Determining language variant in microblog messages

Gustavo Laboreiro  
LIACC / SAPO Labs  
Faculdade de Engenharia  
Universidade do Porto  
gustavo.laboreiro@fe.up.pt

Matko Bošnjak  
mbosnjak@fe.up.pt

Luís Sarmento  
SAPO Labs  
Universidade do Porto  
sarmento@birdwing.net

Eduarda Mendes Rodrigues  
DEI  
Faculdade de Engenharia  
Universidade do Porto  
eduardamr@acm.org

Eugénio Oliveira  
LIACC / DEI  
Faculdade de Engenharia  
Universidade do Porto  
eco@fe.up.pt

ABSTRACT

It is difficult to determine the country of origin of the author of a short message based only on the text. This is an even more complex problem when more than one country uses the same native language. In this paper, we address the specific problem of detecting the two main variants of the Portuguese language — European and Brazilian — in Twitter micro-blogging data, by proposing and evaluating a set of high-precision features. We follow an automatic classification approach using a Naïve Bayes classifier, achieving 95% accuracy. We find that our system is adequate for real-time tweet classification.

1. INTRODUCTION

In this work, the specific problem we are trying to solve is identifying Portuguese users in Twitter. The official language in Portugal is Portuguese. That is also the official language in other countries, such as Brazil. However, due to cultural differences, several words evolved towards different meanings in these countries. In order to semantically understand the text, we first need to identify the nationality of the author. For example, “Nivea” is a German skin-care company very popular in Portugal, but in Brazil it is a common proper name. The text “I love Nivea” could be interpreted either as a statement of brand recognition and approval, or as a personal expression of affection for another person.

Hong et al. [10] placed Portuguese as the third most used language on Twitter in 2011, with a percentage of about 9% of all messages. Portuguese is spoken mainly in Portugal and Brazil, with Brazil having approximately 20 times the population of Portugal. This shows how asymmetric the problem is: choosing a random Tweet in Portuguese, there is a 95% chance of it originating in Brazil — considering equal Twitter usage in both countries. Actually, Brazil is the second most represented country in this microblog system1.

Since only a small fraction of users of social networks add usable information to their profiles, such as location [3], identifying the nationality of the author becomes an intricate problem. Even when present, this information can be too broad (e.g. “Europe”), too specific (e.g. the street name), contain spelling errors, ambiguous names (e.g. “Lagos” — in Portugal or in Nigeria?), or present misleading information (e.g. “Hogwarts”). Thus, we are left with inferring the location of the users through indirect methods. Several solutions have been considered.

Time zones could help in locating an user on the globe. For example, by observing the regular inactive hours of users that could match their sleeping time. Unfortunately, the number of messages required to create a meaningful user profile would limit this classification to the users that post frequently enough. Portugal and Brazil are too close geographically to allow for conclusive decisions regarding a vast percentage of users. Also, automatic posting systems, such as those from online news, have posting patterns that are less dependent on the time of day [13]. Finally, when analysing some events — for example, a World Cup soccer game — the difference in time zones is less meaningful, as the users are reacting at the same moment.

Fink et al. used references to geographic features or locations [6], but since Portugal is a destination often chosen by emigrants, particularly from other Portuguese-speaking countries [12], by itself this solution is insufficient in identifying the nationality of the author.

Social network information (friends and followers) is capable of providing us with an accurate location [7]. This was not used in our work because, to the extent that is possible, we wanted to rely only on information that comes with the tweets, and the network of the user does not.

We opted to tackle this problem by looking more deeply at the language used to write the messages. In many cases there is a strong affinity between languages and geographic locations, but to be truly efficient in guessing the location of a person, we need to be able to differentiate between variants of the same language.

We wish to perform this variant detection work as a classi-
fication task, where our scientific contribution lies in the set of features we propose. We distinguish between the analysis of the features (the present work) and the automatic updating of the vocabulary that assists a few of them. Since the latter task is dependent on the success of the former, we use manually constructed lists, and defer the maintenance of those referred lists to subsequent work.

