Overview of the M-WePNaD Task: Multilingual Web Person Name Disambiguation at IberEval 2017

Soto Montalvo[1], Raquel Martínez[2], Víctor Fresno[2], Agustín D. Delgado[2], Arkaitz Zubiaga[3], Richard Berendsen[4]

[1] URJC, soto.montalvo@urjc.es
[2] NLP&IR Group, UNED, {raquel,vfresno,agustin.delgado}@lsi.uned.es
[3] University of Warwick, a.zubiaga@warwick.ac.uk
[4] Luminis Amsterdam, richard.berendsen@luminis.eu

Abstract. Multilingual Web Person Name Disambiguation is a new shared task proposed for the first time at the IberEval 2017 evaluation campaign. For a set of web search results associated with a person name, the task deals with the grouping of the results based on the particular individual they refer to. Different from previous works dealing with monolingual search results, this task has further considered the challenge posed by search results written in different languages. This task allows to evaluate the performance of participating systems in a multilingual scenario. This overview summarizes a total of 18 runs received from four participating teams. We present the datasets utilized and the methodology defined for the task and the evaluation, along with an analysis of the results and the submitted systems.

Keywords: person name disambiguation on the web, document clustering, multilingualism, web search

1 Introduction

It is increasingly usual for people to turn to Internet search engines to look for information about people. According to Google Trends, three out of the top 10 Google Searches in 2016 were linked to person names[5]. However, person names tend to be ambiguous and hence a search for a particular name likely includes results for different individuals. In these cases, a list of individuals included in the results along with a breakdown of different individuals would come in handy for the user who is looking for a particular individual. This task was first introduced in the WePS (Web People Search) campaigns[6], and attracted substantial interest in the scientific community, as manifested in a number of shared tasks that tackled it, particularly the WePS-1, WePS-2, WePS-3 campaigns [1–3]. These campaigns provided several annotated corpora becoming a referent for

[5] https://trends.google.com/trends/topcharts#geo&date=2016

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

this problem and allowing a comparative study of the performance of different systems. However, all those campaigns presented a monolingual scenario where the query results were written in only one language.

Despite the multilingual nature of the Web[7], existing work on person name disambiguation has not considered yet search results written in multiple languages. The objective of M-WePNaD task is centered around a multilingual scenario where the results for a query, as well as each individual, can be written in different languages.

The remainder of this paper is organized as follows. Section 2 presents the task. Next, Section 3 describes the datasets we released for training and testing, and we briefly discuss the differences between the two sets. Section 4 briefly describes the evaluation measures. Section 5 summarizes the proposed approaches of the participants. Section 6 presents and discusses the results. Finally, conclusions are presented in Section 7.

2 Task Description

The M-WePNaD task is a person name disambiguation task on the Web focused on distinguishing the different individuals that are contained within the search results for a person name query. The person name disambiguation task can be defined as a clustering problem, where the input is a ranked list of n search results, and the output needs to provide both the number of different individuals identified within those results, as well as the set of pages associated with each of the individuals.

The heterogeneous nature of web results increases the difficulty of this task. For instance, some web pages related to a certain individual could be professional sites (e.g. corporation web pages), while others may contain personal information (e.g. blogs and social profiles) and both kinds of web pages could have very little vocabulary in common. Particularly, [4] concluded that the inclusion of content from social networking platforms increases the difficulty of the task.

While previous evaluation campaigns had been limited to monolingual scenarios, the M-WePNaD task was assessed in a multilingual setting, considering the realistic scenario where a search engine returns results in different languages for a person name query. For instance, web pages with professional information for an individual who is not a native English speaker may be written in English, while other personal web pages may be written in their native language. Celebrities who are known internationally are also likely to have web pages in different languages.

We compiled an evaluation corpus called MC4WePS [5], which was manually annotated by three experts. This corpus was used to evaluate the performance of multilingual disambiguation systems, enabling also evaluation for different document genres as the corpus includes not only web pages but also social media posts. The corpus was split into two parts, one for training and one for testing.

[7] The most used language on the Web is English, followed by Chinese and Spanish.

