WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task

Javier Artiles
UNED NLP & IR group
Madrid, Spain
javart@bec.uned.es

Julio Gonzalo
UNED NLP & IR group
Madrid, Spain
julio@lsi.uned.es

Satoshi Sekine
CS Dept., New York University
New York, USA
sekine@cs.nyu.edu

ABSTRACT

The second WePS (Web People Search) Evaluation campaign took place in 2008-2009 with the participation of 19 research groups from Europe, Asia and North America. Given the output of a Web Search Engine for a (usually ambiguous) person name as query, two tasks were addressed: a clustering task, which consists of grouping together web pages referring to the same person, and an extraction task, which consists of extracting salient attributes for each of the persons sharing the same name. This paper presents the definition, resources, methodology and evaluation metrics, participation and comparative results for the clustering task.

General Terms

H.3.5 [INFORMATION STORAGE AND RETRIEVAL]: Online Information Systems, Web-based Service; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.2.7 [Artificial Intelligence]: Natural Language Processing

Keywords

Web Search, Web People Search, Text Clustering, Metasearch Engines, Evaluation

1. INTRODUCTION

Searching the Web for names of people can be a challenging task when a single name is shared by many people. This ambiguity has recently become an active research topic and, simultaneously, a relevant application domain for Web search services. Zoominfo.com, Spock.com, 123people.com are examples of sites which perform web people search, although with limited disambiguation capabilities.

A study of the query log of the AllTheWeb and Altavista search sites gives an idea of the relevance of the people search task: 11-17% of the queries were composed of a person name with aditional terms and 4% were identified simply as person names [28]. According to the data available from 1990 U.S. Census Bureau, only 90,000 different names are shared by 100 million people [5]. As the amount of information in the WWW grows, more of these people are mentioned in different web pages. Therefore, a query for a common name in the Web will usually produce a list of results where different people are mentioned.

This situation leaves to the user the task of finding the pages relevant to the particular person he is interested in. The user might refine the original query with additional terms, but this usually implies filtering out relevant documents in the process. In some cases, the existence of a predominant person (such as a celebrity or a historical figure) makes it likely to dominate the ranking of search results, complicating the task of finding information about other people sharing her name.

The disambiguation of person names in Web results is usually compared to two other Natural Language Processing tasks: Word Sense Disambiguation (WSD) [1] and Cross-document Coreference (CDC) [6]. Most of early research work on person name ambiguity focuses on the Coreference problem or uses methods found in the WSD literature. It is only recently that the web name ambiguity has been approached as a separate problem and defined as a NLP task - *Web People Search* - on its own [5, 4].

Therefore, it is useful to point out some crucial differences between WSD, CRC and Web People Search:

- WSD tipically concentrates in the disambiguation of common words (nouns, verbs, adjectives) for which a relatively small number of senses exist, compared to the hundreds or thousands of people that can share the same name. Word senses in dictionaries, often have subtle differences which made them hard to distinguish in practice, while person name ambiguity can be considered as a homograph-level ambiguity.
- WSD can rely on dictionaries to define the number of possible senses for a word. In the case of name ambiguity no such dictionary is available, even though in theory there is an exact number of people that can be accounted as sharing the same name.
- The objective of CDC is to reconstruct the coreference chain for every mention a person. In Web person name disambiguation it suffices to group the documents that contain at least one mention to the same person.

The first Web People Search (WePS-1) Evaluation [4] gathered 17 research teams. The task was to cluster search results for a given name according to the different people that share this name. For this purpose a large dataset was collected and manually annotated. This testbed has now become the de-facto standard benchmark for the task and is being used beyond the WePS exercise [13, 25, 24, 15].

The second WePS Evaluation has benefited from the experience acquired in WePS-1: we have explored evaluation

WWW2009, April 20-24, 2009, Madrid, Spain.
.

metrics to find the best-suited measure for the task [3], we have improved the methodology to build the test set, and we have expanded the evaluation campaign with a new attribute extraction task [27]. In WePS-2, 19 research teams from around the world have participated in one or both of the following tasks: (i) clustering web pages to solve the ambiguity of search results, and (ii) extracting 18 kinds of attribute values for target individuals whose names appear on a set of web pages.

In this paper we present an overview of the WePS 2 clustering task, including: a description of the datasets, the methodology to produce our gold-standard, the evaluation metrics and the campaign design. We also provide an overview of the partipating systems in section 3. The results of the evaluation are presented and discussed in section 4, and we end with some concluding remarks in Section 5.

2. EXPERIMENTAL METHODOLOGY

2.1 Data

The data distributed to participants was divided in development and test data.

2.1.1 Development data

This data was handed to participants so they could develop and test their systems before processing the evaluation test set. It consists of the corpora and clustering gold standard previously used for the WePS-1 campaign [4], and it was built basically with the same methodology used in WePS-2. The WePS-1 data includes 47 ambiguous names and up to 100 manually clustered search results for each name. The number of clusters per name has a large variability (from 1 up to 91 different people sharing the same name) even for the 10 names extracted from Wikipedia biographies. We assumed that names with a Wikipedia entry would be less ambiguous (because a celebrity tends to monopolize search results), but our Wikipedia names had an average of 23 clusters per name in the WePS-1 training data and twice this amount in the WePS-1 test data. Note that average ambiguity is not very informative for this task, because its distribution does not seem to follow a binomial distribution: a name corresponding with just two people seems as likely as a name with 30 or a name with 90.

