal to $O(max(l\cdot {n}, l\cdot {m})) = O(l\cdot {n})$. Since the number of terms in the vocabulary will always be equal or greater than the number of coincident terms ($n \ge m$), and as the number of terms in the document will always be much higher than the number of language varieties ($l<<n$), we can determine the cost as lineal with respect to the number of terms in the document $O(n)$. With respect to the number of features needed to represent a document, we showed in Table 6 the considerable reduction of the proposed low dimensionality representation.

Robustness

In order to analyse the robustness of the low dimensionality representation to different languages, we experimented with the development set of the DSLCC corpus from the Discriminating between Similar Languages task BIBREF1 . The corpus consists of 2,000 sentences per language or variety, with between 20 and 100 tokens per sentence, obtained from news headers. In Table 9 we show the results obtained with the proposed representation and the two distributed representations, Skip-gram and SenVec. It is important to notice that, in general, when a particular representation improves for one language is at cost of the other one. We can conclude that the three representations obtained comparative results and support the robustness of the low dimensionality representation.

Conclusions

In this work, we proposed the LDR low dimensionality representation for language variety identification. Experimental results outperformed traditional state-of-the-art representations and obtained competitive results compared with two distributed representation-based approaches that employed the popular continuous Skip-gram model. The dimensionality reduction obtained by means of LDR is from thousands to only 6 features per language variety. This allows to deal with large collections in big data environments such as social media. Recently, we have applied LDR to the age and gender identification task obtaining competitive results with the best performing teams in the author profiling task at the PAN Lab at CLEF. As a future work, we plan to apply LDR to other author profiling tasks such as personality recognition.