Noname manuscript No.
(will be inserted by the editor)

MC4WEPS: A Multilingual Corpus for Web People Search Disambiguation

Soto Montalvo[1] · **Raquel Martínez**[2] · **Leonardo Campillos**[3] · **Agustín D. Delgado**[2] · **Víctor Fresno**[2] · **Felisa Verdejo**[2]

the date of receipt and acceptance should be inserted later

Abstract This article introduces the MC4WEPS corpus, a new resource for evaluating Web People Search Disambiguation tasks, and describes its design, collection and annotation process, the agreement between the different annotators, and finally introduces a baseline evaluation. This corpus is built by compiling multilingual search engines results where the queries are person names. Proper noun disambiguation is an open problem in natural language ambiguity resolution and, specifically, resolving the ambiguity of person names in Web search results is still a challenging problem. However, state-of-the-art approaches have been evaluated only with monolingual web page collections. The MC4WEPS corpus aims to provide the research community with a reference corpus for the task of disambiguating search engine results where the query is a person name shared by homonymous individuals. The features of this new corpus stand out from existing corpora for the same task, namely multilingualism and inclusion of social networking websites. These characteristics make it more representative of a real search scenario, especially for evaluating person name disambiguation in a multilingual context. The article also includes detailed information about the format and the availability of the corpus.

Keywords Corpus linguistics · Multilingual · Annotation · People Name Disambiguation

1 Introduction

Searching for information about people on the Web is a common practice. An estimate on the data collected from a major commercial search engine that about $2 \sim 4\%$ of the daily Web queries are in the form of just personal

[1] URJC, Madrid, Spain
[2] NLP&IR Group, UNED, Madrid, Spain
[3] LLI-UAM, Spain

names [28]. This amount is even higher considering all queries that contain personal names: between 11 and 17% of Web queries include a person name [1]. However, one of the major difficulties when searching for people on the Web is the fact that different people have the same name.

Resolving the ambiguity of person names in Web searches is a challenging problem and is a growing area of interest for Natural Language Processing (NLP) and Information Retrieval (IR) communities. According to the description provided by the Web People Search (WePS) Task in SemEval 2007 [2], the problem of disambiguating person names in a Web search scenario can be defined as follows: given a query consisting of a person name in addition to search engine results 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 lies in estimating the number of different homonymous individuals and grouping the pages of the same individual in the same cluster. This task is therefore addressed as a clustering problem. It should not be confused with *entity linking* (EL), where the goal 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 difficulty of disambiguating person names resides in the fact that many people share the same name. For instance, given a query for *Tom Mitchell*, the top 100 results returned by Google contain 37 different namesakes [24]. It is estimated that in United States alone, the 300 most common male names are used by more than 114 million people [31]. The problem of disambiguating person names has had an impact on the Internet and has promoted the development of vertical search engines that are specialized in Web people searches (e.g. `spokeo.com`, `123people.com` or `zoominfo.com`). In addition, many Web People Name Disambiguation approaches have been developed. All the methods proposed in WePS-1, WePS-2 and WePS-3 campaigns are presented in [2, 3, 1], respectively; and some examples of the most recent are [11, 10, 6, 8, 24, 21, 15, 33].

Along with state-of-the-art approaches, language resources have played an essential role in research on NLP and a wide range of language technologies. Especially, the corpora provide a material basis and a test bed for building NLP systems [32]. A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety [23]. The traditional written corpora for linguistics research were created primarily from printed text, but with the growth of the World Wide Web as an information resource, they are increasingly being used as training data in NLP tasks [20]. Thus, the Web offers enormous possibilities for corpus development [18] and, in the case of the Web People Search Disambiguation task, it is the natural source of information to build new benchmark collections.

Web pages referring to the same individual are of different natures, which makes the clustering process difficult. For example, some pages may be professional sites, while others may be blogs containing personal information. In addition, social networking services usually cause search engines to return several profiles belonging to different individuals sharing the same name. These social pages often introduce noisy information and make the state of the art algorithms break down [6]. Due to these problems, the users have to refine the queries with additional terms. This task gets harder when the person name is shared by a celebrity or 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. For example, a Web search about a *George Bush* other than the former U.S. president can return many pages about the former president, which may be problematic [34]. Consequently, it may be necessary to search once more to find web pages about the target person that could be hidden among the numerous unrelated ones. Finally, in some cases, web pages referring to a specific individual may also be written in different languages, which increases the difficulty of finding the desired results.

This paper introduces the MC4WEPS (*Multilingual Corpus for WEb People Search*) corpus, a multilingual resource for the Web People Search Disambiguation task, and describes its characteristics, structure and format. Even though there are several benchmark collections publicly available to train and evaluate the algorithms that deal with this task, these collections only cover some of the aspects that can be found in a real search scenario. For example, individuals with web pages of different natures, or who share a celebrity's name. However, none of the existing corpora also cover, at the same time, the other two requirements we have considered in this collection: multilingualism and a realistic number of social networking web pages. The MC4WEPS corpus provides an evaluation scenario which addresses all these criteria.

The article is structured as follows: Section 2 provides a background on other Web corpus collections for the Web People Search Disambiguation task. In Section 3 we describe how the corpus was constructed and the degree of inter-annotator agreement. In Section 4 we compare the new corpus with other reference corpora for this task. Section 5 presents a baseline evaluation of the corpus with some state-of-the-art algorithms. Finally, the paper ends with some conclusions in Section 6.

2 Related Work

Supervised and corpus based approaches in NLP research have evolved since annotated corpora began to be collected. Despite being numerous, most corpora were created for specific research projects and are not publicly available. In spite of this, several corpora are available for evaluating the performance of person name disambiguation systems. We shall describe these corpora briefly below.

Three corpora were created for the Web People Search campaigns. WePS[1] is a competitive evaluation campaign that proposes several tasks, including resolution of disambiguation on the Web data. In particular, WePS-1, WePS-2

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

and WePS-3 campaigns provided an evaluation framework consisting of several annotated data sets composed of English person names.

In essence, the WePS task that was proposed to the SemEval-2007 participants (WePS-1) was the following: systems receive a set of web pages (which are the result of a Web search for a person name), and these have to be clustered in as many sets as there are entities sharing the name [2]. WePS-1 provided a training corpus and a test corpus. For the training corpus, the compilers selected person names from different sources in order to provide different ambiguity scenarios. The corpus is made up of 49 person names and the top 100 search results written in English for each name from the Yahoo! search engine. Nevertheless, some person names had fewer than 100 web pages, since some URLs returned by the search engine were often no longer available. Out of those 49 person names, 32 were collected from the Web03 corpus (see [22] for more details). Furthermore, 7 person names were extracted from the English Wikipedia with a view to including prominent individuals (e.g. popular or historical figures), which ensured lower ambiguity. Finally, 10 names were randomly selected from the Program Committee listing of the Computer Science conference ECDL-06. These 10 names were potentially less ambiguous, because computer science scholars usually have a stronger presence on the Internet than individuals from other professional fields. For its part, the WePS-1 test corpus is composed of 30 English person names and the top 100 search results written in English for each name from the Yahoo! search engine. Out of these 30 person names, 10 were extracted from the English Wikipedia; another 10 names from participants of the Association for Computational Linguistics conference 2006 (ACL'06); and the last 10 names were collected from the U.S. Census.

