The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task

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

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

Satoshi Sekine
Computer Science Department
New York University, USA
sekine@cs.nyu.edu
nlp.cs.nyu.edu/sekine

Abstract

This paper presents the task definition, resources, participation, and comparative results for the Web People Search task, which was organized as part of the SemEval-2007 evaluation exercise. This task consists of clustering a set of documents that mention an ambiguous person name according to the actual entities referred to using that name.

1 Introduction

Finding information about people in the World Wide Web is one of the most common activities of Internet users. Person names, however, are highly ambiguous. In most cases, the results for a person name search are a mix of pages about different people sharing the same name. The user is then forced either to add terms to the query (probably losing recall and focusing on one single aspect of the person), or to browse every document in order to filter the information about the person he is actually looking for.

In an ideal system the user would simply type a person name, and receive search results clustered according to the different people sharing that name. And this is, in essence, the WePS (Web People Search) task we have proposed to SemEval-2007 participants: systems receive a set of web pages (which are the result of a web search for a person name), and they have to cluster them in as many sets as entities sharing the name. This task has close links with Word Sense Disambiguation (WSD), which is generally formulated as the task of deciding which sense a word has in a given context. In both cases, the problem addressed is the resolution of the ambiguity in a natural language expression. A couple of differences make our problem different. WSD is usually focused on open-class words (common nouns, adjectives, verbs and adverbs). The first difference is that boundaries between word senses in a dictionary are often subtle or even conflicting, making binary decisions harder and sometimes even useless depending on the application. In contrast, distinctions between people should be easier to establish. The second difference is that WSD usually operates with a dictionary containing a relatively small number of senses that can be assigned to each word. Our task is rather a case of Word Sense Discrimination, because the number of "senses" (actual people) is unknown *a priori*, and it is in average much higher than in the WSD task (there are 90,000 different names shared by 100 million people according to the U.S. Census Bureau).

There is also a strong relation of our proposed task with the Co-reference Resolution problem, focused on linking mentions (including pronouns) in a text. Our task can be seen as a co-reference resolution problem where the focus is on solving inter-document co-reference, disregarding the linking of all the mentions of an entity inside each document.

An early work in name disambiguation (Bagga and Baldwin, 1998) uses the similarity between documents in a Vector Space using a "bag of words" representation. An alternative approach by Mann and Yarowsky (2003) is based on a rich feature space of automatically extracted biographic information. Fleischman and Hovy (2004) propose a Maximum Entropy model trained to give the probability that

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 64–69,
Prague, June 2007. ©2007 Association for Computational Linguistics

two names refer to the same individual [1].

The paper is organized as follows. Section 2 provides a description of the experimental methodology, the training and test data provided to the participants, the evaluation measures, baseline systems and the campaign design. Section 3 gives a description of the participant systems and provides the evaluation results. Finally, Section 4 presents some conclusions.

2 Experimental Methodology

2.1 Data

Following the general SemEval guidelines, we have prepared trial, training and test data sets for the task, which are described below.

2.1.1 Trial data

For this evaluation campaign we initially delivered a trial corpus for the potential participants. The trial data consisted of an adapted version of the WePS corpus described in (Artiles et al., 2006). The predominant feature of this corpus is a high number of entities in each document set, due to the fact that the ambiguous names were extracted from the most common names in the US Census. This corpus did not completely match task specifications because it did not consider documents with internal ambiguity, nor it did consider non-person entities; but it was, however, a cost-effective way of releasing data to play around with. During the first weeks after releasing this trial data to potential participants, some annotation mistakes were noticed. We preferred, however, to leave the corpus "as is" and concentrate our efforts in producing clean training and test datasets, rather than investing time in improving trial data.

2.1.2 Training data

In order to provide different ambiguity scenarios, we selected person names from different sources:

US Census. We reused the Web03 corpus (Mann, 2006), which contains 32 names randomly picked from the US Census, and was well suited for the task.

