CS839 Project Stage 1 Report

Sreejita Dutta, Deepanshu Gera, Rahul Jayan

I. Entity type

We extracted company names from Bloomberg News articles.
Examples:
1) Coca-Cola Co
2) Amerigroup Corp
3) Time Warner Cable Inc
4) Amazon
The company names in the documents were manually marked up using the tags <markup> and </markup>

II. Training set and test set

Training set contains **201** documents and Testing set contains **100** documents.
Total no. of mentions marked up : **2291**
No. of mentions in training set (I) : **1549**
No. of mentions in testing set (J) : **742**

III. Preprocessing rules

After creating the possible ngrams, we added preprocessing rules to prune out n-grams that would most likely never be company names.

1. If the starting letter of each word is not capitalized.
2. If the n-gram is a known country, city ,continent, USA state or stock exchange.
3. If the n-gram contains a known designation like “Chief Executive Officer” or any other obvious red flags like “Ministry”, “Council”, “Week”, “Year” etc.
4. If the n-gram contains special characters like “(“ , “)” etc.
5. If the n-gram starts with an article like An or A, or a word like And or This.
6. Unigram pruning: ethnicities, articles, prepositions, month, day of the week, prefixes and suffixes only,etc.

IV. Feature Engineering

Features were added incrementally throughout the entire training process.

1. Index of the starting word inside that document
2. Index of the ending word inside that document
3. Does it contain a known company suffix or prefix? Example Inc., Co., Corp.
4. If it is a starting substring of an n-gram (in that document) which has a known company suffix or prefix? E.g. Microsoft is a starting substring of Microsoft Corp.
5. If it is a substring of an n-gram (in that document) which has a known company suffix or prefix? E.g. Corp is a substring of Microsoft Corp.
6. If it is part of the filename (which is essentially the news headline) ?
7. Does “based” appear upstream? For example, “Seattle based XYZ”.
8. Does “ ‘s” appear downstream?

9. Does “shares” appear upstream/downstream ? This is because a lot of sentences have “shares of XYZ company”
10. Do known designations (like CEO , CFO etc.) appear upstream or downstream?
11. Does “said” or “told” appear upstream ? This is specifically to distinguish the person’s name from company name.

When a feature deals with tokens immediately upstream or downstream, we only looked at 3 words before and 3 words after.
After feature generation (using process_data.ipynb), test_pruned_sr.csv and trained_pruned_sr.csv were generated. These were used by the ML classifiers.

V. Training Classifiers

We considered the following machine-learning algorithms for our information extraction task :

- Decision Trees (ml-decisiontree.ipynb)
- Random Forests (ml-randomforest.ipynb)
- Support Vector Machines (ml-svm.ipynb)
- Linear Regression (ml-linear-regression.ipynb)
- Logistic Regression (ml-logistic-regression.ipynb)

Support Vector Machines (M- classifier) the classifier that we selected after performing 10-fold cross validation on training set for ***the first time***.
Precision: **0.86187**
Recall: **0.8013**
F1: **0.83**

Support Vector Machines (X-classifier) was the classifier we finally settled on.

The final Training set accuracy of all the algorithms are :-

Classifier	Precision	Recall	F1
SVM	**0.935**	**0.7426**	**0.83**
Decision Tree	0.8866	0.8580	0.872
Random Forest	0.87755	0.8657	0.8715
Linear Regression	0.8831	0.8774	0.8802
Logistic Regression	0.8561	0.8802	0.8679

VI. Test set accuracy (Prediction on set J)

Since we achieved the target precision and recall using Support Vector Machines, we applied this classifier to the test set and obtained the following accuracy:
Precision: **0.8452**
Recall: **0.7520**
F1: **0.795**

Unfortunately, we were not able to reach precision > 0.90 (by 0.6) on the test set J. There are several potential explanations on why this task is hard.
Some company names can be difficult to distinguish from other entity types . For e.g., J.C. Penny is a person name and also a company; Bedford is a company and also a city. Also, company names can have complex structures and it is difficult to distinguish partial names from full names.