PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **16** (2004), 207-228

PLISKA
STUDIA MATHEMATICA
BULGARICA

A STATISTICAL APPROACH FOR MULTILINGUAL DOCUMENT CLUSTERING AND TOPIC EXTRACTION FROM CLUSTERS

Joaquim Silva, João Mexia, Carlos A. Coelho, Gabriel Lopes

This paper describes a statistics-based methodology for document unsupervised clustering and cluster topics extraction.

For this purpose, multiword lexical units (MWUs) of any length are automatically extracted from corpora using the LiPXtractor extractor — a language independent statistics-based tool.

The MWUs are taken as base-features to describe documents. These features are transformed and a document similarity matrix is constructed. From this matrix, a reduced set of features is selected using an approach based on Principal Component Analysis. Then, using the Model Based Clustering Analysis software, it is possible to obtain the best number of clusters. Precision and Recall for document-cluster assignment range above 90%.

Most important MWUs are extracted from each cluster and taken as document cluster topics.

Results on new document classification will just be mentioned.

1. Introduction

We aimed at developing a computational approach for automatically separating documents from multilingual corpora into clusters. We required no prior knowledge about document subject matters or language. So, we cleaned every keyword

2000 *Mathematics Subject Classification:* 62H30

Key words: Cluster analysis, Applied statistics, Document Clustering, Text Mining, Topics Extraction.

that might influence the behaviour of our methodology and might bias the unsupervised clustering method proposed. Since we want a language independent system, no morpho-syntactic information is used. Just statistical methods are applied.

Available software for clustering usually needs a matrix of objects characterized by a set of features. In order to obtain those features, we used the LiPXtractor to automatically extract MWUs from corpora. This extractor is composed by three tools: the LocalMaxs algorithm, the Fair Dispersion Point Normalization and the Symmetric Conditional Probability (SCP) cohesion measure ([2] and [1]). The MWUs extracted from corpora, such as *Journal officiel des Communautés européenes*, *Common Customs Tariff*, *hazardous waste*, *segurança social*, etc., provide important information about the text content. However, as there are thousands of MWUs extracted from a few hundred of documents, those MWUs can not be used as direct features for document clustering; they are used as *base-features* to obtain a reduced set of new features to cluster documents.

The number of clusters is usually unknown by the user. Therefore, we use the Model-Based Clustering Analysis software to automatically obtain the most likely correct number of clusters according to the input data.

The most informative and discriminating MWUs correspond to cluster topics, as we will show.

This paper is organized as follows: features extraction is explained in section 2; data transformations to approximate to Normal distribution, clustering and summarization are presented in sections 3 and 4; section 5 presents and discusses results obtained; related work appears in section 6 and conclusions are drawn in section 7.

2. Extracting Multiword Features from the Corpus

Three tools working together, are used for extracting MWUs from any corpus: the LocalMaxs algorithm, the Symmetric Conditional Probability (SCP) statistical measure and the Fair Dispersion Point Normalization (FDPN). A full explanation of these tools is given in [1]. However, a brief description is presented here for paper self-containment.

Thus, let us consider that an n-gram is a string of words in any text. For example the word *president* is an 1-gram; the string *President of the Republic* is a 4-gram. LocalMaxs is based on the idea that each n-gram has a kind of "glue" or cohesion sticking the words together within the n-gram. Different n-grams usually have different cohesion values. One can intuitively accept that there is

We use the notation $(w_1 \dots w_n)$ or $w_1 \dots w_n$ to refer an n-gram of length n.

a strong cohesion within the n-gram (*Giscard d'Estaing*) i.e. between the words *Giscard* and *d'Estaing*. However, one cannot say that there is a strong cohesion within the n-gram (*or uninterrupted*) or within the (*of two*). So, the $SCP(.)$ cohesion value of a generic bigram $(x\ y)$ is obtained by

$$SCP(x\ y) = p(x|y)\,.\,p(y|x) = \frac{p(x\ y)}{p(y)}\,.\,\frac{p(x\ y)}{p(x)} = \frac{p(x\ y)^2}{p(x)\,.\,p(y)} \tag{1}$$

where $p(x\ y)$, $p(x)$ and $p(y)$ are the probabilities of occurrence of bigram $(x\ y)$ and unigrams x and y in the corpus; $p(x|y)$ stands for the conditional probability of occurrence of x in the first (left) position of a bigram in the text, given that y appears in the second (right) position of the same bigram. Similarly $p(y|x)$ stands for the probability of occurrence of y in the second (right) position of a bigram, given that x appears in the first (left) position of the same bigram.

However, in order to measure the cohesion value of each n-gram of any size in the corpus, the FDPN concept is applied to the $SCP(.)$ measure and a new cohesion measure, $SCP\_f(.)$, is obtained.

$$SCP\_f(w_1 \ldots w_n) = \frac{p(w_1 \ldots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \ldots w_i).p(w_{i+1} \ldots w_n)} \tag{2}$$

where $p(w_1 \ldots w_n)$ is the probability of the n-gram $w_1 \ldots w_n$ in the corpus. So, any n-gram of any length is "transformed" in a pseudo-bigram that reflects the *average cohesion* between each two adjacent contiguous sub-n-gram of the original n-gram. Now it is possible to compare cohesions from n-grams of different sizes.

