Natural Language Processing

HW3 Information Extraction Report

1. What I done in my code:

I reuse the code from hw2, but since the sentence in this assignment is much longer and the probability is sparser, so I use log on the probability. So instead of multiply all the probabilities, I add all the log(probability). The probability 0 should be set to –inf, because log(0)=-inf.

I use +1 smoothing in this assignment.

I add some features to my code, there are 6 features:

1) Long words, which is the length of word is bigger than 9.

2) All digit word.

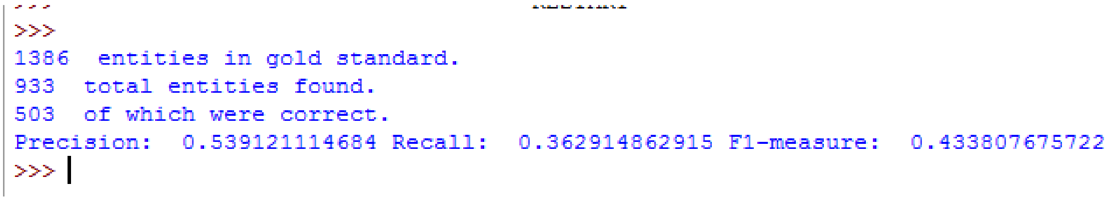
3) All upper case words.

4) Digit and numeric only words.

5) Words with character that is not digit or number.

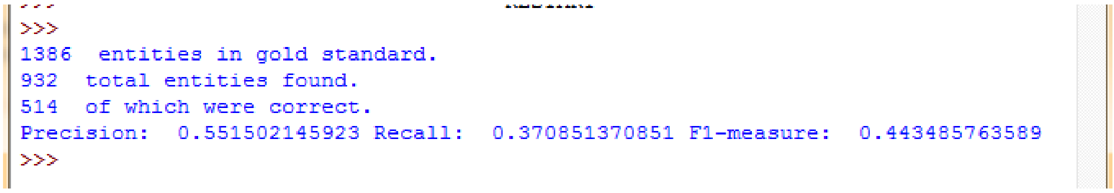
The features has priorities, if that's a long word and also all upper cases, it will be marked as LongWord.

I use 90% of the training file as my training set, and I use 10% as my test set. The result I got only use log without features is F1=0.43



```
>>>
1386  entities in gold standard.
933  total entities found.
503  of which were correct.
Precision:  0.539121114684 Recall:  0.362914862915 F1-measure:  0.433807675722
>>> |
```

The result I got with log and features is F1=0.44



```
>>>
1386  entities in gold standard.
932  total entities found.
514  of which were correct.
Precision:  0.551502145923 Recall:  0.370851370851 F1-measure:  0.443485763589
>>>
```

2. Result analysis:

If we don't use log and features, the result is really bad. Only use features or only use logs could both achieve a F1 measure bigger than 0.4. If we use both, the F1 measure will be better than only using one of it.

If we just use log, the f1 value could be 0.43. If we add features to the code, the result could be 0.44. In this situation, the features wouldn’t help much. Some features even decrease the F1 measure like “words bigger than 5”, the too-long word length is too small to be set to 5, most words would fit in this category.

So the main problem is the sparse of the probability and the too-long sentences which make the whole probability of a sentence too trivial. Both the log and features could help to solve this problem.