The WePS-2 campaign [3] proposed the two following tasks: (1) Clustering web pages to solve the ambiguity of search results, and (2) Extracting 18 attribute type values for target individuals whose names appear on a set of web pages. This campaign provided two types of data: a development corpus and a test corpus. The development corpus consists of the corpora and clustering gold standard previously used for the WePS-1 campaign (49 PERSON NAMES). Whereas the test corpus consists of 30 person names and it is compiled similarly as WePS-1 corpora. The WePS-2 corpus collected 30 person names and the top 150 search results from the Yahoo! search engine (using the name as a quoted query and searching only for pages written in English). In some cases, some pages from the search results were not included in the final corpus because they could not be downloaded or were not available when creating the corpus. In addition, those documents that did not contain at least one occurrence of the person name were removed. In this case, 10 out of these person names were randomly sampled from the list of biographies in the English Wikipedia, another 10 names where randomly extracted from the list of Programme Committee members for the annual meeting of the Association for Computational Linguistics (ACL'08), and the last 10 person names were randomly composed by using frequent names and surnames in the U.S. Census.

The WePS-3 campaign [1] proposed a task that merged the problems proposed in the two previous WePS campaigns, where the system must return both the documents and the attributes for each different set of people sharing a given name. In WePS-3 the amount of test data was increased, both in number of documents and in person names. The WePS-3 corpus gathered 300 person names and the top 200 search results written in English for each name from the Yahoo! search engine. To obtain the names, the campaign organizers followed similar procedures to those used in WePS-2, i.e. randomly extracting person names from the US Census, Wikipedia and Computer Science conferences program committees. But in addition to that, the WePS-3 task included names for which at least one person had one of the following occupations: attorney, corporate executive or realtor. Fifty names were extracted from each of these sources to make a total of 300 names.

We have revised these three corpora and we have seen that all of them are monolingual, and have limited profiles from social networking pages. In addition, WePS-2 organizers did not take these kind of pages into account for the evaluation. Thus, they do not cover all aspects occurring in the current context of Web People Search Disambiguation tasks.

For their part, Bekkerman and McCallum attempted to approach the problem of finding Web occurrences of a group of people [5]. That is, they brought up the question of how a social network needs to be leveraged if we want to look for several people who are related in some way. In order to test their proposal, they developed a corpus for the people name disambiguation task, but with some people names related in some way. They extracted 12 person names that appeared in headers of email messages collected by participants in a research project. All of the individuals were likely to be present on the Web. Then they collected and hand-labeled a data set of over 1000 web pages retrieved from Google queries on the 12 personal names, which were queried in quotation marks. In 10 out of 12 cases the person names were heavily ambiguous, and all the web pages were written in English language.

Nonetheless, and as we stated before, a more realistic scenario for evaluating person name disambiguation tasks requires gathering data from social networking sites, because a person name queried in a Web search engine usually returns social network profiles. For this reason, [6] the ECIR2012 corpus[2] provides a data set that, basically, differs from the WePS corpora because it contains a significant amount of social networking profiles. This corpus was created by selecting queries from query logs of a people search engine. The logs were collected between September 2010 and February 2011 and contain queries, associated clicks, and browser cookies (for user identification). In order to select ambiguous queries, they required queries to have clicks to at least three different profiles within one social media platform. In addition, they required that at least seven searches were performed, with clicks to at least two search engines or social media platforms. The document set was constructed by retrieving 20 documents (profiles) from each of five large social media plat-

[2] http://ilps.science.uva.nl/resources/ecir2012rdwps

forms (e.g., Facebook, LinkedIn, Twitter) and 50 documents from three major Web search engines (Google, Yahoo! and Bing). The URLs were used to delete repeated web pages, and those documents that did not contain the searched person name were ignored. The ECIR2012 corpus was eventually made up of 33 ambiguous names and a total of 3,487 web pages. The number of search results varied, ranging from 27 to 164, depending on the person name. This corpus poses new challenges for disambiguation methods that had proved to be very effective previously, due to the fact that social media profiles are textually sparse. However, the documents included in this corpus were mainly written in Dutch, so it might only be useful for studying person name disambiguation in a monolingual scenario. Moreover, considering that social documents were retrieved from large social media platforms and not directly from a search engine, this corpus does not represent a real-word scenario for this task, such as it is defined in WePS campaigns.

Finally, multilingual web pages in the major world languages are common nowadays, and person name queries often return results in more than one language. It would therefore be desirable for multilingual corpora to be available for tackling the Web People Search Disambiguation task in this scenario. As far as we know, the only multilingual corpus available is a resource developed to evaluate the language independence of an unsupervised method for disambiguating names of people, places and organizations [25]. The corpus was created by locating large news corpora for each of the four languages being considered (Bulgarian, English, Spanish, and Romanian). Subsequently, they identified the Named Entities in the news documents automatically. In order to facilitate evaluation, they created ambiguities in the data by conflating names that are largely unambiguous. For example, they took all occurrences of *Bill Clinton* and all occurrences of *Tony Blair* and made their names ambiguous by replacing them with *Bill Clinton-Tony Blair*. The corpus contains 16 person names, but only one of them is present in documents of different languages. We consider this corpus a bit artificial, since person name ambiguity was forced by the developers of the corpus, i.e. it was not natural. Regarding multilingualism, the person names involved were not shared between the documents of different languages, so the disambiguation task was basically monolingual. Moreover, this corpus is made up of news from the Web search engine scenario, and consequently the social network profiles were not present.

This brief survey shows that, despite there being some corpora available for evaluating the different proposals for disambiguating person names on the Web, they are basically monolingual. Furthermore, except from ECIR2012, they do not contain or process a significant percentage of social network pages, which are usually returned by search engines when searching for information with a person name query. In this paper we present a new corpus in order to provide a more current and realistic scenario for training and evaluating those systems that deal with this task. The two main features of this corpus are: (1) The inclusion multilingual results; and (2) Social networking profiles as kept as they were retrieved by the search engine. The next section of this article provides a detailed description of the pipeline used to create the corpus and its characteristics.

3 Design of the MC4WEPS Corpus

Here we will explain the details of the compilation of this new corpus. We developed the MC4WEPS corpus by mainly considering two design aspects: degrees of ambiguity, and multilingualism. Two new contributions of our work are that, first, we have designed the corpus by including person names with several degrees of ambiguity and, second, the corpus contains both monolingual and multilingual search results in different degrees.

3.1 Compilation Criteria

We carried out the compilation and annotation of the corpus throughout 2014. The first step in our approach involved identifying different sets of seed person names, which ensured variety in terms of ambiguity and languages. Depending on the search results, these initial names were changed until a sufficiently varied corpus was collected.

The person names queries were defined according to the following variables:

- *Ambiguity*: queries could be *non ambiguous*, *ambiguous*, or *very ambiguous.* We classified a person name as *very ambiguous* when the results of the search engine corresponded to more than 10 different individuals in a range between 100 and 300 web pages. Otherwise the name was considered *ambiguous*, unless all the results corresponded only to one individual (these cases were *non ambiguous*).
- *Language*: the results could be *monolingual* (all the web pages were written in the same language) and *multilingual* (there were web pages 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.

Regarding ambiguity, the candidate person names for the queries included celebrities, international researchers, politicians, prominent professionals of a linguistic community and popular names in the main languages considered.

