CS6714 Project Part 2

Named Entity Disambiguation/Named Entity Linking

Adam Yi <z5231521@cse.unsw.edu.au>

Overview

I used learning2rank for this task by training a Gradient Boosting Decision Tree. Specifically, for each mention/candidate pair, a 25-dimensional feature vector is extracted, mapping to a boolean value (0 or 1) indicating whether the candidate is the correct entity or not. For pre-processing, inverted index is built for entire wikipedia corpus, mention docs. I also built separate inverted index when only considering tokens of entity type NUM, ORG, GPE, and PERSON respectively. All indexes are built with lemmas of tokens.

Feature Space

The feature vector contains the following variables:

1. n_nums: number of digit characters (0-9) in mention string
2. n_nums_2: number of digit characters (0-9) in candidate title
3. len_mention: number of tokens in mention string
4. len_title: number of tokens in candidate title
5. all_caps: boolean value (0/1) indicating whether mention string is all capitalized
6. n_caps: number of capitalized letters in mention string
7. ttfidf: wikipedia corpus TF-IDF score of mention string tokens
8. etfidf: wikipedia corpus TF-IDF score of mention string entities
9. atf: weighted sum of wikipedia corpus TF-IDF score of all tokens in mention doc, based on distance to mention string
10. stf: wikipedia corpus TF-IDF score of mention sentence tokens
11. ntf: TF-IDF score of all keywords (nouns and numbers) of mention sentence and adjacent sentences in wikipedia corpus
12. antf: weighted sum of wikipedia corpus TF-IDF score of keywords (nouns and numbers) in mention doc, based on distance to mention string
13. bm25: BM25 score of mention sentence in candidate doc
14. tbm25: BM25 score of candidate title in mention doc
15. title_tfidf: Sum of IDF of mention string tokens that also appear in candidate title
16. title_rtfidf: Sum of TF-IDF score of candidate title tokens in mention doc
17. root_title_tfidf: TF-IDF score of the root token in candidate title’s DPT (dependency parse

tree)

18. tfs: Cosine similarity of TF-IDF vectors of candidate doc and mention doc
19. ttfs: Cosine similarity of average TF vector of tokens in mention string and candidate title across wikipedia corpus
20. match_words: number of exact lemma matches at beginning of mention string and candidate title
21. all_match: boolean value (0/1) indicating whether mention string lemmas completely match candidate title lemmas

22-25 is the weighted TF-IDF score of mention doc tokens of entity type NUM, ORG, GPE, and PERSON respectively, based on the distance to mention string.

Learning Params

```
param = {
    'learning_rate': 0.1,
    'n_estimators': 50,
    'max_depth': 5,
    'min_child_weight': 1,
    'gamma': 0,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'objective': 'rank:pairwise',
}
```

Evaluation

95.1% accuracy for dev dataset. 81.9% accuracy for dev2 dataset.