CS137 Information Retrieval (Spring 2013): Project #3 Coreference Resolution

Chuan Wang

March 9, 2013

1 Machine-Learning Package

I use the Mallet MaxEnt classifier in the cluster.

2 Features Extracted

1. Sentences distance
2. i-pronoun feature: true if i is a pronoun, false otherwise
3. j-pronoun feature: true if j is a pronoun, false otherwise
4. String match feature: true if the string of i matches the string of j, false otherwise.
5. Definite noun phrase feature: true if j is a definite noun phrase, false otherwise
6. only one capitalized letter
7. Demonstrative noun phrase feature: true if j is a demonstrative noun phrase (starting with this, that, these, those), false otherwise
8. Number agreement feature: true if i and j agree in number, false otherwise
9. Semantic class agreement feature: i and j are in agreement if the semantic class of one is the parent of that of the other, or if they are the same
10. Both-Proper-Names Feature: true if both i and j are proper nouns, false otherwise. 11. Alias Feature: true if one is the alias of the other, false otherwise. Match of last name of person names, part of organization name
12. Appositive feature: true if j is in apposition to i, false otherwise
13. pronoun string match: if i or j is pronoun, return whether they string match or not.
14. proper noun match: if i or j is proper noun, return whether they string match or not.
15. tree distance: according to the parsed tree, return tree path length from i to j.
16. both pronoun: if i and j both are pronoun, return true.
17. GPE-match: if type of i and j is GPE, check if i and j are fuzzy match

(using stemmer).
18. j-depth: return the tree depth of j.
19. return the pos tag of the lowest descendant of this tree that dominates i and j.

3 Feature analysis

It turns out that the tree distance, appositive, pronoun string match and proper noun string match features bring the most improvements.

4 Performance on test set

precision = 0.410915934755 recall = 0.410144020038 f1 = 0.410529614541

5 Performance on develop set

precision = 0.371739130435 recall = 0.428213689482 f1 = 0.397982932506