We focus on features that are simple to process, work with minimal available text, and that, within the scope of microblogging messages, reflect differences between the European and Brazilian variants of Portuguese. These features cover common expressions, named entities, common grammatical variations, URL and writing styles. We train classification models on datasets containing up to 100 Twitter messages from each of 1400 Portuguese and 1400 Brazilian users. Although not manually annotated, our largest dataset is 56 times larger than the one used by Carter et al. [1] (5000 messages distributed across 5 languages). We show that our proposed features have outperformed the traditional n-gram approach in accuracy (80%) achieving up to 95%, when using a Naïve Bayes classifier. We also show that our proposed features can be analysed faster than n-grams while still producing more accurate results.

In the next section we present related work on language identification. In section 3 we explain our approach for language variance identification. In Section 4 we describe the generation of our dataset and the experiment performed. Our results are then introduced and analysed, and our conclusions are drawned in Section 7. We conclude with the future steps for our work.

2. RELATED WORK

Extensive work [11, 9] has been done in the area of automatic language identification. The most frequent approaches use character n-grams together with Naïve Bayes, or Markov Models. One of the most popular approaches is a method of profile ranking [2] in which the n-gram profile of a language is built, and further language classification relies on simple matching of profiles of texts of interest and the previously built reference profiles. All of the above methods attain high accuracy on moderate and large sized texts. However, the accuracy tends to drop slightly on shorter text passages.

A recent study by Gottron and Lipka [8] shows that Naïve Bayes, trained on 5-grams of Reuters news text in 10 different languages, and applied to the headlines of the news is able to reach 99% accuracy. The same authors achieved 81% accuracy on single-word queries. However, since news titles and news content usually have a strong affinity, high accuracy is not surprising.

Vatanen et al. [20] compared profile ranking [2] with Naïve Bayes using different smoothing methods. The models were trained on the Universal Declaration of Human Rights in 50 languages and tested on its short cut-outs (ranging from 5 to 21 characters). They attained precision of 72.5% and recall of 72.3%, and compared it to Google’s AJAX language API, with its 53.1% precision and 28.0% recall.

Distinguishing very similar languages on web pages has been identified as a difficult problem. Martins and Silva [16] have shown that their system, using a modified approach of Cavnar and Trenkle [2], is unable to distinguish between Brazilian and European Portuguese due to insignificant language differences. da Silva and Lopes [4] succeeded with 98% precision in distinguishing those two variants of Portuguese using n-grams with dimensionality reduction and Quadratic Discrimination Score. They used “formal writing documents” as a corpus (e.g. official Brazilian government documents and Portuguese news), with average length of 99 lines. The work by Ljubesic et al. [15] confirmed that distinguishing variants of languages is possible. They differentiated between Croatian and Serbian, using second-order Markov Models paired with a list of forbidden words. They reached precision and recall above 99%. However, Martins and Ljubesic [16, 15] used a controlled corpus of larger documents, and we are interested in short, microblog style texts.

News and web pages exhibit different linguistic properties than microblogs, that are a distinct kind of text, in part, due to their short format (e.g., 140 characters in Twitter). Tang et al. [19] showed that, from a linguistic and sentiment analysis perspective, there is a difference between microblog texts and other balanced corpus (e.g. news articles). This difference validates the use of different language processing methods for microblog messages. Hong et al. [10] pointed to cross-language differences in the adoption of Twitter entities such as URL, hashtags, mentions, replies and retweets. Furthermore they showed different semantic properties of languages.

Carter et al. [1] used the approach of Cavnar and Trenkle [2] to distinguish between 5 languages on Twitter. Each language was trained over n-grams of 1000 tweets in the corpus using additional prior knowledge. Prior knowledge consisted of the languages of web pages referenced in the tweet, the language guessed of previous posts, the language used by mentioned users, the language used in the previous post in the current conversation, and the language used in other tweets containing the same tag. They achieved accuracy of 92.4% without prior knowledge and 97.4% with it.