Wikipedia. Another seven names were sampled from a list of ambiguous person names in the English Wikipedia. These were expected to have a few predominant entities (popular or historical), and therefore a lower ambiguity than the previous set.

ECDL. Finally, ten additional names were randomly selected from the Program Committee listing of a Computer Science conference (ECDL 2006). This set offers a scenario of potentially low ambiguity (computer science scholars usually have a stronger Internet presence than other professional fields) with the added value of the *a priori* knowledge of a domain specific type of entity (scholar) present in the data.

All datasets consist of collections of web pages obtained from the 100 top results for a person name query to an Internet search engine [2]. Note that 100 is an upper bound, because in some occasions the URL returned by the search engine no longer exists.

The second and third datasets (developed explicitly for our task) consist of 17 person names and 1685 associated documents in total (99 documents per name in average). Each web page was downloaded and stored for off-line processing. We also stored the basic metadata associated to each search result, including the original URL, title, position in the results ranking and the corresponding snippet generated by the search engine.

In the process of generating the corpus, the selection of the names plays an important role, potentially conditioning the degree of ambiguity that will be found later in the Web search results. The reasons for this variability in the ambiguity of names are diverse and do not always correlate with the straightforward census frequency. A much more decisive feature is, for instance, the presence of famous entities sharing the ambiguous name with less popular people. As we are considering top search results, these can easily be monopolized by a single entity that is popular in the Internet.

After the annotation of this data (see section 2.1.4.) we found our predictions about the average ambiguity of each dataset not to be completely accurate. In Table 1 we see that the ECDL-06 average ambiguity is indeed relatively low (except for the documents for "Thomas Baker" standing as the most ambiguous name in the whole training). Wikipedia names have an average ambiguity of 23,14 entities

[1] For a comprehensive bibliography on person name disambiguation refer to http://nlp.uned.es/weps

[2] We used the Yahoo! API from Yahoo! Search Web Services (http://developer.yahoo.com/search/web/).

Name	entities	documents	discarded
Wikipedia names			
John Kennedy	27	99	6
George Clinton	27	99	6
Michael Howard	32	99	8
Paul Collins	37	98	6
Tony Abbott	7	98	9
Alexander Macomb	21	100	14
David Lodge	11	100	9
Average	23,14	99,00	8,29
ECDL-06 Names			
Edward Fox	16	100	36
Allan Hanbury	2	100	32
Donna Harman	7	98	6
Andrew Powell	19	98	48
Gregory Crane	4	99	17
Jane Hunter	15	99	59
Paul Clough	14	100	35
Thomas Baker	60	100	31
Christine Borgman	7	99	11
Anita Coleman	9	99	28
Average	15,30	99,20	30,30
WEB03 Corpus			
Tim Whisler	10	33	8
Roy Tamashiro	5	23	6
Cynthia Voigt	1	405	314
Miranda Bollinger	2	2	0
Guy Dunbar	4	51	34
Todd Platts	2	239	144
Stacey Doughty	1	2	0
Young Dawkins	4	61	35
Luke Choi	13	20	6
Gregory Brennan	32	96	38
Ione Westover	1	4	0
Patrick Karlsson	10	24	8
Celeste Paquette	2	17	2
Elmo Hardy	3	55	15
Louis Sidoti	2	6	3
Alexander Markham	9	32	16
Helen Cawthorne	3	46	13
Dan Rhone	2	4	2
Maile Doyle	1	13	1
Alice Gilbreath	8	74	30
Sidney Shorter	3	4	0
Alfred Schroeder	35	112	58
Cathie Ely	1	2	0
Martin Nagel	14	55	31
Abby Watkins	13	124	35
Mary Lemanski	2	152	78
Gillian Symons	3	30	6
Pam Tetu	1	4	2
Guy Crider	2	2	0
Armando Valencia	16	79	20
Hannah Bassham	2	3	0
Charlotte Bergeron	5	21	8
Average	5,90	47,20	18,00
Global average	10,76	71,02	26,00

Table 1: Training Data

per name, which is higher than for the ECDL set. The WEB03 Corpus has the lowest ambiguity (5,9 entities per name), for two reasons: first, randomly picked names belong predominantly to the long tail of unfrequent person names which, *per se*, have low ambiguity. Being rare names implies that in average there are fewer documents returned by the search engine (47,20 per name), which also reduces the possibilities to find ambiguity.

