TALN 2004, Session Poster, Fès, 19–21 avril 2004

Extracting Named Entities. A Statistical Approach

Joaquim Silva (1), Zornitsa Kozareva (2), Veska Noncheva (2),
Gabriel Lopes (1)
(1) Departamento de Informática, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa
Quinta da Torre, 2725 Monte da Caparica, Portugal
{jfs,gpl}@di.fct.unl.pt
(2) Faculty of Mathmatics and Informatics
Plovdiv, Bulgaria
{zkozareva@hotmail.com} {nonchev@plovdiv.techno-link.com}

Résumé - Abstract

Les entitées nomées et plus généralement les multi-mots sont des ressources importantes pour plusieurs applications. Cependant, les métodes d'extraction automatique, indépendentes de la langue, de multi-mots, ne nous donnent pas des données 100% fiables. Dans ce papier nous proposons premièrement une méthode pour selectioner entités nomées d'entre les multi-mots extraits automatiquement et, deuxièmement, une méthode de groupement des entités nomées non-supervisionée et indépendente de la langue, en utilisant de la statistique. La deuxième phase de groupement rends l'évaluation humaine plus simple. Les traits utilisés pour le groupement sont décrits et motivés. L'analyse faite pour le groupement nous a permis d'obtenir différents groupes d'entités nomées. La méthode a été appliquée sur le bulgare et l'anglais. La précision obtenue pour certains groupes a été très haute. D'autres groupes doivent être encore rafinés. Par ailleurs, les traits discrimants appris pendant la phase de groupement nous permettent de classifier de nouvelles entités nomées.

Named entities and more generally Multiword Lexical Units (MWUs) are important for various applications. However, language independent methods for automatically extracting MWUs do not provide us with clean data. So, in this paper we propose a method for selecting possible named entities from automatically extracted MWUs, and later, a statistics-based language independent unsupervised approach is applied to possible named entities in order to cluster them according to their type. Statistical features used by our clustering process are described and motivated. The Model-Based Clustering Analysis (MBCA) software enabled us to obtain different clusters for proposed named entities. The method was applied to Bulgarian and English. For some clusters, precision is very high; other clusters still need further refinement. Based on the obtained clusters, it is also possible to classify new possible named entities.

Mots-clefs – Keywords

Entités Nommées, Unités Multi-mots, Groupement, Classification
Named Entities, Multiword Units, Clustering, Classification

1 Introduction

This paper aims at proposing a methodology for clustering named entities. A language independent method is explained for extracting relevant multiword units (MWUs) (Silva *et al.*, 1999) in section 2. As it is shown in section 3, from these MWUs, possible named entities are filtered using simple heuristics. Attributes used for clustering them are described in section 4. Clustering methodology and results are presented in section 5. A method for classifying new named entities is shown in section 6, and related work and conclusions in section 7.

2 Extracting MWUs 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 *et al.*, 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. However, for determining the cohesion of an n-gram, we need to transform it into a pseudo-bigram. So, identical weight is given to every possible contiguous bigram in the n-gram by calculating the arithmetic mean show in equation 1. This corresponds to the FDPN concept applied to the $SCP(.)$, and then, a new measure, $SCP\_f(.)$, is obtained (Silva *et al.*, 1999).

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

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 multi-words. 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 in this languages usually have no long non-capital words with low probability, a second filter was applied by calculating the following value for each MWU:

$$minPL(w_1 \dots w_n) = \min_i(PL(w_i)) \qquad \text{where} \qquad PL(w_i) = \frac{freq(w_i)}{N \times length(w_i)} \ . \tag{2}$$

And $freq(w_i)$ 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 $minPL(w_1 \dots w_n)$ 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 $PLStdDev$.

$$Permanency(w_1 \dots w_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(w_i)}{f^*(w_i)} \tag{3}$$

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.

$$PLStdDev(w_1 \dots w_n) = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(PL(w_i) - \overline{PL(w)})^2} \qquad \overline{PL(w)} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} PL(w_i) \ . \tag{4}$$

Equation 2 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.

$$z_{k,i} = \frac{x_{k,i} - x_{k,\cdot}}{std(x_k)} \qquad x_{k,\cdot} = \frac{1}{l}\sum_{i=1}^{i=l} x_{k,i} \qquad std(x_k) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(x_{k,i} - x_{k,\cdot})^2} \ . \tag{5}$$

$z_{k,i}$ is the standardized value for the i-th element of the attribute k ($x_{k,i}$); $x_{k,\cdot}$ 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.

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.