Additionally this data includes the Web03 corpus [18] which features a more diverse number of documents for each name and a lower average ambiguity.

2.1.2 Test data

The WePS-2 test data consists of 30 datasets (Table 1), each one corresponding to one ambiguous name. As in WePS-1, three different sources where used to obtain the names:

Wikipedia. Ten names where randomly sampled from the list of biographies in the English Wikipedia. Unlike the WePS-1 dataset, in this occasion our hypothesis of lower ambiguity for names in the Wikipedia has a correspondance with the data: as we can see in the results of the manual annotation (table 1), out of these ten datasets, six contain less than ten different people and three of them are dominated by only one person.

ACL'08. Another ten names where randomly extracted from the list of Programme Committee members for the annual meeting of the Association for Computational Linguistics (ACL'08). These cases present a different type of ambiguity scenario, where we know in advance that at least one of the people mentioned should be a Computer Science scholar.

US Census. Using the lists of first and last names in the 1990 US Census[1] data, we composed 10 random names. In order to avoid extremely rare or unexistent name combinations, we weighted the probability of choosing a name according to its frequency in the Census. The result is a set of fairly ambiguous names, with an average of 30 different people mentioned in each dataset.

For each name we obtained the top 150 search results from an Internet search engine[2] (using the name as a quoted query and searching only for pages written in English). All the information obtained from the search results (snippets, position in the ranking, document title, original URL) was stored and distributed to the participants as part of the datasets.

Some web pages from the search results were not included in the final corpus. In some cases, pages could not be downloaded or were not available at the time the corpus was created. We also filtered out documents that did not contain at least one occurrence of the person name. Finally, in order to simplify the preprocessing task for participants, only HTML pages were included in the datasets.

2.1.3 Manual Annotation

Each dataset was annotated independently by two different assessors. Each one was requested to manually group the search results so that each group contained all and only those documents referring to one of the individuals sharing the ambiguous name. Once the first 100 document were grouped, the annotation stopped. A web application (fig. 1) was developed to ease the annotation process, allowing the annotator to quickly browse and edit the clusters as well as leaving comments regarding specific clusters or documents.

It was allowed to assign the same document to more than one cluster whenever necessary; for instance, a web page with results from Amazon might be a list of books written by two different authors sharing the same name; in this case, the annotator was supposed to assign the page to two different clusters.

In some cases, it was not possible to decide which individual was mentioned in a page, and the page was discarded from the dataset. These are the most frequent reasons to discard a page:

- Unclear pages: in general pages that don't offer enough information. This is often the case for Facebook, Linkedin and other public profiles from social networks. For instance, a facebook public profile may just contain the name and a statement such as "I like movies and chocolate". This information is compatible with virtually any cluster, and is therefore not useful to resolve ambiguity.

- Genealogy pages: These documents are simply too complex to treat: they are too long and contain large genealogy trees which are hard to compare with other web pages.

[1] http://www.census.gov/genealogy/names/

[2] We used the Web Search Service API provided by Yahoo! (http://developer.yahoo.com/search/)

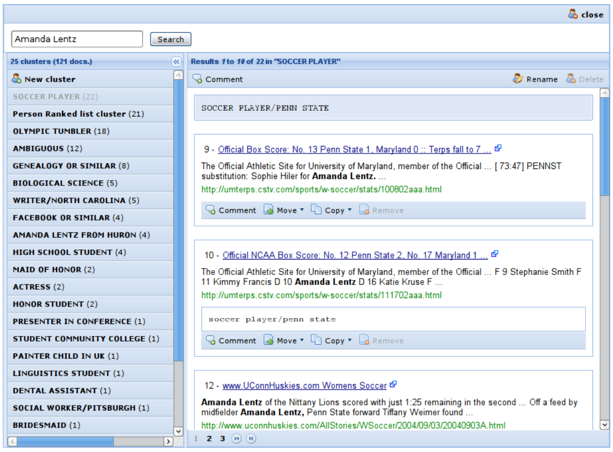
Name	**entities**	**documents**	**discarded**
Wikipedia names			
Bertram Brooker	1	128	30
David Tua	1	134	36
Franz Masereel	3	126	26
Herb Ritts	2	127	31
James Patterson	4	133	33
Jason Hart	22	130	38
Louis Lowe	24	100	25
Mike Robertson	39	123	35
Nicholas Maw	1	135	36
Tom Linton	10	135	41
Average	10,70	127,10	33,10
ACL'08 names			
Benjamin Snyder	28	95	40
Cheng Niu	7	100	7
David Weir	26	128	33
Emily Bender	19	120	31
Gideon Mann	2	95	6
Hao Zhang	24	100	13
Hoi Fang	21	90	28
Ivan Titov	5	101	28
Mirella Lapata	2	91	1
Tamer Elsayed	8	101	18
Average	14,20	102,10	20,50
Census names			
Amanda Lentz	20	121	46
Helen Thomas	3	127	27
Janelle Lee	34	93	37
Jonathan Shaw	26	126	46
Judith Schwartz	30	124	39
Otis Lee	26	118	40
Rita Fisher	24	109	13
Sharon Cummings	30	113	29
Susan Jones	56	110	30
Theodore Smith	54	111	43
Average	30,30	115,20	35,00
Global average	18,64	114,42	29,42

Table 1: Test Data

- Pages in languages other than English: these are documents which were incorrectly tagged by the Yahoo language identifier filter, and are mostly in Chinese (e.g. in *Hao Zhang*, *Feng Hui* datasets), Arabic (*Tamer Elsayed*), Norwegian or Finnish (*Ivan Titov*).