Participants had nearly two months to develop their systems making use of the training corpus. Afterwards, the test corpus was released, whereupon participants ran their systems and sent the results back to the task organizers. The organizers also provided the performance scores for different baseline approaches. Participants were restricted to the submission of up to five different result sets. In this overview we present the evaluation of these submissions, which we list in two different rankings.

3 Data Sets

The MC4WePS corpus was collected in 2014, issuing numerous search queries and storing those that met the requirements of ambiguity and multilingualism.

Each query includes a first name and last name, with no quotes, and searches were issued in both Google and Yahoo. The criteria to choose the queries took into account:

- Ambiguity: non-ambiguous, ambiguous or highly ambiguous names. A person's name is considered highly ambiguous when it has results for more than 10 individuals. Cases with 2 to 9 individuals were considered ambiguous, while those with a single individual were deemed non-ambiguous.
- Language: results can be monolingual, where all pages are written in the same language, or multilingual, where pages are written in more than one language. Additionally, for each cluster of pages belonging to the same individual, we considered whether the results were monolingual or multilingual. This was due to the fact that even though the results for a person name query are multilingual, the clusters for each different individual could be monolingual or multilingual.

The MC4WePS dataset contains search results of 100 person names with a number of search results between 90 and 110 each. It is worth noting that different person names in the corpus have different degrees of ambiguity; in addition a web page can be multilingual, and not all the content in the corpus are regular HTML web pages, but also other kinds of documents are included, such as social media posts or pdf documents. A detailed description of the corpus can be found in [5].

There can be overlaps between clusters as a search result could refer to two or more different individuals with the same name, for instance social profile pages with lists of different individuals with the same name. When a search result does not belong to any individual or the individual cannot be inferred, then this is annotated as “Not Related” (NR). For each query, these search results are grouped as a single cluster of NR results in the gold standard annotations.

The MC4WePS corpus was randomly divided into two parts: training set (65%) and test set (35%).

3.1 Training Set

We provided participants with a single set for training, which includes 65 different person names, randomly sampled from the entire dataset. The list of names and their characteristics can be seen in Table 1. The second part of the table contains in the last row the average values for the different data of the whole training set.

Table 1: Characteristics of the M-WePNaD training set. #Webs represents the number of search results of each person name; #Individuals the number of clusters related to some individual (i.e. number of clusters, not counting the Not Related (NR) one); %Social refers the percentage of social web pages (identified by their URL-domain); Top Language the most frequent language of the web pages according to the annotators–ES, EN, and FR mean Spanish, English and French, respectively–; %WebsDL the percentage of web pages written in different language than the most frequent one; and %NRs the percentage of web pages annotated as NR.

Person Name	#Webs	#Individuals	%Social	Top Language	%WebsDL	%NRs
adam rosales	110	8	9.09	EN	0.91	10.0
albert claude	106	9	10.38	EN	13.21	24.53
álex rovira	95	20	23.16	EN	43.16	6.32
alfred nowak	109	15	3.67	EN	30.28	66.06
almudena sierra	100	22	12.0	ES	1.0	63.0
amber rodríguez	106	73	11.32	EN	9.43	10.38
andrea alonso	105	49	9.52	ES	6.67	20.95
antonio camacho	109	39	24.77	EN	29.36	46.79
brian fuentes	100	12	7.0	EN	2.0	3.0
chris andersen	100	6	5.0	EN	26.0	2.0
cicely saunders	110	2	7.27	EN	1.82	10.91
claudio reyna	107	5	7.48	EN	4.67	2.8
david cutler	98	37	15.31	EN	0.0	19.39
elena ochoa	110	15	8.18	ES	10.0	4.55
emily dickinson	107	1	3.74	EN	0.0	0.93
francisco bernis	100	4	4.0	EN	50.0	29.0
franco modigliani	109	2	2.75	EN	38.53	1.83
frederick sanger	100	2	0.0	EN	0.0	5.0
gaspar zarrías	110	3	4.55	ES	2.73	0.0
george bush	108	4	2.78	EN	25.0	13.89
gorka larrumbide	109	3	4.59	ES	9.17	32.11
henri michaux	98	1	3.06	EN	7.14	1.02
james martin	100	48	5.0	EN	2.0	14.0
javi nieves	106	3	4.72	ES	3.77	1.89
jesse garcía	109	26	6.42	EN	31.19	16.51

