Web People Search

- Results of the first evaluation and the plan for the second -

Javier Artiles
UNED NLP&IR group
Ciudad Universitaria, s. n.
28040 Madrid , Spain
+34 91 398 8106
javart@bec.uned.es

Satoshi Sekine
New York University
715 Broadway, 7th floor
New York, NY 10003
+1-212-998-3175
sekine@cs.nyu.edu

Julio Gonzalo
UNED NLP&IR group
Ciudad Universitaria, s. n.
28040 Madrid , Spain
+34 91 398 7922
Julio@lsi.uned.es

ABSTRACT

This paper presents the motivation, resources and results for the first Web People Search task, which was organized as part of the SemEval-2007 evaluation exercise. Also, we will describe a survey and proposal for a new task, "attribute extraction", which is planned for inclusion in the second evaluation, planned for autumn, 2008.

Categories and Subject Descriptors

H.3.5 [INFORMATION STORAGE AND RETRIEVAL]: Online Information Systems, *Web-based Service*

General Terms

Performance, Design, Experimentation, Security, Human Factors, Languages.

Keywords

Disambiguation, person names, attributes of people, information extraction.

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/she 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 conducted at SemEval-2007 (Artiles et al. 2007). The participating systems receive a set of web pages for a person name, and they have to cluster them into different entities.

2. The First Evaluation

The first evaluation was conducted in early 2007 and the results were reported at the SemEval-2007 workshop. Please refer to (Artiles et al. 07) and the participant's papers for details.

2.1 Data

In order to provide different ambiguity scenarios, we selected person names from different sources as seen in Table 1. For each name, which was randomly selected from the sources, a collection of web pages is obtained from the 100 top results using Yahoo! API. Given this set of approximately 100 documents, two annotators work on manual clustering of the documents according to the actual entity referred to. The differences are resolved by a meta-annotator (one of the organizers).

Table 1 Training and test data

Training			Test		
source	Av. entity	Av. doc.	source	Av. entity	Av. doc.
Wikipedia	23.14	99.00	Wikipedia	56.50	99.30
ECDL06	15.30	99.20	ACL06	31.00	98.40
WEB03*	5.90	47.20	Census	50.30	99.10
Total av.	10.76	71.20	Total av.	45.93	98.93

2.2 Results

16 teams submitted their results. Each participant tries to create clusters as similar as possible to the clusters created by the annotators. The results were measured by F-measure, based on purity and inverse purity. 7 systems outperformed one of the simplest baselines (each name belongs to one cluster), and the best system achieved 0.75 F-measure. However, the score is well below the human performance of 0.91 and 0.98 for the two annotators.

Almost all the systems worked as follows. First, the text part is extracted from the HTML page by a tool or simple rules. Then the text is processed by NLP and/or Web tools, such as a stemmer, POS tagger, chunker, Named Entity (NE) tagger, coreference, fact extraction, extraction of e-mail and URL, or link analyzer. Using the values of these features as a vector of the documents, the similarity is calculated (mostly by TD/IDF weighted cosine metric, but there are variations), and then clustering (mostly agglomerative clustering, but there are variations) was conducted. In order to determine the clusters, a threshold is needed; this was mostly tuned using the training data. It was notably discussed by many participants that an NE tagger is one of the most important technologies, as well as the threshold tuning for clustering.

Table 1 shows a strange disparity in the number of entities between the training data and test data. We don't really know the cause unless we analyze the search engine. But it has to be noted that this has an undesirable effect on threshold tuning. Fixing this is one of the motivations for holding the second evaluation. Another motivation is the observation that the most rational clue to solve the problem was found to be the attributes of people. This will be discussed in the next section.

WWW 2008, April 21--25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04

3. The Second evaluation

With the new challenges to be solved, we are planning to hold the second evaluation in 2008. In this evaluation, we will have an additional subtask, "attribute extraction" for people on the Web pages. It was noticed by the systems and the annotators that the attributes, such as birth date, spouse name, occupation and so on, are very important clues for the disambiguation. We believe it is the right direction to study such problem and try to implement technologies to identify such attributes. We will make this evaluation an independent subtask.

