Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 51-58 recibido 11-04-14 revisado 30-06-14 aceptado 02-07-14

An Unsupervised Algorithm for Person Name Disambiguation in the Web*

Algoritmo no Supervisado para Desambiguación de Nombres de Personas en la Web

Agustín D. Delgado, Raquel Martínez, Víctor Fresno
Universidad Nacional de Educación a Distancia (UNED)
Juan del Rosal, 16, 28040 - Madrid
{agustin.delgado, raquel, vfresno}@lsi.uned.es

Soto Montalvo
Universidad Rey Juan Carlos (URJC)
Tulipán, S/N, 28933 - Móstoles
soto.montalvo@urjc.es

Resumen: En este trabajo presentamos un sistema no supervisado para agrupar los resultados proporcionados por un motor de búsqueda cuando la consulta corresponde a un nombre de persona compartido por diferentes individuos. Las páginas web se representan mediante n-gramas de diferente información y tamaño. Además, proponemos un algoritmo de clustering capaz de calcular el número de clusters y devolver grupos de páginas web correspondientes a cada uno de los individuos, sin necesidad de entrenamiento ni umbrales predefinidos, como hacen los mejores sistemas del estado del arte en esta tarea. Hemos evaluado nuestra propuesta con tres colecciones de evaluación propuestas en diferentes campañas de evaluación para la tarea de Desambiguación de Personas en la Web. Los resultados obtenidos son competitivos y comparables a aquellos obtenidos por los mejores sistemas del estado del arte que utilizan algún tipo de supervisión.
Palabras clave: aprendizaje no supervisado, clustering, n-gramas, búsqueda de personas en la web

Abstract: In this paper we present an unsupervised approach for clustering the results of a search engine when the query is a person name shared by different individuals. We represent the web pages using n-grams, comparing different kind of information and different length of n-grams. Moreover, we propose a new clustering algorithm that calculates the number of clusters and establishes the groups of web pages according to the different individuals, without the need of any training data or predefined thresholds, as the successful state of the art systems do. Our approach is compared with three gold standard collections compiled by different evaluation campaigns for the task of Web People Search. We obtain really competitive results, comparable to those obtained by the best approaches that use annotated data.
Keywords: unsupervised learning, clustering, n-grams, web people search

1 Introduction

Resolving the ambiguity of person names in web search results is a challenging problem becoming an area of interest for Natural Language Processing (NLP) and Information Retrieval (IR) communities. This task can be defined informally as follows: given a query of a person name in addition to the results of a search engine for that query, the goal is to cluster the resultant web pages according to the different individuals they refer to. Thus, the challenge of this task is estimating the number of different individuals and grouping the pages of the same individual in the same cluster.

The difficulty of this task resides in the fact that a single person name can be shared by many people. This problem has had an impact in the Internet and that is why several vertical search engines specialized in web people search have appeared in the last years, e.g. `spokeo.com`, `123people.com` or `zoominfo.com`. This

* The authors would like to thank the financial support for this research to the Spanish research project Holopedia funded by the Ministerio de Ciencia e Innovación under grant TIN2010-21128-C02 and by UNED Project (2012V/PUNED/0004).

task should not be mixed up with *entity linking* (EL). The goal of EL is to link name mentions of entities in a document collection to entities in a reference knowledge base (typically Wikipedia), or to detect new entities.

The main difficulties of clustering web pages referring to the same individual come from their possible heterogeneous nature. For example, some pages may be professional sites, while others may be blogs containing personal information. To overcome these difficulties the users have to refine the queries with additional terms. This task gets harder when the person name is shared by a celebrity or by a historical figure, because the results of the search engines are dominated by that individual, making the search of information about other individuals more difficult.

WePS[1] (Web People Search) evaluation campaigns proposed this task in a web searching scenario providing several corpora for evaluating the results of their participants, particularly WePS-1, WePS-2 and WePS-3 campaigns. This framework allows to compare our approach with the state of the art systems.

