Cluster Analysis of Named Entities

Zornitsa Kozareva[1], Joaquim Silva[2], Pablo Gamallo[2], and Gabriel Lopes[2]

[1] Faculty of Mathematics and Informatics, Plovdiv University, 236 Bulgaria blvd., Plovdiv, Bulgaria
[2] Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2725 Monte da Caparica, Portugal

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.

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 filter out possible named entities (section 3). Clustering attributes are described in section 4. Results are presented in section 5, and conclusions are drown in section 6.

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) [6]. 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 with different sizes: $(n+1)$, n and $(n-1)$. 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 a new measure, $SCP\_f(.)$, is obtained [6].

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

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 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:

$$minPL(w_1 \dots w_n) = \min_i (PL(w_i)) \qquad PL(w_i) = \tfrac{freq(w_i)}{N \, . \, length(w_i)} \ . \tag{2}$$

$freq(w_i)$ 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.

4 Attributes

Proper features were needed for clustering the filtered named entities. The best features found after some tests were *Permanency* and *PLStdDev*.

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

$f(w_i)$ is the frequency of the word w_i, while $f^*(w_i)$ indicates the frequency of the same word in the corpus but considering 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.

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

$$\overline{PL(w_i)} = \frac{1}{n}\sum_{i=1}^{n} PL(w_i) \ . \tag{5}$$

Equations (2) give $PL(.)$. $PLStdDev$ 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. For equal discriminant power, standardization was applied to both features.

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

For English and Bulgarian, MBCA proposed 5 clusters according to the VVV model (Variable volume, Variable shape and Variable orientation; details in [3]). Now we present three elements randomly taken from each cluster both for English and Bulgarian.

Cluster 1: *HUMANITARIAN AID*, *ANNUAL REPORT*, *SECTOR UNDERSTANDING ON EXPORT CREDITS*.

Cluster 2: *Media Markets*, *White Cement Committee*, *Management Committee*.

Cluster 3: *Vega Cueva*, *Herrenbuck Herrenstuck Hex*, *Glatzen Harstell*.

Cluster 4: *Health and Social Services*, *Northern Ireland Office Crown*, *Bayer France and Bayer Spain*.

Cluster 5: *Secretary-General of the United Nations*, *Republic of Trinidad and Tobago*, *Department of Tourism and Transport*.

Cluster 1: *Dobromir Krystew Atanasow*, *Simeon Zahariew Simeonow*, *Diliana Kirilowa Ignatowa*.

Cluster 2: *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 3: *Emil Georgiew Mihow i Walentin Minchew (Emil Georgiew Mihow and Walentin Minche)*, *Boris i Stefan Hadjiew*, *Konstantin Petrow Mochikow i Kiril Iwanow Okow*.

Cluster 4: *Diakowa ot Sliwen (Diakowa from Sliwen)*, *Pechew ot Warna*, *Ugyrchin i Iablanica (Ugyrchin and Iablanica)*.

Cluster 5: *Sweta Nedelia*, *Dolno Kozarewo*, *Georgi Todorow Jilow.*

5.2 Discussion

Table 1. Evaluation of English clusters

Cluster	Example	Total	Prec. (%)	Rec. (%)
1	HUMANITARIAN AID	287	97	65
2	White Cement Committee	342	22	88
3	Vega Cueva	117	90	80
4	Health and Social Services	134	29	63
5	Department of Tourism and Transport	120	50	98

Table 2. Evaluation of Bulgarian clusters

Cluster	Example	Total	Prec. (%)	Rec. (%)
1	Dobromir Krystew Atanasow	286	100	79
2	DYRJAWNA SOBSTWENOST	450	94	84
3	Emil Georgiev Mihow i Walentin Minchew	44	100	94
4	Diakowa ot Sliwen	40	25	55
5	Sweta Nedelia	180	49	

Clusters were proposed by MBCA considering VVV model the *best model*, for both languages. 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 3 for English (90% precision, table 1) and clusters 1 and 3 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 1; those having high *PLStdDev* values are in cluster 3. 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: the English one has documents from the European Parliament containing individual person names, names of institutions, etc. in French, German and other languages while the Bulgarian comes from tax legislation from Bulgaria and picks husband and wife names. 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 cities 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 enterprize, this would count as failure.

6 Related Work and Conclusions

Recently, some Machine Learning approaches such as [4,1] 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 [2], or symbolic dependent such as [5]. 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. The possible named entities were filtered and clustered using just two attributes. This methodology were 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 be also classified, concerning the clusters obtained. However, lack of space prevent us from describing here our classification method for that purpose.

References

1. Carreras X., Màrquez L., Padró L. 2002. Named Entity Extraction Using AdaBoot. *Proceedings of CONLL-2002*, pages 167-170. Taipei, Taiwan.
2. Carreras X., Màrquez L., Padró L. 2003. Entity Extraction Recognition For Catalan Using Spanish Resources. *In Proceedings of the 10th Conference of the EACL 2003*. Budapest, Hungary.
3. Fraley, C., Raftery, A. E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The computer Journal*, 41, pages 578-588.
4. McNamee P., Mayfield J. 2002. Entity Extraction Without Language-Specific Resources. *Proceedings of CONLL-2002*, pages 183-186. Taipei, Taiwan.
5. 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.
6. 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. *Lectures Notes in Artificial Intelligence, Springer-Verlag*, volume 1695, pages 113-132.