2.1.3 Test data

For the test data we followed the same process described for the training. In the name selection we tried to maintain a similar distribution of ambiguity degrees and scenario. For that reason we randomly extracted 10 person names from the English Wikipedia and another 10 names from participants in the ACL-06 conference. In the case of the US census names, we decided to focus on relatively common names, to avoid the problems explained above.

Unfortunately, after the annotation was finished (once the submission deadline had expired), we found a major increase in the ambiguity degrees (Table 2) of all data sets. While we expected a raise in the case of the US census names, the other two cases just show that there is a high (and unpredictable) variability, which would require much larger data sets to have reliable population samples.

This has made the task particularly challenging for participants, because naive learning strategies (such as empirical adjustment of distance thresholds to optimize standard clustering algorithms) might be misleaded by the training set.

2.1.4 Annotation

The annotation of the data was performed separately in each set of documents related to an ambiguous name. Given this set of approximately 100 documents that mention the ambiguous name, the annotation consisted in the manual clustering of each document according to the actual entity that is referred on it.

When non person entities were found (for instance, organization or places named after a person) the annotation was performed without any special rule. Generally, the annotator browses documents following the original ranking in the search results; after reading a document he will decide whether the mentions of the ambiguous name refer to a new entity or to a entity previously identified. We asked the annotators to concentrate first on mentions that strictly contained the search string, and then to pay attention to the co-referent variations of the name. For instance "John Edward Fox" or "Edward Fox Smith" would be valid mentions. "Edward J. Fox", however, breaks the original search string, and we do not get into name variation detection, so it will be considered valid only if it is co-referent to a valid

Name	entities	documents	discarded
Wikipedia names			
Arthur Morgan	19	100	52
James Morehead	48	100	11
James Davidson	59	98	16
Patrick Killen	25	96	4
William Dickson	91	100	8
George Foster	42	99	11
James Hamilton	81	100	15
John Nelson	55	100	25
Thomas Fraser	73	100	13
Thomas Kirk	72	100	20
Average	56,50	99,30	17,50
ACL06 Names			
Dekang Lin	1	99	0
Chris Brockett	19	98	5
James Curran	63	99	9
Mark Johnson	70	99	7
Jerry Hobbs	15	99	7
Frank Keller	28	100	20
Leon Barrett	33	98	9
Robert Moore	38	98	28
Sharon Goldwater	2	97	4
Stephen Clark	41	97	39
Average	31,00	98,40	12,80
US Census Names			
Alvin Cooper	43	99	9
Harry Hughes	39	98	9
Jonathan Brooks	83	97	8
Jude Brown	32	100	39
Karen Peterson	64	100	16
Marcy Jackson	51	100	5
Martha Edwards	82	100	9
Neil Clark	21	99	7
Stephan Johnson	36	100	20
Violet Howard	52	98	27
Average	50,30	99,10	14,90
Global average	45,93	98,93	15,07

Table 2: Test Data

mention.

In order to perform the clustering, the annotator was asked to pay attention to objective facts (biographical dates, related names, occupations, etc.) and to be conservative when making decisions. The final result is a complete clustering of the documents, where each cluster contains the documents that refer to a particular entity. Following the previous example, in documents for the name "Edward Fox" the annotator found 16 different entities with that name. Note that there is no *a priori* knowledge about the number of entities that will be discovered in a document set. This makes the task specially difficult when there are many different entities and a high volume of scattered biographical information to take into account.

In cases where the document does not offer enough information to decide whether it belongs to a cluster or is a new entity, it is discarded from the evaluation process (not from the dataset). Another common reason for discarding documents was the absence of the person name in the document, usually due to a mismatch between the search engine cache and the downloaded URL.