The most successful state of the art systems have addressed this problem with some kind of supervision. This work proposes a data-driven method for this task with the aim of eliminating the elements of human annotation involvement in the process as much as possible. The main contribution of this work is a new unsupervised approach for resolving person name ambiguity of web search results. It is based on the use of capitalized n-grams to represent the pages that share the same person name, and also in an algorithm that decides if two web pages have to be grouped using a threshold that only depends on the information of both pages.

The paper is organized as follows: in Section 2 we discuss related work; Section 3 details the way we represent the web pages and our algorithm; in Section 4 we describe the collections used for evaluating our approach and we show our results making a comparison with other systems; the paper ends with some conclusions and future work in Section 5.

2 Related Work

Several approaches have been proposed for clustering search results for a person name query. The main differences among all of them are the features they use to represent the web pages and the clustering algorithm. However, the most successful of them have in common that they use some kind of supervision: learning thresholds and/or fixing manually the value of some parameters according to training data.

Regarding the way of representing a web page, the most popular features used by the most successful state of the art approaches are Name Entities (NE) and Bag of Words (BoW) weighted by TF-IDF function. In addition to such features, the systems usually use other kind of information. Top systems from WePS-1 and WePS-2 campaigns, CU_COMSEM (Chen and Martin, 2007) and PolyUHK (Chen, Yat Mei Lee, and Huang, 2009), distinguish several kind of tokens according to different schemes (URL tokens, title tokens, ...) and build a feature vector for each sort of tokens, using also information based on the noun phrases appearing in the documents. PolyUHK also represents the web pages with n-grams and adds pattern techniques, attribute extraction and detection when a web page is written in a formal way. A more recent system, HAC_Topic (Liu, Lu, and Xu, 2011), also uses BoW of local and global terms weighted by TF-IDF. It adds a topic capturing method to create a Hit List of shared high weighted tokens for each cluster obtaining better results than WePS-1 participants. IRST-BP system (Popescu and Magnani, 2007), the third in WePS-1 participant ranking, proposes a method based in the hypothesis that appropriated n-grams characterize a person and makes extensive use of NE and other features as temporal expressions. PSNUS system (Elmacioglu et al., 2007) use a large number of different features including tokens, NE, hostnames and domains, and n-gram representation of the URL links of each web page. (Artiles, Amigó, and Gonzalo, 2009a) studies, using also the collections WePS-1 and WePS-2, the role of several features as NE, n-grams or noun phrases for this task reformulating this problem as a classification task. In their conclusions, they claim that using NE does not provide substantial improvement than using other combination of features that do not require linguistic processing (snippet tokens, n-grams, ...). They also present results applying only n-grams of different length, claiming that n-grams longer than 2 are not effective, but bigrams improves the results of tokens. On the other hand, the WePS-3 best system, YHBJ (Chong and Shi, 2010), uses information extracted manually from Wikipedia adding

[1] http://nlp.uned.es/weps/

to BoW and NE weighted by TF-IDF.

Regarding the clustering algorithms, looking at WePS campaigns results, the top ranked systems have in common the use of the Hierarchical Agglomerative Clustering algorithm (HAC) described in (Manning, Raghavan, and Schütze, 2008). Different versions of this algorithm were used by (Chen and Martin, 2007; Chen, Yat Mei Lee, and Huang, 2009; Elmacioglu et al., 2007; Liu, Lu, and Xu, 2011; Balog et al., 2009; Chong and Shi, 2010).

The only system that does not use training data, DAEDALUS (Lana-Serrano, Villena-Román, and González-Cristóbal, 2010), which uses k-Medoids, got poor results in WePS-3 campaign. In short, the successful state of the art systems need some kind of supervised learning using training data or fixing parameters manually. In this paper we explore and propose an approach to address this problem by means of data-driven techniques without the use of any kind of supervision.

3 Proposed Approach

We distinguish two main phases in this clustering task: web page representation (Sections 3.1 and 3.2) and web page grouping (Sections 3.3 and 3.4).

