

Project Report Stage Two

Hongyi Wang, Hao Fu, Miao Yang

March 2, 2017

1 Name of All Team Members

- Hongyi Wang
- Hao Fu
- Miao Yang

2 Entity Extraction

2.1 Entity Type

In the stage two of this project, we planed to extract the **name of university** from the text files we collected in stage one. The features we extract from all mentions contain:

- length of mentions (i.e. the string)
- if the word "university" contains in the mention
- if the word "state" contains in the mention
- if a name of state (e.g. Wisconsin, California, and etc) contains in the mention
- if dash (i.e. -) contains in the mention
- if the mention contain number
- if all letters in the mention are capital (e.g. MIT)

The labels in our extracted data instances are "True" (i.e. the mention is a university name) or "False" (i.e. the mention isn't a university name)

2.2 Entity Examples

We give several examples of mentions here:

MIT, Rochester Institute of Technology, University of California–Riverside, UCR, ASU-Temple, CUA, and etc. The those foregoing mentions were all labeled True in our data set.

Here are also several False labeled instances:

School's, Undergraduate, NCAA-division, Institution, and etc.

3 About the Data Set (Mentions)

In our data set, we have entirely 1322 instances (i.e. mentions or data points). In addition, we have 931 mentions, which were extracted from 206 documents from our text files in set I (we call it dev set in the following), and 391 mentions, which were extracted from 102 documents from the text files in set J (we call it test set in the following).

4 Performance on Different Models and Model Selection

4.1 How We Implement Machine Learning

For dev set, the training set is shuffled at the beginning, instead of implementing simple cross validation, we implement k -fold cross validation, i.e. split the whole training set into k subsets, and then use $k - 1$ of k subsets of training set to train the model, and test the trained model on the remaining set. We tried SVM with Gaussian kernel, decision tree, random forest, logistic regression, and linear regression (set 0.5 as threshold, with no Lasso or Ridge regularization). The results on different models are given below:

- length of mentions (i.e. the string)
- **SVM:** Test set accuracy: 0.9693 Precision: 0.9848 Recall: 0.9557 F₁: 0.97
- **Decision Tree:** Test set accuracy: 0.9719 Precision: 0.9849 Recall: 0.9606 F₁: 0.9726
- **Random Forest:** Test set accuracy: 0.9770 Precision: 0.9899 Recall: 0.9655 F₁: 0.9776
- **Logistic Regression:** Test set accuracy: 0.9719 Precision: 0.9746 Recall: 0.9569 F₁: 0.9713
- **Linear Regression:** Test set accuracy: 0.9668 Precision: 0.9847 Recall: 0.9507 F₁: 0.9674

Due to the result, obviously the result from random forest is the best. Thus we select the result from random forest.

4.2 Any Rule-based Post-processing?

Clearly, since the foregoing results are nearly "perfect" we do not need any rule-based post-processing for our information extraction work.