CS838 Data Science project progress report - Stage 2

Jinman Zhao
jzhao237@wisc.edu

Bin Guo
bguo23@wisc.edu

Di Wu
dwu73@wisc.edu

Entities to be extracted

Entity type: Company names.
Examples: Google, Amazon, Royal Dutch Shell, P&G
Counterexamples: Amazon Pay, Apple Watch, Tim Cook, CEO, the United States
Total number of mentions: 1502

Table 1: Document number and mention number

Document set	Document number	Mention number
I (training set)	250	1135
J (testing set)	100	367

Learning steps

We take out all n-grams (n=1,2,3,4) from text documents and try to predict whether they are company names or not.

For machine learning classifiers, we used following features.

1. Number of words, int
2. Number of characters, int
3. All capitalized, boolean
4. All initial capitalized, boolean
5. No previous word initial capitalized, boolean
6. No after word initial capitalized, boolean
7. Contain common ends (Inc. Corporation, etc), boolean
8. Contain common previous word (competitor, acquired, etc.), boolean
9. Contain hyphen, boolean
10. Contain ampersand, boolean
11. Is at sentence start, boolean

For rule-based post-processing, we used following rules.

1. Black list of proper nouns: geographical terms (country, American city, continent), months (January to December).
2. Black list of function words and pronoun: of, a, its, their, etc,

Table 2: Test on different Classifiers with set I

Classifier	Precision	Recall	F1
Logistic Regression	0.81	0.11	0.19
SVM	0.93	0.1	0.18
Linear Regression	0.1	0.01	0.02
Decision Tree	0.33	0.71	0.45
Random Forest	0.44	0.37	0.40

Table 3: Classifier and the results

Classifier		Precision	Recall	F1	Evaluate on
M	Decision Tree	0.31	0.12	0.17	I
X	Decision Tree	0.4	0.77	0.53	J
Y	Decision Tree	0.46	0.72	0.56	J

Discussion

The final precision is not so great for only 46%. Here are some major issues with the classification for companies.

1. It is tricky to distinguish company names with people names, location name. We blacklisted some country and city names, but there are some building names that we can not filter out. Also some people's names that showed up in the text document.
2. There are many product names that are very similar to the company names in format.
    a. Some products that we are familiar with is easy to filter by eyeballing them. But it is easily mixed up with company name features by machine.
    For example, Amazon Prime is a product name but not a company name.

b. Some other strange product names are even harder to tell if we never heard of them before.

For example, Mosaic Company has some products named: MicroEssentials, K-Mag and Pegasus.

Other thoughts and ideas

- When improving the recall with relatively high precision, we can add in a rule to see if the single word instance shows up in classified positive multiple words instances.
  For example:
  {Mosaic} can be identified as positive because {The Mosaic Company} is already labeled positive.
- Add more black lists with some famous names from these companies.