3.1 Feature Selection

The aim of this phase is to extract relevant information that could identify an individual. Several of the state of the art systems use word n-grams to represent the whole or part of the information of a web page. Our main assumption is that co-occurrences of word n-grams, particularly of capitalized words, could be an effective representation in this task. We assume the main following hypotheses:

(i) Capitalized n-grams co-occurrence could be a reliable way for deciding when two web pages refer the same individual. Capitalized n-grams usually are NE (organizations and company names, locations or other person names related with the individual) or information not detected by some NE recognizers as for example, the title of books, films, TV shows, and so on. In a previous study with WePS-1 training corpus using the Stanford NER[2] to annotate NE, we detected that only 55.78 % of the capitalized tokens were annotated as NE or components of a NE by the NER tool. So the use of capitalized tokens allows increase the number of features in connection to the use of only NE. We also compared the n-gram representation with capitalized tokens and with NE. We found that 30.97 % of the 3-grams composed by capitalized tokens were also NE 3-grams, and 25.64 % of the 4-grams composed by capitalized tokens were also NE 4-grams. So also in the case of n-grams the use of capitalized tokens increases the number of features compared to the use of only NE.

(ii) If two web pages share capitalized n-grams, the higher is the value of n, the more probable the two web pages refer to the same individual. We define "long enough n-grams" as those compose by at least 3 capitalized tokens.

Thus, a web page W is initially represented as the sequence of tokens starting in uppercase, in the order as they appear in the web page. Notice that some web pages could not be represented with this proposal because all their content was written in lowercase. In the case of the collections that we describe in Section 4.1, 0.66 % of the web pages are not represented for this reason.

3.2 Weighting Functions

We test the well known TF and TF-IDF functions, and z-score (Andrade and Medina, 1998). The z-score of an n-gram a in a web page W_i is defined as follows:

$$z\text{-score}(a, W_i) = \frac{TF(a,W_i)-\mu}{\sigma}$$

where $TF(a, W_i)$ is the frequency of the n-gram a in W_i; μ is the mean frequency of the n-gram a in the background set; and σ is the deviation of the n-gram a in the background set. In this context the background set is the set of web pages that share the person name. This score gives an idea of the distance of the frequency of an n-gram in a web page from the general distribution of this n-gram in the background set.

3.3 Similarity Functions

To determine the similarity between two web pages we try the cosine distance, a widely measure used in clustering, and the weighted Jaccard coefficient between two bags of n-grams defined as:

$$W.Jaccard(W_i^n, W_j^n) = \frac{\sum_k min(m(t_{k_i}^n, i), m(t_{k_j}^n, j))}{\sum_k max(m(t_{k_i}^n, i), m(t_{k_j}^n, j))}$$

where the meaning of $m(t_{k_i}^n, i)$ is explained in Section 3.4. Since weighted Jaccard coefficient

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

needs non-negative entries and we want the cosine similarity of two documents to range from 0 to 1, we translate the values of the z-score so that they are always non-negative.

3.4 Algorithm

The algorithm $UPND$ (Unsupervised Person Name Disambiguator) can be seen in Algorithm 1.

$UPND$ algorithm receives as input a set of web documents with a mention to the same person name, let be $\mathcal{W} = \{W_1, W_2, \ldots, W_N\}$, and starts assigning a cluster C_i for each document W_i. $UPND$ also receives as input a pair of positive integer values r_1 and r_2, such that $r_2 \geq r_1$, specifying the range of values of n in the n-grams extracted from each web document. In each step of the algorithm we assign to each web page W_i a bag of n-grams $W_i^n = \{(t_1^n, m(t_1^n, i)), (t_2^n, m(t_2^n, i)), \ldots, (t_{k_i}^n, m(t_{k_i}^n, i))\}$, where each t_r^n is a n-gram extracted from W_i and $m(t_r^n, i)$ is the corresponding weight of the n-gram t_r^n in the web page W_i, being $r \in \{1, 2, \ldots, k_i\}$. In Algorithm 1 the function $setNGrams(n, \mathcal{W})$ in line 6 calculates for each web page in the set $\mathcal{W}$ its bag of n-grams representation. $Sim(W_i^n, W_j^n)$ in line 9 refers to the similarity between web pages W_i and W_j.