We found that, in many cases, different entities were mentioned using the ambiguous name within a single document. This was the case when a document mentions relatives with names that contain the ambiguous string (for instance "Edward Fox" and "Edward Fox Jr."). Another common case of intra-document ambiguity is that of pages containing database search results, such as book lists from Amazon, actors from IMDB, etc. A similar case is that of pages that explicitly analyze the ambiguity of a person name (Wikipedia "disambiguation" pages). The way this situation was handled, in terms of the annotation, was to assign each document to as many clusters as entities were referred to on it with the ambiguous name.

2.2 Evaluation measures

Evaluation was performed in each document set (web pages mentioning an ambiguous person name) of the data distributed as test. The human annotation was used as the gold standard for the evaluation.

Each system was evaluated using the standard *purity* and *inverse purity* clustering measures Purity is related to the *precision* measure, well known in Information Retrieval. This measure focuses on the frequency of the most common category in each cluster, and rewards the clustering solutions that introduce less noise in each cluster. Being C the set of clusters to be evaluated, L the set of categories (manually annotated) and n the number of clustered elements, purity is computed by taking the weighted average of maximal precision values:

$$\text{Purity} = \sum_i \frac{|C_i|}{n} \text{max Precision}(C_i, L_j)$$

where the precision of a cluster C_i for a given category L_j is defined as:

$$\text{Precision}(C_i, L_j) = \frac{|C_i \bigcap L_j|}{|C_i|}$$

Inverse Purity focuses on the cluster with maximum recall for each category, rewarding the clustering solutions that gathers more elements of each category in a corresponding single cluster. Inverse Purity is defined as:

$$\text{Inverse Purity} = \sum_i \frac{|L_i|}{n} \max \text{Precision}(L_i, C_j)$$

For the final ranking of systems we used the harmonic mean of purity and inverse purity $F_{\alpha=0,5}$. The F measure is defined as follows:

$$\text{F} = \frac{1}{\alpha \frac{1}{\text{Purity}} + (1-\alpha)\frac{1}{\text{Inverse Purity}}}$$

$F_{\alpha=0,2}$ is included as an additional measure giving more importance to the inverse purity aspect. The rationale is that, for a search engine user, it should be easier to discard a few incorrect web pages in a cluster containing all the information needed, than having to collect the relevant information across many different clusters. Therefore, achieving a high inverse purity should be rewarded more than having high purity.

2.3 Baselines

Two simple baseline approaches were applied to the test data. 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.

2.4 Campaign design

The schedule for the evaluation campaign was set by the SemEval organisation as follows: (i) release task description and trial data set; (ii) release of training and test; (iii) participants send their answers to the task organizers; (iv) the task organizers evaluate the answers and send the results.

The task description and the initial trial data set were publicly released before the start of the official evaluation.

The official evaluation period started with the simultaneous release of both training and test data, together with a scoring script with the main evaluation measures to be used. This period spanned five weeks in which teams were allowed to register and download the data. During that period, results for a given task had to be submitted no later than 21 days after downloading the training data and no later than 7 days after downloading the test data. Only one submission per team was allowed.

Training data included the downloaded web pages, their associated metadata and the human clustering of each document set, providing a development test-bed for the participant's systems. We also specified the source of each ambiguous name in the training data (Wikipedia, ECDL conference and US Census). Test data only included the downloaded web pages and their metadata. This section of the corpus was used for the systems evaluation. Participants were required to send a clustering for each test document set.

Finally, after the evaluation period was finished and all the participants sent their data, the task organizers sent the evaluation for the test data.

3 Results of the evaluation campaign

29 teams expressed their interest in the task; this number exceeded our expectations for this pilot experience, and confirms the potential interest of the research community in this highly practical problem. Out of them, 16 teams submitted results within the deadline; their results are reported below.