2.1. Improved LocalMaxs Algorithm

After having proposed the LocalMaxs algorithm ([2] and [1]), we present here an improved version of that algorithm:

Definition 1. *Let $W = w_1 \ldots w_n$ be an n-gram and $g(.)$ a cohesion generic function. And let: $\Omega_{n-1}(W)$ be the set of $g(.)$ values for all contiguous $(n-1)$-grams contained in the n-gram W; $\Omega_{n+1}(W)$ be the set of $g(.)$ values for all contiguous $(n+1)$-grams which contain the n-gram W, and let $leng(W)$ be the length (number of words) of n-gram W. We say that*

W is a MWU if and only if,

Roughly we can say that known statistical cohesion / association measures such as $Dice(.)$, $MI(.)$, χ^2, etc. seems to be "tailored" to measure just 2-grams. However, by applying FDPN to those measures, it is possible to use them for measuring the cohesion values of n-grams for any value of n [1].

$$\begin{aligned} &for\ \forall x \in \Omega_{n-1}(W), \forall y \in \Omega_{n+1}(W) \\ &\quad (leng(W) = 2 \wedge g(W) > y) \quad \vee \\ &\quad (leng(W) > 2 \wedge g(W) > \frac{x+y}{2}) \ . \end{aligned}$$

Then, for n-grams with $n \geq 3$, LocalMaxs algorithm elects every n-gram whose cohesion value is greater than the average of two maxima: the greatest cohesion value found in the contiguous $(n-1)$-grams contained in the n-gram, and the greatest cohesion found in the contiguous $(n+1)$-grams containing the n-gram. This version of the algorithm elects all the MWUs extracted by the LocalMaxs proper algorithm where only the n-grams corresponding to local maxima cohesion values were extracted ([2] and [1]). However, in the LocalMaxs proper algorithm, if *Supreme Court* was elected as a MWU, then *European Supreme Court* would not (see [2] and [1]); now, most cases like this are solved by this improved version. LiPXtractor is the MWUs extractor that uses $SCP\_f(.)$ cohesion function as $g(.)$ in LocalMaxs algorithm.

2.2. The Number of Features

Since we want to cluster documents, we must build a matrix of documents characterized in accordance with the smallest possible set of variables and convey that matrix to clustering software. In order to test our approach, we used a multilingual 1 330 423 words corpus (Sub-Eur-Lex-II) with 339 documents. LiPXtractor algorithm extracted 121 305 MWUs from that corpus. Obviously, we cannot use such a high number of features for distinguishing such a small number of objects (339 documents). However, these MWUs (base-features) provide the basis for building a new and reduced set of features.

2.3. Reducing the Number of Features

Let us take the following extracted MWUs: *nomenclature of the Common Customs Tariff*, *Common Customs Tariff* and *uniform application of the nomenclature*. For document clustering purposes, there is some redundance in these MWUs, since, for example, whenever *nomenclature of the Common Customs Tariff* is in a document, *Common Customs Tariff* is also in the same document and it may happen that, *uniform application of the nomenclature* is also in that

LocalMaxs has been used in other applications with other statistical measures, as it is shown in [1]. However, for Information Retrieval (IR) purposes, very interesting results were obtained by using $SCP\_f(.)$, in comparison with other measures [2].

This is part of the European Legislation in Force corpus: htp://europa.eu.int/eur-lex.

document. Let us show how these redundancies can be used to reduce the number of features.

Thus, according to Principal Components Analysis (PCA), often the original m correlated random variables (features) $X_1, X_2, \ldots, X_m$ can be "replaced" by a subset $Y_1, Y_2, \ldots, Y_k$ of the m new *uncorrelated* variables (components) $Y_1, Y_2, \ldots, Y_m$, each one being a linear combination of the m original variables, i.e., those k *principal components* provide most of the information of the original m variables [5, pages 340-350]. The original data set, consisting of l measurements of m variables, is reduced to another data set consisting of l measurements of k principal components. Principal components depend solely on the covariance matrix $\vec{\Sigma}$ (or the correlation matrix $\vec{\rho}$) of the original random variables $X_1, X_2, \ldots, X_m$. Now we state $MW_1, MW_2, \ldots, MW_p$ as being the original p variables (MWUs) of the Sub-Eur-Lex-II corpus. Then, for a reduced set of new variables (principal components) we would have to estimate the associated covariance matrix of the variables $MW_1, \ldots, MW_p$. So, let the sample covariance matrix $\vec{MW}$ be the estimator of $\vec{\Sigma}$.

$$\boldsymbol{MW} = \begin{bmatrix} MW_{1,1} & MW_{1,2} & \ldots & MW_{1,p} \\ MW_{1,2} & MW_{2,2} & \ldots & MW_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ MW_{1,p} & MW_{2,p} & \ldots & MW_{p,p} \end{bmatrix} \tag{3}$$

where $MW_{i,k}$ estimates the covariance $Cov(MW_i, MW_k)$. $\vec{MW}$ can be seen as a similarity matrix between MWUs. Unfortunately, due to this matrix huge size ($121\,305 \times 121\,305$), we cannot obtain principal components using available software. Moreover it is unlikely that PCA could achieve the reduction we need: from 121 305 original features to $k < 339$ (the number of documents) new features (principal components).

2.4. Geometrical Representations of Document Similarity

We can associate to the jth document, the vector $\vec{d}_j^T = [x_{1,j}, \ldots, x_{p,j}]$ where $x_{i,j}$ is the original numbers of occurrences of the ith MWU in the jth document.

From now on we will use p for the number of MWUs of the corpus and n for the number of documents.

These numbers of occurrences can be transformed, as we will see in Sect. 2.5..

Then, we can have a smaller (339×339) covariance matrix

$$\vec{\mathrm{S}} = \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,n} \\ S_{1,2} & S_{2,2} & \dots & S_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1,n} & S_{2,n} & \dots & S_{n,n} \end{bmatrix} \tag{4}$$

where the generic element in matrix $\vec{S}$ is given by

$$S_{j,l} = \frac{1}{p-1} \sum_{i=1}^{i=p} (x_{i,j} - x_{\cdot,j})(x_{i,l} - x_{\cdot,l}) \tag{5}$$

where $x_{\cdot,j}$, meaning the average number of occurrences per MWU in the jth document, is given by

$$x_{\cdot,j} = \frac{1}{p} \sum_{i=1}^{i=p} x_{i,j} \ . \tag{6}$$

Then $\vec{S}$ will be a matrix of similarities between documents.