Once both assesors had annotated the full dataset, they met, discuss their annotations, and produce a single consensuated manual tagging of the data, which is used as WePS-2 gold standard. The cause of most disagreements was a different interpretation of which facts constitute sufficient evidence to merge a specific page with a given cluster. Frequently there are borderline cases where many interpretations are possible, and it is therefore difficult to establish a general annotation policy.

Occupations are usually a good hint for the task. For example, in the *Benjamin Snyder* dataset we found three documents that mention people living in Boston. One of them is an MIT student in engineering, while other is a lawyer. It is therefore reasonable to assume that they are different people. The third document, however, describes a wine aficionado. But this is a hobby rather than an occupation, and it could be compatible with the previous ones. One of the annotators decided that there wasn't enough information to either create a new cluster for the document or to add it to one of the existing cluster, and therefore discarded the document. The second annotator, however, decided that as there is no information linking the wine aficionado with the other two profiles, it was appropiate to consider it a a separate person. After the discussion, this second option prevailed.



SOCCER PLAYER (22)
Person Ranked list cluster (21)
OLYMPIC TUMBLER (18)
AMBIGUOUS (12)
GENEALOGY OR SIMILAR (8)
BIOLOGICAL SCIENCE (5)
WRITER/NORTH CAROLINA (5)
FACEBOOK OR SIMILAR (4)
AMANDA LENTZ FROM HURON (4)
HIGH SCHOOL STUDENT (4)
MAID OF HONOR (2)
ACTRESS (2)
HONOR STUDENT (2)
PRESENTER IN CONFERENCE (1)
STUDENT COMMUNITY COLLEGE (1)
PAINTER CHILD IN UK (1)
LINGUISTICS STUDENT (1)
DENTAL ASSISTANT (1)
SOCIAL WORKER/PITSBURGH (1)
BRIDESMAID (1)

SOCCER PLAYER/PENN STATE

9 - Official Box Score: No. 13 Penn State 1, Maryland 0 :: Terps fall to 7 ...
The Official Athletic Site for University of Maryland, member of the Official ... [73:47] PENNST substitution: Sophie Hiler for **Amanda Lentz.** ...
http://umterps.cstv.com/sports/w-soccer/stats/100802aaa.html

10 - Official NCAA Box Score: No. 12 Penn State 2, No. 17 Maryland 1 ...
The Official Athletic Site for University of Maryland, member of the Official ... F 9 Stephanie Smith F 11 Kimmy Francis D 10 **Amanda Lentz** D 16 Katie Kruse F ...
http://umterps.cstv.com/sports/w-soccer/stats/111702aaa.html

soccer player/penn state

12 - www.UConnHuskies.com Womens Soccer
Amanda Lentz of the Nittany Lions scored with just 1:25 remaining in the second ... Off a feed by midfielder **Amanda Lentz,** Penn State forward Tiffany Weimer found ...
http://www.uconnhuskies.com/AllStories/WSoccer/2004/09/03/20040903A.html

Figure 1: Clustering annotation GUI

Other conflictive cases are simply derived from the task complexity: sometimes a page provides a large amount of text and the key information is not visible at a first glance. In datasets where there is a high ambiguity, the annotator has to keep all the details of the people he finds, and compare those to the new evidence offered by each new page; eventually this leads to human errors. The strategy of having two independent annotations and then a discussion round helped detecting this kind of errors.

In comparison with WePS-1, the new dataset has a much lower ambiguity: in average, there are 18,64 different people per name, but the predominant person for a given name owns half of the documents. Again, note that averages are not particularly informative, because there are many extreme cases. For instance, there are 3 (out of 30) names with only one person, and 6 cases with 30 or more different people sharing a single name.

There is also a higher number of discarded documents compared to WePS-1. This is due to the more conservative tagging guidelines that prevented many documents from being grouped with insufficient information.

2.2 Evaluation metrics

In this section we explain the shortcomings of the standard clustering metrics (Purity and Inverse Purity) used for WePS-1 and describe the evaluation metrics adopted in WePS-2.

2.2.1 Lessons learned in WePS-1

In WePS-1, systems where evaluated using the standard clustering metrics Purity and Inverse Purity [4], and the manual annotation as gold standard. During this first evaluation, Paul Kalmar (one of the participants) noticed[3] that it was possible to use a cheat system to obtain a maximal value of Inverse Purity together with a high score of Purity.

