Rich entity recognition in English text

Authors
Rishav Bhowmick, Michael Heilman, Kemal Oflazer, Behrang Mohit, Noah Smith

Institutions
Carnegie Mellon University in Qatar, Doha, Qatar
Carnegie Mellon University, Pittsburg, PA, USA

Entity type recognition is used as a pre-processing step in common applications like summarization of text, classifying documents or automatic answering of questions posed in natural language. Here, ‘entity’ refers to concrete and abstract objects identified by proper and common nouns. Entity recognition focuses on detecting instances of types like person, location, organization, and so on. For example, an entity recognizer would take as input:

George Washington was the first President of the United States of America.

and output:

<noun.person> George Washington </noun.person> was the first <noun.person> President </noun.person> of the <noun.location> United States of America </noun.Location>.

The task can be performed using machine learning techniques to train a system that recognizes entities with performance comparable to a human annotator. Challenges like the lack of a large annotated training data corpus, impossible nature of listing all entity types, and ambiguity in language make this problem hard. There are existing entity recognizers which perform this task but with fair performance. One of the ways adopted to improve the performance of an existing entity recognizer is feature engineering. We initially find out which of the existing features, used in the recognizer, affect the performance most strongly. We accomplish this by adding and removing one or more features at a time from the feature list. We then use the training data to train a model and test to find out which set of features are important. The evaluation metric involves finding the precision, recall and f-score (which is the harmonic mean of precision and recall). As a next step, we add new features like word clusters and bigram word features to find out any improvements. Word clusters help when the training data does not have some words, but words belonging to the same cluster are present in the training data. This helps tagging unseen words in the test set. We also experiment with varying the size of the training data to find out how it affects the performance. Additionally, we look into Wikipedia as a source of additional features for the training data. Wikipedia has an elaborate internal link structure that can provide vital information about the category of a word. This category can be linked to a broader-sensed entity type.