Escoufier and L'Hermier [3] proposed an approach, based on PCA, to derive geometrical representations from similarity matrices. Since $\vec{S}$ is symmetric we have $\vec{S} = P\vec{\Lambda}P^T$, with $\vec{P}$ orthogonal ($\vec{P} = [\vec{e}_1, \dots, \vec{e}_n]$, the matrix of normalized eigenvectors of $\vec{S}$) and $\vec{\Lambda}$ diagonal. The principal elements of $\vec{\Lambda}$ are the eigenvalues $\lambda_1, \dots, \lambda_n$ of $\vec{S}$ and $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n \geq 0$. Thus $\vec{S} = \vec{Q}\vec{Q}^T$ with

$$\vec{Q} = \vec{P}\vec{\Lambda}^{1/2} \ . \tag{7}$$

The elements of the ith line of $\vec{Q}$ will be the coordinates of the point associated with the ith document. We may consider only the coordinates corresponding to the leading eigenvalues. Then, to assess how much of the total information is carried out by the first k components, i.e. the first k columns of $\vec{Q}$, we may use

$$PTV(k) = \frac{\sum_{j=1}^{j=k} \lambda_j}{\sum_{j=1}^{j=n} \lambda_j} \ . \tag{8}$$

So, by taking the first k columns of matrix $\vec{Q}$ such that $PTV(k)$ equals, say 0.85 or more, we can reduce the initial large number of features to $k \leq n$ new features (components). However, considering the 339 documents of the corpus, if

From now on when we replace an index by a dot, a mean value has been obtained.
PTV are initials for cumulative Proportion of the Total Variance.

we use the original number of occurrences of the ith MWU in the jth document ($x_{i,j}$) to obtain "similarities" (see equation (5)), we need the first 129 components to provide 0.85 of the total information, i.e. $PTV(129) = 0.85$. Although it corresponds to 0.11% of the initial 121 305, large numbers of components must be avoided in order to minimize computational effort of the clustering process. So, to reduce this number, we need to "stimulate" similarities between documents, and therefore, the original occurrences of the MWUs in documents must be transformed.

2.5. Transformed Occurrences

As referred above, the geometrical representation may be obtained from transformed occurrences. The technique we used has four phases. In the first phase we standardize in order to correct document heterogeneity. This heterogeneity is measured by the variation between the number of occurrences (frequency) of the different MWUs inside each document. This variation may be assessed by

$$V(D_j) = \frac{1}{p-1}\sum_{i=1}^{i=p}(x_{i,j} - x_{\cdot,j})^2 \qquad j = 1, \ldots, n \tag{9}$$

where $x_{\cdot,j}$ is given by (6). The standardized values will be

$$z_{i,j} = \frac{(x_{i,j} - x_{\cdot,j})}{\sqrt{V(D_j)}} \qquad i = 1, \ldots, p;\ \ j = 1, \ldots, n\ . \tag{10}$$

In the second phase we evaluate the variation between documents for each MWU. Although, each MWU associated variation must reflect how much the MWU occurrences vary in different documents, due to document content, not due to document size. Therefore we use normalized values to calculate this variation:

$$V(MW_i) = \frac{1}{n-1}\sum_{j=1}^{j=n}(z_{i,j} - z_{i,\cdot})^2 \qquad i = 1, \ldots, p\ . \tag{11}$$

These values are important since we found that, generally, the higher $V(MW_i)$, the more information is carried out by the ith MWU. On the other hand, it was observed that MWUs constituted by long words, usually are more informative from an Information Retrieval / Text Mining point of view (e.g. *agricultural products* or *communauté economique européenne* are more informative than *same way*, *plus au moins* or *reach the level*). Thus, in a third phase we define *weighted occurrences* as

$$x^*_{i,j} = x_{i,j} \cdot V(MW_i) \cdot AL(MW_i) \qquad i = 1, \ldots, p;\ \ j = 1, \ldots, n \tag{12}$$

where $AL(MW_i)$ is the average length of the words in the ith MWU. This is measured by the average number of characters per word.

Lastly, in the fourth phase we carry out a second standardization considering the *weighted occurrences*. This is for correcting document size heterogeneity, since we do not want that the document size affects its relative importance. Thus

$$z^*_{i,j} = \frac{(x^*_{i,j} - x^*_{\cdot,j})}{\sqrt{V(D^*_j)}} \qquad i = 1, \ldots, p\,; \;\; j = 1, \ldots, n \tag{13}$$

$$\text{where} \qquad V(D^*_j) = \frac{1}{p-1} \sum_{i=1}^{i=p} (x^*_{i,j} - x^*_{\cdot,j})^2 \qquad j = 1, \ldots, n \;\; . \tag{14}$$

These standardizations are *transformed occurrences* and are used to obtain the similarity matrix between documents, whose generic element is given by

$$S_{j,l} = \frac{1}{p-1} \sum_{i=1}^{i=p} (z^*_{i,j} - z^*_{\cdot,j})(z^*_{i,l} - z^*_{\cdot,l}) \qquad j = 1, \ldots, n\,; \;\; l = 1, \ldots, n \;\; , \tag{15}$$

or simply

$$S_{j,l} = \frac{1}{p-1} \sum_{i=1}^{i=p} z^*_{i,j} \, . \, z^*_{i,l} \qquad j = 1, \ldots, n\,; \;\; l = 1, \ldots, n \;\; , \tag{16}$$

As $z^*_{\cdot,j} = z^*_{\cdot,l} = 0$, because both are means of standardized values.

2.6. Non-informative MWUs

Some high-frequency MWUs appearing in most documents written in the same language are not informative from a Text Mining point of view, e.g., locutions such as *Considérant que* (*having regard*), *and in particular*, or other expressions which are incorrect MWUs, such as *of the* or *dans les* (*in the*). Although these expressions are useless to identify document topics, they are informative for distinguishing different languages. As a matter of fact they occur in most documents of the same language, and their associated variation (see equation (11)) is usually high or very high, i.e., they are relevant to "approximate" documents of the same language for calculating similarities between documents (see equations (12), (13) and (16)).

