PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **18** (2007), 341–360

PLISKA
STUDIA MATHEMATICA
BULGARICA

USING COVARIANCE AS A SIMILARITY MEASURE FOR DOCUMENT LANGUAGE IDENTIFICATION IN HARD CONTEXTS

Joaquim Ferreira da Silva Gabriel Pereira Lopes

Existing Language Identification (LID) approaches achieve 100% precision in most common situations, dealing with sufficiently large documents, written in just one language. However, there are many situations where text language is hard to identify and where current LID approaches do not provide a reliable solution. One such situation occurs when it is necessary to discriminate the correct variant of the language used in a text. In this paper, we present a fully statistics-based LID approach which is shown to be correct for common texts and maintains its robustness when classifying hard LID documents. For that, character sequences were used as base features. The Discriminant Ability of each sequence, in each training situation, is measured and used to filter out less important character sequences. Document similarity measure, based on the covariance concept, was defined. In the training phase, document clusters are built in a reduced k uncorrelated dimensions space. In the classification phase the Quadratic Discriminant Score decides which cluster (language) must be assigned to the documents one needs to classify.

1. Introduction

Multi-language access to multi-language information require crawled web pages to be categorised by language, by topic, ... Machine Translation also requires

2000 *Mathematics Subject Classification:* C2P99
Key words: Statistical Applications

Language Identification. So, in this paper, we will address text categorisation by language in circumstances harder than those that are commonly reported. The problems addressed occurred in the framework of the European Specific Support Action ASTROLABIUM, contract MOBI-CT-2003-003344, related to Scientific Mobility in Europe.

The LID problem can be seen as an instance of a more general problem: classifying objects using attributes. For this purpose different kinds of attributes have been used: particular characters [11], particular words [13, 7], word classes [9], particular character n-grams [2, 1, 4], particularly shaped words from images [12].

The Small Word Technique (SWT), proposed by Ingle, [7], is representative of the group of approaches using particular words. SWT is based on the intuition that common (function) words such as determiners, conjunctions and prepositions are good clues for guessing the language. Function words are often short and highly frequent. SWT uses a corpus for training, and for each language, words appearing more than a frequency threshold and having five characters or less are retained. So, given a new document to classify, the document is tokenized. Tokens appearing in the short word list are assigned their probabilities and tokens not in the list are assigned a minimum probability. The probability that a document is written in a given language is taken as the product of the probabilities of each token to belong to that language. SWT technique reaches 100% precision when documents at stake are large enough (see [6]).

The Trigram Technique (TT), described by Cavnar and Beesley, [2, 1], determines the frequencies of three letters strings (trigrams) in training corpus, as character trigrams are not equally probable for every language. For example, words ending in *-ck* are more likely to be English words than French ones. But words ending in *-ez* are more likely to be French. For an implementation of TT (see [6]), during the training phase, trigram statistics were collected from the ECI-CDROM[1] corpus. For each language, from a group of ten languages, texts were aggregated till one million characters were counted. Each resulting text was tokenized using the space (blank) character as the unique token separator, and an underscore was added before and after each token in order to easily identify initial and terminal trigrams. For each language, all trigrams appearing more than 100 times were retained as attributes. Trigram probabilities were calculated for the languages at stake. The probability that a new document is written in a given language is taken as the product of the probabilities of each selected trigram in the document to belong to that language. Selected trigrams not occurring in that

[1] see http://www.cogsci.ed.ac.uk/elsnet/eci.html for information on how to obtain this data.

document are assigned a minimal probability. Most probable language is chosen as the language the document is written in. TT and SWT reach same precision level.

Dunning's approach, [4], which involves n-gram statistics and Markov models, is reported to perform very well — 99.9 % of accuracy in discriminating two languages: English and Spanish.

However, none of these approaches presents results for contexts where the document language is hard to identify. One of these situations occurs when there is a need for correctly identifying the variant of the language (e.g., Brazilian or European Portuguese) a document is written in. This is the main problem we will address in this paper. Another hard context occurs with touristic information web pages (containing words from two or more languages, being each language represented in considerable percentages). That is the case of the two types of texts addressed to English readers: 1) restaurant menus, containing lots of Portuguese dish names, and 2) small advertisements, having no small functional words or a rather limited number of them, as exemplified by: `Hotel Real Palácio - Lisbon Coast - Portugal` and `Estalagem da Cegonha - Stork Inn - Vilamoura - Algarve - Portugal`.

Both advertisements were extracted from www.portugalvirtual.pt. They are addressed to English speaking people but the number of words in English is lower than the number of Portuguese words. These advertisements, examples of hard context documents, would certainly be a good challenge for the approaches mentioned before. In fact, taking into account the SWT, the classification of first advertisement is undecidable since there are no highly frequent small words in the text. Second advertisement would be wrongly classified as Portuguese. In fact, `'da'`, the only frequent small word present is a Portuguese one. In what concerns character trigrams approach (TT), trigrams `alá`, `lác`, `áci`, having low frequency would point to Portuguese. But trigrams `on_` and `st_` where blank character (token separator) has been replaced by the underscore, would push the classification into English. As a consequence, correctness of classification would not achieve 100% and deterioration would certainly occur. Considering Dunning's approach, [4], it is hard to predict what would be the classification results after training with such hard context documents. About Newman's approach, [11], we predict that its technique based on the presence of particular characters may be not sufficient to deal with this type of documents, where the main language discriminators does not lay on special characters.