Table 1: (continued)

Person Name	**#Webs**	**#Individuals**	**%Social**	**Top Language**	**%WebsDL**	**%NRs**
john harrison	109	50	15.6	EN	11.01	19.27
john orozco	100	9	11.0	EN	4.0	20.0
john smith	101	52	10.89	EN	0.0	10.89
joseph murray	105	47	7.62	EN	0.95	20.0
julián lópez	109	28	4.59	ES	6.42	1.83
julio iglesias	109	2	2.75	ES	14.68	0.92
katia gerreiro	110	8	10.91	EN	26.36	0.0
ken olsen	100	41	5.0	EN	0.0	6.0
lauren tamayo	101	8	11.88	EN	3.96	10.89
leonor garcía	100	53	9.0	ES	3.0	12.0
manuel alvar	109	4	3.67	ES	0.92	34.86
manuel campo	103	7	3.88	ES	0.0	2.91
maría dueñas	100	5	6.0	ES	14.0	0.0
mary lasker	103	3	1.94	EN	0.0	15.53
matt biondi	106	12	10.38	EN	9.43	5.66
michael bloomberg	110	2	6.36	EN	0.0	1.82
michael collins	108	31	1.85	EN	0.0	13.89
michael hammond	100	79	20.0	EN	1.0	11.0
michael portillo	105	2	4.76	EN	7.62	0.95
michel bernard	100	5	0.0	FR	39.0	95.0
michelle bachelet	107	2	8.41	EN	16.82	4.67
miguel cabrera	108	3	5.56	EN	0.93	3.7
miriam gonzález	110	43	11.82	ES	29.09	5.45
olegario martínez	100	38	12.0	ES	15.0	10.0
oswald avery	110	2	7.27	EN	9.09	3.64
palmira hernández	105	37	8.57	ES	20.95	60.95
paul erhlich	99	9	4.04	EN	16.16	7.07
paul zamecnik	102	6	1.96	EN	2.94	6.86
pedro duque	110	5	4.55	ES	4.55	12.73
pierre dumont	99	39	10.1	EN	41.41	15.15
rafael matesanz	110	6	7.27	EN	44.55	2.73
randy miller	99	52	12.12	EN	0.0	33.33
raúl gonzález	107	32	4.67	ES	10.28	1.87
richard rogers	100	40	13.0	EN	9.0	16.0
richard vaughan	108	5	4.63	ES	7.41	5.56
rita levi	104	2	1.92	ES	47.12	1.92
robin lópez	102	10	12.75	EN	1.96	13.73
roger becker	103	29	4.85	EN	13.59	18.45
virginia díaz	106	40	11.32	ES	17.92	16.04
william miller	107	40	7.48	EN	0.0	37.38
AVERAGE	104.69	19.95	7.66	14.88	-	12.19

3.2 Test Set

The test corpus consists of 35 different person names, whose characteristics can be seen in Table 2. The last row shows the average values for the different data of the whole test set.

3.3 Comparing Training and Test Sets

As can be seen in Table 1 and Table 2, the training and test sets have a comparable average composition with regard to the percentages of NR search results and social web pages. The two sets are also similar in terms of the distribution of the most common language for a given person name. In the test set, the percentages are 28.57% ES, 68.57% EN, and 2.86% FR; whereas in the training set they are 30.76% ES, 67.69% EN, and 1.55% FR.

The main difference between both sets lies in the degree of ambiguity of the person names. Based on the threshold defined by Montalvo et al. [5] that determines search results pertaining to more than 10 individuals are very ambiguous, the training set contains 54% ambiguous person names and 46% very ambiguous names; on the other hand, the test set contains 40% ambiguous person names and 60% very ambiguous names. This means that the test set is less balanced when it comes to the ambiguity of the names than the training set; the test set contains more very ambiguous names.