So, it seems that either they should be removed to distinguish topics in documents written in the same language, or they should be kept for distinguishing documents of different languages. To solve this problem, we use the following

criterion: the MWUs having at least one extremity (the leftmost or the rightmost word) that exists in at least 90% of the documents we are working with, are removed from the initial set of MWUs. We follow that criterion since these expressions usually begin or end with words occurring in most documents of the same language, e.g., *of*, *les*, *que*, etc.. As we will see in Subsect. 3.3., the documents and MWUs with which the system is working, depends on the node of the clustering tree.

To summarize, in this section we obtained a matrix where a small set of components classifies a group of documents. This matrix will be used as input for clustering. For this purpose, the matrix of document similarity ($\vec{S}$) (see matrix (4)) was calculated. Its generic element is given by equation (16). Then, from $\vec{S}$, $\vec{Q}$ was obtained by (7) and a new matrix ($\vec{C}$) corresponds to the first k columns of $\vec{Q}$, such that $PTV(k) \geq 0.85$. Finally, $\vec{C}$ will be conveyed to clustering software.

Considering the initial 121 305 MWUs for the 339 documents of the Sub-Eur-Lex-II corpus, we obtained $PTV(3) = .79.5$; $PTV(4) = .874$ and $PTV(5) = .912$. Then according to the criterion previously explained we selected the first 4 components (columns).

3. Clustering Documents

We need to split documents into clusters. However we do not know how many clusters should be obtained. Moreover, though we have obtained k features (components) to evaluate the documents, we do not know neither the composition of each cluster, nor its volume, shape and orientation in the k-axes space.

3.1. The Model-Based Cluster Analysis

Considering the problem of determining the structure of clustered data, without prior knowledge of the number of clusters or any other information about their composition, Fraley and Raftery [4] developed the Model-Based Clustering Analysis (MBCA). By this approach, data are represented by a mixture model where each element corresponds to a different cluster. Models with varying geometric properties are obtained through different Gaussian parameterizations and cross-cluster constraints. Partitions (clusters) are determined by the EM (expectation-maximization) algorithm for maximum likelihood, with initial agglomerative hierarchical clustering (see [4] for details). This clustering methodology is based on multivariate normal (Gaussian) mixtures. So the density function associated

to cluster c has the form

$$f_c(\vec{x}_i|\vec{\mu}_c,\vec{\Sigma}_c) = \frac{exp\{-\frac{1}{2}(\vec{x}_i-\vec{\mu}_c)^T\vec{\Sigma}_c^{-1}(\vec{x}_i-\vec{\mu}_c)\}}{(2\pi)^{\frac{p}{2}}\left|\vec{\Sigma}_c\right|^{\frac{1}{2}}} \quad . \tag{17}$$

Clusters are ellipsoidal, centered at the means $\vec{\mu}_c$; $\vec{x}_i$ is an element of cluster c. The covariance matrix $\vec{\Sigma}_c$ determines other geometric characteristics. This clustering methodology is based on the parameterization of the covariance matrix in terms of eigenvalue decomposition in the form $\vec{\Sigma}_c = \lambda_c\vec{D}_c\vec{A}_c\vec{D}_c^T$, where $\vec{D}_c$ is the orthogonal matrix of eigenvectors, determining the orientation of the principal components of $\vec{\Sigma}_c$. $\vec{A}_c$ is the diagonal matrix whose elements are proportional to the eigenvalues of $\vec{\Sigma}_c$, determining the shape of the ellipsoid. The volume of the ellipsoid is specified by the scalar λ_c. Characteristics (orientation, shape and volume) of distributions are estimated from the input data, and can be allowed to vary between clusters, or constrained to be the same for all clusters. Considering our application, input data is given by the first k columns of matrix $\vec{Q}$ (see equations (7) and (8)), that is matrix $\vec{C}$.

MBCA subsumes the approach with $\vec{\Sigma}_c = \lambda\vec{I}$, long known as `k-means`, where sum of squares criterion is used, based on the assumption that all clusters are spherical and have the same volume (see Table 1). However, in the case of `k-means`, the number of clusters has to be specified in advance — an information the user usually can not provide. Moreover, considering many applications, real clusters are far from spherical in shape. Therefore we have chosen MBCA for clustering documents. Then, function `emclust` has been used with `S-PLUS` package, which is available for `Windows` and `Linux`.

During the cluster analysis, `emclust` shows the Bayesian Information Criterion (BIC), a measure of evidence of clustering, for each "pair" *model-number of clusters*. These "pairs" are compared using BIC: the larger the BIC, the stronger the evidence of clustering (see [4]). The problem of determining the number of clusters is solved by choosing the *best model*. Table 1 shows the different models used during the calculation of the *best model*. Models must be specified as a parameter of the function `emclust`. However, usually there is no prior knowledge about the model to choose. Then, by specifying all models, `emclust` gives us BIC values for each pair *model-number of clusters* and proposes the *best model* which indicates which cluster must be assigned to each object (document).