[3] Discussion forum of Web People Search Task 2007 (Mar 23th 2007) http://groups.google.com/group/web-people-search-task—semeval-2007/

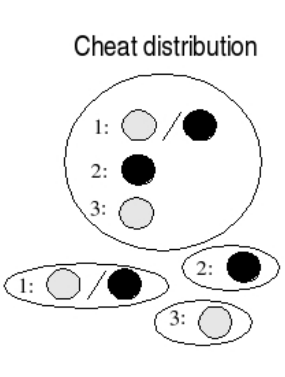


Figure 2: Output of a cheat system

Purity		I.Purity		$F_{\alpha=0,5}$	
S4	0,81	**Cheat S**	1,00	S1	0,79
S3	0,75	S14	0,95	**Cheat S**	0,78
S2	0,73	S13	0,93	S3	0,77
S1	0,72	S15	0,91	S2	0,77
Cheat S	0,64	S5	0,90	S4	0,69
S6	0,60	S10	0,89	S5	0,67
S9	0,58	S7	0,88	S6	0,66
S8	0,55	S1	0,88	S7	0,64
S5	0,53	S12	0,83	S8	0,62
S7	0,50	S11	0,82	S9	0,61
S10	0,45	S2	0,82	S10	0,6
S11	0,45	S3	0,80	S11	0,58
S12	0,39	S6	0,73	S12	0,53
S13	0,36	S8	0,71	S13	0,52
S14	0,35	S9	0,64	S14	0,51
S15	0,30	S4	0,60	S15	0,45

Table 2: WEPS-1 systems ranking according to *Purity*, *InversePurity* **and** $F_{\alpha=0,5}$

The key is overlapping clusters: the WePS clustering task allows systems to put a document in several clusters if it refers to multiple people. The cheat system exploits this feature, producing a cluster of size one for each document, plus a cluster containing all the documents. For example, in Figure 2 three elements are shown in the cheat and the correct cluster distributions[4]. Inverse purity is, by definition, perfect in the cheat system output, since all the documents can be found in one cluster. This would usually correspond to a low purity for this noisy cluster. But at this point, by duplicating each document and generating a new singleton cluster, we add many clusters with maximal purity, and hence the average purity is substantially increased. In summary, the cheat system produces a clustering which is useless, but gets a high score.

The results of the cheat system on the WePS-1 testbed show that it is competitive enough to be on the top of the real systems ranking (table 2).

2.2.2 *WePS-2: B-Cubed*

In an extensive study of the different families of clustering metrics [3] we found that B-Cubed [6] metrics are the only ones that satisfy four intuitive formal constraints on evaluation metrics for the clustering problem. We then extended the original B-Cubed definition to handle overlapping clustering.

BCubed metrics independently compute the precision and recall associated to each item in the distribution. The precision of one item represents the amount of items in the same cluster that belong to its category. Analogously, the recall of one item represents how many items from its category appear in its cluster.

B-Cubed metrics need to be extended to represent the correctness of the relation between two items in an overlapping clustering, where the multiplicity of item occurrences in clusters and categories must be taken into account. For instance, if two items share two categories and share just one cluster, then the clustering is not capturing completely the relation between both items. On the other hand, if two items share three clusters but just two categories, then the clustering is introducing more information than necessary.

These new aspects can be measured in terms of *precision* and *recall* between two items.

$$\text{Mult. Precision}(e,e') = \frac{\text{Min}(|C(e)\cap C(e')|, |L(e)\cap L(e')|)}{|C(e)\cap C(e')|}$$

$$\text{Mult. Recall}(e,e') = \frac{\text{Min}(|C(e)\cap C(e')|, |L(e)\cap L(e')|)}{|L(e)\cap L(e')|}$$

e and e' are two items, $L(e)$ the set of categories and $C(e)$ the set of clusters associated to e. Note that Multiplicity Precision is defined only when e, e' share some cluster, and Multiplicity Recall when e, e' share some category. This is enough to define Bcubed extensions. Multiplicity Precision is used when two items share one or more clusters, and it is maximal (1) when the number of shared categories is lower or equal than the number of shared clusters, and it is minimal (0) when the two items do not share any category. Reversely, Multiplicity Recall is used when two items share one or more categories, and it is maximal when the the number of shared clusters is lower or equal to the number of shared categories, and it is minimal when the two items do not share any cluster.

The next step is integrating multiplicity precision and recall into the overall BCubed metrics. For this, the original Bcubed definitions are used, but replacing the *Correctness* function with multiplicity precision (for Bcubed precision) and multiplicity Recall (for Bcubed recall). Then, the extended Bcubed precision associated to one item will be its averaged multiplicity precision over other items sharing some of its categories; and the overall **extended Bcubed precision** will be the averaged precision of all items. The **extended BCubed recall** is obtained using the same procedure. Formally:

$$\text{Pre. BCubed} = \text{Avg}_e[\text{Avg}_{e'.C(e)\cap C(e')\neq\emptyset}[\text{Mult. precision}(e,e')]]$$

$$\text{Recall BCubed} = \text{Avg}_e[\text{Avg}_{e'.L(e)\cap L(e')\neq\emptyset}[\text{Mult. recall}(e,e')]]$$

The harmonic mean ($F_{\alpha=0,5}$) of B-Cubed precision and recall was used for the ranking of systems.

