Cluster Analysis and Classification of Named Entities

Joaquim F. Ferreira da Silva
Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
jfs@di.fct.unl.pt

Zornitsa Kozareva
Faculty of Mathematics and Informatics, Plovdiv University
236, Bulgaria blvd., Plovdiv, Bulgaria
zkozareva@hotmail.com

José Gabriel Pereira Lopes
Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
gpl@di.fct.unl.pt

Abstract

This paper presents a statistics-based and language independent unsupervised approach for clustering possible named entities. We describe and motivate the features and statistical filters used by our clustering process. Using the Model-Based Clustering Analysis software we obtained different clusters of named entities. The method was applied to Bulgarian and English. For some clusters, precision is close to 100%; this helps human validation and saves time. Other clusters still need further refinement. Based on the obtained clusters, it is possible to classify new named entities.

1 Introduction

Language independent extraction of multiword units (MWUs) as proposed in section 2, gives rise to a huge number of MWUs, not all named entities. In this paper we describe how to statistically filter out possible named entities (section 3). Clustering attributes are described in section 4. Clustering results are presented in section 5. A Classifier for new named entities is shown in section 6 and conclusions are drown in section 7.

2 Extracting Multiwords 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) (Silva & Lopes, 1999). Thus, let us take an *n*-gram as a string of *n* words in any text. So, isolated words are 1-grams and the string *President of the Republic* is a 4-gram. One can intuitively accept that there is a strong cohesion within the 4-gram *United Nations General Assembly*, but not in the 4-gram *of that but not*. LocalMaxs algorithm is based on the idea that a MWU should be an *n*-gram whose cohesion is higher than any (*n*-1)-gram contained in the *n*-gram; and should also be higher than the cohesion of all the (*n*+1)-grams containing that *n*-gram. Thus, LocalMaxs needs to compare cohesions of *n*-grams having different sizes: (*n*+1), *n* and (*n*-1) and sharing all but one word in the borders, as we are interested on sequential *n*-grams. Then FDPN concept is applied to the *SCP(.)* measure in order to "transform" every *n*-gram of any length (*n*) in a pseudo-bigram, and then a new measure, *SCP_f(.)*, is obtained (Silva & Lopes, 1999).

$$SCP\_f(w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{Avp} \quad (1)$$

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(w_1 \dots w_i) \cdot p(w_{i+1} \dots w_n) \quad (2)$$

where $p(w_1 \dots w_j)$ is the probability of the *n*-gram $w_1 \dots w_j$ in the corpus. So, *SCP_f(.)* reflects the *average cohesion* between any two adjacent contiguous sub-*n*-gram of the original *n*-gram.

3 Filtering MWUs

For testing our approach we used an English corpus with 10,506,267 words and a Bulgarian corpus with 4,110,838 words. LocalMaxs extracted 207,088 MWUs from the first corpus and 164,655 MWUs from the second. These MWUs include named entities among other multiwords. After separating those MWUs whose first and last words start with a capital letter, the number of MWUs decreased to 50,558 for English and 11,498 for Bulgarian. Since named entities usually have no long non-capital words with low probability, a second filter was applied by calculating the following value for each MWU:

$$\min Pl(w_1 \dots w_n) = \min_i (PL(w_i)) \quad (3)$$

$$PL(w_i) = \frac{freq(w_i)}{N \cdot length(w_i)} \quad (4)$$

where $freq(w_i)$ is the frequency of the *i*-th non-capital word of the MWU in the corpus; *N* stands for the corpus size we

are working with, and *length(w_i)* is the number of characters of word w_i. Then, MWUs having *minPL($w_1 \ldots w_n$)* greater than a threshold were taken as good named entities, since they have no long non-capital words with medium or low probability. The threshold found seemed to be the same for both languages: 0.0053 (Kozareva et al., 2004).

4 Attributes

Proper features were needed for clustering filtered named entities. As shown in (Kozareva et al., 2004), the best features found were *Permanency* and *PLStdDev*.

$$Permanency(w_1 \ldots w_n) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{f(w_i)}{f^*(w_i)} \qquad (5)$$

where $f(w_i)$ is the frequency of the word (w_i) in the corpus, while $f^*(w_i)$ is the frequency of the same word but taking all occurrences of case insensitive forms (ex: *life, Life, LIFE*). This feature helps to distinguish names of persons (where *Permanency* is close to 1, as they occur written the same way) from other types of named entities. The second attribute is based on the standard deviation concept taking the probability and the length of words in the named entity.