3.4 Format and Distribution

The datasets are structured in directories. Each directory corresponds to a specific search query that matches the pattern "name-lastname", and includes the search results associated with that person name. Each search result is in turn stored in a separate directory, named after the rank of that particular result in the entire list of search results. A directory with a search result contains the following files:

– The web page linked by the search result. Note that not all search results point to HTML web pages, but there are also other document formats: pdf, doc, etc.
– A metadata.xml file with the following information:
  - URL of search result.
  - ISO 639-1 codes for languages the web page is written in. It contains a comma-separated list of languages where several were found.
  - Download date.
  - Name of annotator.
– A file with the plain text of the search results, which was extracted using Apache TiKa (https://tika.apache.org/).

Figure 1 shows an example of the metadata file for a search result for the person name query *Julio Iglesias.*

The access to training and test sets was restricted to registered participants. The blind version of the test dataset did not include the ground truth files.[8]

[8] The MC4WEPS corpus is available at `http://nlp.uned.es/web-nlp/resources`.

Table 2. Characteristics of the M-WePNaD test set. #Webs represents the number of search results of each person name; #Individuals the number of clusters related to some individual (i.e. number of clusters, not counting the Not Related (NR) one); %Social refers the percentage of social web pages (identified by their URL-domain); Top Language the most frequent language of the web pages according to the annotators– ES, EN, and FR mean Spanish, English and French, respectively–; %WebsDL the percentage of web pages written in different language than the most frequent one; and %NRs the percentage of web pages annotated as NR.

Person Name	**#Webs**	**#Individuals**	**%Social**	**Top Language**	**%WebsDL**	**%NRs**
agustin gonzalez	99	45	8.08	ES	6.06	6.06
albert barille	99	1	2.02	EN	47.47	11.11
albert gomez	105	50	11.43	EN	11.43	28.57
alberto angulo	108	49	7.41	ES	7.41	6.48
alberto granado	107	2	3.74	ES	17.76	0.93
aldo donelli	110	4	8.18	EN	30.0	14.55
almudena ariza	110	6	12.73	ES	10.0	4.55
alvaro vargas	100	50	24.0	EN	34.0	8.0
amanda navarro	102	50	7.84	EN	41.18	28.43
david robles	100	58	8.0	ES	35.0	7.0
didier dupont	109	34	26.61	FR	50.46	45.87
edward heath	103	8	1.94	EN	8.74	12.62
hendrick janssen	104	19	7.69	EN	55.77	74.04
jacques cousteau	109	2	5.5	EN	4.59	0.92
john williams	102	44	18.63	EN	12.75	16.67
jorge fernandez	107	28	4.67	ES	0.0	5.61
jose ortega	108	40	11.11	EN	7.41	19.44
joseph lister	109	12	8.26	EN	5.5	5.5
liliana jimenez	90	31	23.33	ES	48.89	61.11
marina castano	100	5	5.0	ES	0.0	2.0
mario gomez	100	18	3.0	ES	42.0	1.0
mark davies	105	60	17.14	EN	0.0	19.05
mary leakey	110	2	5.45	EN	0.0	3.64
michael hastings	100	19	5.0	EN	0.0	7.0
michelle martinez	105	49	13.33	EN	22.86	7.62
norah jones	101	1	6.93	EN	1.98	0.99
peter kirkpatrick	106	35	7.55	EN	6.6	25.47
peter mitchell	110	60	30.0	EN	0.0	21.82
rafael morales	100	47	8.0	ES	12.0	18.0
richard branson	100	3	6.0	EN	0.0	2.0
rick warren	99	5	9.09	EN	0.0	0.0
ryan gosling	103	2	6.8	EN	0.0	0.97
thomas klett	98	33	8.16	EN	29.59	42.86
tim duncan	103	3	3.88	EN	26.21	2.91
william osler	106	5	3.77	EN	4.72	23.58
AVERAGE	103.63	25.14	9.72	-	16.58	15.32

```
<?xml version="1.0" encoding="UTF-8"?>
<tns:Annotation_Corpus
       xmlns:tns="http://www.example.org/metadata-corpus"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.example.org/metadata-corpus
             metadata-corpus.xsd">
   <tns:url>http://es.wikipedia.org/wiki/Julio_Iglesias</tns:url>
   <tns:language>ES</tns:language>
   <tns:downloadDate>2013-06-08</tns:downloadDate>
   <tns:annotator>CF</tns:annotator>
</tns:Annotation_Corpus>
```

Fig. 1. Example of metadata for a web page search result for the query *Julio Iglesias*.