As for the languages, we mainly focused the queries on names in English and Spanish, which are two of the top three languages used on the Web, according to the statistics presented at Internet World Stats[3]. The most used language on the Web is English, followed by Chinese and Spanish. The Chinese language does not use the Latin alphabet and therefore falls outside the scope

[3] http://www.internetworldstats.com/stats7.htm

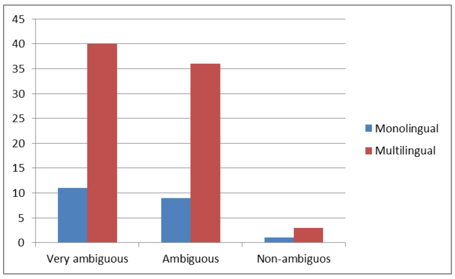


Fig. 1 Composition of the corpus according to ambiguity and language of results.

of this corpus. Therefore, the main languages selected in this corpus coincide with the top Latin alphabet languages used on the Web. On the other hand, we consider that every name, independent of its language of provenance, could essentially generate a bilingual ambiguity, because in current web context the results of web searches are usually in the same language as the query, as well as in English. In this sense, in order to build the corpus, we begin by selecting queries in Spanish (French or other language) person names, English names, and a combination of both (the first name in one language and the surname in the other).

The distribution of names according to their origin language are as follows: 36 Spanish names; 34 English names; 8 French names; 2 Italian names; 1 Portuguese name; 1 Basque name; 1 Catalan name, and 17 mixed names with first and last name in different language. Notice that the origin language of a name does not necessarily imply that the language of the web pages referring to that name must be the same.

The MC4WEPS corpus is made up of web sites downloaded from web searches on 100 different person names. The results returned by the search engine included all kind of web resources: hypertext documents, images, sounds files, and other files that are common on the Web (e.g. PDF or PPT).

Figure 1 shows the composition of the corpus according to the degree of ambiguity of query and the languages of the results. As can be seen, the results for the person name queries are mainly multilingual, and have a certain degree of ambiguity (51% of person names are very ambiguous, 45% ambiguous and 4% non-ambiguous).

Non ambiguous results tended to refer to celebrities from the media (e.g. *Norah Jones*), or outstanding individuals (e.g. *Emily Dickinson*). Ambiguous results were monolingual when they corresponded to renowned people, but only in the heart of a linguistic community (e.g. *Manuel Campo*, a Spanish journalist). On the contrary, multilingual results were generally sites devoted to distinguished international persons (e.g. *Jacques Cousteau*). Lastly, very ambiguous results were usually combinations of popular names and surnames (e.g. *John Smith* and *Michael Collins*).

There are 30 different languages in the MC4WEPS corpus, but only two of them are seen very frequently in the results. Figure 2 shows the predominant

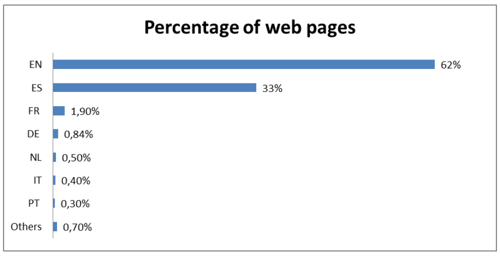


Fig. 2 Presence of the main languages in MC4WEPS corpus.

languages, i.e. those appearing on more than twenty web pages. We use the standardized nomenclature ISO 639 to classify the languages (in particular, the 2-letter ISO 639-1 nomenclature). There are a total of seven of the most common languages (FR, DE, IT, PT, EN, ES and NL), although English and Spanish predominate. The least frequently occurring languages in the corpus (with less than twenty web pages) are the following: HR, VI, HU, ID, OC, UK, DA, IS, GL, EL, ZH, NO, CS, SK, RU, TR, SV, EU, RO, JA, FI, CA and PL.

In accordance with the design decisions, the predominant languages in the corpus are English and Spanish. On the other hand, due to the fact that all person names in the new corpus are either English or European names, the presence of Asian languages in the corpus is limited. Even so, the corpus does contain some pages written in Asian languages and also in languages of Cyrillic alphabet such as Russian or Ukrainian, because these pages include the query name in Latin alphabet.

Regarding the multilingualism of person names, the number of names with web pages written in more than two languages represent 46% of the names in the corpus; whereas 35% of the names contain web pages in two languages, and the rest of the names only have pages in one language (19%).

Tables 1, 2, and 3 present the 100 person names found in the corpus. Each table has seven columns: the first one shows whether the query results are monolingual or multilingual; the second column shows whether the clusters of different individuals who share the name are monolingual or multilingual; the third column presents the person names; the fourth and fifth columns show the number of web pages for each person name classified into related (R) and non-related pages (NR), respectively (R and NR concepts are described in Section 3.2); and the last two columns show the number of individuals whose web pages are all written in the same language (Mono) and in different languages (Multi), respectively. The information for non-ambiguous names is presented in Table 1, Table 2 contains very ambiguous names, and finally, Table 3 presents the information of ambiguous names.

As the tables show, 96% of person names are ambiguous, being 51% of them very ambiguous, and only 4% non-ambiguous. These percentages agree with the objective of building a corpus with person names of different degrees of ambiguity. Furthermore, for each type of ambiguity, around 75% of the

Table 1 List of non-ambiguous people names (4% of the names of the collection).

Language of the results	Language of the clusters	Person Name	Web Pages		Individuals	
			R	NR	Mono	Multi
Monolingual (1%)	Monolingual (1%)	Emily Dickinson	106	1	1	0
Multilingual (3%)	Multilingual (3%)	Albert Barillé	88	11	0	1
		Henri Michaux	97	1	0	1
		Norah Jones	10	1	0	1

names have web pages in different languages, which was the other aim when designing the corpus. In the multilingual cases, some individuals have web pages in different languages, but others have all their corresponding pages written in the same language. In this sense, the corpus features a varied degree of multilingual names. This characteristic can make the task of disambiguation more difficult, because although the corpus is multilingual, the clusters that represent the different individuals are not always multilingual.

3.2 Compilation and Annotation Procedures

Each query was composed of a name and a surname (e.g. *Julio Iglesias*). We do not use quotation marks in the queries in order to increase the recall obtaining more web pages with similar names about the same individual, for example the names *George Bush* and *George W. Bush*. From here on we refer to each of these queries as AMB-QUERY. Mozilla Firefox navigator was used for all the queries, and Google and Yahoo were the selected search engines. These search engines were also used with the following configuration of the advanced search option: any language; any region; anytime; and Safesearch to "off". Annotators observed than Google filtered the results by the language of the country of the IP more than Yahoo did, even though advanced search was used with both `google.com` and `yahoo.com`.

For each AMB-QUERY, the first 110 web sites were saved. Pages retrieved from external search engines (e.g. Google Books or Images, Yahoo Shopping, Ask, Google or Yahoo News) were discarded. We also ruled out sponsored links, advertisements, repeated results, and pages with errors in downloading or whose content was unavailable because it had expired.

Each web page was downloaded and stored for off-line processing. We also stored the basic metadata associated with each search result, including the original URL, the ISO 639-1 language code of the site contents (e.g. ES, 'Spanish', or EN, 'English'), the date of downloading, and the name of the annotator. These metadata were stored in XML format. Figure 3 is an example of said metadata.

Five linguists gathered the data. Each one performed searches for 20 different AMB-QUERIES, downloaded the documents, and tagged the metadata. Then, they checked the results and classified the sites according to the indi-

Table 2 List of very ambiguous people names (51% of the names of the collection).