$$PLStdDev(w_1 \ldots w_n) = \sqrt{\frac{1}{n}(PL(w_i) - \overline{PL(w)})^2} \qquad (6)$$

$$\overline{PL(w)} = \frac{1}{n}\sum_{i=1}^{i=n} PL(w_i) \qquad (7)$$

Equation (4) gives *PL(.)*. *PLStdDev* is useful for distinguishing named entities such as *Republic of Bulgaria* from others like *Bulgarian Parliament*. These named entities have different variation on the probability and length of their words. Here, we present the standardization to assign the same discriminant power to every attribute.

$$z_{k,i} = \frac{x_{k,i} - x_{k,.}}{std(x_k)} \qquad (8) \qquad x_{k,.} = \frac{1}{l}\sum_{i=1}^{i=l} x_{k,i} \qquad (9)$$

$z_{k,i}$ is the standardized value for the *i*-th element of the attribute *k* ($x_{k,i}$); $x_{k,.}$ is the mean value of the elements for the same attribute and *l* is the number of elements; $std(x_k)$ holds as the standard deviation of the same set:

$$std(x_k) = \sqrt{\frac{1}{l}\sum_{i=1}^{i=l}(x_{k,i} - x_{k,.})^2} \qquad (10)$$

5 Clustering Named Entities

Then, having a matrix of named entities characterized by previous 2 features, clustering is done using Model Based Clustering Analysis (MBCA) software. Different models are "simulated" for the input matrix, and the most likely model is proposed by this approach. Due to limitations imposed by the heavy clustering calculations done by MBCA, we clustered just a representative 1000 elements sample from the initial set of named entities for each corpus.

5.1 Results of the Clustering

As shown in (Kozareva et al., 2004), for each corpus (English and Bulgarian), MBCA proposed 5 clusters. We present three elements randomly taken from each cluster.

Cluster e1: *HUMANITARIAN AID, ANNUAL REPORT, SECTOR UNDERSTANDING ON EXPORT CREDITS.*

Cluster e2: *Media Markets*, *White Cement Committee*, *Management Committee*.

Cluster e3: *Vega Cueva*, *Herrenbuck Herrenstuck Hex*, *Glatzen Harstell*.

Cluster e4: *Health and Social Services*, *Northern Ireland Office Crown*, *Bayer France and Bayer Spain*.

Cluster e5: *Secretary-General of the United Nations*, *Republic of Trinidad and Tobago*, *Department of Tourism and Transport.*

Cluster b1: *Dobromir Krystew Atanasow*, Simeon Zahariew Simeonow, Diliana Kirilowa Ignatowa.

Cluster b2: *MINERALNA WODA OT WODOIZTOCHNIK TK-1 (MINERAL WATER FROM WATER TANK TK-1)*, *ZAKANATA S PRESTAPLENIE TRIABWA DA SE RAZGRANICHAWA (THREATENING WITH CRIME SHOULD BE DISTINGUISHED*, *DYRJAWNA SOBSTWENOST (STATE PROPERTY).*

Cluster b3: *Emil Georgiew Mihow i Walentin Minchew (Emil Georgiew Mihow and Walentin Minche)*, *Boris i Stefan Hadjiew*, *Konstantin Petrow Mochikow i Kiril Iwanow Okow.*

Cluster b4: *Diakowa ot Sliwen (Diakowa from Sliwen)*, *Pechew ot Warna*, *Ugyrchin i Iablanica (Ugyrchin and Iablanica).*

Cluster b5: *Sweta Nedelia*, *Dolno Kozarewo*, *Georgi Todorow Jilow*.

5.2 Discussion

Cluster	Total	Precision (%)	Recall (%)
e1	287	97	65
e2	342	22	88
e3	117	90	80
e4	134	29	63
e5	120	50	98

Table 1: Evaluation of English clusters

Cluster	Total	Precision (%)	Recall (%)
b1	286	100	79
b2	450	94	84
b3	44	100	94
b4	40	25	55
b5	180	49	24

Table 2: Evaluation of Bulgarian clusters

Clusters were proposed by MBCA considering VVV (Variable volume, Variable shape and Variable orientation) the best model for both languages; details in (Fraley & Raftery, 98). This shows that clusters are not always spherical and have not the same volume. Person names tend to have frozen writing, which is detected by *Permanency* attribute: cluster *e3* for English 90% precision, table 1 and clusters *b1* and *b3* for Bugarian (100% precision, table 2). However, for Bulgarian person names, those having no small and frequent words, that is low *PLStdDev* values, were put in cluster *b1*; those having high *PLStdDev* values are in cluster *b3*. So, for person names, we have 1 cluster for English and 2 for Bulgarian. This is due to the very different nature of the corpora (kozareva et al., 2004). Wrong written named entities are rare events corresponding to low *Permanency* values: cluster 1 for English (97% precision) and cluster 2 for Bulgarian (94% precision). The results of the other clusters require future work. They tend to have institutions and city names: clusters 2, 4 and 5 for English with 22%, 29% and 50% precision respectively, and clusters 4 and 5 for Bulgarian with 25% and 49% precision.