[4] The clustering is represented as follows: (i) elements in the clustering are identified by the numbers, (ii) the colored circles indicate to which class each element belongs in the gold standard (iii) more than one colored circle next to an element means that the element belongs to multiple classes in the gold standard

$$F = \frac{1}{\alpha\frac{1}{\text{B-Cubed precision}} + (1-\alpha)\frac{1}{\text{B-Cubed recall}}}$$

We also include the results for $F_{\alpha=0,2}$. This additional measure rewards a better recall while still considering the precision aspect. This parametrization of the F measure captures the intuition that filtering out a few noisy documents from the relevant cluster (i.e. having a problem of precision) is more acceptable to the user than having to inspect all other clusters in search for missing information (i.e. having a problem of recall in the relevant cluster).

2.3 Baselines

As in WePS-1, two simple baseline approaches were applied to the test data (fig. 3. The *ALL-IN-ONE* baseline provides a clustering solution where all the documents are assigned to a single cluster. This has the effect of always achieving the highest score in the *inverse purity* measure, because all classes have their documents in a single cluster. On the other hand, the *purity* measure will be equal to the *precision* of the predominant class in that single cluster. The *ONE-IN-ONE* baseline gives another extreme clustering solution, where every document is assigned to a different cluster. In this case *purity* always gives its maximum value, while *inverse purity* will decrease with larger classes.

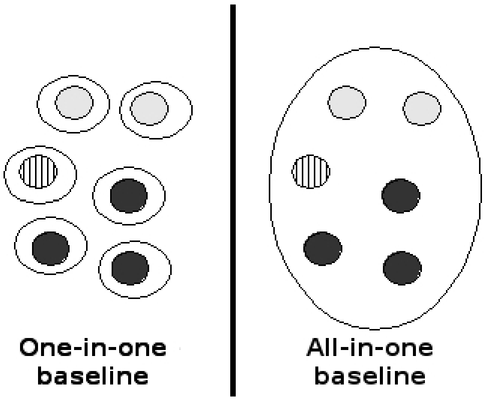


Figure 3: Baseline systems

A third baseline consists of a simple clustering system. It consists of a hierarchical agglomerative clustering algorithm with single linkage. Similarity is calculated using cosine, documents are represented by a bag of words and weighted with tf*idf. We evaluated two variations of this baseline, one that uses a BoW of tokens in the document (*HAC-TOKENS*) and other that uses bigrams (*HAC-BIGRAMS*). English stopwords and bigrams containing a stopword are removed in each case. A fixed similarity threshold is obtained from the training data and then applied to the test data.

The selection of an appropiate threshold (or other clustering stop criteria) is a challenging issue [21] in the WePS task. In order to provide an upper bound of the results that can be achieved with the previous baseline system and a perfect threshold selection, we have evaluated the results obtained when the best threshold is selected for each topic (i.e. each person name) - *BEST-HAC-TOKENS* and *BEST-HAC-BIGRAMS*-. This is not a baseline, but it gives us an insight on the importance of the clustering threshold, and how much it can improve the results for simple baseline systems like *HAC-TOKENS* and *HAC-BIGRAMS*.

Finally, we have also included the WePS-1 cheat system in the results, in order to verify that the new metrics detect and penalize this non-informative baseline.

2.4 Campaign design

The schedule for the evaluation campaign was set as follows: (i) release of the task description, development data sets and scoring program; (ii) release of the test data sets; (iii) participants sent their answers to the task organizers[5]; (iv) the task organizers evaluate the answers and send the results.

The task description was released before the start of the official evaluation. The training data and evaluation scripts were already available since they were produced and published during the WePS-1 campaign. Three months were given for the participants to develop their systems.

The official evaluation period started with the release of the test datasets. These datasets included the search results metadata and HTML documents for each dataset. Participants had one week to run their systems and submit up to five different sets of answers to the organizers.

Once the evaluation period finished, the answers were evaluated by the organizers and the detailed results submitted to each team. The ranking with all the submitted runs was then made public, as well as the gold standard for the test data.

3. PARTICIPANTS

In this section we summarize some of the common traits found in the systems developed by the participants.

All systems have some preprocessing stage where HTML documents are converted into plain text (Java HTML Parser[6] and Beautiful Soup[7] were among the most popular tools used for this purpose). The next steps generally involve tokenization, stop-word removal and, in some cases, sentence detection (e.g. [8, 12, 23, 11]). Porter stemming was performed by some teams [7, 10, 16, 19, 29], although it is not clear to which extent it affects the clustering results (some of the top systems use it and some don't).

The most commonly used feature is the bag of words, although in some cases it was restricted to sentences where the ambiguous name occurs within a window of words around [12, 23, 14]. Named entity recognition (NER) is usually presented as an important feature in order to obtain accurate clustering results. Indeed it is the second most commonly used feature [12, 20, 14, 11, 22, 9][8]. Surprisingly, three out of the top four systems in the ranking didn't use NER [8, 7, 12, 23]. Apparently NER is not necessary to build a competitive system, although it may still be a valuable source of information. In most cases, features were weighted using simple tf*idf functions; other measures used are the gain ratio [17], Kullback-Leibler divergence [19] and self information [23].

Bigrams were used by [8, 23] and seem to provide a good tradeoff between precision and recall. Among the most sparse

[5] Participants where allowed to submit up to five runs.