Language of the results	Language of the clusters	Person Name	Web Pages		Individuals	
			R	NR	Mono	Multi
Monolingual (11%)	Monolingual (11%)	David Cutler	79	19	37	0
		John Smith	90	11	52	0
		Jorge Fernández	101	6	28	0
		Ken Olsen	94	6	41	0
		Mark Davies	85	20	60	0
		Michael Collins	93	15	31	0
		Michael Hammond	89	11	79	0
		Michael Hastings	93	7	19	0
		Peter Mitchell	86	24	60	0
		Randy Miller	66	33	52	0
		William Miller	67	40	40	0
Multilingual (40%)	Multilingual (34%)	Agustín González	93	6	41	4
		Albert Gomez	75	30	49	1
		Álex Rovira	89	6	12	8
		Alfred Nowak	37	72	12	3
		Almudena Sierra	37	63	19	3
		Álvaro Vargas	92	8	39	11
		Amber Rodríguez	95	11	65	8
		Antonio Camacho	58	51	29	10
		David Robles	93	7	49	9
		Didier Dupont	59	50	20	14
		Elena Ochoa	105	5	13	2
		Hendrick Janssen	27	77	17	2
		James Martin	86	14	46	2
		Jesse García	91	18	22	4
		John Harrison	88	21	39	11
		José Ortega	87	21	37	3
		Joseph Lister	103	6	9	3
		Joseph Murray	84	21	46	1
		Julián López	107	2	26	2
		Liliana Jiménez	35	55	26	5
		Mario Gómez	99	1	17	1
		Matt Biondi	100	6	8	4
		Michelle Martínez	97	8	48	1
		Miriam Gonzalez	104	6	37	6
		Peter Kirkpatrick	79	27	31	4
		Pierre Dumont	84	15	31	8
		Raúl González	105	2	29	3
		Richard Rogers	84	16	30	10
		Roger Becker	84	19	24	5
		Amanda Navarro	73	29	49	1
		Brian Fuentes	97	3	9	3
		John Williams	85	17	43	1
		Olegario Martínez	90	10	36	2
		Thomas Klett	56	42	26	7
	Monolingual (6%)	Alberto Angulo	101	7	49	0
		Andrea Alonso	83	22	49	0
		Leonor García	88	12	53	0
		Palmira Hernández	41	64	37	0
		Rafael Morales	82	18	47	0
		Virginia Díaz	89	17	40	0

Table 3 List of ambiguous people names (45% of the names of the collection).

Language of the results	Language of the clusters	Person Name	Web Pages		Individuals	
			R	NR	Mono	Multi
Monolingual (9%)	Monolingual (9%)	Frederick Sanger	95	5	2	0
		Manuel Campo	100	3	7	0
		Marina Castaño	98	2	5	0
		Michael Bloomberg	108	2	2	0
		Richard Branson	98	2	3	0
		Rick Warren	99	0	5	0
		Ryan Gosling	102	1	2	0
		Mary Lasker	87	16	3	0
		Mary Leakey	106	4	2	0
Multilingual (36%)	Multilingual (34%)	Albert Claude	80	26	8	1
		Alberto Granado	106	1	1	1
		Aldo Donelli	94	16	2	2
		Almudena Ariza	105	5	4	2
		Chris Andersen	98	2	5	1
		Cicely Saunders	98	12	1	1
		Edward Heath	90	13	6	2
		Francisco Bernis	71	29	1	3
		Franco Modigliani	107	2	0	2
		Julio Iglesias	108	1	1	1
		Katia Guerreiro	110	0	6	2
		Manuel Alvar	71	38	3	1
		María Dueñas	100	0	3	2
		Michael Portillo	104	1	1	1
		Michelle Bachelet	102	5	1	1
		Oswald Avery	106	4	1	1
		Paul Erhlich	92	7	7	2
		Paul Zamecnik	95	7	3	3
		Pedro Duque	96	14	3	2
		Rafael Matesanz	107	3	5	1
		Richard Vaughan	102	6	3	2
		Rita Levi	102	2	1	1
		Tim Duncan	100	3	2	1
		William Osler	81	25	4	1
		Gaspar Zarrías	110	0	2	1
		George Bush	93	15	2	2
		Gorka Larrumbide	74	35	2	1
		Jacques Cousteau	108	1	1	1
		Javi Nieves	104	2	1	2
		Adam Rosales	99	11	7	1
		Claudio Reyna	104	3	4	1
		John Orozco	80	20	8	1
		Lauren Tamayo	90	11	7	1
		Miguel Cabrera	104	4	1	2
	Monolingual (2%)	Robin López	88	14	10	0
		Michel Bernard	5	95	5	0

```
<?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. 3 Example of metadata for a web page of the *Julio Iglesias* name.

viduals they referred to. Thus, for each AMB-QUERY, a gold standard was created by specifying the following information:

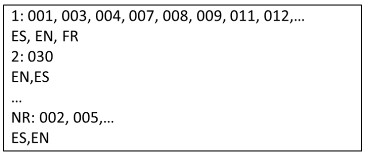
- The different individuals that were found. Each different individual was represented with a number that also identified the cluster.
- The identifier of each downloaded document that corresponded to each cluster. The web site identifier was the position of the website in the ranking results.
- The language code of the cluster, taking into account that there could be more than one language.

In many cases, a cluster of non-related (NR) pages was also created for each AMB-QUERY. A web page was considered non-related in the following cases:

- Social network web pages where the individual was not clearly identifiable (this assessment could be slightly subjective).
- Web pages where the identity of the individual was unclear, and could only be identified through the web page links.
- Web sites of streets, public places such as airports or libraries, foundations, organisms or institutions with the name of a person, such as *Paul Ehrlich Foundation* or *Jacques Cousteau Society*.
- Web pages referring to person names with the same words as the query, but in reverse order, such as *Avery Oswald* for the query *Oswald Avery*.
- Web pages referring to different person names made up of the first name of the query and the surname, for example, people with the names *Paula Jiménez* and *Liliana Viola* for the search *Liliana Jiménez*.

Figure 4 shows a simplified version of the gold standard for the name *Pedro Duque*. Cluster number 1 refers to the famous Spanish astronaut followed by the list of his corresponding sites identifiers 001, 003, 004, etc., and the language codes of these sites. Cluster number 2 refers to a Spanish manager in the automotive sector, etc., and the results number 002, 005,... are non-related pages for that name.

On the other hand, in recent years the number of different social networks has increased and search engines return results from social media platforms when searching information with a person name query. This fact poses new challenges for disambiguation methods that had previously been shown to be