To decide when two web pages refer the same individual we propose a threshold γ. This threshold takes into account two factors: the number of n-grams shared by the web pages and the size of both web pages. For each pair of web pages represented as bag of n-grams, let be W_i^n and W_j^n, we define the following thresholds:

$$\gamma_{max}(W_i^n, W_j^n) = \frac{min(k_i, k_j) - shared(W_i^n, W_j^n)}{max(k_i, k_j)}$$

$$\gamma_{min}(W_i^n, W_j^n) = \frac{min(k_i, k_j) - shared(W_i^n, W_j^n)}{min(k_i, k_j)}$$

where k_i and k_j are the number of n-grams of W_i and W_j respectively, and $shared(W_i^n, W_j^n)$ is the number of n-grams shared by those web pages, i.e. $shared(W_i^n, W_j^n) = |W_i^n \cap W_j^n|$. Notice that $shared(W_i^n, W_j^n)$ is superiorly limited by $min(k_i, k_j)$.

These thresholds hold two desirable properties: (i) The more n-grams are shared by W_i and W_j, the lower the threshold is, so the clustering condition of the algorithm is less strict. (ii) It avoids the penalization due to big differences between the size of the web pages.

γ_{min} benefits the grouping of those web pages that are subsets of other bigger web pages. However, this can lead to a mistake when a small web page is similar to part of other bigger one, but that belongs to different persons. Then, we try to balance this effect by including also γ_{max}. The final threshold is the arithmetic mean of the previous functions:

$$\gamma_{avg}(W_i^n, W_j^n) = \frac{\gamma_{max}(W_i^n, W_j^n) + \gamma_{min}(W_i^n, W_j^n)}{2}$$

what avoids giving advantage to web pages according to their size. We tested these three threshold and γ_{avg} shows a behavior more independent of the size of the n-grams, the similarity functions and the weighting functions.

Thus, two web pages W_i and W_j refer to the same person if $Sim(W_i^n, W_j^n) \geq \gamma_{avg}(W_i^n, W_j^n)$, so $C_i = C_i \cup C_j$ (lines 9, 10 and 11).

The algorithm has three input parameters: $\mathcal{W}$, the set of web pages with the same person name, and r_1 and r_2 that allows the algorithm to iterate this process for r_1-grams to r_2-grams.

This algorithm is polynomial and has a computational cost in $\mathcal{O}(N^2)$, where N is the number of web pages.

Algorithm 1 $UPND(\mathcal{W}, r_1, r_2)$

Require: Set of web pages that shared a person name $\mathcal{W} = \{W_1, W_2, ..., W_N\}$, $r_1, r_2 \geq 1$ such that $r_2 \geq r_1$
Ensure: Set of clusters $\mathcal{C} = \{C_1, C_2, ..., C_l\}$

```
1:  for n = 1 to N do
2:      C_i = {W_i}
3:  end for
4:  C = {C_1, C_2, ..., C_N}
5:  for n = r_1 to r_2 do
6:      setNGrams(n, W)
7:      for i = 1 to N do
8:          for j = i + 1 to N do
9:              if Sim(W_i^n, W_j^n) ≥ γ_avg(W_i^n, W_j^n) then
10:                 C_i = C_i ∪ C_j
11:                 C = C \ {C_j}
12:             end if
13:         end for
14:     end for
15: end for
16: return C
```

4 Experiments

In this section we present the corpora of web pages, the experiments carried out and the results.

4.1 Web People Search Collections

WePS is a competitive evaluation campaign that proposes several tasks including resolution of disambiguation on the Web data. In particular, WePS-1, WePS-2 and WePS-3 campaigns provide an evaluation framework consisting in several annotated data sets composed of English person names.

