FINAL REPORT - TEXTJOINER IMPROVEMENTS

Tejaswinee Sohoni
Net ID: tss995
Email ID: tejaswineesohoni2015@u.northwestern.edu
EECS 349 - Northwestern University

Automatically extracting information from the web is called Web Information Extraction. On demand Web Information Extraction systems allow users to search the web for textual queries, for instance, “Nobel Laureates from Austria”. This is done by specifying the query as a relation. However, such systems have to make a trade-off between precision and recall. Therefore, a new approach has been proposed by researchers in Prof. Downey’s group, in which queries are considered as conjunctions and disjunctions of multiple contexts instead of a single context. This offers high precision as well as high recall. This approach has been implemented in a system called TextJoiner.

However, the existing TextJoiner system does not take into account every mention of a particular entity in a piece of text. It misses out on the mentions of an entity where it is referred to with pronouns or equivalent words. This makes the task of coreference resolution interesting to the TextJoiner system. Coreference resolution means finding all the expressions that refer to the same entity in a text. For instance, in the sentence,

“I would like to spend some time with Mary because only she can answer my questions.”, John said.

I, my and John refer to one entity, and she and Mary refer to another. If the TextJoiner system can understand this difference, it can discover more sentences about the entity being queried, and hence can extract more information out of the text.

As a solution to this task, I used the Stanford Coreferencing Library. The output I got from this library is a chain of coreferences where each chain indicates a