Yiran FEI

Andrew ID: yiranf

September 24, 2014

HW1 - Report

11-693 Software Method

Requirement

The Named Entity Recognition is used to detect the Gene names in text-based sentences. Given a set of formatted sentences containing some gene name entities, this system should recognize the possible gene names in these sentences and output them to a file.

Architecture

This system is based on the UIMA framework and core algorithm is based on a free name entity recognizer system called lingpipe (http://alias-i.com/lingpipe/index.html).

UIMA is composed of 3 parts: the collection reader, analysis engine and CAS consumer. In this system, the consumer is just a printer which can write CAS information to the file system and the AE is a NE recognizer. The data come from the collection reader line by line, then get through the Analysis Engine and finally the selected data are printed out to the output file.

Algorithm

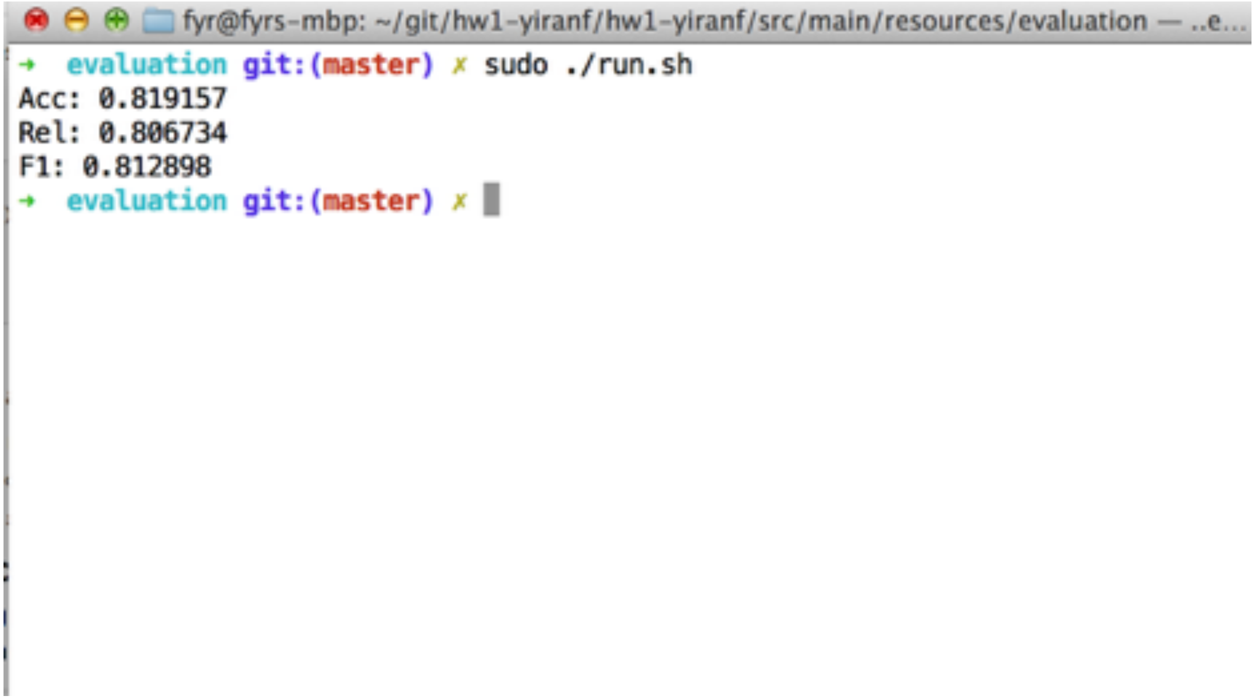
The AE engine is based on the lingpipe. lingpipe provides many ways to recognize the name entity, including dictionary-based recognize and statistical-based recognizer. Because there's a trained model of gene name tag provided by the lingpipe and according to the test, it works quiet well with high accuracy and recall ratio, I choose to use this to implement the AE engine. In detail, I use a method called Confidence Named Entity Chunking (http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html) and test the accuracy with different parameters.

During the training period of the training, a gene dataset from NCBS is used. This dataset can be retrieved freely in: http://alias-i.com/lingpipe/demos/models/pos-en-bio-genia.HiddenMarkovModel .

Evaluation

My evaluation program is implemented by C++. It reads the sample.out first, store all results from it and read the data to be tested.

Using this tool, I configure the parameter of the algorithm and find that when the evaluation confident value is set to 0.63 (which means if some name got a score higher than 0.63, it will be outputted the the results), the F1 value can be 0.81289, which is currently the best during my tests:


fyr@fyrs-mbp: ~/git/hw1-yiranf/hw1-yiranf/src/main/resources/evaluation — ..e...

```
➜  evaluation git:(master) ✗ sudo ./run.sh
Acc: 0.819157
Rel: 0.806734
F1: 0.812898
➜  evaluation git:(master) ✗ 
```

Reference

All the codes of hw1 are written by myself or modified from the SDK example of UIMA and lingpipe. These modified codes are under the protection of Apache License and Alias-i Royalty Free License.