4 Evaluation Measures

We use three metrics for the evaluation: Reliability (R), Sensitivity (S) and their harmonic mean $F_{0.5}$ [6]. These metrics generalize the B-Cubed metrics [7] when there are overlapping clusters, as it is the case with the MC4WePS corpus. In particular, Reliability extends B-Cubed Precision and Sensitivity extends B-Cubed Recall.

5 Overview of the Submitted Approaches

Thirteen teams signed up for the M-WePNaD task, although only four of them managed to participate in the task on time, submitting a total of 18 runs.

In what follows, we analyze their approaches from two perspectives: search result representation (including whether or not translation resources were used), and the clustering algorithms.

- The ATMC_UNED team [8] presented four runs that have in common the use of clustering algorithms able to estimate the number of clusters with no need of information from training data. Three of the four runs use the ATC algorithm [9], an algorithm that works in two phases: a phase of cluster creation followed by a phase of cluster fusion. Run 4 uses the ATCM algorithm [10], which identifies those features written the same way in several languages (called *comparable features*) and gives them a special role when comparing search results written in different languages without the need of translation resources. The author explores four different representation approaches: the textual features of the document with no translation (Run 1), a translated version of the document (Run 2), a halfway approach that uses the original document's textual features in the phase of cluster creation and uses a translation tool to translate the centroid features in the fusion cluster phase (Run 3), and an approach that combines the original document's textual features in addition to a representation based on using only the *comparable features* of web pages written in different language. On the other hand, none of the four approaches identifies and groups the not related search results, so all of them get worse results when considering all the web pages. Regarding the treatment of overlapping clusters, none of the four approaches deals with them, so a web page can only be in one cluster. Finally, the author applies an heuristic described in [11] to treat in a special way the web pages from social media platforms and people search engines.

- The LSI_UNED team's approach [12] is mainly based on the application of word embeddings to represent the documents retrieved by the search engine as vectors. Then, these vectors are used in a clustering algorithm (developed by the authors and adapted to the characteristics of the corpus), where a similarity threshold γ determines when the grouping is carried out. To obtain the word embeddings, first they performed a removal of stopwords and extracted the named entities, using pre-trained word embeddings for representation. The tools they used were the Stanford Named Entity Recognizer[9] and ConVec[10], a publicly available collection of word vectors generated from Wikipedia concepts. To obtain the document representation, the authors calculated the average vector of all the vectors corresponding to the words within the document. They calculated the similarity between each document and the rest of the documents related to the same person name by means of cosine similarity. The similarity weight associated to each document was the average of the similarity between that document, and the rest of documents related to the same person name. Finally, the authors considered that all the documents with a similarity weight above a specific γ threshold should be gathered in the same initial cluster. This team initially submitted four runs corresponding to different values of γ ($\gamma_1 = 0 : 70$, $\gamma_2 = 0 : 75$, $\gamma_3 = 0 : 80$, and $\gamma_4 = 0 : 85$). Finally, a fifth run was also evaluated using a different configuration of the system, in which all the words within the documents (except stop words) were considered in order to represent them, and not only named entities, as in the previous runs. None of their submitted runs deals with the multilingual nature of the task nor the overlap between clusters.
- The Loz_Team [13] submitted five runs that experimented with different settings a hierarchical agglomerative clustering (HAC) algorithm using the Euclidean distance as a similarity measure. They tested three different ways of representing the content: (1) a binary representation capturing presence or not of each word, (2) a weighted representation capturing the number of occurrences of each word, and (3) a TF-IDF metric. They also tested two different stoppage criteria, namely $k = 5$ and $k = 15$. With these different settings, the authors tested the following five combinations: (1) weighted representation + $k = 5$, (2) binary representation + $k = 15$, (3) TF-IDF + $k = 15$, (4) binary representation + $k = 5$, and (5) TF-IDF + $k = 5$. As in the previous team, none of the five runs developed by this team tackled the challenges posed by the multilingual nature of the dataset or the overlap between clusters.