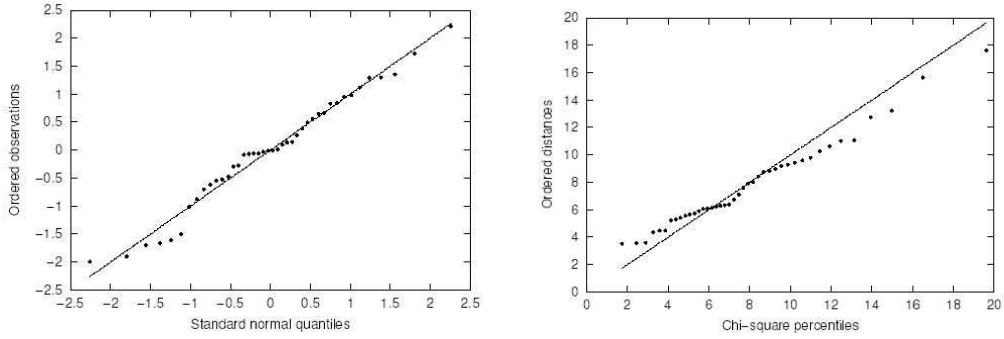
3.2. Assessing Normality of Data. Data Transformations

MBCA works based on the assumption of normality of data. Then, Gaussian distribution must be checked for the univariate marginal distributions of the

Table 1: Parameterizations of the covariance matrix $\vec{\Sigma}_c$ in the Gaussian model and their geometric interpretation

Ref.	$\vec{\Sigma}_c$	Distribution	Volume	Shape	Orientation
EI	$\lambda\vec{I}$	Spherical	Equal	Equal	
VI	$\lambda_c\vec{I}$	Spherical	Variable	Equal	
EEE	$\lambda\vec{D}\vec{A}\vec{D}^T$	Ellipsoidal	Equal	Equal	Equal
VVV	$\lambda_c\vec{D}_c\vec{A}_c\vec{D}_c^T$	Ellipsoidal	Variable	Variable	Variable
EEV	$\lambda\vec{D}_c\vec{A}\vec{D}_c^T$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_c\vec{D}_c\vec{A}\vec{D}_c^T$	Ellipsoidal	Variable	Equal	Variable

documents on each component. For this purpose, each columns of the matrix $\vec{C}$ is standardized, ordered and put on y axis of the QQ-plot. Then, standardized normal quantiles are generated and put on x axis of the QQ-plot (see [5, pages 146-162] for details). Fig. 1(a) represents the QQ-plot for the 2th component, assessing the normality of data of cluster 1.1 . This QQ-plot is representative, since most of the components for other clusters produced similar QQ-plots. Most



(a) (b)

Figure 1: QQ-plot of data for the 2nd component of cluster 1.1 (a); Chi-square plot of the ordered distances for data in cluster 1.1 (b)

In Sect. 5. we will deal with specific clusters.

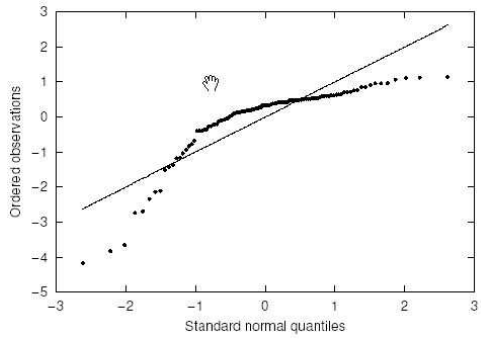


Figure 2: QQ-plot of data for the 2nd component of cluster 2.

of the times, if QQ-plots associated to the components are straight (univariate distributions are normal), the joint distribution of the k dimensions (components) are multivariate normal. However, multivariate normality must be tested. Then, a Chi-square plot is constructed for each cluster. Thus, the square distances are ordered from smallest to largest as $d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(m)}$, where $d^2_{(j)} = (\vec{x}_j - \overline{\vec{x}}_c)^T \vec{S}_c^{-1} (\vec{x}_j - \overline{\vec{x}}_c)$. Vector $\vec{x}_j$ is the jth element (document) of cluster c; $\overline{\vec{x}}_c$ is the means vector for the k dimensions of that cluster, and $\vec{S}_c^{-1}$ is the inverse of the estimator of the cluster covariance matrix. Then the pairs $(d^2_{(j)}, \chi^2_k((j-1/2)/m))$ are graphed, where m is the number of elements of the cluster and $\chi^2_k((j-1/2)/m)$ is the $100(j-1/2)/m$ percentile of the Chi-square distribution with k (the number of components, (see equation 8)) degrees of freedom. The plot should resemble a straight line. A systematic curved pattern suggests lack of normality. The plots of the figures 1(a) and 1(b) does not show systematic curved patterns. However, since lack of normality were suggested for the distributions associated to the components of the clusters 2 (see figure 2), some transformation were made to approximate to normality.

So, let y be an arbitrary element of a given column of matrix $\vec{C}$ we want to transform. We considered the following slightly modified family of power transformations from y to $y^{(\lambda)}$ [6]:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} . \tag{18}$$

Power transformations are defined only for positive variables. However this is not restrictive, because a single constant can be added to each element in the column

if some of the elements are negative. Thus, given the elements $y_1, y_2, \ldots, y_m$ in a given column, the Box-Cox [6] solution for the choice of an appropriate power λ for that column is the one which maximizes the expression

$$l(\lambda) = -\frac{m}{2} \ln \left[\frac{1}{m} \sum_{j=1}^{j=m} (y_j^{(\lambda)} - \overline{y^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{j=m} \ln y_j \tag{19}$$

$$\text{where} \qquad \overline{y^{(\lambda)}} = \frac{1}{m} \sum_{j=1}^{j=m} y_j^{(\lambda)} \tag{20}$$

and m is the number of elements of the column; $y_j^{(\lambda)}$ is defined in (18). So, every element y of the ith column of matrix $\vec{C}$ associated to cluster 2 was transformed from y to $y^{(\lambda)}$ according to (18) where $\lambda = \hat{\lambda}_i$, i.e., the value of λ maximizing $l(\lambda)$. This new matrix was conveyed to clustering software.

3.3. Sub-clusters

As we will see in Sect. 5., our approach organized Sub-Eur-Lex-II corpus in 3 main clusters: Portuguese, French and English documents. However, we can distinguish different subjects in different documents of the same cluster. So, a hierarchical tree of clusters is built as follows: let us consider that every cluster in the tree is a node. For every node, non-informative MWUs are removed (see Subsect. 2.6.) from the set of MWUs contained in the documents of that node (a subset of the original MWUs), in order to obtain a new similarity matrix between documents (see equation (16)). Then, the first k columns of the new matrix $\vec{Q}$ are taken (see equations (7) and (8)) to form the matrix $\vec{C}$ of this cluster. Then, from this matrix ($\vec{C}$), new clusters are proposed by MBCA.