```
1: 001, 003, 004, 007, 008, 009, 011, 012,…
ES, EN, FR
2: 030
EN,ES
…
NR: 002, 005,…
ES,EN
```

Fig. 4 Simplified version of the gold standard for *Pedro Duque* name.

very effective [6]. Specifically, the text of social media profiles can be short, noisy, and strongly dependent on the context, complicating the task of extracting good textual features that can help in the disambiguation process. Therefore, our proposal of a new corpus contains a significant and more realistic amount of social web pages from different social networks (e.g. Facebook, Linkedin, Twitter, or Google+), where some of the pages may contain more textual information related to the individual than others pages, but all the social pages are returned by the search engine as relevant for a person name query, and consequently, all of them must be dealt with in the disambiguation process.

Pages from social networks presented additional difficulties to annotators. These types of sites frequently showed results of several persons with the same name, since some typical social pages may contain a list of these different individuals in the network with the links to access each person's profile. These listings of homonymous individuals had to be assigned to other previously identified clusters, which was not straightforward task. This was mainly due to the fact that the cluster corresponding to an individual contains not only pages referring to him/her, but also social networking pages related to him/her and homonymous individuals. Therefore, these listing pages were included in more than one cluster, namely in each cluster corresponding to a different homonymous individual. For example, if a listing page for the name *Pierre Dumont* included information about two different people corresponding to cluster 1 and 2 for this search, this listing page was included both in cluster 1 and 2. Similarly, telephone listings and yellow-pages posed the same problem. In these cases, individuals were more difficult to recognize, because their personal data is scarce. Indeed, individuals could not always be identified on pages from social networks (especially, if a picture was not available).

Table 4 shows information about the distribution of social networks web pages in the MC4WEPS corpus. It shows the percentage of social networking pages and the percentage of individuals (clusters) with pages from social network websites for each person name. Only two person names do not contain social networking pages, which indicates that these types of pages have a high presence on the Web, and 75% of person names in the collection have less than 10% of social network pages. Regarding individuals, all the individuals of 7 person names have social network pages (e.g. *Alberto Granado*), and 25 person names have social networking pages in 50% or more of their individuals (e.g. *Adam Rosales*). In this aspect, this corpus represents a current real scenario better than other state-of-the-art corpora do.

Finally, the MC4WEPS corpus was revised in order to detect format and inconsistency mistakes during the manual annotation process. Different kinds of mistakes were found and corrected:

- The gold standard contains a non-existing web page ID in the corpus.
- The gold standard does not contain an existing web page ID in the corpus.
- A cluster includes languages that do not appear in their corresponding web pages.
- Languages annotated in web pages do not appear in their corresponding clusters.
- The gold standard includes a web page as non-related but it also appears in some clusters.

After correcting those annotation errors the corpus and gold standard were considered finished.

3.3 Inter-Annotator Agreement

In the last few years, the reliability of corpora annotation has acquired an increasing importance and has become a key requirement for creating a usable annotated corpus. The data annotated by a single person may be prone to bias and error, which results in unreliable annotations. For this reason, the recent trend in corpus development is to have more than one annotator annotate corpus independently [7]. In accordance with this trend, one of our objectives was to find out whether the annotators reached a satisfying level of agreement when they performed the same coding task.

In corpus statistics, corpus reliability is measured by a coefficient of agreement. The Kappa statistics [9] is a coefficient of agreement for nominal scales, which measures the proportion of observed agreement to agreement by chance and the maximum agreement attainable to chance agreement considering pairwise agreement. An extension of the Kappa statistics was proposed later [13] for measuring agreement in ordinal scale data.

In recent years the Kappa coefficient has been used in computational linguistics as a standard, because of its simplicity and robustness. The equation for computing the Kappa coefficient is the following:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the proportion of times that the annotators agree and $P(E)$ is the proportion of times that we would expect them to agree by chance. The values of K are constrained to the interval [-1, 1]. A K value of 1 means perfect agreement, whereas a K value of 0 means that agreement is equal to chance, and finally, a K value of -1 means complete disagreement.

Usually, the data for paired ratings on a 2-category nominal scale are displayed in a 2×2 contingency table. We calculated a contingency table (Table 5) to measure the agreement between the annotators in the MC4WEPS corpus, where:

Table 4 Presence of pages from social networks in the collection.

Person Name	Social web pages	Clusters with social web pages	Person Name	Social web pages	Clusters with social web pages
Adam Rosales	7.27%	66.67%	Ken Olsen	5%	9.52%
Agustín González	7.07%	15.22%	Lauren Tamayo	9.9%	77.78%
Albert Barillé	2.02%	50%	Leonor García	7%	9.26%
Albert Claude	10.38%	20%	Liliana Jiménez	21.11%	43.75%
Albert Gomez	9.52%	19.61%	Manuel Alvar	2.75%	40%
Alberto Angulo	5.56%	12%	Manuel Campo	2.91%	12.5%
Alberto Granado	2.8%	100%	María Dueñas	5%	80%
Aldo Donelli	7.27%	60%	Marina Castaño	4%	66.67%
Álex Rovira	21.05%	71.43%	Mario Gomez	3%	15.79%
Alfred Nowak	2.75%	12.5%	Mark Davies	17.14%	26.23%
Almudena Ariza	10.91%	100%	Mary Lasker	0.97%	25%
Almudena Sierra	10%	34.78%	Mary Leakey	4.55%	100%
Álvaro Vargas	22%	43.14%	Matt Biondi	9.43%	53.85%
Amanda Navarro	6.86%	13.73%	Michael Bloomberg	5.45%	66.67%
Amber Rodríguez	10.38%	14.86%	Michael Collins	0.93%	6.25%
Andrea Alonso	8.57%	14%	Michael Hammond	19%	22.5%
Antonio Camacho	22.94%	40%	Michael Hastings	4%	15%
Brian Fuentes	6%	38.46%	Michael Portillo	3.81%	66.67%
Chris Andersen	4%	42.86%	Michel Bernard	0%	0%
Cicely Saunders	5.45%	100%	Michelle Bachelet	6.54%	66.67%
Claudio Reyna	6.54%	66.67%	Michelle Martínez	11.43%	14%
David Cutler	13.27%	23.68%	Miguel Cabrera	5.56%	50%
David Robles	6%	10.17%	Miriam Gonzalez	10.91%	25%
Didier Dupont	22.94%	57.14%	Norah Jones	4.95%	100%
Edward Heath	1.94%	22.22%	Olegario Martínez	10%	10.26%
Elena Ochoa	7.27%	50%	Oswald Avery	6.36%	66.67%
Emily Dickinson	3.74%	50%	Palmira Hernández	7.62%	18.42%
Francisco Bernis	4%	20%	Paul Erhlich	3.03%	30%
Franco Modigliani	1.83%	33.33%	Paul Zamecnik	1.96%	14.29%
Frederick Sanger	0%	0%	Pedro Duque	3.64%	33.33%
Gaspar Zarrías	3.64%	33.33%	Peter Kirkpatrick	6.6%	13.89%
George Bush	1.85%	20%	Peter Mitchell	28.18%	36.07%
Gorka Larrumbide	3.67%	50%	Pierre Dumont	9.09%	15%
Hendrick Janssen	6.73%	5%	Rafael Matesanz	5.45%	28.57%
Henri Michaux	2.04%	50%	Rafael Morales	7%	12.5%
Jacques Cousteau	3.67%	66.67%	Randy Miller	10.1%	13.21%
James Martin	5%	6.12%	Raul González	3.74%	12.12%
Javi Nieves	2.83%	50%	Richard Branson	6%	75%
Jesse García	6.42%	18.52%	Richard Rogers	13%	29.27%
John Harrison	14.68%	25.49%	Richard Vaughan	3.7%	66.67%
John Orozco	9%	60%	Rick Warren	7.07%	80%
John Smith	10.89%	18.87%	Rita Levi	0.96%	33.33%
John Williams	16.67%	33.33%	Robin Lopez	10.78%	63.64%
Jorge Fernández	3.74%	13.79%	Roger Becker	3.88%	13.33%
José Ortega	9.26%	19.51%	Ryan Gosling	6.8%	100%
Joseph Lister	7.34%	53.85%	Thomas Klett	7.14%	17.65%
Joseph Murray	5.71%	12.5%	Tim Duncan	2.91%	75%
Julián López	10.34%	3.67%	Virginia Díaz	9.43%	19.51%
Julio Iglesias	1.83%	33.33%	William Miller	6.54%	14.63%
Katia Guerreiro	9.09%	100%	William Osler	3.77%	33.33%