[6] http://htmlparser.sourceforge.net/

[7] http://www.crummy.com/software/BeautifulSoup/

[8] The most used tool was Stanford's NER software (http://nlp.stanford.edu/ner/).

features we find systems that used hyperlinks [12, 9, 8, 17], email addresses, phone numbers, dates [17], variations of the ambiguous name [9], etc.

Hierarchical Agglomerative Clustering (HAC) is the most popular clustering algorithm, although the choice of linkage varies (e.g. single link in [8, 20], group average in [12]). In cases where a hierarchical algorithm like HAC was used, the number of clusters in the output was usually determined by a fixed similarity threshold. This threshold determines how close two elements (documents or clusters) must be in order to be grouped together. Two teams [17, 29] used a relatively novel clustering algorithm: Fuzzy ants clustering [26]. This algorithm determines by itself the number of clusters without the need of a similarity threshold. The similarity measure between documents was commonly handled using the cosine of feature vectors.

Only four teams considered duplication of documents, but this feature didn't play an important role on the WePS-2 data.

4. RESULTS OF THE EVALUATION

We were contacted from 32 teams expressing their interest in the clustering task. Out of them, 17 teams submitted a total of 75 different runs.

4.1 Results and discussion

Table 3 presents the results of the 17 participants and the 3 baseline systems. F values are macro-averaged, i.e., F is computed for every test case, and then averaged over all test cases. When a team submitted multiple runs, we have chosen the run with best score as the team representative in the ranking.

The results according to B-Cubed metrics are shown in Table 3. It is worth noticing that

- There are subtle differences, but no major ranking swaps between Bcubed and Purity-Inverse Purity rankings (see Table 4). The exception is the cheat system baseline, which is no longer one of the best systems according to the new metrics.

- The first three teams have similar performance in terms of $F_{0,5}$. Out of them, UVA_1 has the most balanced result (0,85 precision, 0,80 recall), and ITC-UT_1 is more precision-oriented (0,93 precision, 0,73 recall), therefore it gets penalized in the $F_{0,2}$ measure.

- The same team (PolyUHK, which was CU-COMSEM in WePS-1) obtained the best score both in WePS-1 and WePS-2.

- Five teams fall below the ALL_IN_ONE baseline. Although all of them reach higher precision values than the baseline, their combined F measure cannot compensate the perfect recall of that baseline. Note that the ALL_IN_ONE baseline has a BCubed precision which is higher than expected (0,43), because in average half of the documents in every test set belong to one single person. In the WePS-1 dataset documents were more evenly distributed among the clusters, and therefore the ONE_IN_ONE baseline gave better results than the ALL_IN_ONE baseline strategy.

- The *HAC-TOKENS* and *HAC-BIGRAMS* baselines obtain high precision but a poor recall, which situates them in the middle of the ranking. Bigrams improve the recall while mantaining a high precision, and so achieves a better combined score. Note that these baseline systems might be oriented towards precision because the WePS-1 dataset had, in average, very small clusters.

- The upper bound systems *BEST-HAC-TOKENS* and *BEST-HAC-BIGRAMS* obtain excellent results, a bit better than the three best teams. This seems to be an indication that the best scoring systems are doing a good job, because they nearly match the behaviour of Oracle systems which know which is the best clustering threshold for each test instance. On the other hand, this result also suggests that improving the selection of the clustering threshold may lead to competitive results even with a naive clustering approach.

		Macro-averaged Scores			
		F-measures		B-Cubed	
rank	run	$\alpha =_{,5}$	$\alpha =_{,2}$	Pre.	Rec.
	BEST-HAC-TOKENS	,85	,84	,89	,83
	BEST-HAC-BIGRAMS	,85	,83	,91	,81
1	PolyUHK	,82	,80	,87	,79
2	UVA_1	,81	,80	,85	,80
3	ITC-UT_1	,81	,76	,93	,73
4	XMEDIA_3	,72	,68	,82	,66
5	UCI_2	,71	,77	,66	,84
6	LANZHOU_1	,70	,67	,80	,66
7	FICO_3	,70	,64	,85	,62
8	UMD_4	,70	,63	,94	,60
	HAC-BIGRAMS	,67	,59	,95	,55
9	UGUELPH_1	,63	,75	,54	,93
10	CASIANED_4	,63	,68	,65	,75
	HAC-TOKENS	,59	,52	,95	,48
11	AUG_4	,57	,56	,73	,58
12	UPM-SINT_4	,56	,59	,60	,66
	ALL_IN_ONE	,53	,66	,43	1,00
	CHEAT_SYS	,52	,65	,43	1,00
13	UNN_2	,52	,48	,76	,47
14	ECNU_1	,41	,44	,50	,55
15	UNED_3	,40	,38	,66	,39
16	PRIYAVEN	,39	,37	,61	,38
	ONE_IN_ONE	,34	,27	1,00	,24
17	BUAP_1	,33	,27	,89	,25

Table 3: Official team ranking using B-Cubed measures

4.2 Robustness of F results with different α values

The grouping thresholds chosen by the clustering systems imply a trade-off choice between BCubed Recall and Precision; systems tend to achieve better results according to one metric at the cost of the other. Therefore, the system ranking can suffer drastic changes depending on the α parameter chosen for the F combining function, given that it determines the relative weight assigned to Precision and Recall. This phenomenon is discussed in detail in [2]. In that paper, the UIR (*Unanimous Improvement Ratio*) measure is proposed in order to check what extent the detection of a system improvement is biased by the metric weighting scheme (i.e. the α parameter in our case).