[9] https://nlp.stanford.edu/software/CRF-NER.shtml

[10] https://github.com/ehsansherkat/ConVec

- The PanMorCresp_Team [14] submitted four runs based on the HAC algorithm. For all runs, the files with the plain text version of the search result are used. For each query, vector representations of the text are generated independently. The text is split into tokens on blank characters. The tokens are lowercased. Some runs use additional token normalization. Next, words that occur only once in the collection are removed. After creating the vocabulary, binary document vectors are created, indicating the presence or absence of words in a document. The cosine similarity is used to compute similarities between document vectors. None of the runs use any translation or other language-specific decisions. None of the runs try to detect non-related search results. The four runs then investigate the effect of other typical choices one has to make when employing HAC. First, token normalization. Run 3 and Run 4 eliminate punctuation. Second, which words to use in the vocabulary; besides effectiveness, computational efficiency plays a role here. Run 1 and Run 2 include the 4,000 most frequent terms. Run 3 and Run 4 remove stop words and include the 7,500 most frequent remaining terms. Third, how to compute cluster similarities. Run 1 uses complete linkage, Run 2 uses the average similarity between documents in both clusters, and Run 3 and Run 4 use single linkage. Fourth, how to define the stopping criterion. Run 1 makes the stopping criterion depend on the query. It computes the average similarity between documents and divides this by a factor n. On the training set, this parameter was tuned to $n = 2$. Run 2, Run 3, and Run 4 use a global stopping criterion. Run 2 and Run 3 tune a minimal similarity threshold on the training corpus. For Run 3 the resulting threshold was 0.65; for Run 2 it is not given. Run 4 uses an exact number of clusters as a stopping criterion.

6 Results and Discussion

We produced two different rankings of the participants after evaluating all the submissions:

- Evaluation results by not considering the Not Related results. This means that all the results of this kind and the corresponding cluster were not taken into account.
- Evaluation results considering all web results. This means that all the results and the clusters were taken into account.

Table 3 shows the results without considering Not Related results in the evaluation, whereas Table 4 shows the results considering all the pages. Both tables contain two baselines ONE IN ONE and ALL IN ONE. ONE IN ONE returns each search result as a singleton cluster, while ALL IN ONE returns only one cluster that includes all the search results. Note that these baselines are independent of the document representation.

The results obtained by the ATMC_UNED team runs overcome the results obtained by the baselines and the rest of the participants, showing the potential of the ATC and ATMC algorithms over the rest, particularly HAC algorithms.

Table 3. Evaluation results not considering Not Related results.

System and run number	R	S	$F_{0.5}$
ATMC_UNED - run 3	0.80	0.84	0.81
ATMC_UNED - run 4	0.79	0.85	0.81
ATMC_UNED - run 2	0.82	0.79	0.80
ATMC_UNED - run 1	0.79	0.83	0.79
LSI_UNED - run 3	0.59	0.85	0.61
LSI_UNED - run 4	0.74	0.71	0.61
LSI_UNED - run 2	0.52	0.93	0.58
LSI_UNED - run 5	0.52	0.92	0.5
PanMonCresp_Team - run 4	0.53	0.87	0.57
LSI_UNED - run 1	0.49	0.97	0.56
Baseline - ALL-IN-ONE	0.47	0.99	0.54
Loz_Team - run 3	0.57	0.71	0.52
Loz_Team - run 5	0.51	0.83	0.52
Loz_Team - run 2	0.55	0.65	0.50
Loz_Team - run 4	0.50	0.81	0.50
PanMorCresp_Team - run 3	0.53	0.82	0.47
Loz_Team - run 1	0.50	0.76	0.46
PanMorCresp_Team - run 1	0.80	0.51	0.43
Baseline - ONE-IN-ONE	1.0	0.32	0.42
PanMorCresp_Team - run 2	0.50	0.65	0.41

The results obtained by all their runs are quite similar. However, Run 1 uses features from the original content of the web pages and gets worse results with respect to Run 2 and Run 3, which use a machine translation tool. Run 4 compares the web pages written in different language with their *comparable features* and gets similar results than Run 2 and Run 3 without the need of translation resources. The main advantage of this last approach is that it avoids additional preprocessing steps dedicated to translating the web pages, which is desirable in problems which have to be solved in real time. The ATMC_UNED team has not proposed any method to group not related web pages, so their Sensitivity and the F-measure results are worse when considering them in the evaluation.