Table 5 Data for paired ratings in a 2 × 2 contingency table.

First partition	Second partition Pair in same cluster	 Pair in different cluster	Total
Pair in same cluster	a	b	p_1
Pair in different cluster	c	d	q_1
Total	p_2	q_2	N

- a is the number of pairs in the same group in the first partition and in the second,
- b is the number of pairs in the same group in the first partition, but in different in the second,
- c is the number of pairs in different groups in the first partition, but in the same in the second,
- d is the number of pairs in different groups in the first partition and in the second,
- $N = a + b + c + d$.

Considering the contingency table, equations 2 and 3 show how *P(A)* and *P(E)* are computed, respectively:

$$P(A) = \frac{a+d}{N} \tag{2}$$

$$P(E) = \frac{(a+b)*(a+c)}{N} + \frac{(c+d)*(b+d)}{N} \tag{3}$$

Before computing the agreement between the annotators according to the Kappa coefficient, we considered two issues that have been neglected in the computational linguistics literature [12]: (1) There are two main ways of computing *P(E)*, which reflect different conceptualizations of the problem; and (2) K is affected by skewed distributions of categories (the *prevalence problem*) and by the degree to which the annotators disagree (the *bias problem*).

The two methods of computing *P(E)* are: the expected agreement, according to whether the distribution of proportions over the categories is considered by the annotators to be equal [13,29], or not [9]. In practice, the two computations produce very similar outcomes in most cases, especially for the highest values of K. However, they can indeed result in different values of K, which can lead to contradictory conclusions on inter-coder agreement. The other problem affecting the K coefficient is related with the asymmetry in the distributions of categories and the degree to which the annotators disagree. That is, for a fixed *P(A)*, the values of K vary substantially in the presence of prevalence, bias, or both.

Taking into account the previous problems associated with the Kappa coefficient, we have measured the inter-annotator agreement, not only by calculating K, but also with other measures that have previously been used to compare clustering solutions [16,27]: the Rand Index, the Jaccard Coefficient, and the Folkes and Mallows index. These indexes are based on the distribution of the pairs of texts. For all these measures the values are ≥ 0 and ≤ 1, therefore, high values indicate high agreement between the annotators. In the following we present the equations for each measure, considering again the data of the contingency table:

- The Rand Index [26] calculates the fraction of elements that were correctly classified (or, vice versa, misclassified) out of all the elements.

$$R = \frac{a+d}{N} \tag{4}$$

- The Jaccard Coefficient [17] measures the proportion of agreements in a set of n comparisons. It counts the number of pairs that belong to the same cluster in both partitions divided by the number of pairs that are included in at least one of the two partitions.

$$J = \frac{a}{a+b+c} \tag{5}$$

- The Fowlkes and Mallows index [14] measures the geometric mean of the proportion of pairs that belong to the same cluster in both partitions, relative to the number of pairs that belong to the same cluster for at least one partition.

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \tag{6}$$

Each annotator of the MC4WEPS corpus cross validated the results of two other person names that were searched and tagged by other annotators. Therefore, 10 out of the 100 AMB-QUERIES were cross validated. Next, a third linguist (the coordinator of the annotation task) performed a final cross-validation of 10 AMB-QUERIES, selecting 2 AMB-QUERIES from the different sets of previous 10 cross validated AMB-QUERIES, in order to perform a final cross-validation. The idea was to confirm that the different annotations were consistent. The result of this final validation was set as the baseline to compare with the other gold standards that were cross validated. Table 6 summarizes all the agreement values computed, with the different measures used, and cross-validated query results.

Some guidelines to classify the strength of agreement depending on the magnitude of Kappa have been proposed [19]: the extent of agreement can be qualified as *Poor* ($K < 0$), *Slight* ($0 \leq k \leq 0.21$), *Fair* ($0.21 \leq k \leq 0.40$), *Moderate* ($0.41 \leq k \leq 0.60$), *Substantial* ($0.61 \leq k \leq 0.80$), *Almost perfect* ($0.81 \leq k \leq 1.00$). A Kappa value between 40% and 60% indicates a moderate level of agreement, while value ranges of (60% - 80%) and (80% - 100%) indicate substantial and almost perfect levels of agreement respectively. Due to the arbitrary nature of these benchmarks, another benchmark scale have been proposed [13]. The Fleiss' scale collapsed the Landis-Koch benchmark into three

Table 6 Agreement indexes for the 10 cross validated queries.

Person Name	*K*	*R*	*Jaccard*	*FM*
Elena Ochoa	0.94	0.97	0.95	0.97
Mario Gómez	0.84	0.92	0.87	0.93
María Dueñas	1.0	1.0	1.0	1.0
David Cutler	0.81	0.97	0.71	0.83
Olegario Martínez	0.74	0.93	0.65	0.79
Hendrick Janssen	0.41	0.75	0.66	0.80
Miguel Cabrera	0.85	0.96	0.96	0.98
Henri Michaux	0.19	0.86	0.86	0.93
Paul Zamecnik	0.36	0.69	0.62	0.78
Pedro Duque	0.78	0.9	0.86	0.93

ranges: Kappa values of 40% or less are labeled as *Poor*; values in the 40%-75% range represent an *Intermediate to Good* extent of agreement; and finally, values in the 75%-100% range indicate an *Excellent* extent of agreement.

With regard to the agreement values in our corpus samples, 60% of names have an *Excellent* extent of agreement within the Kappa and the Fleiss' scale. These names also obtained good values of agreement by applying the rest of measurements. All of these names are ambiguous (50% very ambiguous, and the remaining 50% ambiguous).

The extent of agreement is *Intermediate to good* for 20% of the names (e.g. *Olegario Martínez*), and in this case these names are ambiguous. The agreement values for the rest of measures are again good; only the values of Jaccard Coefficient fall slightly. All the previous results indicate that the degree of agreement between the annotators is not directly related with the degree of ambiguity.

Finally, there are two names where the agreement is *Poor*: *Henri Michaux* (non-ambiguous) and *Paul Zamecnik* (ambiguous). Surprisingly, the value of Kappa is very low for the non-ambiguous name *Henri Michaux*, despite the fact that all the web pages should be in the same cluster (all search results referred the same individual). This low Kappa value is due to the cluster of nonrelated (NR) pages, which we also included when calculating the agreement. Two annotators only differed in eight NR web pages (one linguist associated a web page with the cluster of the individual or with the cluster of NR pages, and the other linguist annotated the contrary). Nevertheless, we considered pairs of web pages when we computed the agreement. This produced many different pairs between the annotators, and a low kappa value. Indeed, the three names where almost all the web pages are grouped in one cluster have the lowest Kappa value of agreement (*Hendrick Janssen*, *Henri Michaux* and *Paul Zamecnik*). This fact confirms the problems previously described that affect this coefficient. Even so, the name *Henri Michaux* has good agreement values with the rest of measurements (between 0.86 and 0.93). The agreement is a bit lower for the name *Paul Zamecnik* (between 0.62 and 0.78), but it is substantially better than the Kappa value (0.36).

In summary, due to the good agreement values between the annotators of the MC4WEPS sample, we consider that it is a reliable corpus for the task of Web People Search Disambiguation.

3.4 Availability

The MC4WEPS corpus is available to researchers on the public resources[4] web page of the Natural Language Processing and Information Retrieval Group at UNED.

