Task Definition of the Clustering Subtask for WePS-2

Input/output

Systems receive as input the top documents retrieved by a web search engine using a person names as query. We provide the full contents of each document, as well as the original document ranking data (document rank, URL, snippet, estimated total results for the search, etc).

The output of must be a clustering of the web pages, where each cluster is assumed to contain all (and only those) pages that refer to the same individual.
It might be the case that different people with the same name may be mentioned simultaneously in the same document (this happens frequently in genealogies, Wikipedia pages and lists of authors). If this is the case, the document should appear in all the necessary clusters.

For example, in the manual clustering for "John Kennedy" we find that the document "5" mentions the entity "0" (35th President of the United States) and also the entity "3" (first son of John Fitzgerald Kennedy). Consequently the document "5" is added to both clusters.

```
<clustering name="John Kennedy">

    <entity id="0">
        <doc rank="0" />
        <doc rank="1" />
        <doc rank="4" />
        <doc rank="5" />
        ...
    </entity>

    <entity id="1">
        <doc rank="2" />
        <doc rank="42" />
        <doc rank="46" />
    </entity>

    <entity id="3">
        <doc rank="5" />
        <doc rank="9" />
        ...
    </entity>
    ...
</clustering>
```

The XML format conventions for search results metadata and clustering solutions follows the ones used in WePS-1.

Development data

Data for the development of participant's systems will be composed of the corpus and clustering gold standard developed for first WePS evaluation. For more details please refer to the WePS website (http://nlp.uned.es/weps/weps-1-data/).

Test data

Test data is composed of 30 ambiguous names: 10 name sets from the 1990 US Census, 10 from participants in ACL'08 and 10 from Wikipedia. Each name is made of two tokens, a first name and a last name. Wikipedia and ACL'08 names are extracted randomly from lists of person names (respectively http://en.wikipedia.org/wiki/Category:Human_name_disambiguation_pages and http://www.acl2008.org). Names from 1990 US Census are created by combining a first and a last name, each one extracted with the probability indicated by the census data (http://www.census.gov/genealogy/names/).

For each name a web search has been performed using Yahoo! API. The top results metadata stored in an XML file for each name. This metadata includes the estimated total number of results and the title, snippet and URL of each web result.

Around 100 documents have been downloaded from the top ranked search results. In some cases documents have been removed from the corpus. These documents will still appear in the XML file that describes the original search results, but won't be evaluated. For instance, non-HTML documents have been removed, with the exception of plain text documents. These are explicitly differentiated in the file extension (html or txt).

Each document in the corpus has been checked to contain the search string (either in the raw HTML doc or after removing tags and diacritics). Documents where the search string cannot be found have been omitted.

Finally, documents have been saved using UTF-8 character encoding, although this is not a 100% safe process when dealing with web data and in some cases corrupted character might remain.

Gold standard

The gold standard for the test data is the result of the manual clustering of the documents by human annotators. The annotator is presented with the ranked list of search results for a person name, and asked to separate them in groups where each document refers to the same individual. Documents with mentions to multiple individuals with the same name are allowed to be placed in as many clusters as necessary.

When the person name is used to refer to other type of named entity (location, organization, etc) is treated normally. A clustering solution might contain clusters for companies or places named after the ambiguous person name.

Individuals mentioned with variations of the ambiguous person name will be also taked into account for the clustering process. In order to avoid confusion we define person name variation as a name that contains all the tokens in the original ambiguous name. Of course there will be cases were the person will be mentioned using only the first name, the last name or another type of anaphoric reference. When these anaphoric mentions are linked to a person name variation then the information can be used for the clustering.

Let's take, for instance, the case of the John Kennedy document set. All documents have been checked to contain the original search string "John Kennedy", nevertheless variations of the ambiguous name occur frequently.
In one document we read the following sentences:

John Fitzgerald Kennedy (May 29, 1917 - November 22, 1963) was the thirty-fifth President of the United States, serving from 1961 until his assassination in 1963.
Kennedy was assassinated on November 22, 1963, in Dallas, Texas.
[...]
The coconut which was used to scrawl a rescue message given to Solomon Islander scouts is still at the **John F. Kennedy Library**.
[...]
Joseph Kennedy, Sr. was a leading McCarthy supporter.

The first sentence contains a valid name variation and the information associated to this name occurence allows us to stablish that the document should appear in the cluster for the 35th US President.
In the second sentence, "Kennedy" is not considered a name variation. But because it's an anaphoric reference to the name variation in the previous sentence, we consider the information in this sentence for the clustering.
The third sentence contains another name variation, and it's context indicates that the document should also appear in the cluster for the Johnn Kennedy Library.
Finally the name in the fourth sentence is not a valid variation and it is not linked to one, thus it is ignored.

Documents might be discarded during the manual clustering process for one of this two reasons:
- The document contains the ambiguous name, but it does not refer to nobody or nothing in particular. This can happen in spam web pages, or pages generated automatically from name lists.
- There is not enough evidence to decide where to put it in the clustering solution.

Evaluation methodology

Systems output for the test data will be compared to a human gold standard.
Based on the experience in WePS-1 and the feedback we received from participants we have decided to use the extended B-Cubed clustering measure (Amigó et al. 2008) which overcomes the limitations of standard clustering measures when dealing with overlapped clusters. For comparability with WePS-1 we will provide also an evaluation using Purity, Inverse Purity measures.

All documents (except for those explicitly discarded) are required to be included in the systems clustering solution. Documents missing in the systems output will be aggregated in a new cluster and evaluated as a part of the clustering solution.