In these experiments we use WePS-1 (Artiles, Gonzalo, and Sekine, 2007) test corpus composed by 30 English person names and the top 100 search results from Yahoo! search engine; WePS-2 (Artiles, Gonzalo, and Sekine, 2009b) containing 30 person names and the top 150 search results from Yahoo! search engine; and WePS-3 (Artiles et al., 2010) with 300 person names and the top 200 search results from Yahoo!

4.2 Results and Discussion

We present our results for all the corpora comparing them with the state of the art systems. The figures in the tables are macro-averaged, i.e., they are calculated for each person name and then averaged over all test cases. The metrics used in this section are the BCubed metrics defined in (Bagga and Baldwin, 1998): BCubed precision (BP), BCubed recall (BR) and their harmonic mean $F_{0,5}(BP/BR)$. (Artiles, 2009) showed that these metrics are accurate for clustering tasks, particularly for person name disambiguation in the Web.

We use the Wilcoxon test (Wilcoxon, 1945) to detect statistical significance in the differences of the results considering a confidence level of 95 %. In order to compare our algorithm with the WePS better results using the Wilcoxon test, the samples consist in the pairs of values $F_{\alpha=0,5}(BP/BR)$ of each system for each person name.

In order to evaluate our representation approach we first run our algorithm representing the web pages with the n-grams considering all the tokens. Table 1 shows the results of $UPND$ algorithm representing the web pages with 4-grams ($UPND(\mathcal{W}, 4, 4)$) and 3-grams ($UPND(\mathcal{W}, 3, 3)$). Previous experiments using bigrams showed that they are less suitable for this approach. For the representation of $\mathcal{W}$ we discard those n-grams that only appear in one document. The figures shows that, in general, the results obtained with 4-grams outperform those with 3-grams. Weighted Jaccard similarity seems to be more independent of the weighting functions than Cosine. On the other hand, most of the times Cosine gets its best scores when it is applied with z-score. Notice that Jaccard obtains an improvement of the Recall results, whereas Cosine gets better Precision results. The significance test comparing the best scores for Jaccard and Cosine (TF with Jaccard, z-score with Cosine) shows that there are not significant differences. In this case the representation with all 4-grams obtains high Precision scores, whereas the representation with 3-grams increase Recall but with too low Precision scores.

Then we carried out the same experiments but representing the web pages with capitalized n-grams. Table 2 shows these results. In this case, the figures shows that, in general and contrary to the previous experiments, it is not obvious which size of n works the best. The significance test comparing the best scores for each size of n: 4-grams with z-score and Jaccard, and 3-grams with z-score and Cosine shows that there are not significant differences. Thus, given than the representation with 3-grams is less expensive that the one with 4-grams we selected the former. Focussing on 3-grams, the significance test comparing the best scores for Jaccard and Cosine (TF with Jaccard, z-score with Cosine) shows that only with the WePS-3 data set there is a significant difference in favor of z-score+Cosine.

Since we consider that in this task is more relevant Precision than Recall, as we want to have groups of mostly true positives (web pages of the same individual), we select the combination of z-score as weighting function and cosine as similarity function as the most suitable combination for our algorithm. Therefore we use it in the following experiments.

Finally, comparing the selected representation with all the n-grams (4-grams, z-score, cosine) with the selected one for capitalized n-grams (3-grams, z-score, cosine) the significance test shows that only there is a significance difference with WePS-1 data set in favor of the representation with all the n-grams. Thus, we consider that the representation only with capitalized n-grams is competitive, since it obtains comparable results to those obtained with all the n-grams, with the advantage of being more efficient both in space and time.