As part of their work on latent user attributes, Rao et al. [17] studied if they could determine if Indian Twitter users, communicating in English, were from the north or south region of India based only on their messages. With 200 users in each category, they achieved their best result (77.1% accuracy) when using their sociolinguistic-feature model. These features relate to forms of expression, such as the use of smileys, ellipses, upper-cased words, “LOL”, “OMG”, and so on. Classification was done using SVM.

3. METHODOLOGY

We aim to answer the following research questions: i) how can we classify microblog users according to their variant of Portuguese (specifically European vs. Brazilian), based solely on the contents of their messages; and ii) how do our classifiers perform as we increase the number of messages per user examined?

To this end, and knowing that our ultimate goal is to identify users from (or relevant to) the Portuguese Twitterosphere, we outline our approach in the following steps. First we create two sets of microblog users — one for Portugal and another for Brazil — and then proceed to sample messages from each user. We then train a classifier based on the texts written by a number of users, expecting to be able to predict the nationality of the remaining ones. The classification is supported by several groups of features that relate to six different strategies in language variant identification.

3.1 User selection

We wanted to create a large set of Portuguese and Brazil-
ian users for our study, both to capture the variety in language and style that exists in microblogs, and to provide statistical significance to the results. Unfortunately, it would take too long to manually annotate the required number of users in a proper way. We opted to use automatic annotations, employing filters over the stated location and the social network of each user.

To improve our confidence on our annotation, we added a second source of information. Gonzalez et al. show that users tend to have geographically close followers [7]. We believe that these links are, to a degree, a user’s connection to their community, and do not change, in essence, as easily as the geographic location of the user. Thus, we have greater confidence in a user that claims to reside in the same country as their friends, and should repudiate accounts that show contradictory information.

3.2 Features

We defined six feature groups: n-gram basic features and a set of five feature groups that we refer to as our proposed features. Some of our proposed features make use of pre-compiled lists of words, names or short expressions. These lists are meant to give an indication of the suitability of the identification strategy employed. We make these lists available at http://labs.sapo.pt/resources/.

3.2.1 N-gram based features

In our work, n-grams set a baseline against which all other groups of features are compared.

The n-grams feature group covers all sequences of one, two and three characters in the message. N-grams are frequently used in text classification due to their simplicity and distinct results [2], but they can also raise the dimensionality of the feature space and extend the processing time. To avoid this problem we consider only n-grams that occur more than 0.01 × (total number of messages) times.

3.2.2 Stylistic

The style in which messages are written can help distinguish one author from another. We define a set of stylistic features based on work by Sousa-Silva [18], since it is possible that cultural or social factors can influence the style employed by Twitter users from two different countries.

This feature group is divided into 4 segments: i) Quantitative markers, ii) Marks of Emotion, iii) Punctuation and iv) Accents. Of these, Marks of Emotion is expected to be the most relevant, since they showed the highest precision distinguishing authors [18].

Interjections are a way of textually expressing emotion, and given the orality influence in microblogs, they can be frequent. Through observation, we noticed that Brazilians tend to express emotion using different onomatopoeias than Portuguese (e.g. the laugh “kkk” is exclusive of Brazil).

To identify interjections we look at words with 5 or more letters, having a low ratio of distinct characters to the word length (≤ 0.5). In this way we can recognise simple interjections (e.g. “hehehe”) as well as more complex ones that do not follow a pattern (e.g. “hushaushusha”).

3.2.3 Entities

With the help of native speakers of each language variant, we put together two lists containing names that we consider likely to be mentioned in Social Media. Some of these names are idiomatic (e.g. fans of the Brazilian soccer club Fluminense are some times called “bambi”). They include regions, cities, politicians, soccer players, singers, actors, political parties, soccer clubs, big national companies, and similar entities that are frequently discussed online. The lists contain approximately 130 entries for each variant, including variations.

3.2.4 Word tokens

We want to determine the frequency of words exclusive to one of the two language variants. To this end we used GNU Aspell2 and its two dictionaries for both Portuguese variants. We also took note of the words we found in neither dictionary and used them as features, since they could be vernacular mentions or entity names. In particular, we assume that words composed only of consonants are abbreviations. We rely on this heuristic instead of employing a list of abbreviations that is difficult to keep updated.

