

HLT/COE Research in Cross-document Coreference Resolution

James Mayfield, Paul McNamee, Christine Piatko, Clayton Fink and Tim Finin

MERC researchers are working with the new Human Language Technology Center of Excellence (HLT/COE) to find new ways to extract information from free text in languages such as English and Arabic for inclusion in a knowledge base. One of the key problems that must be solved to make such knowledge base population practical is to determine when the same person or organization is mentioned in two different documents.

We have developed a host of new techniques to help solve this problem. Three of the main tasks that we must carry out are search space reduction (avoiding asking low-likelihood questions like “is ‘George W. Bush’ likely to refer to the same person as ‘Britney Spears’?”), featurization (identifying evidence that supports or refutes such pairings), and categorization (making a decision about whether a given pair of entities should be deemed to be coreferent).

For search space reduction, we developed techniques to reduce the search space by three orders of magnitude without significant negative impact on the results. For featurization, we divided our features into six broad classes: character level, document level, metadata, knowledge base ontology, knowledge base instances, and semantic match. The feature space can also be divided into those features that provide evidence for coreference, and those that provide evidence against coreference (some do both). We tried to ensure that both types of features were well represented in our feature space. For categorization, we successfully used Support Vector machines to combine our many features into a single coreference decision.

To evaluate our work, we recently completed our first entry in the NIST-sponsored Automatic Content Extraction (ACE) evaluation, performing cross-document coreference resolution tasks in both English and Arabic. Our results showed that our approach to featurization and categorization works well when sufficient training data is available. While no single group of features that we used was necessary for good performance, the use of all features produced significantly improved performance over a baseline of simple string matching.

In the future, we will use the results of the ACE evaluation to design direct evaluation of a knowledge base that has been populated from text. We will apply our techniques to sponsor data that may differ in many characteristics from the text typically found in unclassified collections. Finally, we will extend the knowledge bases we populate to include temporal qualification (identifying *when* an extracted fact was true), certainty and contradiction (assessing how likely it is that extracted information is correct, and handling contradictory information in the knowledge base), and provenance (maintaining links from extracted information to the source documents that support it).