		Macro-averaged Scores			
		F-measures			
rank	run	$\alpha =,_5$	$\alpha =,_2$	Pur	Inv_Pur
	BEST-HAC-TOKENS	,90	,89	,93	,88
	BEST-HAC-BIGRAMS	,90	,87	,94	,86
1	PolyUHK	,88	,87	,91	,86
2	UVA_1	,87	,87	,89	,87
3	ITC-UT_1	,87	,83	,95	,81
	CHEAT_SYS	,87	,94	,78	1,00
4	UMD_4	,81	,76	,95	,72
5	XMEDIA_3	,80	,76	,91	,73
6	UCI_2	,80	,84	,75	,89
7	LANZHOU_1	,80	,78	,85	,77
8	FICO_3	,80	,76	,90	,73
	HAC-BIGRAMS	,78	,64	,96	,67
9	UGUELPH_1	,74	,84	,64	,95
10	CASIANED_4	,73	,77	,72	,83
	HAC-TOKENS	,71	,64	,96	,60
11	AUG_4	,69	,68	,79	,68
12	UPM-SINT_4	,67	,70	,69	,74
	ALL_IN_ONE	,67	,79	,56	1,00
13	UNN_2	,64	,59	,80	,57
14	ECNU_1	,53	,56	,60	,63
15	PRIYAVEN	,53	,49	,71	,48
16	UNED_3	,51	,48	,71	,48
17	BUAP_1	,37	,30	,89	,27
	ONE_IN_ONE	,34	,27	1,00	,24

Table 4: Team ranking using Purity measures

UIR is able to combine two evaluation metrics without assigning any a priori relative relevance. This measure considers three variables: (1) the number of topics $T_{\forall m.a \geq b}$ in which System a improves or equals b according to both metrics, (2) the number of topics $T_{\forall m.b \geq a}$ in which System b improves or equals a for both metrics and (3) the total number of topics T in the testbed.

$$UIR(a,b) = \frac{T_{\forall m.a \geq b} - T_{\forall m.b \geq a}}{T}$$

Consecuently, the more one metric is improved at the cost of the other $(T - T_{\forall m.a \geq b} - T_{\forall m.b \geq a})$, the more UIR decreases. The theoretical foundations of MIR are reported in [2]. We have considered MIR=0,25 as threshold for significance: if $UIR(a,b) \geq 0,25$, then a is consistently better than b regardless of how we weight the contributions of precision and recall. This threshold represents that the diference between cases of quality increase (for both metrics) and quality decrease is bigger than the 25% of the total number of topics. The justification for this threshold is described in [2].

Table 4.2 shows the results of applying UIR to the WePS-2 systems. The third column represents the set of systems that are improved by the corresponding system with UIR>0.25. The fourth column represents the *reference system*, defined as, given a system a, the system that improves a with maximum UIR. It represents the system with which a should be replaced in order to improve results without sacrificing any partial evaluation metric. Finally, the last column represents the UIR between the system and its reference.

This table adds new insights into the evaluation process. First of all, note that, although the three top-scoring systems have a similar performance in terms of F (0,82, 0,81 and 0,81), PolyUHK is consistently the best according to UIR (it is the reference for 10 systems). In the most extreme case, UIR(PolyUHK,PRIYAVEN)=1, which means that PolyUHK improves both precision and recall of PRIYAVEN for all test cases in the dataset.

Note also that, although the ALL_IN_ONE baseline is better than five systems according to F, it is not better than any of them according to UIR. In fact, only the ONE_IN_ONE baseline is able to improve some system (BUAP_2).

5. CONCLUSIONS

The WePS-2 campaign has maintained the high level of participation achieved in its first edition. This campaign has also featured more robust clustering evaluation measures and a more efficient annotation process, based on the experience acquired in WePS-1. The results of the evaluation have also been analyzed using a novel approach (Unanimous Improvement Ratio [2]) that tackles the bias introduced by metric weighting schemes. The datasets annotated for this evaluation (now part of WePS testbed), as well as an updated evaluation script, have been made publicly available[9].

In parallel with the clustering task, WePS-2 also included a new person Attribute Extraction task [27]. The extraction of biographical attributes can be a valuable source of information for accurate document clustering, but also an important aid for real users to browse the clustering results. In a forthcoming edition of the WePS campaign we hope to address the relations between both tasks in an integrated way. This should provide new insights into the challenges faced by WePS systems.

6. ACKNOWLEDGMENTS

This work has been partially supported with a grant from the Spanish government (project Text-Mess) and with a grant from the European Comission (project TrebleCLEF). We are also indebted to Roberto Coromina and Douglas Briceno, our human assessors, for their thorough work, and to the WePS steering committee members for their helpful suggestions along the evaluation campaign.

7. REFERENCES

[1] E. Agirre and P. Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications.* Springer, 2006.

[2] E. Amigó, J. Gonzalo, and J. Artiles. Combining evaluation metrics via the unanimous improvement ratio and its application in weps clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[3] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2008.