Precision and recall values were calculated on the basis of majority of specific type of named entities clustered. So, if in the cluster of person names occurred a name of an enterprise, this would count as failure.

6 Classifying New Named Entities

Although we have just clustered a representative sample of the initial set of named entities the remaining elements must be also classified, concerning the clusters obtained. Beside that, we must be able to classify new named entities that did not occur in our corpus.

6.1 The Discriminant Quadratic Score

So, every unclassified element must be part of any class represented by one of the already formed clusters, or be clearly out of those clusters. During the process of classifying new proper names we used the Discriminant Quadratic Score in order to indicate "how close" a named entity represented by the vector $\vec{y}$ is to a class *i*.

$$d_i^Q(\vec{y}) = -\frac{1}{2}\ln|\vec{\Sigma}_i| - \frac{1}{2}(\vec{y}-\vec{\mu}_i)^T\vec{\Sigma}_i^{-1}(\vec{y}-\vec{\mu}_i) + \ln p_i \quad (11)$$

Covariance matrix $\vec{\Sigma}_i$ is associated with the attributes that characterize the elements from the class *i*, and it is estimated by the covariance matrix $\vec{S}_i$ taking the elements (named entities) of cluster *i*. The generic covariance matrix among the attributes of a cluster is given by:

$$\vec{S} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,k} \\ S_{1,2} & S_{2,2} & \cdots & S_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1,k} & S_{2,k} & \cdots & S_{k,k} \end{bmatrix} \quad (12)$$

Where *k* is the number of attributes characterizing the elements of the cluster; in our case *k* is 2 and

$$S_{l,j} = \frac{1}{n-1}\sum_{i=1}^{i=n}(e_{l,i} - e_{l,.})(e_{j,i} - e_{j,.}) \quad (13)$$

where $e_{l,i}$ is the value of the named entity *i* for the attribute *l*, and *n* is the number of elements in the cluster. The mean value $e_{l,.}$ of the attribute *l* in the cluster is given by

$$e_{l,.} = \frac{1}{n}\sum_{i=1}^{i=n} e_{l,i} \quad . \quad (14)$$

For Discriminant Quadratic Score, $d_i^Q(\vec{y})$, contribute the following factors, being $\vec{y}$ still the vector that represents an element to be classified: the logarithm of the determinant of the covariance matrix associated with class *i*, represented by cluster *i*; the logarithm of the probability of an element (named entity) belonging to class *i* (this value is estimated by the logarithm of the number of the named entities of cluster *i* divided by the number of named entities of all clusters; and the *Mahalanobis distance* between $\vec{y}$ and the vector of means of class *i,* ($\vec{\mu}_i$), represented by the vector of means of cluster *i*, ($\vec{c}_i$). This last factor (Mahalanobis distance) is very important, and the lower it is, the higher the score $d_i^Q(\vec{y})$. So, let $\vec{y}$ be a vector that represents an element to be classified, and π_r a class represented by cluster *r* that contains named entities (vectors $\vec{e}_1, \vec{e}_2, \ldots \vec{e}_n$). Then $\vec{y}$ belongs to π_r if and only if

$$d_r^Q(\vec{y}) = \max_i d_i^Q(\vec{y}) \wedge d_r^Q(\vec{y}) \geq \min_j d_r^Q(\vec{e}_j) \quad (15)$$

where *i=1,2,…g*; *g* is the number of classes (clusters), and *j=1,2,…n*, where *n* is the number of named entities of the cluster. Equation (15) describes a criterion for classifying new named entities. This corresponds to the *Minimum Total Probability of Misclassification Rule for Normal Populations* criterion, but with an extra condition we present here: $d_r^Q(\vec{y}) \geq \min_j d_r^Q(\vec{e}_j)$. This condition sets that an element $\vec{y}$ belongs to class *r* if its Quadratic Score is also higher or equal than the Quadratic Score of all elements of cluster *r*. This prevents a very "distant" and "strange" element to be classified as a member of any class represented by the clusters.