LID symbolic approaches such as the one proposed in [9], are based on grammatical classes which makes them language dependent approaches, as they need

to know the grammar or other morphosyntactic information of the languages to be identified. This is a limitation since the *learning process* may be not possible if a specific grammar or language information is unknown or poorly known.

The problem of classifying small touristic documents, as mentioned before, was addressed in [3]. For this particular class of documents we obtained then 95.95% precision. Those results were obtained using the same approach we will present in this paper, but assuming particular working hypotheses that will be explained later in section 4..

The only attempt we know that tried to identify variants of the same language, was addressed in [10], as a practical necessity for web crawlers, and showed how inadequate are current LID approaches for solving this problem. In fact, SWT or TT approaches would certainly present poor results, since function words or frequently occurring sequences of characters — used as discriminant tools by SWT and TT — are basically the same for variants of the same language.

Given the promising results we had obtained by using the approach proposed in [3] to solve the *touristic documents* hard context, we tried to apply it to discriminate variants of the same language. First results obtained, assuming the same kind of working hypotheses were disappointing and required deeper analysis. The consequences of that work are explained in this paper, where we present a fully statistical LID approach, improving our earlier approach [3]. This improvement enables the discrimination of language variants, and maintains approach's ability to solve other specific LID hard context problems that had already been solved prior to the improvement we will address in this paper. As we will show, we obtained 98.4 % accuracy in discriminating two variants of Portuguese: European and Brazilian Portuguese.

Our approach uses the Quadratic Discrimination Score to decide which cluster (language) must be assigned to the document we want to classify. Space properties of the clusters are based on a document similarity measure which is calculated using character n-grams, with n ranging from 2 to 8. In the training phase the probabilities of all n-grams of characters are determined. The selection of the character n-grams that will be used for classification purposes takes into account the Discriminant Ability of each character sequence, a measure we had also developed.

So, in section 2 we will explain the training phase of our supervised approach. Section 3 presents the classification phase. Results will be presented in section 4, showing that our method is as robust as any other method for non problematic contexts and demonstrating additionally that it handles well documents written in languages that are hard to identify. Finally, at section 5, we will conclude and

elaborate on future work.

2. The Training Phase

2.1. Selecting Discriminant Character N-grams

Different character sequences differ on their degree of representativeness of each language in a corpus. For example, 4-gram `'der#'`[2] , is typical of German and Dutch documents, but not of French or Portuguese documents. Probability of string `'der#'` in different documents written in the same language is stable. It is relatively high and not much different in German or Dutch documents, but it is very low or close to zero in French or Portuguese documents. Similarly, sequences as `'van#de'`, `'ly#'`, `'ng#'`, `'un'` and `'o#'` have strong discriminant ability for LID purposes too. But, most sequences in the corpus have weak or very weak discriminant power. That is the case of `'100%'`, `'.#BO'`, `'rg#'`, `'95'`, `'rnh'`, etc. In fact the strings in the last group are not typical of any language we know; they may or may not occur in any document no matter the language it is written in. Average probability of each of these sequences is low and tends to change very little from document to document and from language to language.

In what follows, we will build necessary knowledge to present our measure of discriminant ability that will characterise every character sequence. Let s be any n-gram of characters in the training corpus and let $avp(s, l)$ be the average probability of s in documents of language l, that is,

$$avp(s, l) = \frac{1}{\|\mathcal{D}_l\|} \sum_{d \in \mathcal{D}_l} p(s, d) \tag{1}$$

where $\|\mathcal{D}_l\|$ is the number of documents in language l, in the corpus; and probability of s in document d, $p(s, d)$, is determined using equation (2),

$$p(s, d) = \frac{f(s, d)}{\|d\|} \tag{2}$$

where $f(s, d)$ stands for the frequency of sequence s in document d and $\|d\|$ denotes the number of unigrams, that is, the number of characters in d, including blank characters. Let $avp(s, .)$ be the average of those average probabilities for all languages in the corpus:

$$avp(s, .) = \frac{1}{\|\mathcal{L}\|} \sum_{l \in \mathcal{L}} avp(s, l) \tag{3}$$

[2] In all character n-grams used in this paper, the symbol `'#'` denotes the blank (space) character.

being $\mathcal{L}$ the set of languages in the corpus and $\|\mathcal{L}\|$ the size of $\mathcal{L}$. Then, based on the variance concept, let $VA(s)$ be the variation of that average probability along all different languages in the corpus. In other words, $VA(s)$ measures the variation of the average probability of s in documents of the same language, considering all different languages in the corpus.

$$VA(s) = \frac{1}{\|\mathcal{L}\|-1} \sum_{l \in \mathcal{L}} (avp(s,l) - avp(s,.))^2 \quad . \tag{4}$$

So, $VA(.)$ must be high for sequences having good discriminant ability, as its average probability in documents of the same language varies very little but has high variation when we consider all languages at stake. Now let $vp(s,l)$ be the variation of the probability of sequence s in the documents of language l:

$$vp(s,l) = \frac{1}{\|\mathcal{D}_l\|-1} \sum_{d \in \mathcal{D}_l} (p(s,d) - p(s,.))^2 \tag{5}$$

$$\text{being} \qquad p(s,.) = \frac{1}{\|\mathcal{D}_l\|} \sum_{d \in \mathcal{D}_l} p(s,d) \quad . \tag{6}$$