The results obtained by the LSI_UNED team overcome the results obtained by the baselines but are lower than results obtained by the ATMC_UNED team. Going into detail, using all the words in the documents (Run 5) is under the run that only considers named entities and uses the same threshold (Run 2). This implies that the addition of all the possible words in the documents introduces more noise than valuable information. On the other hand, in general, if using named entities and increasing the threshold value, the Reliability increases while the Sensibility decreases. Finally, considering all web pages can be seen as a more difficult task, but the number of unrelated web pages in the corpus is small

Table 4. Evaluation results considering all the pages.

System and run number	R	S	$F_{0.5}$
ATMC_UNED - run 3	0.79	0.74	0.75
ATMC_UNED - run 4	0.78	0.75	0.75
ATMC_UNED - run 1	0.78	0.73	0.74
ATMC_UNED - run 2	0.82	0.69	0.73
LSI_UNED - run 3	0.59	0.81	0.60
LSI_UNED - run 2	0.52	0.92	0.59
LSI_UNED - run 5	0.52	0.90	0.59
LSI_UNED - run 1	0.49	0.97	0.58
LSI_UNED - run 4	0.74	0.66	0.58
Loz_Team - run 1	0.49	0.73	0.58
PanMorCresp_Team - run 4	0.52	0.86	0.58
Baseline - ALL-IN-ONE	0.47	1.0	0.56
Loz_Team - run 5	0.50	0.80	0.54
Loz_Team - run 3	0.56	0.66	0.53
Loz_Team - run 4	0.49	0.78	0.52
Loz_Team - run 2	0.54	0.61	0.50
PanMorCresp_Team - run 3	0.53	0.81	0.50
PanMorCresp_Team - run 2	0.49	0.62	0.43
PanMorCresp_Team - run 1	0.79	0.46	0.40
Baseline - ONE-IN-ONE	1.0	0.25	0.36

and hence the results are quite similar between these two settings which use a threshold-based clustering approach.

The PanMorCresp_Team Run 1 and Run 2 perform about equally well regardless of whether or not related pages are taken into account in the evaluation. Run 1 achieves good Reliability, which fits well with the fact that complete linkage was used. This comes at the cost of a low Sensitivity. For Run 2, the picture is reversed. Run 3 and Run 4 obtain a higher score than Run 1 and Run 2. Punctuation removal, stop word removal and the larger vocabulary may play a role in this. In addition, HAC single linkage was used in both of these runs. Run 4 is the best of the PanMorCresp_Team runs; the only difference with regard to Run 3 is that it uses a fixed number of clusters (9) as a stopping criterion. The score of Run 4 beats both baselines and is on par with scores obtained with the other approaches save the scores obtained by the ATMC_UNED runs.

Out of the five runs submitted by the Loz_Team, only Run 1 managed to outperform the ALL-IN-ONE baseline. The rest of the runs only managed to outperform the ONE-IN-ONE baseline, performing worse than the ALL-IN-ONE baseline. One of the main reasons why these approaches did not perform as well may be due to the fact that the multilingualism and overlaps between clusters have not been considered, posing a significant limitation for this task. Their best performing approach (Run 1) uses a weighted representation of words, which shows that considering the frequency of words in documents leads to better performance than the sole use of a binary representation capturing the presence or not of words as well as TD-IDF. They also found that considering five clusters as the stopping criterion, instead of 15, leads to an increased Sensitivity score, which is however at the expense of a little drop in Reliability.