Table 3 shows the results of $UPND$ with WePS-1 test, WePS-2 and WePS-3 corpora in addition to the top ranking systems of the campaigns, and also the results obtained by

		WePS-1			WePS-2			WePS-3		
		BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)
4-grams										
W. Jaccard	TF	0.86	0.75	**0.79**	0.90	0.72	**0.79**	0.62	0.57	**0.54**
	z-score	0.85	0.75	**0.79**	0.9	0.73	**0.79**	0.61	0.58	**0.54**
	TF-IDF	0.86	0.75	**0.79**	0.90	0.72	**0.79**	0.62	0.57	**0.54**
Cosine	TF	0.90	0.70	0.78	0.95	0.63	0.74	0.70	0.47	0.52
	z-score	0.89	0.71	0.78	0.95	0.67	0.77	0.69	0.50	0.53
	TF-IDF	0.90	0.69	0.77	0.95	0.57	0.7	0.72	0.44	0.51
3-grams										
W. Jaccard	TF	0.58	0.87	0.68	0.68	0.89	0.76	0.36	0.81	0.45
	z-score	0.57	0.88	0.67	0.68	0.89	0.75	0.35	0.82	0.45
	TF-IDF	0.58	0.87	0.68	0.68	0.89	0.76	0.36	0.81	0.45
Cosine	TF	0.69	0.8	**0.73**	0.78	0.81	0.78	0.46	0.66	**0.49**
	z-score	0.66	0.83	0.72	0.78	0.84	**0.8**	0.44	0.71	**0.49**
	TF-IDF	0.7	0.79	**0.73**	0.78	0.76	0.75	0.48	0.63	**0.49**

Table 1: Results of $UPND$ algorithm for WePS test data sets using all the n-grams.

		WePS-1			WePS-2			WePS-3		
		BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)
4-grams										
W. Jaccard	TF	0.89	0.67	**0.76**	0.95	0.69	**0.79**	0.68	0.51	0.53
	z-score	0.89	0.67	**0.76**	0.93	0.69	**0.79**	0.67	0.52	**0.54**
	TF-IDF	0.89	0.67	**0.76**	0.95	0.69	**0.79**	0.68	0.51	0.53
Cosine	TF	0.93	0.63	0.75	0.96	0.60	0.72	0.74	0.44	0.51
	z-score	0.92	0.65	**0.76**	0.96	0.63	0.75	0.73	0.46	0.52
	TF-IDF	0.93	0.63	0.74	0.96	0.59	0.71	0.74	0.44	0.51
3-grams										
W. Jaccard	TF	0.72	0.78	0.73	0.81	0.83	**0.81**	0.46	0.70	0.50
	z-score	0.70	0.79	0.73	0.8	0.84	**0.81**	0.45	0.72	0.50
	TF-IDF	0.72	0.78	0.73	0.81	0.83	**0.81**	0.46	0.70	0.50
Cosine	TF	0.78	0.73	0.74	0.85	0.76	0.79	0.56	0.59	**0.52**
	z-score	0.76	0.76	**0.75**	0.85	0.79	**0.81**	0.54	0.62	**0.52**
	TF-IDF	0.78	0.75	**0.75**	0.86	0.75	0.79	0.57	0.57	0.52

Table 2: Results of $UPND$ algorithm for WePS test data sets using capitalized n-grams.

HAC_Topic system in the case of WePS-1. We include the results obtained by three unsupervised baselines called ALL_IN_ONE, ONE_IN_ONE and Fast AP. ALL_IN_ONE provides a clustering solution where all the documents are assigned to a single cluster, ONE_IN_ONE returns a clustering solution where every document is assigned to a different cluster, and Fast AP applies a fast version of Affinity Propagation described in (Fujiwara, Irie, and Kitahara, 2011) using the function TF-IDF to weight the tokens of each web page, and the cosine distance to compute the similarity.

Our algorithm $UPND$ outperforms WePS-1 participants and all the unsupervised baselines described before. HAC_Topic also outperforms the WePS-1 top participant systems and our algorithm. This system uses several parameters obtained by training with the WePS-2 data set: token weight according to the kind of token (terms from URL, title, snippets, ...) and thresholds used in the clustering process. Note that WePS-1 participants used the training corpus provided to the campaign, the WePS-1 training data, so in this case the best performance of HAC_Topic could be not only due to the different approach, but also because of the different training data set.