If sequence s is typical of language l, then $vp(s,l)$ must be low, meaning that the probability of s is relatively constant in the documents of language l, that is s is *faithful* to l. If s is typical of a language or group of languages, then its $vp(.,.)$ value for the rest of the languages must be relatively low too, since the probability of s in the documents of those other languages are zero or close to zero; that may be measured by

$$AV(s) = \frac{1}{\|\mathcal{L}\|} \sum_{l \in \mathcal{L}} vp(s,l) \quad . \tag{7}$$

So, sequences with high discriminant ability usually have low $AV(.)$ values. Finally, we define the Discriminant Ability of a sequence s by

$$DA(s) = \frac{VA(s)}{AV(s)} \tag{8}$$

where $VA(s)$ and $AV(s)$ are given respectively by equations 4 and 7. Thus, by equation 8, highly discriminant sequences must have high $DA(.)$ values. Examples of $DA(.)$ values obtained during the training phase with a corpus having 235 typical documents, distributed by 19 European languages, with 11 to 14 documents

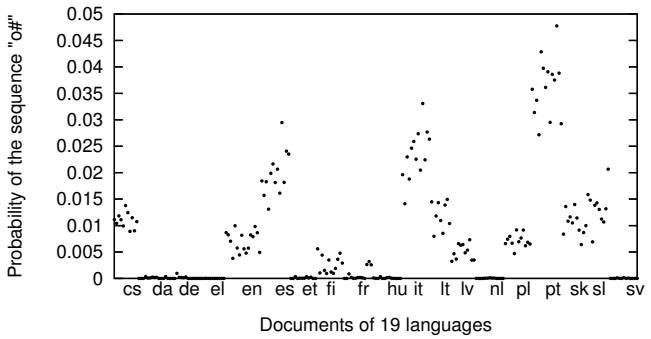

Figure 1: Probability of a strongly discriminant sequence in different documents of 19 different languages; each point represents a single document.

per language, are: $DA(\text{`}van\#d\text{'}) = 21.14$; $DA(\text{`}ung\text{'}) = 17.31$; $DA(\text{`}o\#\text{'}) = 13.7$; $DA(\text{`}d\#f\text{'}) = 0.990$ and $DA(\text{`}008\text{'}) = 0.0010$. For a better understanding of the Discriminant Ability, let us look at figures 1 and 2. In these figures, the tags representing the languages are listed between parenthesis: Czech (cs); Danish (da); German (de); Greek (el); English (en); Spanish (es); Estonian (et); Finnish (fi); French (fr); Hungarian (hu); Italian (it); Lithuanian (lt); Latvian (lv); Dutch (nl); Polish (pl); Portuguese (pt); Slovak (sk); Slovene (sl) and Swedish (sv). Figure 1 shows the probability of the bigram 'o#' in the documents of the training corpus we used. By this figure we can see that probabilities of sequence 'o#' are similar for all Czech (cs) documents, close to an average probability of 0.01. It also shows that, although the probability of 'o#' is very similar in Danish documents, its average probability is much lower than in the Czech case. We can see that there is considerable variation of average probabilities of bigram 'o#' when we take into account 19 target languages; this means a high value for $VA(\text{`}o\#\text{'})$. Besides, figure 1 presents relatively *compact clouds*, which means that the probabilities of the sequence are relatively constant in documents of the same language; so $AV(\text{`}o\#\text{'})$ is low. The combination of these factors (high $VA(\text{`}o\#\text{'})$ and low $AV(\text{`}o\#\text{'})$) implies a high Discriminant Ability ($DA(\text{`}o\#\text{'}) = 13.7$). In fact, bigram 'o#' is typical for some Latin and Slavic languages (Portuguese, Spanish and Italian, Czech, Polish, Slovak and Slavonian) but rare for others (French, Hungarian among others), thus making it a very discriminant sequence.

On the other hand, figure 2 shows a relatively low Discriminant Ability sequence: the trigram 'd#f'. In fact, the average probability of the sequence 'd#f' considering all different languages presents a very low variation, that is, a low

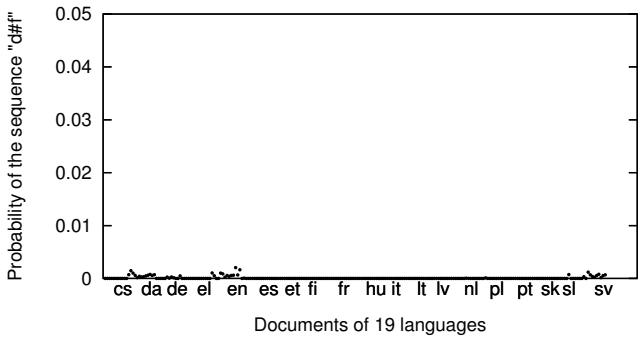

Figure 2: Probability of a weakly discriminant sequence in different documents of 19 different languages; each point represents a single document.

$VA(\text{`}d\#f\text{'})$. This is the main factor that implies a low Discriminant Ability value ($DA(\text{`}d\#f\text{'}) = 0.990$).

$DA(.)$ must not be confused as a particular case of the variance-to-mean ratio (VMR), which is given by the variance divided by the mean considering a probability distribution. In fact, $DA(.)$ is given by the variance of the *local* average probabilities, divided by the average of the *local* variances (see equation 8). Furthermore, contrary to VMR, by equations 4 and 7 we can see that the numerator and the denominator of equation 8 work with different distributions.

