CS839 – Data Science

Project Stage – 1 Named Entity Recognition

Team Members :
Varun Batra (vbatra@wisc.edu)
Vibhor Goel (vgoel5@wisc.edu)
Adarsh Kumar (kumar92@wisc.edu)

Entity: Locations
We have chosen location to be our enity. We have marked streets, cities, states, countries, continents. We have refrained from marking museums, stadiums, buildings in the locations.

Dataset: We decided to choose New York Times Articles Dataset. The number of mentions are listed below.

Document Count	
Train Set	Test Set
216	100

Mention Count	
Train Set	Test Set
1025	800

Methodology:

1. First, we randomly chose ~350 documents from the New York Times Articles Dataset.
2. Then, we marked the occurrences of all the locations in the text documents. We made sure that all three of us consistently marked the locations based on a common understanding. For example, New York is marked as a place, but Metropolitan Museum of Art was not marked.
3. We randomly chose 200 documents as training documents and 100 as test documents.
4. We tokenized all the documents to generate location information about all set of unigrams, bigrams and trigrams. Also, the context information is recorded by storing the neighboring 10 words.
5. Next, we performed a set of pre-processing rules, like removing some stopwords, common verbs, digits, lowercases.
6. Next, we extracted a set of 23 features like, ‘Is_prev_location_descriptor’, ‘is_previous_title’, ‘say_synonym’, ‘location_based’, distances from verb, ‘token_length’,etc. We also used word2vec for the tokens as a feature.

7. Finally, we performed the classification using various techniques and the results are as listed below. We used Decision tree as the initial classifier as we went about improving the feature vector. Once done, we tried other classifiers and found Random Forest to work the best.

    a. Decision Tree

Dataset	Precision	Recall	F-1 Score
Train	81.35	79.11	80.21
Test	79.61	77.066	78.31

    b. Linear Regression

Dataset	Precision	Recall	F-1 Score
Train	90.90	70.65	79.51
Test	87.36	65.46	74.84

    c. SVM

Dataset	Precision	Recall	F-1 Score
Train	88.16	80.97	84.41
Test	84.69	78.93	81.71

    d. Logistic Regression

Dataset	Precision	Recall	F-1 Score
Train	91.97	80.97	86.12
Test	87.07	78.13	82.36

    e. Random Forest

Dataset	Precision	Recall	F-1 Score
Train	94.26	80.43	86.80
Test	91.19	78.66	84.46

8. Post Processing rules: To get rid of some of the common false negatives, we added a small whitelist to further enhance our results consisting of common countries. The whitelist was - ['Britain', 'France', 'Saudi Arabia', 'USA', 'Europe', 'Australia's', 'Europe's', 'Canada's', 'Britain's', 'France's']. The recall improved by 4%, while precision reduced slightly.

    a. Random Forest

Dataset	Precision	Recall	F-1 Score
Test	90.34	83.6	86.84

9. As we can see, we were able to meet the 90+% precision and 60+% recall requirement using Random Forest without any blacklist/whitelist/post processing.