$UPND$ obtains significative better results than the WePS-1 top participant results, and HAC_Topic obtains significative better results than it according to the Wilcoxon test. $UPND$ obtains significative better results than IRST-BP system (the third in the WePS-1 ranking), also based on the co-ocurrence of n-grams.

Regarding WePS-2 we add in Table 3 two oracle systems provided by the organizers. The oracle systems use BoW of tokens (ORACLE_1) or bigrams (ORACLE_2) weighted by TF-IDF, deleting previously stop words, and later applying HAC with single linkage with the best thresholds for each person name. We do not include the results of the HAC_Topic system since it uses this

	System	BP	BR	$F_{0.5}$(BP/BR)
WePS-1	(+) HAC_Topic	0.79	0.85	**0.81** †
	(-) ***UPND (all-4g)***	0.89	0.71	0.78 •
	(-) ***UPND (cap-3g)***	0.76	0.76	0.75 •
	(+)(*) CU_COMSEM	0.61	0.83	0.70 †
	(+)(*) PSNUS	0.68	0.73	0.70 †
	(+)(*) IRST-BP	0.68	0.71	0.69 †
	(+)(*) UVA	0.79	0.50	0.61 †
	(+)(*) SHEF	0.54	0.74	0.62 †
	(-) ONE_IN_ONE	**1.00**	0.43	0.57 •
	(-) Fast AP	0.69	0.55	0.56 †
	(-) ALL_IN_ONE	0.18	**0.98**	0.25 •
WePS-2	(+) ORACLE_1	0.89	0.83	**0.85** •
	(+) ORACLE_2	0.91	0.81	**0.85** •
	(+)(*) PolyUHK	0.87	0.79	0.82
	(+)(*) ITC-UT_1	0.93	0.73	0.81
	(-) ***UPND (cap-3g)***	0.85	0.79	0.81 •
	(+)(*) UVA_1	0.85	0.80	0.81
	(-) ***UPND (all-4g)***	0.95	0.67	0.77 •
	(+)(*) XMEDIA_3	0.82	0.66	0.72 †
	(+)(*) UCI_2	0.66	0.84	0.71 †
	(-) ALL_IN_ONE	0.43	**1.00**	0.53 •
	(-) Fast AP	0.80	0.33	0.41 †
	(-) ONE_IN_ONE	**1.00**	0.24	0.34 •
WePS-3	(+)(*) YHBJ_2	0.61	0.60	**0.55**
	(-) ***UPND (cap-3g)***	0.54	0.62	0.52 •
	(+)(*) AXIS_2	0.69	0.46	0.50 †
	(-) ***UPND (all-4g)***	0.44	0.71	0.49 •
	(+)(*) TALP_5	0.40	0.66	0.44 †
	(+)(*) RGAI_AE_1	0.38	0.61	0.40 †
	(+)(*) WOLVES_1	0.31	0.80	0.40 †
	(-)(*) DAEDALUS_3	0.29	0.84	0.39 †
	(-) Fast AP	0.73	0.30	0.38 †
	(-) ONE_IN_ONE	**1.00**	0.23	0.35 •
	(-) ALL_IN_ONE	0.22	**1.00**	0.32 •

Table 3: Result of $UPND$ and the top state of the art systems with WePS corpora: (+) means system with supervision; (-) without supervision and (*) campaign participant. Significant differences between $UPND$ and other systems are denoted by (†); (•) means that in this case the statistical significance is not evaluated.

data set for training their algorithm.

The significance test shows that the top WePS-2 systems PolyUHK, UVA_1 and ITC-UT_1 obtain similar results than $UPND(cap-3g)$, however they use some kind of supervision. The results of all these systems are the closest to the oracle systems, which know the best thresholds for each person name.

In the case of WePS-3, the organizers did not consider for evaluation the whole clustering solution provided by the systems like in previous editions, but only checks the accuracy of the clusters corresponding to two selected individuals per person name. In this case, the first two systems YHBJ_2 and $UPND(cap-3g)$ do not have significant differences in their results. Notice that YHBJ_2 system makes use of concepts extracted manually from Wikipedia. $UPND$ also obtains significative better results than DAEDALUS_3, the only one participant that does not use training data.

