CS137 Information Extraction (Spring 2013): Project #4 Relation Extraction

Chuan Wang

April 6, 2013

1 Machine-Learning Package

I use the Mallet MaxEnt classifier in the cluster.

2 Features Extracted

2.1 word feature

1. HM1: head word of M1
2. HM2: head word of M2
3. HM12: combination of HM1 and HM2
4. WBNULL: when no word in between
5. WBFL: the only word in between when only one word in between
6. WBF: first word in between when at least two words in between
7. WBL: last word in between when at least two words in between
8. WBO: other words in between except first and last words when at least three words in between
9. BM1F: first word before M1
10. BM1L: second word before M1
11. AM2F: first word after M2
12. AM2L: second word after M2

2.2 Entity Type

13. ET12: combination of mention entity types

2.3 Overlap

14. MB: number of other mentions in between
15. WB: number of words in between

2.4 base phrase chunking

16. CPHBNULL when no phrase in between
17. CPHBFL: the only phrase head when only one phrase in between
18. CPHBF: first phrase head in between when at least two phrases in between
19. CPHBL: last phrase head in between when at least two phrase heads in between
20. CPHBO: other phrase heads in between except first and last phrase heads when at least three phrases in between
21. CPHBM1F: first phrase head before M1
22. CPHBM1L: second phrase head before M1
23. CPHAM2F: first phrase head after M2
24. CPHAM2F: second phrase head after M2

2.5 Parse Tree

25. PTP: path of phrase labels (removing dupli- cates) connecting M1 and M2 in the parse tree

3 Feature analysis

Also the paper's result has demonstrated that the NP chunk feature is the most salient one. But my experiment did not prove this. So I guess the problem is the practical implementation. How we convert the full parsed tree to the shallow chunk representation is crucial to the final performace.There're two main problems which causes the downfall of the performance. First, I tried to find the lowest subtree which dominates the two mentions and find NPs between (before or after them), then I discovered that the phrase head I found is not always between the two mention. It may be come before or after the mentions. This approximation may cause the downfall of the performance.
Second problem is how we define the NP phrase head, the method I tried is to treat the first noun in the NP phrase as head. However, it turns out that under many cases, the first noun is not the head. May be the last noun would be better?

4 Performance on test set

precision = 0.707547169811 recall = 0.195822454308 f1 = 0.306748466258

5 Performance on develop set

precision = 0.679738562092 recall = 0.199616122841 f1 = 0.308605341246