We created two lists containing expressions more popular in one of the language variants. We identified over 200 regarding Brazil, and near 80 used in Portugal, accounting for variations such as the lack of accents, common abbreviations and misspellings. For example, this list contains words such as “billion” (“bilião” in Portugal, “billion” in Brazil), “bus” (“autocarro” in Portugal, “ônibus” in Brazil), and expressions such as “pisar na bola” (literally “to step on the ball”, but in Brazil it means to act against one’s expectations).

Finally, we also count the number of times monetary references occur for both Portugal (Euro) and Brazil (Real).

3.2.5 Grammar

Portuguese and Brazilians in some situations can show different preferences in grammar. This is noticeable in the verbal forms and the grammatical persons used. For example, in Brazil the gerund is used more often, and the second person of the singular “você” can be used in both formal and informal situations. In Portugal the infinitive is more frequent, and the second person of the singular “tu” is used in informal conversation.

Also, in Brazil the object pronoun frequently precedes the verb (e.g. “me escreve” — “[to me] write”), while in Portugal the opposite is more common (e.g. “escreve-me” — “write [to me]”). We use JSpell3, to detect these patterns, and use a feature as a flag when one is detected.

Another distinction is the absence of the definite article before a possessive. In Brazil it is often omitted, as it is optional in the language; as in “meu texto” — “my text”. In Portugal the article is frequently used: “o meu texto”.

3.2.6 URL

The last group of features is extracted from URL mentioned in the tweets. We maintain two types of features related to URL, after expansion from shortening services. The first type relates to the Top Level Domain (TLD). The national TLD for Portugal and Brazil are respectively “pt” and “br”. Here, we simply count the number of URL that we find, in each message, having each of these TLD.

However, there are many websites outside of these TLD which are significantly more popular in one of the countries. For example, the hostname for the Brazilian TV station Rede Globo is “redeglobo.globo.com”. To address this,

2http://aspell.net/
we use the second type of features that counts the number of times that we find each hostname in all URL in the message.

4. EXPERIMENTAL SETUP

The tests were made in two steps. In the first step we generate the features that describe messages in the dataset. In the second step we use a classifier to test the adequacy of the features using 5-fold cross validation.

Our goal is not to contrast different classification techniques for optimising classification accuracy, but to assess the usefulness of the sets of features we propose. Thus, the choice of classifier was secondary in our work. We opted for Naïve Bayes\(^3\) due to its simplicity and because it is commonly used as a baseline. It also enables us to process large amounts of features quickly.

4.1 Dataset generation

Our dataset was created by sampling Portuguese and Brazilian users and collecting their tweets. To ensure the sample was unbiased, we used the TREC Tweets2011 corpus\(^5\), which provides a random sample of Twitter (approximately 16 million messages) from 2011-01-24 to 2011-02-08.

In June 2011 we gathered information about all the users in the Tweets2011 corpus. In February 2012 we re-examined the users in this corpus, and discarded those that had changed the location in their profile.

We filtered out all Portuguese and Brazilian users by matching their free-form location field in Twitter against “Portugal”, “Brasil” (the native spelling for “Brazil”) or one of the top 10 cities (municipalities) for each country\(^6\), excluding “Porto” (in Portugal), that is frequently mistaken for “Porto Alto” (in Brazil). From the remaining users, we further excluded those from whom we could not retrieve more than 100 messages (excluding retweets).

In order to enable the social network filtering, we also retrieved information from the accounts that were following each of the relevant users. We selected all users with more than 10 followers, where the ratio of followers from their country exceeds those from the other country by a factor greater than 3. These conditions resulted from experimentation, and allow for an acceptable balance in the number of users in each set (2768 users in Brazil and 1455 users in Portugal), and we consider them strict enough to satisfy our labelling using an unsupervised approach. From each of these sets we randomly selected 1400 users to use in our experiments.

