CS 839: Data Science

Project Stage One

Ankit Jain(ajain64) | Ankit Maharia(maharia) | Prerak Mall(pmall)

APPROACH

In this stage of the project we decided to extract the entity "Name" from news articles related to Politics. Below are some of the examples:

Donald J. Trump, Hillary Clinton, Barack Obama, Jakob Segel, Douglas Selvage, Bernie Sanders etc.

We decide to mark these entity types using custom tags that would help us to identify correct positive labels. Below is the snapshot of tagged entities:

1. He first told <Entity>Sean Hannity</Entity> last month that he…
2. Some experts say Mr. <Entity>Trump</Entity> will find it impossible to square…
3. Mr. <Entity>Clinton</Entity> signed a statute declaring...

Initially a total of 1982 names were marked which were distributed across 300 documents.

Out of those 300 documents, 200 documents were added to Set I and 100 were added to Set J.

The dataset configuration is described below:

Set	Documents	Number of Marked Entities
I	200	1250
J	100	732
B	300	1982

METHODOLOGY

Following steps were employed to build the end to end pipeline:

1. Data Cleaning Phase: In order to take care of the all the special characters like ., ’s, [,] and other punctuation symbols.

2. Data Pruning Phase: To eliminate the obvious negative entities (stop-words / places) and reduce the processing needed to be done by the model/classifier.

3. Feature Definition: Build the features for the data points to be passed for the training.

4. CV Tests: Executed multiple classifiers on the feature vectors of the cleaned pre-processed data in order to check the precision, recall and F1 values after cross validation on the Set I dataset.

5. The following depicts the initial performance of the best classifier (M):

   Set I (Based on 5-fold Cross Validation):
   Best Classifier Model: Decision Tree (Based on F-Score)
   Precision: 0.6132312
   Recall: 0.582285
   F Score: 0.59665

6. Code Review: While addressing the subpar classification performance, we encountered a bug with respect to certain non-ASCII characters (Double Quotes, Single Quotes etc.) which were causing an error in cleaning due to Unicode mismatch. This resulted in many false negatives.
   Additionally, features based on general grammar semantics like occurrence of capital case letters, pronouns, articles, verbs were incorporated in the model.
   Also, since our entity were names, a feature pertaining to the occurrence of occupation and salutation/title in the proximity, for instance, Dr., Mr., President, Senator etc. was added.
   All these modifications helped us decrease the number of false positives and increase the true positive.

7. These modifications resulted in the following improvements (Classifier X):

   Set I (Based on 5-fold Cross Validation):
   Best Classifier: Decision Tree
   Precision: 0.927261
   Recall: 0.65274
   F Score: 0.76572

8. Actual Testing: Finally, in order to test our model, we trained it on Set I and then tested using **Set J with the Decision Tree model** and were able to achieve the following results: (Classifier Y)

   Classifier: Decision Tree
   Precision: 0.916030
   Recall: 0.655737
   F Score: 0.76433
   Classifier Y is same to X

PS: NO RULE BASED POST PROCESSING WAS USED IN ACHIEVING THE ABOVE RESULTS (NO WHITELIST/BLACKLIST)

APPENDIX

1. Domain agnostic features used:
    a. Word starts with capital letter.
    b. All the words are capital.
    c. Next/Previous word is capital.
    d. Ends with ‘s.
    e. Ends/Starts with quotes.
    f. Previous word ends with a punctuation other than “.” or “?”.
    g. Previous word ends with a punctuation “.” or “?”.
    h. Features based on articles, pronouns, verbs etc.
2. Domain related features considered were as follows:
    a. Since the data was taken from a news source (New York Times) that contained formal written English, we used a feature to track titles like Mr., Mrs., Miss. etc. occurring before the word.
    b. Since the dataset was pertaining political domain, we curated a list of occupations in this field and used that as a feature for our classification by looking ahead up to 3 positions from the current word.
3. For the preprocessing step, we ensured the following things:
    a. No positive marked label should be pruned.
    b. On noticing high frequency of country names in the dataset we incorporated an unbiased and generalized solution of pruning out major countries and cities.
4. To better maintain the code changes and keep scaling the tagged data set, we also developed multiple testcase that would help debug various issues stage by stage in the pipeline.
5. One of the major takeaway from this stage was that features largely depend on the problem domain and the underlying dataset should be taken into consideration while defining the learning model.