In addition, we built two partitions of the corpus. First, we split the MC4WEPS corpus into training/test sets for use with supervised methods. 80% of the names were included in the training set, and the other 20% in the test set. In order to ensure that the training set has names with different degrees of ambiguity, we selected 80% of the non-ambiguous, ambiguous and very ambiguous names. We also used the same criteria for the language of the results, and we selected not only English and Spanish names, but also names of other origins, as well as mixed names (with names and surnames of different origin). We called this first partition MC4WEPS-80-20. Likewise, we created a second corpus partition. We split the MC4WEPS corpus into training/test sets with the same size relation (80-20%), but now selecting the names in a random way. Therefore, the proportionalities of the different aspects between the two sets are not ensured as in the previous case. This new partition was called MC4WEPS-80-20-Random. These versions of the corpus are also available at the same web site as the original corpus.

Regarding the copyright issue, there is no easy way of determining whether the content of a particular page is copyrighted, nor it is feasible to ask hundreds of potential copyright holders for usage permission. All the contents of the MC4WEPS corpus were publicly available on the Web and all the contents of pages from social networks were obtained without being logged into any social website, in order to avoid downloading private contents.

Even though the results of the searches are mainly HTML web pages, 1.24% are in one of these other formats: XHTML, PDF, SHTML, TXT, and XHT. Of these formats, the predominant one is PDF. An XML document with metadata is associated with each document, regardless of its format.

4 Exploring the MC4WEPS corpus

In this section we will compare the MC4WEPS corpus with state-of-the-art corpora and report the differences between them. This comparison will help to better understand the particular characteristics of the corpus we developed.

We make a comparison according to different aspects such as the size of the corpora regarding the number of person names, the different levels of ambiguity contained in the corpora, the presence of different languages on the

[4] http://nlp.uned.es/web-nlp/resources

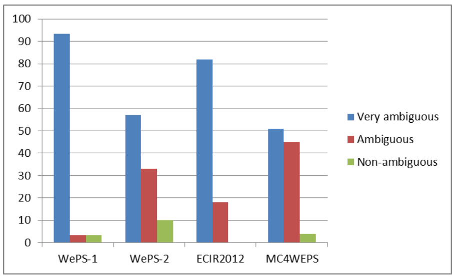


Fig. 5 Percentages of different degrees of ambiguity in several corpora.

web pages within the corpora, and also the presence of social network pages, among others.

In terms of the number of different person names, the largest corpus is WePS-3 (300 person names), followed by the MC4WEPS corpus (100 person names). The other corpora are considerably smaller. The ECIR2012 corpus contains 33 person names, whereas WePS1 and WePS2 corpora contain 79 person names, respectively.

With respect to ambiguity, we classified each name in the new corpus as *very ambiguous* (more than ten individuals share the name), *non ambiguous* (the name belongs only to one individual) and *ambiguous* (less than ten individuals and more than one share the name). We considered the same criteria for the rest of the corpora, obtaining the results showed in Figure 5. This figure does not show the information about the WePS-3 corpus because its Gold Standard only contains clusters for two individuals for each person name. Accordingly, we cannot compute the percentages of ambiguity for this corpus.

The MC4WEPS corpus is the only resource where the percentage of ambiguous and very ambiguous person names is more balanced. The rest of the corpora contain mainly very ambiguous names; indeed, all of those found in the ECIR2012 corpus are ambiguous. We think that although there are many very ambiguous names in a real search scenario, less ambiguous queries are frequent as well. Thus, when creating evaluation corpora for people name disambiguation on the Web, we think it makes sense to consider both very ambiguous and ambiguous person name queries.

The only really multilingual resource is the MC4WEPS corpus, which contains web pages in different languages. The other corpora are monolingual, with web pages only in English, or Dutch in the case of the ECIR2012 corpus. We are living in a multilingual world and the amount of non-English information that is globally accessible is growing continuously. Nowadays, users of search engines make a query in one language and can retrieve documents in more than one language. In this sense, the corpus developed is a reliable data set of Web searches.

Concerning the content of social networks, two of the four corpora (ECIR2012 and MC4WEPS) contain a significant amount of these web pages. WePS-1 test corpus contains 1.52% of social web pages, whereas test WePS-2 and WePS-3 corpora contain 2.41% and 2.76% of social web pages, respectively. As we can see these percentages are low compared to a common scenario and, moreover, a lot of these social pages were discarded in the final evaluation because there was not enough evidence to decide where to put them in the clustering solution. However, in the ECIR2012 corpus, 46.09% of the web pages were taken from social networks, whereas the MC4WEPS corpus contains 7.26% of these types of pages. The ECIR2012 corpus has more pages from social network sites due to the procedures used to collect them. The rankings of several search engines were used for each query, and then the social networking web pages of each ranking were gathered. Furthermore, the social networks for the queries (Twitter, Linkedin, Facebook, Myspace and Hyves) were predetermined. In our case, we only used the ranking obtained from Google or Yahoo! and we considered every social network that appeared in the first 100 search engine results. In a certain way, the development of ECIR2012 corpus seems more artificial. In our case, the MC4WEPS corpus was developed following the definition of the Web People Search Disambiguation task. Thus, we did not add other web pages to the ranking. In addition, the MC4WEPS corpus contains a greater variety of social networks (13 in total). The most frequent were Facebook, LinkedIn and Twitter, which are also predominant in the ECIR2012 corpus, together with Hyves, which is a Dutch social network similar to Facebook.

We also analyzed the degree to which famous individuals sharing an ambiguous name with ordinary people monopolized the top of the ranking. Documents relevant to famous people frequently outnumber those relevant to ordinary people [30]. We considered that an individual is more prominent than the rest when their corresponding cluster contains at least 25% of the total web pages for the person name. In the MC4WEPS corpus, 76% of the names are shared by prominent people. In the other corpora, the percentages of names shared by at least one prominent person is as follows: 43.33% of the names in WePS-1; 83.33% of the names in WePS-2; and 42.42% of the names in ECIR2012.

Finally, another different aspect of the MC4WEPS corpus with respect to the rest of the corpora is that the compiled search engine results are not only web pages in HTML format, but also other types of documents with different formats.

In summary, the MC4WEPS corpus has several features that make it more comprehensive and useful than the existing state-of-the-art corpora for testing new approaches to the Web People Search Disambiguation task.

5 MC4WEPS corpus evaluation

The best participants in the WePS evaluation campaigns [2,3,1] used different approaches based on Hierarchical Agglomerative Clustering (HAC) algorithm for grouping web search results. Therefore, to evaluate the corpora MC4WEPS-80-20 and MC4WEPS-80-20-Random we chose HAC algorithm with single-link technique as a baseline. For each corpus, we used the training

Table 7 Baseline corpus evaluation with HAC clustering algorithm.

	BP	BR	$F_{0.5}$
MC4WEPS-80-20	0.64	0.86	0.67
MC4WEPS-80-20-Random	0.57	0.82	0.61

set to learn the threshold that the HAC algorithm requires, and then we applied the algorithm to the test set. We represented the web pages by means of 1-grams, with tf-idf weighting schema and using cosine similarity. Table 7 presents the results. We evaluated the results with B-Cubed metrics [4]. These metrics are considered more suitable than other ones to measure the performance of People Name Disambiguation systems [3] and are the official metrics in WePS campaigns. We used B-Cubed precision (BP), B-Cubed recall (BR) and their harmonic mean ($F_{0.5}$).