3.1 Results and discussion

Table 3 presents the macro-averaged results obtained by the sixteen systems plus the two baselines on the test data. We found macro-average [3] preferable to micro-average [4] because it has a clear interpretation: if the evaluation measure is F, then we should calculate F for every test case (person name) and then average over all trials. The interpretation of micro-average F is less clear.

The systems are ranked according to the scores obtained with the harmonic mean measure $F_{\alpha=0,5}$ of

[3] Macro-average F consists of computing F for every test set (person name) and then averaging over all test sets.

[4] Micro-average F consists of computing the average P and IP (over all test sets) and then calculating F with these figures.

		Macro-averaged Scores			
		F-measures			
rank	team-id	$\alpha =_{,5}$	$\alpha =_{,2}$	Pur	Inv_Pur
1	CU_COMSEM	,78	,83	,72	,88
2	IRST-BP	,75	,77	,75	,80
3	PSNUS	,75	,78	,73	,82
4	UVA	,67	,62	,81	,60
5	SHEF	,66	,73	,60	,82
6	FICO	,64	,76	,53	,90
7	UNN	,62	,67	,60	,73
8	*ONE-IN-ONE*	*,61*	*,52*	*1,00*	*,47*
9	AUG	,60	,73	,50	,88
10	SWAT-IV	,58	,64	,55	,71
11	UA-ZSA	,58	,60	,58	,64
12	TITPI	,57	,71	,45	,89
13	JHU1-13	,53	,65	,45	,82
14	DFKI2	,50	,63	,39	,83
15	WIT	,49	,66	,36	,93
16	UC3M_13	,48	,66	,35	,95
17	UBC-AS	,40	,55	,30	,91
18	*ALL-IN-ONE*	*,40*	*,58*	*,29*	*1,00*

Table 3: Team ranking

purity and inverse purity. Considering only the participant systems, the average value for the ranking measure was $0,60$ and its standard deviation $0,11$.

Results with $F_{\alpha=0,2}$ are not substantially different (except for the two baselines, which roughly swap positions). There are some ranking swaps, but generally only within close pairs.

The good performance of the *ONE-IN-ONE* baseline system is indicative of the abundance of singleton entities (entities represented by only one document). This situation increases the inverse purity score for this system giving a harmonic measure higher than the expected.

4 Conclusions

The WEPS task ended with considerable success in terms of participation, and we believe that a careful analysis of the contributions made by participants (which is not possible at the time of writing this report) will be an interesting reference for future research. In addition, all the collected and annotated dataset will be publicly available [5] as a benchmark for Web People Search systems.

At the same time, it is clear that building a reliable test-bed for the task is not simple. First of all, the variability across test cases is large and unpredictable, and a system that works well with the names in our test bed may not be reliable in practical, open search situations. Partly because of that, our test-bed happened to be unintentionally challenging for systems, with a large difference between the average ambiguity in the training and test datasets. Secondly, it is probably necessary to think about specific evaluation measures beyond standard clustering metrics such as purity and inverse purity, which are not tailored to the task and do not behave well when multiple classification is allowed. We hope to address these problems in a forthcoming edition of the WEPS task.

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

5 Acknowledgements

This research was supported in part by the National Science Foundation of United States under Grant IIS-00325657 and by a grant from the Spanish government under project Text-Mess (TIN2006-15265-C06). This paper does not necessarily reflect the position of the U.S. Government.

References

Javier Artiles, Julio Gonzalo, and Felisa Verdejo. 2005. A Testbed for People Searching Strategies in the WWW In *Proceedings of the 28th annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 569-570.

Amit Bagga and Breck Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79-85.

Michael B. Fleischman and Eduard Hovy 2004. Multi-document person name resolution. In *Proceedings of ACL-42, Reference Resolution Workshop*.

Gideon S. Mann. 2006. *Multi-Document Statistical Fact Extraction and Fusion* Ph.D. Thesis.

Gideon S. Mann and David Yarowsky 2003. Unsupervised Personal Name Disambiguation In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, pages 33-40.