(Artiles, Amigó, and Gonzalo, 2009a) applied HAC algoritm over n-grams of length 2 to 5 getting similar results of precision than $UPND$ but very low recall scores. This means that applying HAC only over n-grams is not a good choice and $UPND$ takes more advantage of these features.

After all these experiments, we can conclude that our approach gets the best results of all the completely unsupervised approaches. Moreover, the precision scores for all collections are very high and confirm that our approach is accurate to get relevant information for characterizing an individual. We also obtain competitive recall results, what lead to a competitive system that carries out person name disambiguation in web search results without any kind of supervision.

5 *Conclusions and Future Work*

We present a new approach for person name disambiguation of web search results. Our method does not need training data to calculate thresholds to determine the number of different individuals sharing the same name, or whether two web pages refer to the same individual or not. Although supervised approaches have been successful in many NLP and IR tasks, they require enough and representative training data to guaranty consistent results for different data collections, which requires a huge human effort.

The proposed algorithm provides a clustering solution for this task by means of data-driven methods that do not need learning from training data. Our approach obtains very competitive results in all the data sets compared with the best state of the art systems. It is based on getting reliable information for disambiguating, particularly long n-grams composed by uppercase tokens. According to our results, this hypothesis has shown successful, getting high precision values and acceptable recall scores. Anyway, we would like to improve recall results without losing of precision, filter out noisy capitalized n-grams, and build an alternative representation for web pages containing all their tokens in lowercase.

Person name disambiguation has been mainly addressed in a monolingual scenario, e.g. WePS corpora are English data sets. We would like to address this task in a multilingual scenario. Although search engines return their results taking into account the country of the user, with some queries we can get results written in several languages. This scenario has not been considered by the state of the art systems so far.

References

Andrade, M.A. and A. Valencia. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14:600-607.

Artiles, J. 2009. Web People Search. PhD Thesis, UNED University.

Artiles, J., J. Gonzalo, and S. Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64-69. ACL.

Artiles, J., E. Amigó, and J. Gonzalo. 2009a. The Role of Named Entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Artiles, J., J. Gonzalo, and S. Sekine. 2009b. Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Artiles, J., A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Bagga, A. and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pages 79-85. ACL.

Balog, K., J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. de Rijke. 2009. The University of Amsterdam at WePS-2. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Chen, Y. and J. Martin. 2007. CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Named Disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 125-128. ACL.

Chen, Y., S. Yat Mei Lee, and C. Huang. 2009. PolyUHK: A Robust Information Extraction System for Web Personal Names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Elmacioglu, E., Y. Fan Tan, S. Yan, M. Kan, and D. Lee. 2007. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 268-271. ACL.

Fujiwara, Y., G. Irie, and T. Kitahara. 2011. Fast Algorithm for Affinity Propagation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence(IJCAI)- Volume Three*, pages 2238-2243.

Lana-Serrano, S., J. Villena-Román, and J.C. González-Cristóbal. 2010. Daedalus at WebPS-3 2010: k-Medoids Clustering using a Cost Function Minimization. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Liu, Z., Q. Lu, and J. Xu. 2011. High Performance Clustering for Web Person Name Disambiguation using Topic Capturing. In *International Workshop on Entity-Oriented Search (EOS)*.

Long, C. and L. Shi. 2010. Web Person Name Disambiguation by Relevance Weighting of Extended Feature Sets. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Mann, G.S. 2006. Multi-Document Statistical Fact Extraction and Fusion. PhD thesis, Johns Hopkins University, Baltimore, MD, USA. AAI3213760.

Manning, C.D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, USA.

Popescu, O. and B. Magnini. 2007. IRST-BP: Web People Search Using Name Entities In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 195-198. ACL.

Wilcoxon, F. 1945. *Individual Comparisons by Ranking Methods*, 1(6). Biometrics Bulletin.