3.4. Choosing the Best Number of Clusters

As has been said, MBCA calculates the *best model* based on a matrix ($\vec{C}$) which corresponds to the first k columns (components) of matrix $\vec{Q}$. A large number of components means no significant information loss, which is important for a correct clustering (*best model*) to be proposed by MBCA. On the other hand, a large number of components must be avoided, since it takes MBCA to estimate large covariance matrices — during the internal computation for different models — which can be judged to be close to singularity (see [4]). Therefore, as we said before, the following simple criterion is used: the first k components are chosen in such a way that $PTV(k) \geq 0.85$ (see equation (8)).

Then, based on matrix $\vec{C}$ associated to a cluster — we may see the initial set of documents as cluster 0 — MBCA produces a list of BIC values for each model: *VVV (Variable volume, Variable shape, Variable Orientation), EEV, VEV* etc. (see Table 1). Each list may have several local maxima. The largest local maximum over all models is usually proposed as the *best model*. However, a heuristic that works well in practice (for further discussion, see [4]) — and has been followed by us — chooses the number of clusters corresponding to the first decisive local maximum over all the models considered.

4. Summarization

Summarizing a document and summarizing a cluster of documents are different tasks. As a matter of fact, documents of the same cluster have common relevant expressions such as *Common Customs Tariff* or *nomenclature of the Common Customs Tariff*, rather than long sentences which are likely to occur in just one or two documents. Then, summarizing topics seems adequate to disclose the core content of each cluster.

Cluster topics correspond to the most important MWUs in the cluster. Let the cluster from where we want to extract its topics be the "target cluster". Then, in order to extract it, first the MWUs of the *parent node* of the target cluster are ordered according to the value given by $Score(MW_i)$ assigned to the ith MW.

$$Score(MW_i) = V(MW_i) \cdot AL(MW_i) \cdot Thr(MW_i) \qquad \text{where} \tag{21}$$

$$Thr(MW_i) = \begin{cases} 1 & SCP\_f(MW_i) \geq threshold \\ 0 & \text{else} \end{cases} . \tag{22}$$

$V(MW_i)$ and $AL(MW_i)$ have the same meaning as in equation (12). Thus, $Thr(\cdot)$ corresponds to a filter that "eliminates" MWUs whose $SCP\_f(\cdot)$ cohesion value (see Sect. 2.) is lower than $threshold$ — a value empirically set to 0.015. These MWUs, e.g., *in case of*, *and in particular*, etc., are not informative for Information Retrieval or Text Mining purposes. Usually, these MWUs are previously eliminated when selecting the informative MWUs for calculating the covariance matrix; however it may not happen in case of a multilingual set of documents. (see Subsect. 2.6.).

So, the largest $Score(MW_i)$ corresponds to the most important MWU. For example, according with this criterion, the 15 most important MWUs of the initial cluster (the one containing all documents) are the following: *Considerando que (Having regard)*, *considérant que*, *Member States*, *accordance with*, *États membres*

Table 2: Evaluation of the Precision and Recall for each cluster without approximating data to Normal distribution

Cluster	One topic	Real #	Prop. #	Correct #	Prec. (%)	Rec. (%)
1	*Comunidades Europeias*	113	113	113	100	100
2	*Communautés Européennes*	113	113	113	100	100
3	*European Communities*	113	113	113	100	100
1.1	*segurança social*	42	42	40	95.2	95.2
1.2	*resíduos perigosos*	21	20	19	95.0	90.5
1.3	*pauta aduaneira comum*	50	51	49	96.1	98.0
2.1	*sécurité sociale*	42	41	37	90.2	88.1
2.2	*déchets dangereux*	21	21	17	80.9	80.9
2.3	*tarif douanier commun*	50	51	46	90.2	92.0
3.1	*social security*	42	44	41	93.2	97.6
3.2	*hazardous waste*	21	20	19	95.0	95.2
3.3	*Common Customs Tariff*	50	49	47	95.9	94.0

(Member states), *Council Regulation*, *État membre*, *Having regard*, *Communautés européennes (european communities)*, *COMMUNAUTÉS EUROPÉENNES*, *Comunidades Europeias*, *COMUNIDADES EUROPEIAS*, *EUROPEAN COMMUNITIES*, *Tendo em conta (Having regard)*, *European Communities*.

Now, taking for instance the "English documents" as the target cluster, we cannot "guarantee" that *hazardous waste* will be a topic, since not every English document content is about *hazardous waste*. On the other hand, the same topic often appears in different documents, written in different forms, (e.g. *hazardous waste* and *Hazardous waste*). Hence, according to $Score(\cdot)$, the 15 most important MWUs of the target cluster occurring in at least 50% of its documents are put in a list. From this list, the MWUs with $Score(\cdot)$ value not lower than 1/50 of the maximum $Score(\cdot)$ value obtained from that list, are considered topics.

5. Results

Sub-Eur-Lex-II is a multilingual *corpus* with 113 documents per language (Portuguese, French and English). For each document there are two other documents which are translations to the other languages. From table 2 and 3 we can see the hierarchical tree of clusters obtained by this approach.