In order to set up the task, the first challenge is to define what are "the attributes of people". These have to be general enough to cover most people, useful for the disambiguation, and meaningful for the evaluation. We took an empirical method to define them; we extracted possible attributes from the Web pages and created a set of attributes which are frequent and important enough for the evaluation. We looked at 156 documents from the WePS corpus, and annotators extract as many attribute-value pairs as possible. The annotators are instructed to extract attributes of people which can be expressed as "Person's *Attribute* is *Value*". The attribute and the value must be a noun or its equivalent. An attribute and value pair may be expressed in a tabular format, or only the value may be mentioned in a sentence. If the name of the attribute is not explicit in the web page (e.g. "I am a professor" means Person's occupation is professor), then the annotator creates the attribute name. From the 156 documents, the annotators found 123 kinds of attributes; the 6 most frequent attributes are Occupation (116), Work (70), Affiliation (70), Full name (55), Person to work with (41) and Alma Mater (41). The number in parenthesis is the number of pages in which the information was mentioned. Among 123 attributes, there are attributes which are not suitable for the evaluation. For example, domain dependent attributes, such as "Career Points for a basketball player", or an attribute of an attribute value, such as "Birthday of spouse" are not suitable for the evaluation. Also, there are a set of attributes which might be meaningful even if we merge them together, such as "father", "mother", "sibling" as "relatives". By selecting and merging the 123 attributes, we finally made up 16 attribute classes, as shown in Table 2.

The subtask is going to involve extracting the values of those attributes as accurately as possible from Web pages. It would be ideal if we can merge the information for a single person entity from multiple pages, but if we set up such evaluation, it is not easy to handle the effect of clustering mistakes. So, we will evaluate the result for a given page, based on precision, recall and F-measure.

We expect the problem will be solved by a combination of many technologies, such as named entity recognition and classification, text mining, pattern matching, relation discovery, information extraction and more! Conducting this evaluation will definitely give a good opportunity to develop and collect useful resources, such as lists of named entities, such as occupation names and so on, annotation tools or text mining tools. As the outcome, we hope that this evaluation will provide fundamental research opportunities, as well as practical industrial application opportunities in the near future.

Table 2 Sixteen attribute classes

	Attribute class	Freq.	Example
1	Date of birth	21	March 5, 1965
2	Birth place	24	Tokyo, Japan
3	Other name	56	Mister S
4	Occupation	141	Research Associate Professor
5	Affiliation	83	New York University
6	Work	70	Apple Pie Parser
7	Award	26	Best Painter at Elementary School
8	Education	79	PhD, Computer Science
9	Mentor	48	Ralph Grishman
10	Location	63	New York, USA
11	Nationality	5	Japanese
12	Relatives	45	Shigeru Sekine
13	Phone	27	+1-212-998-3175
14	FAX	11	+1-212-995-4123
15	Email	25	sekine@cs.nyu.edu
16	Web site	13	http://nlp.cs.nyu.edu/sekine

4. Conclusion

In this paper, we explained the results of the first Web People Search task, which was conducted in 2007 with considerable success, with 16 participants from all over the world. We are planning to hold the second evaluation in 2008, which includes a new task of "attribute extraction". You can find the tools and datasets of the first evaluation and more details at http://nlp.uned.es/weps.

5. ACKNOWLEDGMENTS

Our thanks to all participants and collaborators of the WePS task.

6. REFERENCES

[1] Artiles, J., Gonzalo, J, Sekine, S. The SemEval-2007WePS Evaluation: Establishing a benchmark for the Web People Search Task. In Proceedings of the Workshop on Semantic Evaluation (SemEval-2007) at ACL07, pages 64-69.

[2] Artiles, J., Gonzalo, J. and Verdejo, F. 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.

[3] 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