CS137 Information Retrieval (Spring 2013): Project #2

Chuan Wang

February 12, 2013

1 Machine-Learning Package

I use the CRF++ package.
url:http://crfpp.googlecode.com/svn/trunk/doc/index.html

2 Features Extracted

2.1 local feature

1. First Word, Case, and Zone
2. Case and Zone of w_{+1} and w_1
3. All capitalized word feature
4. First word of the sentence
5. mixed capitalized word or not
6. only one capitalized letter
7. if end with period
8. if it is all capitalized letter and end with period
9. if previous and next word is one capitalized letter
10. if previous and next word is ending with period
11. contains digit and slash
12. Rare words: If w is not found in FWL (Frequent Word List consists of words that occur in more than 5 different documents.).
13. suffix: fixed length suffix of the entity lists extracted from the training data.

2.2 global feature

InitCaps of Other Occurrences (ICOC)

3 Other resources

Frequent Word List consists of words that occur in more than 5 different documents.
LOC,GPE,ORG list

4 Performance on test set

p = 0.728334956183 r = 0.67815049864 f = 0.70234741784