Table 3: Evaluation of the Precision and Recall for each cluster after approximating the data of cluster 2 to Normal distribution

Cluster	One topic	Real #	Prop. #	Correct #	Prec. (%)	Rec. (%)
1	*Comunidades Europeias*	113	113	113	100	100
2	*Communauts Européennes*	113	113	113	100	100
3	*European Communities*	113	113	113	100	100
1.1	*segurança social*	42	42	40	95.2	95.2
1.2	*resduos perigosos*	21	20	19	95.0	90.5
1.3	*pauta aduaneira comum*	50	51	49	96.1	98.0
2.1	*sécurité sociale*	42	42	41	97.6	97.6
2.2	*déchets dangereux*	21	21	20	95.2	95.2
2.3	*tarif douanier commun*	50	50	49	98.0	98.0
3.1	*social security*	42	44	41	93.2	97.6
3.2	*hazardous waste*	21	20	19	95.0	95.2
3.3	*Common Customs Tariff*	50	49	47	95.9	94.0

5.1. Discussion

The original texts of the Sub-Eur-Lex-II corpus are classified by main topic areas: *Segurança Social*, *Sécurité sociale*, *Social security*, *Gestão dos resíduos e tecnologias limpas*, *Gestion des déchets et technologies propres*, *Waste management and clean technology*, *Classificação Pautal*, *Classment tarifaire* e *Tariff classification*. However, we have removed that information from the documents before extracting the MWUs using LiPXtractor, as we wanted to test our approach for clustering usual documents.

In Tables 2 and 3, column *One topic* means a representative topic obtained for the cluster indicated by column *Cluster*; by *Real #* we mean the real number of documents in the corpus where the topic shown in *One topic* is a main topic; *Prop. #* is the number of documents proposed to belong to the cluster by our approach; *Correct #* is the number of documents correctly proposed; *Prec. (%)* and *Rec. (%)* are Precision and Recall. Precision is given by the value indicated in *Correct #* divided by the number indicated in *Prop. #*; Recall is given by the value indicated in *Correct #* divided by the number in *Real #*.

As we mentioned before, after obtaining matrix $\vec{C}$ with 4 columns (components) characterizing the documents of the initial set (cluster 0), it was taken as input to MBCA. Then clusters 1, 2 and 3 were proposed by this software (function `mclust`) considering VEV-3 (Variable volume, Equal shape, Variable

orientation) the *best model*, (see Table 1). The number of clusters (3) in this level is correct as it corresponds to the real number of languages in the corpus. By the topics presented on these clusters, we can see that their major content is about *European Communities* and the corresponding equivalents in Portuguese and French. On this level of the clustering tree, all the documents were proposed to belong to the correct cluster, that is 100% for Precision and Recall (see table 2).

In order to obtain sub-clusters, new matrices $\vec{C}$ were calculated for clusters 1, 2 and 3 (details on sections 2.4., 2.5. and 2.6.). Then, we needed 7 components for matrix $\vec{C}$ of cluster 1, 4 components for cluster 2 and 8 components for cluster 3. So, the number of components needed on this level (clusters 1, 2 and 3) tends to be greater than for cluster 0. The reason for this difference lies on the fact that there is a very high discriminating power associated to some MWUs of documents of cluster 0, such as *accordance with*, *Considérant que* and *Considerando que*, that is, the value $V(MW_i) \cdot AL(MW_i)$ is very high (equation 12). This factor is important to obtain *weighted occurrences* (equation 12) and the higher it is in general, the stronger is the reduction of the number of components. Although this kind of MWUs (*accordance with*, etc.) are not informative in terms of document topics, they are common to the documents written in the same language, so they are very useful to discriminate languages. However, most approaches consider these expressions as *stop terms* and ignore them since they are "meaningless". But they are not, as we can see.

On the other hand, the highest values for $V(.) \cdot AL(.)$ associated to MWUs of cluster 1, 2 and 3 are found in MWUs such as *social security*, *hazardous waste*, etc., but these values are not as high as those associated to MWUs of cluster 0. This is due to the lower variation of the occurrences of these MWUs (*social security*, etc.) considering all the documents of the cluster. This lower variation gives a lower value for $V(.)$ (equation 12), and then more components are needed to obtain $PTV(.) \geq 0.85$ (see equation 8).

So, from each matrix $\vec{C}$ of clusters 1, 2 and 3, MBCA software proposed 3 sub-clusters (model VEV-3), which is correct considering the human classifications of the documents in groups. We called it 1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 3.1, 3.2 and 3.3. The main topics extracted for each cluster by this approach confirms the content assigned by the human classification. However, Precision and Recall are a little lower for these sub-clusters than in the case of cluster 1, 2 and 3. Although it is higher than 90%, except for the French sub-clusters (see table 2). The 100% values for Precision and Recall on clusters 1, 2 and 3 are achieved due to the fact that besides MWUs such as *accordance with*, *Considérant que* and *Considerando*

que have very high discriminating power, they occour in almost every document of the same language, which is important to a correct assignment of documents to clusters. Although there are many correlated MWUs in the documents of each subcluster 1.1, 1.2 ...3.3, there was no MWUs being common to every document in each sub-cluster. Instead of that, it is possible to find MWUs with some discriminant power, that is, medium values for $V(.) \cdot AL(.)$. Although these MWUs are not many, they are responsible for lower values for Precison and Recall of the sub-clusters (see table 2). However, after transformations on documents data of cluster 2 to approximate to normal distributions, better results were obtained, as we can compare (Precision and Recall values of sub-clusters 2.1, 2.2 and 2.3 in table 2 and 3).

Since our approach is not oriented for any specific language, we believe that different occurrences for the same concept in the three different languages, are the main reason for different Precision and Recall scores comparing "corresponding" clusters and sub-clusters.

The topics extracted from each cluster and sub-cluster agree with the essential content of the documents on the cluster. In fact, either they agree perfectly, as in the case of *Social Security*, *Sécurité sociale*, etc., or they are strongly correlated, such as *waste management and clean technology* versus *hazardous waste*, etc.. However, we can not expect perfect translations from a sub-cluster to another, since this is not a translation approach and the human translations made on these documents are not perfect. In fact, we obtained the topics *Member States* and *États Membres* for French and English in clusters 3 and 2, but the Portuguese equivalent (*Estado-membro*) was not extracted, since it is a unigram, and unigrams are not extracted by LiPXtractor extractor.