Discriminant Ability values are always positive. However, $DA(.)$ values of very weak discriminant sequences are close to zero. In fact, if a sequence s is unable for language discrimination, then its average probability tends to be constant for every language, that is, close to null variation for average probability ($VA(s) \approx 0$); this implies that its Discriminant Ability tends to be null ($DA(s) \approx 0$) since $AV(s)$, the denominator of equation 8, is always greater than zero, in practice. In fact there is only one highly unlikely case where $AV(s) = 0$: for that, the probability of the sequence s in documents of the same language would have to be constant, and this should happen for every language, which is highly unlikely.

Every sequence s having $DA(s) > 0$ has some discriminant ability. But, due to reasons of computational weight and complexity, we need to reduce the number of features to work with, and choose the smallest possible set of sequences, provided that high precision results are obtained. Led by available data, namely by the discriminant ability of '$d\#f$', we have chosen the value 1 as an empirical threshold, T, such that every sequence, obeying inequality $DA(.) > T$, for $T = 1$, would be chosen as a feature for next learning step. This empirical threshold

enabled strong feature reduction. Taking the mentioned training corpus from 19 languages we reduced from 4 410 812 distinct character n-grams, with n ranging from 2 to 8, to 36 877 Discriminant Sequences and obtained 100% precision, as it will be shown in section 4.. Unfortunately, even 36 877 was still a huge number of features to be applied directly in a classification process. So let us describe, in next subsection, the similarity measure we developed in order to build a document similarity matrix. Later, based on Principal Component Analysis, we will further reduce that number of features for classification purposes.

2.2. A Measure for Document Similarity

We determine the similarity between documents d_i and d_j by using the following correlation:

$$Sim(d_i, d_j) = \frac{Cov(d_i, d_j)}{\sqrt{Cov(d_i, d_i)}\sqrt{Cov(d_j, d_j)}} \tag{9}$$

where $Cov(d_i, d_j)$, based on the covariance concept, is given by

$$Cov(d_i, d_j) = \frac{1}{\|\mathcal{S}^*\| - 1} \sum_{s \in \mathcal{S}^*} d(s, d_i)\, d(s, d_j) \tag{10}$$

$$\text{where} \qquad d(s, d_i) = p^*(s, d_i) - p^*(., d_i) \tag{11}$$

and $\mathcal{S}^*$ is the set of Discriminant Sequences of characters determined as shown in last section. $\|\mathcal{S}^*\|$ is the cardinality of $\mathcal{S}^*$; $p^*(s, d_i)$ is given by

$$p^*(s, d_i) = p(s, d_i)\,.\,DA(s) \tag{12}$$

which is the probability of the sequence s in the document d_i weighted by the Discriminant Ability of s defined above (equation 8); $p^*(., d_i)$ is the average of $p^*(., .)$ values for document d_i considering the entire Discriminant Sequences set of the corpus, that is

$$p^*(., d_i) = \frac{1}{\|\mathcal{S}^*\|} \sum_{s \in \mathcal{S}^*} p^*(s, d_i) \quad . \tag{13}$$

2.3. Reducing the Number of Final Features

Above, we have seen how to filter out an important volume of features while retaining just those that might be needed for discrimination. In this section we

will transform this set of base features in a much smaller set of final features. For this purpose, we have chosen to characterise each document using the rest of the documents of the corpus, by building a document similarity matrix, $\vec{S}$, using equation 9 to calculate each matrix cell. Cells's values range from -1 to $+1$. When $Sim(d_i, d_j) = 0$ it means that the two documents are neither similar nor dissimilar. When this value is close to 1, d_i and d_j are very similar. When it is negative, d_i and d_j are dissimilar. By considering the training corpus mentioned above, we obtained a 235×235 matrix where the correlation of document d_i with all other documents is represented in line i of matrix $\vec{S}$. For the reader to have an idea of results obtained, Czech documents similarity among themselves range from 0.95 to 1. Similarity between Czech and Danish documents range from 0.15 to 0.18. Danish and German document similarity range from 0.70 to 0.73, while similarity between Danish and English range from 0.44 to 0.50. Similarity between Danish and Greek is close to 0 (about -0.0021), showing that, in practical terms, there is independence between documents of this pair of languages.

So, by selecting most discriminant character n-grams and by simply using similarity matrix $\vec{S}$ we reduced the initial number of attributes to 235 new attributes, since in this matrix every document is characterised by its similarity to each one of the 235 documents. However, 235 attributes are still too many and this number should be further reduced.

So, columns of similarity matrix $\vec{S}$ may be seen as attributes characterising each document represented by each line. Then, since there are strong similarities (correlations) between many pairs of documents, it is possible to generate a reduced new set of attributes based on this set of 235 ones.

Escoufier and L'Hermier [5] proposed an approach, based on Principal Components Analysis, to derive geometrical representations from similarity matrices. Since $\vec{S}$ is symmetric we have $\vec{S} = \vec{P}\vec{\Lambda}\vec{P}^T$, with $\vec{P}$ orthogonal ($\vec{P} = [\vec{e}_1, \ldots, \vec{e}_n]$ is the matrix of normalised eigenvectors of $\vec{S}$) and $\vec{\Lambda}$ diagonal; see [8] for details. The principal elements of $\vec{\Lambda}$ are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\vec{S}$ and $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n \geq 0$. Thus $\vec{S} = \vec{Q}\vec{Q}^T$ with

$$\vec{Q} = \vec{P}\vec{\Lambda}^{1/2} \ . \tag{14}$$

The elements of the ith line of $\vec{Q}$ will be the coordinates of the point associated with the ith document. In other words, each line of $\vec{Q}$ represents a document (see [5] for details) and columns of $\vec{Q}$ represents new uncorrelated attributes (components). We may consider only the coordinates corresponding to the leading eigenvalues. Equation 15 assesses how much of the total information is given by

the first k components (columns) of $\vec{Q}$ [3]

$$PTV(k) = \sum_{j=1}^{j=k} \lambda_j \,.\, (\sum_{j=1}^{j=n} \lambda_j)^{-1} \quad . \tag{15}$$

where n is the maximum number of components (235 for the 19 languages corpus). So, in order to obtain at least 95% of the total information given by the 235 columns, that is $PTV(k) >= 0.95$, we needed the first 18 columns of matrix $\vec{Q}$, that is $k = 18$, as $PTV(17) = 0.9444$ and $PTV(18) = 0.9519$; see [8] for details about Principal Components Analysis.

