Co-reference Resolution for Person Names

Nguyen Bach & Simon Fung
{nbach, sfung}@cs.cmu.edu

1. Problem Description

In the first phase of projects, some groups encountered several problems related to the co-reference resolution of person names.

1. *Name variants in the same document.* Within a document, variants of the same name usually appear; for example, William Cohen may be referred to as "William," "Dr. Cohen," or Prof. Cohen. These usually refer to the same person, but they must nevertheless be identified.
2. *Identical names or name variants across different documents.* The same name often refers to different people in different documents. In many information extraction tasks, this ambiguity must be resolved. Existing clustering engines such as Clusty and Vivisimo automatically clusters URLs into categories that are intelligently selected from the words and phrases contained in the search results. However, we would like to provide information about each similarly-named indivdual in a format that is more useful for information extraction tasks.

At this point, we feel that a simple heuristic should provide an adequate solution to problem 1, and therefore problem 2 will most likely be the focus of our efforts.

2. Functionality

We aim to build a system that does the following:

1. Given a webpage, identifies variations of the same name.
2. Given a person name, provides a list of popular URLs and additional information (such as occupation, full name) associated with each person by that name.
3. Given a webpage mentioning a person name, provides a list as described in (2), and gives a score to each person in the list that indicates how likely it is the person mentioned in the webpage.

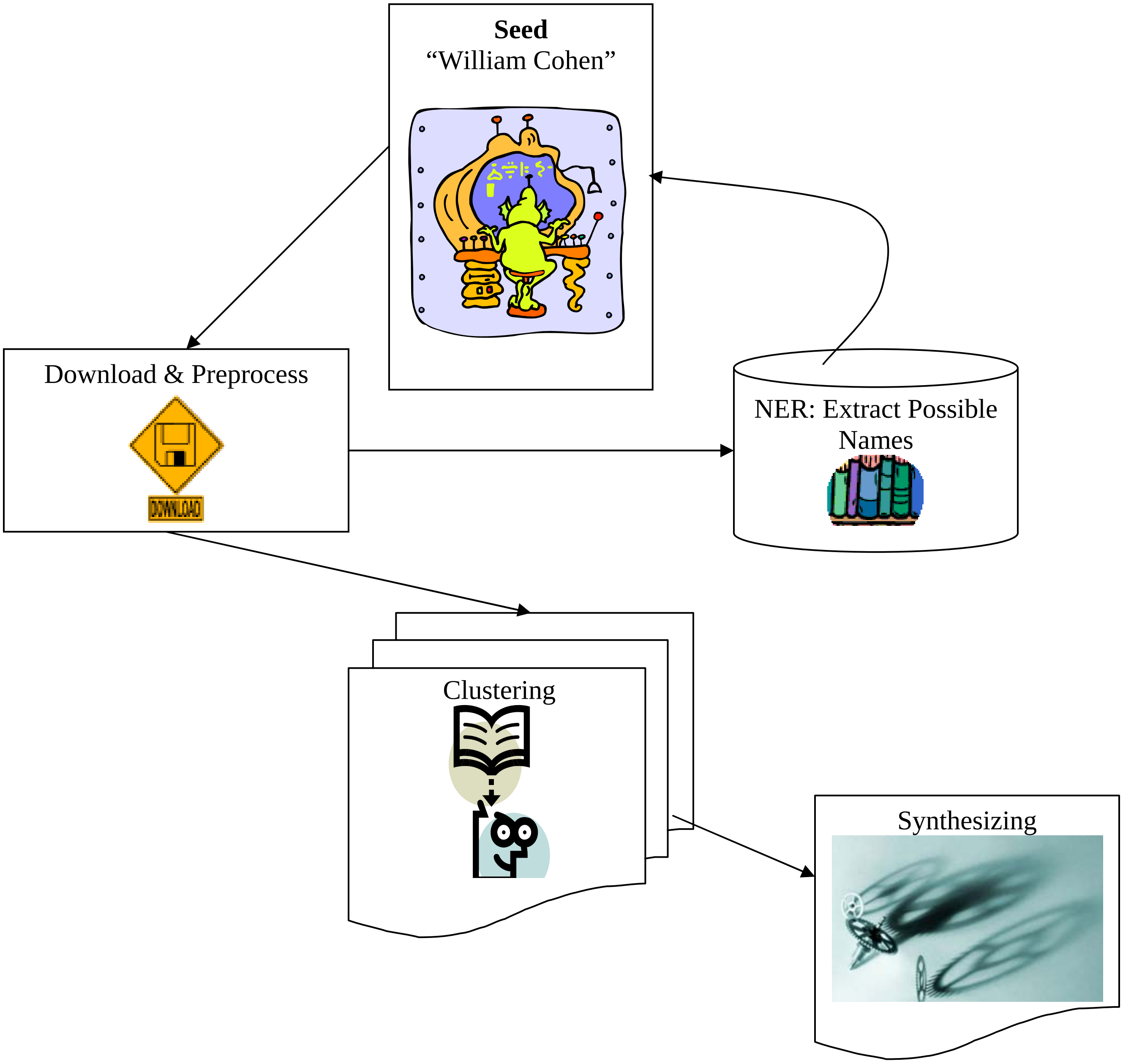
3. System Sketch

The figure in the next page demonstrates the architecture of our system. It is built from various pieces:

- Web Getter: to google, download, and pre-process webpages.
- Named-entity tagger: to detect name, location, date, organization
- Part of speech tagger / Noun phrase chunker: to extract head nouns or pharases
- Topic detector: to discover general topic of a webpages

- Unsupervised cluster tool: to group names in webpages by their professional

These are redundants for a name:
- Gender: the number of hits returns for a query constructed from Name+He and Name+She, choose the one with higher hits.
- Name in Census 1990 data: list of popular female and male name by ranking.
- Web hits: the number of hits returns for a name query from google .
- Semantic feature: based on semantic relatedness of concepts, for example: professors are more likely teaching in universities, and also are Doctors.
- Case and Period: to determine a word and a letter in sentence is a regular word or a part of a possible name.
- Context: context can be of different sizes such as window of words centered surround a name, sentence containing name, group of sentences, or even whole document. Modeling context can be done in many different ways for instance bag of words, set of phrases, set of entities, set of relations, and so on.



4. References

- Gideon S. Mann, David Yarowsky, Unsupervised Personal Name Disambiguation, CoNLL, Edmonton, Canada, 2003

- Fleischman, M. B. and Hovy, E. Multi-Document Person Name Resolution. 42nd Annual Meeting of the Association for Computational Linguistics, Reference Resolution Workshop, Barcelona, Spain. July 2004.

- Yarowsky, D. One Sense Per Collocation. In Proceedings, ARPA Human Language Technology Workshop. Princeton, pp. 266-271, 1993.