Cluster	Example	Total	Prec. (%)	Rec. (%)
e1	HUMANITARIAN AID	287	97	65
e2	Lands Tribunal Law Commission Legal Aid	342	22	88
e3	Vega Cueva	117	90	80
e4	Gatwick and Manchester	134	29	63
e5	Hungary and the Czech Republic	120	50	98
b1	Dobromir Krystew Atanasow	286	100	79
b2	DYRJAWNA SOBSTWENOST	450	94	84
b3	Emil Georgiev Mihow i Walentin Minchew	44	100	94
b4	Diakowa ot Sliwen	40	25	55
b5	Sweta Nedelia	180	49	24

Figure 1: Evaluation of English and Bulgarian clusters

5.1 Results of the Clustering

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

Cluster e1: *HUMANITARIAN AID*, *ANNUAL REPORT*, *SECTOR UNDERSTANDING ON EXPORT CREDITS*.
Cluster e2: *Media Markets*, *Management Committee*, *White Cement 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: *Department of Tourism and Transport*, *Republic of Trinidad and Tobago*, *Secretary-General of the United Nations*.
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

Clusters were proposed by MBCA considering VVV (Variable volume, Variable shape and Variable orientation) the best model for both languages; details in (Fraley and Raftery, 1998). 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). 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 (Silva *et al.*, 2004). Wrong written named entities are rare events corresponding to low Permanency values: cluster *e1* (97% precision) and cluster *b2* (94% precision). The results of the other clusters require future work. They tend to have institutions and city names: clusters *e2*, *e4* and *e5* with 22%, 29% and 50% precision respectively, and clusters *b4* and *b5* 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. 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 named entities 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.

$$d_i^Q(\vec{y}) = -\frac{1}{2}\ln\left|\vec{\Sigma}_i\right| - \frac{1}{2}(\vec{y}-\vec{\mu}_i)^T\vec{\Sigma}_i^{-1}(\vec{y}-\vec{\mu}_i) + \ln p_i \ . \tag{6}$$

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 taking the elements (named entities) of cluster i (see details on (Silva *et al.*, 2004)). For Discriminant Quadratic Score, the most important factor is the Mahalanobis Distance between the named entity $\vec{y}$ and the vector of means of class i, ($\vec{\mu}_i$), represented by the vector of means of cluster i ($\vec{c}_i$) (see details on (Silva *et al.*, 2004)). This lower this factor is, the higher the Quadratic Score. 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 class π_r if and only if

$$d_r^Q(\vec{y}) = \max_i d_i^Q(\vec{y}) \wedge d_r^Q(\vec{y}) \geq \min_j d_r^Q(\vec{e}_j) \tag{7}$$

where $i = 1, 2, \ldots, g$; g is the number of classes (clusters), and $j = 1, 2, \ldots, n$, where n is the number of named entities of the cluster. Equation (7) 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 added: $d_r^Q(\vec{y}) \geq \min_j d_r^Q(\vec{e}_j)$. 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. A detailed explanation about the components of the vector $\vec{y}$, representing the new named entity, is given in (Silva *et al.*, 2004). 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.

7 Related Work, Conclusions and Future Work

Recently, some Machine Learning approaches such as (McNamee and Mayfield, 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 points to 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. However, there are problems to solve: it is not possible to distinguish *São Paulo* as a person name and *São Paulo* as a location name, unless we use semantic information and look at the neighbourhood of the MWU *São Paulo* in the corpus. Besides, $Permanency$ attribute does not work if the language is not alphabetic or if it does not use upper-case for person names. So future work has to be done, concerning other attributes and other kind of informations.

References

CARRERAS X., MÀRQUEZ L., PADRÓ L. (2003). Entity Extraction Recognition For Catalan Using Spanish Resources. *In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003). Budapest, Hungary.*

FERREIRA DA SILVA J., DIAS G., GUILLORÉ S., LOPES G. P. (1999), Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, *Lectures Notes in Artificial Intelligence, Springer-Verlag*, volume 1695, pages 113-132.

FERREIRA DA SILVA J., KOZAREVA Z., LOPES G. P. (2004), Cluster Analysis and Classification of Named Entities. *In Proceedings 4 th International Conference ON Language Resources and Evaluation, May, Lisbon, Portugal* (to be published).

FRALEY C., RAFTERY A. E. (1998),How many clusters? Which clustering method? Answers via model-based cluster analysis, *The computer Journal*, 41, p. 578-588.

KOZAREVA Z., FERREIRA DA SILVA J., GAMALLO P., LOPES G. P. (2004), Cluster Analysis of Named Entities. *In Proceedings of the International Intelligent Information Processing and Web Mining Conference, Zakopane, Poland. 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. *Proceedings of CONLL-2002, pages 183-186*. 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.*