Using this approach, we were able to reduce 4 410 812 initial attributes (distinct character sequences) to 18 uncorrelated final attributes reflecting the initial ones. So, we have achieved to characterise documents in a space of 18 dimensions. And we have got 19 clusters — 235 documents distributed by 19 languages — in a space of 18 axes. This corresponds to the core part of the training phase.

3. The Classification Phase

In the classification phase we will show how this approach identifies the language that new documents are written in.

Since all documents of the training corpus are represented in a space of k dimensions, new documents to classify should be represented in the same k dimensions space. However, since a new document d has sequences of characters, we have to represent it in a k dimensions vector $\vec{v_d}$, corresponding to the same space built during the training phase.

3.1. Representing New Documents in the k Dimensions Training Phase Space

Let d be a new document we want to classify and $\vec{x_d} = [x_1, \ldots, x_m]$ a representation of d, where $x_i = p^*(i,d) = p(i,d) \,.\, DA(i)$ has the same meaning as in equation 12, $m = \|\mathcal{S}^*\|$ denotes the number of Discriminant Sequences of characters learnt during the training phase, and $i = 1, \ldots m$ denotes discriminant sequences of characters. Now let us standardise each element of this vector, in order to be coherent with the training phase documents, since the correlation using no standardised values in equation 9 corresponds to a covariance considering standardised values. So let $\vec{y_d} = [y_1, \ldots, y_m]$ where

$$y_i = \frac{x_i - x_.}{\sqrt{var(X)}} \qquad \text{with} \qquad x_. = \frac{1}{\|\mathcal{S}^*\|} \sum_{s \in \mathcal{S}^*} x_s \tag{16}$$

[3] PTV are initials for cumulative Proportion of the Total Variance.

$$\text{and} \qquad var(X) = \frac{1}{\|\mathcal{S}^*\| - 1} \sum_{s \in \mathcal{S}^*} (x_s - x_\cdot)^2 \ . \tag{17}$$

Considering that document d will be classified according to the training set made by the first corpus, $\vec{y_d}$ has $\|\mathcal{S}^*\|$ elements, that is 36 877 elements. Now let $\vec{s_d}$ be the similarity vector between document d, that is $\vec{y_d}$, and every document in the training corpus:

$$\vec{s_d}^T = \frac{1}{\|\mathcal{S}^*\| - 1} \vec{y_d}^T \vec{Z} \tag{18}$$

where $\vec{Z}$ is a matrix where each column $\vec{z_j}$ is such that $\vec{z_j}^T = [z_1, \ldots, z_m]$ represents the training set document j using standardised values based on the original attributes, just like it was calculated for the new document $\vec{y_d}$, as we have shown in this subsection; again $m = \|\mathcal{S}^*\|$. Notice that, although the calculation of matrix $\vec{Z}$ is heavy, it may be done in the training phase, just once. Now let $\vec{u_d} = [u_1, \ldots, u_n] = \vec{s_d}^T \vec{P} \vec{\Lambda}^{-1/2}$, where $\vec{P}$ and $\vec{\Lambda}$ are the matrices mentioned in equation 14 and n the number of documents (235 in the case of the mentioned corpus). Finally let $\vec{v_d} = [u_1, \ldots, u_k]$ where k is the number of components used in the training phase, that is 18 components for the same corpus. So, $\vec{v_d}$ represents the new document d in the k dimensions space built during the training phase.

3.2. The LID Classifier

So, language identification, is made by assigning the new document to a cluster (language), since every cluster is made by the training corpus documents of the same language represented in that k dimensions space. Our classification approach uses a criterion based on the Minimum Total Probability of Misclassification Rule for Normal Populations [8].

So, let $\vec{v}$ be the vector representing in those k dimensions a new document whose language we want to identify, and π_r the language represented by cluster r. We say that $\vec{v}$ belongs to language π_r if and only if, $d_r^Q(\vec{v}) = \max\limits_i d_i^Q(\vec{v})$, where $d_r^Q(\vec{v})$ is given by equation 19, meaning the Quadratic Discrimination Score or simply Quadratic Score of $\vec{v}$ considering language π_r; $i = 1, 2, \ldots, g$, being g the number of languages.