Finally, we used a specialised tokenizer \([14]\) to process the messages from each user, and expanded the short URL.

After a native speaker of each language variant observed a 5% sample of each dataset, no irregularity was found in the sample of dataset “Brazil”. The sample of dataset “Portugal” was more difficult to evaluate, having 6 cases that precluded a definitive conclusion. One account showed mixed spellings in words, without one of the language variants being dominant. The remaining accounts had insufficient content written in Portuguese to create an informed opinion. In conclusion, these messages are ambiguous.

4.2 Determining the number of messages

In the first experiment we wish to determine how the number of messages available influences our classification. We created several datasets with 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50, 70 and 100 messages per user (MPU).

Messages were selected randomly, and concatenated into a single text, simplifying the later processing of the text with minimal impact.

4.3 Determining the most significant features

To determine the impact of each of our proposed feature groups, we re-ran the experiment using 100 MPU, each time excluding one feature group. We then measured the impact this absence caused in both accuracy and execution time. This is known as a mutilation test.

5. RESULTS

As Figure 1 shows, our proposed features offer nearly continuously increasing performance as more data is available. By contrast, n-gram features start to plateau fairly early (10 MPU), and fail to make use of the extra information. We can also see that our proposed features were more accurate than n-grams when exceeding 4 MPU. When using a small sample from each user (1 or 2 messages), n-grams achieve higher accuracy. Maximum n-gram accuracy was 0.80 (30 MPU), while our proposed features peaked at 0.87 (100 MPU).

In Figure 2 we can observe the total running times for the experiment, as a function of the sample size. This includes both training and testing times. The time cost of both n-grams and our proposed features grow linearly with the number of MPU, but at different rates.

The results of the mutilation tests are displayed in Table 1. Strong emphasis should be put in the Word tokens feature group. When we removed these features, processing time decreased by 2/3, but also incurred an almost 20% accuracy loss. By contrast, our proposed features showed an improvement of almost 9% in accuracy when we excluded the Stylistic feature group, reaching the maximum value we obtained: 95%. To us, this was rather surprising, since previous results had suggested that these features would be adequate for this task \([18]\).

\(^3\)http://search.cpan.org/dist/Algorithm-NaiveBayes/
\(^4\)https://sites.google.com/site/microblogtrack/
\(^5\)2011-guidelines
\(^6\)http://en.wikipedia.org/wiki/List_of_cities_in_Portugal,
http://en.wikipedia.org/wiki/List_of_largest_cities_in_Brazil
Feature groups
Running time (seconds)
10000
n-grams (all)
1000
100
10
proposed
n-grams
1 10 100

Figure 3: Feature space dimension of the proposed features, excluding a feature group and limited and unlimited n-gram feature groups, using 1, 10 and 100 MPU.

6. ANALYSIS

We should first recall that we are comparing an optimised version of n-grams with an unoptimised version of our system. In Figure 3 we represent the potential n-gram feature space dimension and of the feature space we actually used, regarding 1, 10 and 100 MPU. We also show the dimension of the feature space created by our five proposed groups. The number of considered n-grams is almost constant, due to the threshold described in Section 3.2.1. This allows an almost fixed execution time of the n-gram feature group. Without this threshold, the n-grams group would execute slower than our proposed features, since they accumulate many more features. This sort of feature selection could be harmful to our proposed features, since they are precision-oriented rather than recall-oriented (as n-grams). That is, they may appear infrequently, but when present they are very discriminative.

Figure 2: Processing time of three feature groups as a function of the number of messages used.

Table 1: Variations in accuracy and processing time of our proposed features, excluding a feature group in turn, using 100 messages per user.