7 Conclusions

The M-WePNaD shared task on Multilingual Web Person Name Disambiguation took place as part of the IberEval 2017 evaluation campaign. This shared task was the first to consider multilingualism in the person name disambiguation problem, following a series of WePS shared tasks where the corpora were limited to documents in English. The M-WePNaD shared task provided the opportunity for researchers to test their systems on a benchmark dataset and shared task, enabling comparison with one another.

Despite a larger number of teams registering initially for the task, four of them managed to submit results on time, amounting to 18 different submissions. Only two of the four participants, namely the champions and the runners-up, made use of more sophisticated clustering algorithms, whereas the other two relied on the Hierarchical Agglomerative Clustering (HAC) algorithm. Only one of the teams presented an approach that does not require any prior knowledge to fix thresholds, which came from the team that qualified in the top position. We argue that this is a desirable characteristic for web page clustering, owing to the heterogeneous nature of the Web, which poses an additional challenge for learning generalizable patterns.

With respect to the approaches used for web page representation, most of the teams relied on traditional techniques based on bag-of-words and vector space models, with the exception of the runners-up, who used word embeddings.

While the novel aspect proposed in this shared task has been the multilingual nature of the dataset, only one team has proposed approaches that explicitly tackles multilingualism, ATMC_UNED, particularly it has explored three approaches. The results obtained for these approaches slightly outperform the one that does not consider multilingualism. On the other hand, the dataset also included web pages from social media, unlike in previous shared tasks. However, only one of the teams, ATMC_UNED, has taken this into account when developing their system. None of the systems has dealt with unrelated results and overlapping clusters.

Acknowledgments

This work has been part-funded by the Spanish Ministry of Science and Innovation (MAMTRA-MED Project, TIN2016-77820-C3-2-R and MED-RECORD Project, TIN2013-46616-C2-2-R).

References

1. J. Artiles, J. Gonzalo and S. Sekine. (2007). The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
2. J. Artiles, J. Gonzalo and S. Sekine. (2009) Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.
3. J. Artiles, 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.*
4. R. Berendsen, Finding people, papers, and posts: Vertical search algorithms and evaluation, Ph.D. thesis, Informatics Institute, University of Amsterdam (2015). URL: http://dare.uva.nl/document/2/165379
5. S. Montalvo, R. Martínez, L. Campillos, A. D. Delgado, V. Fresno, F. Verdejo. MC4WePS: a multilingual corpus for web people search disambiguation, *Language Resources and Evaluation* (2016). URL: http://dx.doi.org/10.1007/s10579-016-9365-4.
6. E. Amigó, J. Gonzalo, F. Verdejo. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, pp. 643-652. Dublin, Ireland, 2013. URL: http://doi.acm.org/10.1145/2484028.2484081.
7. A. Bagga, B. Baldwin, Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING'98, Association for Computational Linguistics*, Stroudsburg, PA, USA, 1998, pp. 79-85. URL http://dx.doi.org/10.3115/980451.980859.
8. A. Delgado. ATMC team at M-WePNaD task. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017
9. A.D. Delgado, R. Martínez, S. Montalvo and V. Fresno. Person Name Disambiguation in the Web Using Adaptive Threshold Clustering. *Journal of the Association for Information Science and Technology*, 2017. URL: https://doi.org/10.1002/asi.23810.
10. A.D. Delgado: Desambiguacion de nombres de persona en la Web en un contexto multilingüe. PhD Thesis, E.T.S. Ingeniería Informática, UNED, 2017.
11. A.D. Delgado, R. Martínez, S. Montalvo and V. Fresno. Tratamiento de redes sociales en desambiguación de nombres de persona en la web. *Procesamiento del Lenguaje Natural*, 57:117-124, 2016. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5344.
12. A. Duque, L. Araujo and J. Martínez-Romo. LSI UNED at M-WePNaD: Embeddings for Person Name Disambiguation. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017

13. L. Lozano, Jorge Carrillo-de-Albornoz and E. Amigó. UNED Loz Team at M-WePNaD. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017
14. P. Panero, M. Moreno, T. Crespo, Jorge Carrillo-de-Albornoz and E. Amigó. UNED PanMorCrepsTeam at M-WePNaD. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017