$$d_i^Q(\vec{v}) = -\frac{1}{2} \ln \left| \vec{\Sigma}_i \right| - \frac{1}{2} M(\vec{v}, \vec{\mu}_i, \vec{\Sigma}_i^{-1}) + \ln p_i \tag{19}$$

$$\text{where} \qquad M(\vec{v}, \vec{\mu}_i, \vec{\Sigma}_i^{-1}) = (\vec{v} - \vec{\mu}_i)^T \vec{\Sigma}_i^{-1} (\vec{v} - \vec{\mu}_i) \tag{20}$$

and $\vec{\Sigma}_i$ is the covariance matrix associated to the components that characterise the documents of cluster i, that is, language i. This covariance matrix is estimated by the covariance matrix $\vec{E}_i$ which is based on the sample made by the documents (the training documents) of cluster i. So,

$$\vec{E}_i = \begin{bmatrix} E_{1,1} & E_{1,2} & \dots & E_{1,k} \\ E_{1,2} & E_{2,2} & \dots & E_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ E_{1,k} & E_{2,k} & \dots & E_{k,k} \end{bmatrix} \tag{21}$$

where k is the number of components that characterise the documents of the cluster and the generic element in matrix $\vec{E}_i$ is given by

$$E_{l,m} = \frac{1}{n-1} \sum_{j=1}^{j=n} (c_{l,j} - c_{l,.})(c_{m,j} - c_{m,.}) \ , \tag{22}$$

being n the sample size, that is the number of cluster documents and $c_{l,.}$ the average value of the component l for those cluster documents, that is,

$$c_{l,.} = \frac{1}{n} \sum_{j=1}^{j=n} c_{l,j} \ , \tag{23}$$

and $c_{l,j}$ is the value of component l for document j. The means vector associated to language π_i, that is $\vec{\mu}_i$, is estimated by the means vector of cluster i, ($\vec{c}_i$), such that $\vec{c}_i^T = [c_{1,.}, \dots, c_{k,.}]$, being $c_{l,.}$ given by equation 23.

So, the factors that contribute to Quadratic Score value, $d_i^Q(\vec{v})$ are, apart from signs and multiplying constants: the logarithm of the covariance matrix determinant associated to cluster i; the logarithm of the *a priori* probability that a document is written in language i — this value was estimated by the logarithm of the quotient of the number of cluster i documents and the total number of documents in the corpus —; and finally by the most important factor, the Mahalanobis Distance between vector $\vec{v}$, representing the document to classify, and the means vector of cluster i, $\vec{c}_i$ (a estimation of $\vec{\mu}_i$), that is the cluster i centroid representing, say, an *average* document of language π_i, that is: $(\vec{v} - \vec{\mu}_i)^T \vec{\Sigma}_i^{-1} (\vec{v} - \vec{\mu}_i)$. The lower this factor is, the higher the $d_i^Q(\vec{v})$. This distance differs from the Euclidean distance, since, due to the inclusion of $\vec{\Sigma}_i^{-1}$, the scale of the values dispersion for each component along the cluster documents are taken into account to distance calculation.

# Doc's	Doc's Size	Precision
290	>= 6 lines	100%
290	4 lines	98.9%

Table 1: Precision Values for Language Identification in Typical Documents

4. Testing the Approach. Results

We trained our classifier basically using two corpora. First training corpus was made of normal documents with a clearly predominating language. Second training corpus was made of hard language identifying documents, since each document was written in one of the variants of the same language.

First corpus, as mentioned in subsection 2.1., had 235 typical documents distributed by 19 European languages. These documents were taken from the EurLex corpus on European legislation in force. Average document had 10 722 characters (134 lines). The smallest document had 177 characters (2 lines) and the largest one 92 341 characters (1 154 lines). In this corpus it is almost impossible to find a document written in just one language without foreign words. However, there is always a dominant language for every document. In order to assess the performance of our classifier in this easier task, we built 2 different test sets, each one with 290 documents from the same EurLex corpus, distributed by those 19 languages, having 14 to 16 documents per language. So, in the first test set we considered every document having at least 6 screen lines. The second test set was built from the same 290 documents, taking just 4 consecutive screen lines of text randomly extracted from each document. Table 1 shows that precision was 100% for the set where documents had at least 6 screen lines. However, for the case of documents of 4 screen lines, precision was a little lower due to 2 cases: 1 Czech document was wrongly identified as a Slovak one; 1 Portuguese document was wrongly identified as a Spanish one. With such small sized texts, this confusion of most similar language pairs is perfectly acceptable: Czech and Slovak, and Portuguese and Spanish. These results were obtained for the discriminating ability threshold $T = 1$. An example of the Quadratic Score values obtained for a typical Czech document are listed bellow. The highest value obtained was 54.4 for the Czech cluster, enabling the correct identification of the language for that document. Next highest values, -379.2 for Slovak and -381.9 for Slovene, are sufficiently far from the highest one, meaning that there is no doubt about what language the document is written in. Observe also that Slovak and Slovene are the languages most similar to Czech. Immediately after

T	Eur	Brz	Global	$\|\mathcal{S}^*\|$
0.025	94.44%	98.65%	97.83%	121 577
0.050	98.61%	98.31%	98.37%	53 345
0.075	97.22%	97.64%	97.56%	14 040
0.10	97.22%	98.65%	98.37%	8 752
0.25	90.27%	98.65%	97.02%	1 343
0.50	84.72%	98.65%	95.93%	251

Table 2: Precision Values for Portuguese Variant Identification in Documents

are Swedish (−604.4), Danish (−873.0), Estonian (−918.5), Letonian (−948.1), ... Polish, another Slavic language, is significantly distant from Czech, in the tenth position (−1241.2).