<table>
<thead>
<tr>
<th>Feature Group removed</th>
<th>∆ accuracy</th>
<th>∆ time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>-0.16%</td>
<td>-26.83%</td>
</tr>
<tr>
<td>Stylistic</td>
<td>+8.81%</td>
<td>-21.49%</td>
</tr>
<tr>
<td>Grammar</td>
<td>-0.08%</td>
<td>-17.18%</td>
</tr>
<tr>
<td>URL</td>
<td>-0.33%</td>
<td>-16.53%</td>
</tr>
<tr>
<td>Word tokens</td>
<td>-19.37%</td>
<td>-66.51%</td>
</tr>
</tbody>
</table>

We can conclude that combining all features brought little gain in accuracy over our proposed features. The extra overhead in processing time has a minimal impact on accuracy. Compared to our proposed features, the maximum gain was inferior to 3%, when using one message per user, for over 5 times the original processing time. Compared to n-grams, 7.5% better accuracy can be obtained when using all 100 MPU, for nearly 9 times the original processing time.

The almost flat accuracy of n-grams when facing longer text samples, seen in Figure 1, seems to indicate that there is little gain in using a larger training set. In other words, we could say that the relative frequency of n-grams remains stable after a certain number of messages has been observed. To confirm this hypothesis, we compared the most popular n-grams when using 1 and 100 MPU. We used the Kendall tau distance [5] with penalty parameter 1 to measure the ranking similarity of the 200 most frequent n-grams. The similarity measures were normalised using the maximum $\frac{\pi n(n-1)}{2}$, where $n$ is the number of distinct n-grams in the union of both lists being compared. In this way, 0 means that both lists are equal, and 1 indicates that the lists are in total disagreement — e.g. the n-grams are in reverse order, or no common n-gram exists.

We calculated the similarity values for the sets containing 1 message per user, and 100 MPU in each language variant. One thing that we noticed was that the Brazilian variant scored higher (greater variation) in almost every case. The most significant differences were found in 1-grams — 0.258 for Portugal, 0.302 for Brazil. This is due to URL, that can introduce seemingly random characters in the text. Each Kendall tau distance reduces to 0.061 and 0.046 respectively, when excluding URLs. Longer n-grams present greater similarity between the smaller and larger message sets. In the same order as above, similarity for 2-grams measured as 0.042 and 0.047, and for 3-grams 0.094 and 0.109. Thus, we can conclude that 1400 messages (one from each user) is sufficient provide an accurate model of each language variant.

7. CONCLUSION

Our intent was to identify the variant of a language used on microblog messages, as a way to disambiguate the nationality of users. We worked on Portuguese, the official language of Portugal and Brazil, and automatically selected 1400 users referring to each country, based on information extracted from their profile and their social network. After presenting our proposed feature groups that we compared to an n-gram approach, using a Naive Bayes classifier, we tested both approaches using several sample datasets, varying from 1 to 100 messages per user.

Our main conclusion is that lexical differences provide the best discrimination among our proposed features, and shows a promising path for classification improvement.

N-grams required minimal input to generate an adequate model, and thus, when only one or two tweets are available, this feature group provided the best accuracy. This is relevant when, for example, the user writes infrequently, or writes most messages in a foreign language, and thus we need to make a decision based on few messages.

When more messages are available, we can expect higher accuracy from our proposed features. The most accurate classification was obtained by combining Entities, Word tokens, Grammar, and URL features. In this situation we were able to reached 95% accuracy in our experiments.
As for the number of messages form each user, we notice that beyond 10 messages, each additional message contributes less to the classification. With n-grams we see negligible improvements past that point, while our proposed features continue to improve at a lower rate.

8. FUTURE WORK

Some of the features presented here employ lists of vocabulary. The manual generation and maintenance of these lists is a tiresome endeavour, subject to mistakes or omissions that could greatly impact the results. The human factor limits the scalability of the system. For this reason, we intend to explore the process of generating these lists automatically. This could include external data sources, like newspapers or Wikipedia, and/or tweets.

As improvements to the classification process, we wish to test alternate classifiers, comparing their execution times and classification accuracy. We also aim to improve some features, namely in the stylistic group, and determine if they can help in user nationality classification or should be discarded.

Acknowledgments

This work is supported by Fundação para a Ciência e a Tecnologia (FCT) through the REACTION project: Retrieval, Extraction and Aggregation Computing Technology for Integrating and Organizing News (UTA-Est/MAI/0006/2009).

9. REFERENCES