Although human translations made on these documents are good, the number of documents containing an MWU may not be the same for the corresponding MWU in another language. For example, *Communauté économique européenne* is an MWU that occurs in 58 documents, that is more than 50% of 113 documents of cluster 2; so it is elected as a topic (see criterion discribed at the end of subsection 3.4.). Although the Portuguese equivalent *Comunidade económica europeia* occours in 0 documents, its orthografical variant (*Comunidade Económica Europeia*) occours in 55 documents, less than 50% of 113 documents in cluster 1; so it is not elected as topic.

Some MWUs presented as topics are uncompleted topics or even not real topics; for example *nomenclature of the Common* or *Having regard*. However we think that about 80% of these MWUs proposed as topics, can be considered as correct topics or subtopics.

6. Related Work

Some known approaches for extracting topics and relevant information use morpho-syntactic information, e.g., TIPSTER [7]. So, these approaches would need specific morpho-syntactic information in order to extract topics from documents in other languages, and that information might not be available.

In [8], a multi-document summarizer, called MEAD is presented. It generates summaries using cluster topic detection and a tracking system. However, in this approach, topics are unigrams. Though many uniwords have precise meanings, multiword topics are usually more informative and specific than unigrams. For example, *human* is too generic and vague, but *human rights* is much more precise.

Joe Zhou [9] suggests automatic topic-oriented two-word terms, based on mutual information scores for adjacent words. For multiword terms, mutual information score is calculated for non-adjacent words. A threshold is set to decide if multiword terms are important or not. We prefer to avoid thresholds in this phase, by using LocalMaxs approach, since there are relevant terms with lower cohesion scores than other terms with higher scores and then, a threshold may reject a relevant term and elect a non-relevant one.

In [10], an approach for clustering documents is presented. It is assumed that there exists a set of topics underlying the document collection to cluster, since topics are not extracted from the documents. When the number of clusters is not given to this system, it is calculated based on the number of topics. We think that there are two strong limitations on approaches like this one: first, the number of clusters is usually unknown by the user; second, the list of topics may be not complete and it is not dynamic since it must be very difficult to build a list of topics when the documents language or the subjects the documents are about are unknown.

7. Conclusions

By using statistics, it has been possible to overcome some weaknesses in the Computational Linguistics approaches based on simple symbolic methodologies. The strength of our approach lies mainly on language independence. So, this paper presents an unsupervised statistics-based and language independent approach for document clustering and topic extraction. We applied it to a multilingual corpus, using just information extracted from it. No predefined topics, features or descriptors were used. Thousands of MWUs were extracted by LiPXtractor, an extractor based on the improved LocalMaxs algorithm. MWU frequencies were then transformed and aggregated into a small set of new features (components), which — according to the results obtained — showed good document

discriminating power. The best number of clusters was automatically calculated by Model-Based Cluster Analysis and the results obtained led to a rather precise clustering of the documents. Thus, the number of clusters was not left to the user choice, as it might correspond to an *unnatural* clustering. Although we tested this approach on a small corpus (1 330 423 words), the results are encouraging, since about 80% of the proposed topics are sufficiently informative for being taken as correct topics of documents in the obtained document clusters. This lead us to believe that topics, rather than long sentences belonging to just one or two documents, are adequate for defining clusters core content. When distributions suggested lack of normality we made some data transformations and finally, documents were assign to clusters with higher Precision and Recall values: above 90%.

However, further research is required on larger corpora to improve this approach. In future work, we aim to refine the automatic choice of the number of components. We also want to include informative unigrams in order to enrich the base feature set and therefore check if better results are obtained.

Due to the heavy calculation of the clustering process, there are limitations for the input data for clustering software. Therefore, it is not possible to cluster more than a few hundred documents at the same time. However, the number of documents that need to be clustered or classified is much higher. So, in order to solve this problem, we extract a random sample of a few hundred documents from the initial set and use MBCA to cluster them. Then, based on the clusters obtained, we classified the rest of the documents of the initial set. New documents can then be classified or rejected based on the clusters previously determined. The results obtained had a Precision and Recall similar to the results presented. However, lack of space prevent us from describing the classification method used.

REFERENCES

[1] J.F. Silva, G. Dias, S. Guilloré, G.P. Lopes. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. *Lectures Notes in Artificial Intelligence, Springer-Verlag* **1695** (1999), 113–132.

[2] J.F. Silva, G.P. Lopes. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In *Proceedings of the 6th Meeting on the Mathematics of Language, Orlando* (1999), 369–381.

[3] Y. Escoufier,H. L'Hermier. A propos de la Comparaison Graphique des Matrices de Variance. *Biometrischc Zeitschrift* **20** (1978), 477–483.

[4] C. Fraley, A.E. Raftery. How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The computer Journal* **41** (1998), 578–588.

[5] R.A. Johnson, D.W. Wichern. Applied Multivariate Statistical Analysis, second edition. Prentice-Hall, 1988.

[6] G.E.P. Box, D.R. Cox. An Analysis of Transformations, (with discussion).*Journal of the Royal Statistical Society (B)* **26(2)** (1964), 211–252.

[7] Y. Wilks, R. Gaizauskas. Lasie Jumps the Gate. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999, 200–214.

[8] D.R. Radev, J. Hongyan, B. Makgorzata. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *Proceedings of the ANLP/NAACL Workshop on Summarization*, 2000.

[9] J. Zhou. Phrasal Terms In Real-World IR Applications. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999, 225–259.

[10] R.K. Ando, L. Lee. Iterative Residual Rescaling: An Analysis and Generalization of LSI. To appear in the proceedings of SIGIR 2001.

Joaquim Silva
DI, FCT, Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
e-mail: `jfs@di.fct.unl.pt`

João Mexia
DI, FCT, Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal

Carlos A. Coelho
DM, ISA, Universidade Técnica de Lisboa
Tapada da Ajuda, Portugal
e-mail: coelho@isa.pt

Gabriel Lopes
DI, FCT, Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
e-mail: gpl@di.fct.unl.pt