These results show that there is still room for improvement. It is important to note that this baseline was carried out without any translation resources, so the results might improve with a suitable preprocessing with translation resources. On the other hand, supervised approaches could be also applied with the hope of improving the results by making the best use of the training data. Therefore, novel proposals could be applied in order to increase the BP and BR values obtained for this baseline.

The HAC algorithm is highly sensitive to the learned threshold, therefore, if the training corpus contains a higher percentage of non ambiguous and a lower percentage of ambiguous names, the learned threshold will not be sufficiently high, and the algorithm will tend to group web pages of different individuals in the same cluster. The differences between the percentages of ambiguous names and languages in the training and test sets is higher in the MC4WEPS-80-20-Random corpus than in the MC4WEPS-80-20 corpus. For this reason, precision values are worse in experiments carried out using the MC4WEPS-80-20-Random corpus, with respect to those obtained with the other corpus.

6 Conclusions

MC4WEPS is a new corpus designed to be a reliable gold standard for the task of multilingual Web People Search Disambiguation. It will be useful for researchers, freeing them from the burden of having to compile a new data set from scratch with the characteristics of this new corpus. In this way, we provide a data set by which different methods and techniques can be compared to each other. The two main contributions of the MC4WEPS corpus are that it includes multilingual results, and it keeps the social networking profiles retrieved by search engines. These features provide a more realistic scenario for the current web context to train and evaluate the systems for resolving this task. Therefore, we provide two different partitions of the complete corpus, where both are split into training and test sets. In addition, we present a baseline evaluation of the different version of the corpus with one of the most used clustering algorithms for the task of Web People Search Disambiguation.

97% of the person names contained in MC4WEPS corpus are ambiguous, but with different degrees of ambiguity. This makes the MC4WEPS corpus an excellent test bed to check whether the performance of the systems for web people search disambiguation varies depending on the degree of ambiguity.

As stated above, the MC4WEPS corpus is publicly accessible, and we hope other researchers use it. We are already actively using the MC4WEPS corpus in various tasks, and we believe that this use will give us a clearer idea of the strengths and limitations of the corpus.

Acknowledgements We would like to thank the financial support for this research to the Spanish Ministry of Science and Innovation [MED-RECORD Project, TIN2013-46616-C2-2-R], and we would also like to thank the annotators for their work.

References

1. Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., , Amigó, E.: Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In: Third Web People Search Evaluation Forum (WePS-3) (2010)
2. Artiles, J., Gonzalo, J., Sekine, S.: 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), pp. 64–69. ACL (2007)
3. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: Overview of the web people search clustering task. In: Proceedings of the 2nd Web People Search Evaluation Workshop (WePS 2009) (2009)
4. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Anual Meeting of the Association of Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 79–85 (1998)
5. Bekkerman, R., McCallum, A.: Disambiguating web appearances of people in a social network. In: Proceedings of the 14th International World Wide Web Conference (WWW 2005), pp. 463–470 (2005)
6. Berendsen, R., Kovachev, B., Nastou, E.P., de Rijke, M., Weerkamp, W.: Result disambiguation in web people search. In: Proceedings of the 34th European conference on Advances in Information Retrieval (ECIR2012), pp. 146–157 (2012)
7. Bhowmick, P.K., Mitra, P., Basu, A.: An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: Proceedings of the workshop on Human Judgements in Computational Linguistics (COLING 2008), pp. 58–65 (2008)
8. Chen, Y., Lee, S.Y.M., Huang, C.R.: A robust web personal name information extraction system. Expert Systems with Applications **39**, 2690–2699 (2012)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
10. Delgado, A.D., Martínez, R., Fresno, V., Montalvo, S.: An unsupervised algorithm for person name disambiguation in the web. Procesamiento del Lenguaje Natural **53**, 51–58 (2014)
11. Delgado, A.D., Martínez, R., Montalvo, S., Fresno, V.: A data driven approach for person name disambiguation in web search results. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 301–310 (2014)
12. Di, B., Glass, E.M.: Squibs and discussions the kappa statistic: A second look. Computational Linguistics **30**(1), 95–101 (2004)

13. Fleiss, J.L.: Statistical Methods for Rates and Proportions, second edn. Wiley, New York (1981)
14. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. Journal of the American Statistical Association **78**, 553–569 (1983)
15. Gruetze, T., Kasneci, G., Zuo, Z., Naumann, F.: Bootstrapped grouping of results to ambiguous person name queries. In: Proceedings of the 30th International Conference on Data Engineering Workshops (ICDE), pp. 56–61 (2014)
16. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems **17**(2-3), 107–145 (2001)
17. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Socit Vaudoise des Sciences Naturelles **37**, 547–579 (1901)
18. Kilgarriff, A., Grefenstette, G.: Web as corpus: Introduction to the special issue. Computational Linguistics **29**(3), 333–347 (2003)
19. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
20. Liu, V., Curran, J.R.: Web text corpus for natural language processing. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 233–240 (2006)
21. Liu, Z., Lu, Q., Xu, J.: High performance clustering forweb person name disambiguation using topic capturing. In: International Workshop on Entity-Oriented Search (EOS) (2011)
22. Mann, G.S.: Multi-document statistical fact extraction and fusion. Ph.D. thesis, Johns Hopkins University, Baltimore, MD, USA (2006). AAI3213760
23. McEnery, A., Xiao, R., Tono, Y.: Corpus-Based Language Studies: An Advanced Resource Book. London, U.K.: Routledge (2006)
24. Nuray-Turan, R., Kalashnikov, D.V., Mehrotra, S.: Exploiting web querying for web people search. Journal ACM Transactions on Database Systems **37**(1), 1–41 (2012)
25. Pedersen, T., Kulkarni, A., Angheluta, R., Kozareva, Z., Solorio, T.: An unsupervised language independent method of name discrimination using second order co-occurrence features. In: Computational Linguistics and Intelligent Text Processing, *Lecture Notes in Computer Science*, vol. 3878, pp. 208–222. Springer Berlin Heidelberg (2006)
26. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**, 846–850 (1971)
27. Rosell, M., Kann, V., Litton, J.E.: Comparing comparisons: Document clustering evaluation using two manual classifications. In: Proceedings of the International Conference on Natural Language Processing, pp. 207–216 (2004)
28. Shen, D., Walker, T., Zheng, Z., Yang, Q., Li, Y.: Personal name classification in web queries. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08), pp. 149–158 (2008)
29. Siegel, S., Jr., N.J.C.: Nonparametric statistics for the behavioral sciences, second edn. McGraw Hill (1988)
30. Vu, Q.M., Takasu, A., Adachi, J.: Name disambiguation boosted by latent topics from web directories. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '08), pp. 697–703 (2008)
31. Wang, X., Tang, J., Cheng, H., Yu, P.S.: Adana: Active name disambiguation. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM'11), pp. 794–803 (2011)
32. Xiao, R.: The Handbook of Natural Language Processing, chap. Corpus creation, pp. 147–165. CRC Press (2010)
33. Xu, J., Lu, Q., Li, M., Li, W.: Web person disambiguation using hierarchical co-reference model. In: Proceedings of the 16th International Conference CICLing 2015, pp. 279–291 (2015)
34. Yoshida, M., Ikeda, M., Ono, S., Sato, I., Nakagawa, H.: Person name disambiguation by bootstrapping. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10), pp. 10–17 (2010)