[4] J. Artiles, J. Gonzalo, and S. Sekine. 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).* ACL, 2007.

[5] J. Artiles, J. Gonzalo, and F. Verdejo. A testbed for people searching strategies in the www. In *SIGIR*, 2005.

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

System	$F_{0.5}$	Improved systems (UIR > 0.25)	Reference system	UIR for the reference system
PolyUHK (S1)	0,82	S2 S4 S6 S7 S8 S11..S17 B_1	-	-
ITC-UT_1 (S2)	0,81	S4 S6 S7 S8 S11..S17 B_1	S1	0,26
UVA_1 (S3)	0,81	S2 S4 S7 S8 S11..S17 B_1	-	-
XMEDIA_3 (S4)	0,72	S11 S13..S17	S1	0,58
UCI_2 (S5)	0,71	S12..S16	-	-
UMD_4 (S6)	0,70	S4 S7 S11 S13..S17 B_1	S1	0,35
FICO_3 (S7)	0,70	S11 S13..S17	S2	0,65
LANZHOU_1 (S8)	0,70	S11..S17	S1	0,74
UGUELPH_1 (S9)	0,63	S4 S12 S14 S16	-	-
CASIANED_5 (S10)	0,63	S12..S16	-	-
AUG_4 (S11)	0,57	S14.. S17	S3	0,68
UPM-SINT_1 (S12)	0,56	S14 S16	S1	0,71
ALL_IN_ONE_BASELINE (B_{100})	0,53	B_{Cheat}	-	-
UNN_2 (S13)	0,52	S15 S16	S1	0,9
CHEAT_SYS (B_{Cheat})	0,52	-	B_{100}	0,65
ECNU_1 (S14)	0,41	-	S1	0,9
UNED_3 (S15)	0,40	S16	S1	0,97
PRIYAVEN (S16)	0,39	-	S1	1
ONE_IN_ONE_BASELINE (B_1)	0,34	S17	S1	0,29
BUAP_2 (S17)	0,33	-	S6	0,84

Table 5: UIR results for WEPS2 systems using Bcubed Precision and Recall metrics.

[6] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics*. ACL, 1998.

[7] K. Balog, J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. de Rijke. The university of amsterdam at weps2. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[8] Y. Chen, S. Y. M. Lee, and C.-R. Huang. Polyuhk: A robust information extraction system for web personal names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[9] J. Gong and D. Oard. Determine the entity number in hierarchical clustering for web personal name disambiguation. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[10] J. C. González, P. Maté, L. Vadillo, R. Sotomayor, and A. Carrera. Learning by doing: A baseline approach to the clustering of web people search results. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[11] X. Han and J. Zhao. Casianed: Web personal name disambiguation based on professional categorization. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[12] M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. Person name disambiguation on the web by twostage clustering. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[13] D. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse. a web-querying approach to web people search. In *Proc. of Annual International ACM SIGIR Conference*, Singapore, July 20–24 2008.

[14] P. Kalmar and D. Freitag. Features for web person disambiguation. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[15] Z. Kozareva, R. Moraliyski, and G. Dias. Web people search with domain ranking. In *TSD '08: Proceedings of the 11th international conference on Text, Speech and Dialogue*, pages 133–140, Berlin, Heidelberg, 2008. Springer-Verlag.

[16] M. Lan, Y. Z. Zhang, Y. Lu, J. Su, and C. L. Tan. Which who are they? people attribute extraction and disambiguation in web search results. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[17] E. Lefever, T. Fayruzov, V. Hoste, and M. De Cock. Fuzzy ants clustering for web people search. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[18] G. Mann. *Multi-Document Statistical Fact Extraction and Fusion.* PhD thesis, Johns Hopkins University, 2006.

[19] J. Martínez-Romo and L. Araujo. Web people search disambiguation using language model techniques. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[20] R. Nuray-Turan, Z. Chen, D. Kalashnikov, and S. Mehrotra. Exploiting web querying for web people search in weps2. In *Rabia Nuray-Turan Zhaoqi Chen Dmitri V. Kalashnikov Sharad Mehrotra*, 2009.

[21] T. Pedersen and A. Kulkarni. Automatic cluster stopping with criterion functions and the gap statistic. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*

Technology, pages 276–279, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[22] D. Pinto, M. Tovar, D. Vilario, H. Díaz, and H. Jiménez-Salazar. An unsupervised approach based on fingerprinting to the web people search task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[23] L. Romano, K. Buza, C. Giuliano, and L. Schmidt-Thieme. Xmedia: Web people search by clustering with machinely learned similarity measures. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[24] H. Saggion. Experiments on semantic-based clustering for cross-document coreference. In *International Joint Conference on Natural language Processing.*, 2008.

[25] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 499–506, New York, NY, USA, 2008. ACM.

[26] S. Schockaert, M. De Cock, C. Cornelis, and E. Kerre. Clustering web search results using fuzzy ants. volume 22, pages 455–474, 2007.

[27] S. Sekine and J. Artiles. Weps2 attribute extraction task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[28] A. Spink, B. Jansen, and J. Pedersen. Searching for people on web search engines. *Journal of Documentation*, 60:266 – 278, 2004.

[29] P. Venkateshan. Clustering web people search results using fuzzy ant- based clustering. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.