Second training set corpus was made of 100 European Portuguese and 100 Brazilian Portuguese documents. The shortest document had 17 lines; the longest one had 907 lines and the average length was 99 lines. These were formal writing documents, not popular writing ones. This makes the identification of the Portuguese variant, a difficult task even for a human reader who sometimes needs to read more than 8 lines to be sure whether the document is a European or a Brazilian Portuguese one. In this corpus, some *Brazilian* documents contain legislation from Brazilian Ministry of Education; some others contain news from the Newsletter from General Secretariat of Brazilian Republic Presidency. *Portuguese* documents were articles from two newspapers sites: www.publico.pt and www.dn.pt —articles are about politics, sports, economy, ecology, among other topics.

So, in order to train our LID classifier to identify both Portuguese variants, we tried to use the same threshold ($T=1$) used for the first training corpus. However, in this case this threshold limited the Discriminant Sequences set to 10 sequences only; and their $DA(.)$ values varied from 1.03 to 1.77. These low values confirm the very strong similarity between these variants. Obviously, 10 sequences were not enough for a *robust* training. Then we tried other lower thresholds. Table 2 shows precision results obtained for the classification of 369 new documents: 72 European Portuguese and 297 Brazilian Portuguese ones. Each line presents values for different threshold values (T). For $T=0.025$, a large set of Discriminant Sequences was obtained ($\|\mathcal{S}^*\|$ =121 577 sequences); precision for European and Brazilian variants of Portuguese was 94.44% and 98.65%, respectively; 97.83% was the global precision for this threshold, taking into account all documents classified. Considering all different thresholds in table 2, we select $T = 0.050$

as the one to be used in the training and classification phases, that is for this corpus and to classify European and Brazilian Portuguese new documents. The reason for this selection lies on the fact that for $T = 0.05$ the global precision is the highest one and the difference between precisions for European and Brazilian variants (98.61% and 98.31%) is the lowest, for all lines.

By comparison of the 1st and 2nd lines of table 2 we may conclude that, decreasing T in order to increase the size of the Discriminant Sequences set does not mean we will get better results, since precision for European Portuguese decrease from 98.61% to 94.44%. This may be due to the overall complexity associated to the calculation of matrices in the Pricipal Component Analysis, as the number of components (final features) tends to increase when the size of Discriminant Sequences set increases. In fact, 39 components were needed (see equation 15) in the case of $T = 0.05$ and 1 more (40) for $T = 0.025$. Table 2 also shows that, there is not a large variation for precision values when the size of the Discriminant Sequences set varies from 8 752 to 53 345 most discriminant sequences in the corpus. This shows a certain robustness of this approach.

Now, according to Discriminant Ability measure, $DA(.)$, the most discriminant sequences in this second corpus are: 'ct' ($DA('ct') = 1.77$); 'á#' ($DA('á\#') = 1.32$); ...; 'ect' ($DA('ect') = 0.907$); .... These sequences are parts of words showing where both variants are different. In fact, in European Portuguese we write 'projecto' for 'project' instead of 'projeto' in Brazilian Portuguese, for example. Similar phenomena occurs in the UK and USA English variants, with words terminating in 'ize' or 'ise'.

Function words corresponding to sequences such as '#o#' (corresponding in both variants to determiner 'the') or '#por#' (corresponding to preposition 'by'), present very low Discriminant Ability values: ($DA('\#o\#') = 0.016$) — the most discriminant sequence, that is 'ct', is 110 times more discriminant then this function word — and ($DA('\#por\#') = 0.046$). This means, in practical terms, that for distinguishing variants of the same language, discriminant sequences do not correspond to function words. They usually are non-frequent sequences corresponding to sub-words. Then, we may predict that it is unlikely that approaches such as SWT or TT would obtain good results on discriminating variants of the same language, since basically these approaches use the function words of each language (the most frequent ones) as their discriminant tools.

5. Conclusions and Further work

5.1. Conclusions

Language identification (LID) is still a problem for multi-language access to multi-language information on the web when it is necessary to separate web pages content according to the language or variant of the language in which they are written. Same need occurs in applications of Machine Translation.

Existing LID approaches are quite accurate for normal situations. However, usually, their assessment do not include discrimination of variants of the same language, or other hard context documents such as those written in more than one language in considerable percentages. In this paper a fully statistical LID approach is proposed as a contribution to solve LID hard problems.

This approach learns the most discriminant information according to each context. N-grams of characters, that is, sequences of N characters, with N raging from 2 to 8, were used as base-attributes. No character type was excluded, since every sequence of characters may be important. As each character n-gram has its own discriminant ability for each context, we developed the *Discriminant Ability* measure, $DA(.)$, which enabled the selection of a set of discriminant character n-grams (the Discriminant Sequences set) and together with other feature reduction techniques, namely the Principal Component Analysis, enabled a drastic reduction of classification features.

This approach includes two phases. The training phase builds a space of k dimensions, usually less than 40. The number of dimensions tends to grow with the number of languages. However, it may also be large if there is not enough discriminant information concentrated in a few sequences of characters, as it happens for discriminating variants of the same language. In the classification phase, new documents are represented in the same k dimensions space and classified using the Quadratic Discrimination Score.

Two training set corpus were used. The first one was made of common documents distributed by 19 European languages, each one written in a dominant language. Using the acquired knowledge from this training set, over 290 new documents having six or more lines of text were classified and 100 % precision was obtained. Classification of smaller documents having 4 lines of text, led to an error rate of 1.1%, due to confusion between rather similar languages: Slovak and Czech, and Portuguese and Spanish.