6.2 The Vector for the New Candidate

Considering the *Permanency* attribute we mentioned before, we must calculate the corresponding value for the new element we want to classify.

$$newPermanency(w_1 \dots w_n) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{f(w_i)+s}{f^*(w_i)+s} \quad (16)$$

where $f(w_i)$ stands for the frequency of the i-th word. If the word already exists in our corpus then s is set to 0 and we take into consideration the frequency that it already has, otherwise the value of s is set to 1; $f^*(w_i)$ is defined in section 4. The corresponding value for the *PLStdDev* attribute is given by

$$newPlStdDev(w_i \dots w_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \bar{x})^2} \quad (17)$$

where $\bar{x}$ is the mean value of all x_i and

$$x_i = \frac{f(w_i)+s}{N \cdot length(w_i)} \quad . \quad (18)$$

Again, $f(w_i)$ stands for the frequency of the word i in the corpus. If the word i is present in the corpus we take its frequency and the value of s is set to 0, otherwise s is set to 1. N is the size of the corpus we are working with, and $length(w_i)$ is the length of the i-th word. After we obtained the attributes for the new candidate we have standardized them, as we have done before for the elements in the clusters. For obtaining standardized values we used the formula below

$$z_k = \frac{x_k - \overline{O}}{std(O)} \quad (19)$$

in which x_k is the attribute value that is going to be standardized according to the values previously obtained to attribute k in the clustering process. $\overline{O}$ has the same value as $x_{k,.}$ in equation (9), that is the mean value for attribute k obtained during the clustering process. This is done because the standardization of each attribute value, must be made associating the "statistical behaviour" (namely mean and standard deviation) obtained for the same attribute by the representative sample used in the clustering process. The same reasoning is used for *std(O)*, so *std(O)* has the same value as $std(x_k)$ in equation (10). So, we can obtain the standardized values z_1 e z_2 corresponding to the values for the attributes *Permanency* and *PLStdDev* for the named entity to be classified, that is $\vec{y}^T = [z_1, z_2]$. Thus, using this criterion described in equation (15) we are able to classify the new named entity represented by the vector $\vec{y}$.

6.3 Discussion

Although we have done just a few tests on classification, for Bulgarian named entities this classifier showed 100% precision for person names and 60% for institution names and city names. Similar values were obtained for English. Much more tests using other kinds of named entities and other corpora has to be done for a complete assessment of the performance of this classifier.

7 Related Work and Conclusions

Recently, some Machine Learning approaches such as (McNamee & Mayfield, 2002; Carreras et al., 2002) have been used to extract named entities. However, these systems usually require a set of labeled data to be trained on, and this may not be available or be expensive to obtain. Other systems are language oriented such as (Carreras et al., 2003), or symbolic dependent such as (Poibeau et al., 2003). This paper presents an unsupervised statistics-based and language independent approach for clustering named entities. Firstly, thousands of MWUs were extracted from corpora using LocalMaxs algorithm. Possible named entities were filtered and clustered using just two attributes. This methodology was applied on 2 different corpora (English and Bulgarian) and similar results were obtained in both languages for some clusters. The best number of clusters was automatically calculated by Model-Based Cluster Analysis. The results are encouraging, since about 95% of the person names and misspelled expressions were correctly grouped. Although we have just clustered a representative sample of the initial set of named entities, the remaining elements must either be classified or rejected, concerning the clusters obtained.

8 References

- Carreras, X. & Màrquez, L. & Padró, L. (2002). Named Entity Extraction Using AdaBoot. In Proceedings of CONLL-2002, pages 167-170. Taipei, Taiwan.
- Carreras, X. & Màrquez, L. & Padrò, L. (2003). Entity Extraction Recognition for Catalan Using Spanish Resources. In Proceedings of the 10th Conference of the EACL 2003. Budapest, Hungary.
- Fraley, C. & Raftery, A. E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. The computer Journal, 41, (pp 578--588).
- Kozareva, Z. & Silva, J. F. & Gamallo, P. & Lopes, G. P. (2004). Cluster Analysis of Named Entities. In Proceedings of the International Intelligent Information Processing and Web Mining Conference May 17-20, 2004, Zakopane, Poland. In Lecture Notes in Artificial Intelligence LNCS/LNAI. Berlin: Springer-Verlag. 2004 (to be published).
- McNamee, P. & Mayfield, J. (2002). Entity Extraction without Language-Specific Resources. In Proceedings of CONLL-2002, (pp183—186). Taipei, Taiwan.
- Poibeau. T. & INaLCO Named Entity Group. (2003). The Multilingual Named Entity Recognition Framework. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003). Budapest, Hungary.
- Silva, J. F. & Dias, G. & Guilloré, S. & Lopes, G. P. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In Lectures Notes in Artificial Intelligence, Springer-Verlag, volume 1695, (pp113--132).