The second training corpus had 200 European and Brazilian Portuguese documents. We tried to contribute to solve the problem of discriminating variants of the same language: a problem recently addressed in [10], as a practical necessity for web crawlers and a limitation for LID approaches. These authors obtained

poor precision in these cases when they used their *most frequent n-grams* based approach. By applying our approach to this second corpus we reached 98.37% precision when we classified 369 European and Brazilian documents. However, in order to get enough discriminant sequences in this second corpus, we had to significantly reduce the Discriminant Ability threshold to $T = 0.05$, instead of using $T = 1$ as it was successfully used for the first corpus.

We also noticed that for these variants of Portuguese, in practical terms, character sequences corresponding to functional words were not selected as discriminant sequences by our approach. This happened because these classes of words are the same for both variants of Portuguese and occur with similar probability in documents of both variants. Discriminant elements were found mainly on little spelling differences of some words in both variants, at the level of subword strings. So, discriminating variants would certainly become a very difficult task for usual LID approaches such as Small Word Technique (SWT) [7] and Trigram Technique (TT) [2, 1], as these techniques are based on retaining the most frequent words or character sequences of the languages, as discrimination tools.

Thus, we may conclude that discriminant elements depend on each specific context. The results we got show the flexibility of this approach and its ability to capture the effective discriminant elements in each specific context. So, we are confident that we contributed to solve the limitation referred by mentioned authors [10]. Soon we will apply this approach to discriminate variants in other languages, such as UK English and USA English or other cases.

This approach has time-consuming calculations. However, they need to be done just once, during the training phase. The classification task is fast.

5.2. Future work

We have been working on developing a criterion to reject the identification of strange documents considering what was learnt in the training phase. That is, documents whose Quadratic Score is usually very low for all clusters, as they are too distant from the centroid of any cluster. Then, considering that the Quadratic Score $d_i^Q(\vec{v})$ — used in this LID classifier; see equation 19 — is based on Normal distributions, and training space axis are uncorrelated as mentioned before, and considering that Mahalanobis distance, $(\vec{v}-\vec{\mu}_i)^T\vec{\Sigma}_i^{-1}(\vec{v}-\vec{\mu}_i)$, (used in the calculation of $d_i^Q(\vec{v})$) equalises the standard deviation of the different training space axis due to the inclusion of $\vec{\Sigma}_i^{-1}$ (see equation 20), then we may use a Chi-square test to accept or reject the following hypothesis:

$\mathcal{H}_0$: vector $\vec{v}$ belongs to cluster i whose means vector and associated covariance

matrix are μ_i and $\vec{\Sigma}_i$ respectively.

Then, this takes us to a test for a level of significance of α which is given by the following:

$\mathcal{H}_0$ will not be rejected if and only if $(\vec{v} - \vec{\mu}_i)^T \vec{\Sigma}_i^{-1} (\vec{v} - \vec{\mu}_i) <= \chi^2_{DF}(\alpha)$

where DF is the number of degrees of freedom, which is given, in this case, by the k dimensions of the training space.

So, by using a cumulative Chi-square distribution table and the Mahalanobis distance from the vector $\vec{v}$ (representing a document to classify) to the cluster i centroid, we may decide if the document is *too distant* or not, to be considered of the language i. In future work, this decision rule will be included in the LID classifier. Thus, a cluster (language) will be elected as the cluster of the document we want to classify, not only if it presents the highest Quadratic score for that document considering all clusters, but also if its centroid is not too distant from the document.

We have not finished yet all experiments we planned for this rejection criterion, but first results are rather promising. Final results will be presented soon somewhere else.

REFERENCES

[1] K. R. Beesley Language identifier: a computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, (1998), 47–54.

[2] W. B. Cavnar and J. M. Trenkle N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, (1994), 161–175.

[3] J. F. da Silva and J. G. P. Lopes Identification of document language in hard contexts. In *Proceedings of the SIGIR 2006 Workshop on New Directions in Multilingual Information Access*, Seattle, USA, 2006.

[4] T. Dunning Statistical identification of language, 1994.

[5] Y. Escoufier and H.L'Hermier A propos de la comparaison graphique des matrices de variance. *Biometrischc Zeitschrift*, **20(5)** (1978), 477–483.

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

[7] N. C. Ingle A language identification table. *The Incorporated Linguist*, **15(4)** (1976).

[8] R. A. Johnson and D. W. Wichern *Applied Multivariate Statistical Analysis*. Prentice-Hall Intern., 2nd edition, 1988.

[9] R. Lins and P. Gonçalves Automatic language identification of written text. In *Proceedings of the ACM Symposium on Applied Computing, SAC-2004*, 1128–1133.

[10] B. Martins and M. J. Silva Language identification in web pages. In L. M. Liebrock, editor, *The 20th ACM SAC Symposium on Applied Computing. Document Engeneering Track.*, Santa Fé, Novo México, EUA, 2005, 773–777.

[11] P. Newman Foreign language identification – a first step in translation. In *Proc. of the 28th Annual Conference of the American Translators Association*, 1987, 509–516.

[12] P. Sibun and A. L. Spitz Language determination: Natural language processing from scanned document images. In *Proc. of the 4th ACL Conference on Applied Natural Language Processing (13–15 October 1994, Stuttgart)*, 1994.

[13] C. Souten, G. Churcher, J. Hayes, J. Hughes, and S. Johnson Natural language identification using corpus based methods. *Hermes Journal of Linguistics*, **13** (1994), 183–203.

Joaquim Ferreira da Silva, Gabriel Pereira Lopes
CITI / DI / FCT / Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
e-mail: {jfs, gpl}@di.fct.unl.pt