CS839 Project Stage 1

Group members: Zhihan Guo, Jing Liu, Yiwu Zhong

March 10, 2019

1 Entity Type

We extracted the **person names** from The Records of the Grand Historian by Sima Qian. Here are some examples: We extracted the person names **Mengzeng** and **Cheng**, from sentence "Mengzeng enjoyed favour with King Cheng of the Zhou dynasty and is known as Zhai Black Wolf".

2 Total Number

- The number of documents in set I is 200.
- The number of documents in set J is 100.
- The number of mentions in set J is 348.

3 First Time Performance

- The type of the classifier that you selected after performing cross validation on set I *the first time*
  Decision Tree
- Performance on I
  - Precision: 0.8507
  - Recall: 0.8261
  - F1: 0.8382

4 Before Postprocessing Performance

- The type of the classifier that you have finally settled on *before* the rule-based postprocessing step
  Decision Tree
- Performance on J

- Precision = 0.9363
- Recall = 0.8448
- F1 = 0.8882

5 Rule-based Post-processing

If you have done any rule-based post-processing, then give examples of rules that you have used, and describe where can we find all the rules (e.g., is it in the code directory somewhere?).

We added one rule which is under code directory, in main.py: rule_based_post_processing
https://github.com/ScarletGuo/CS839-DataSciences/blob/9beead44a2077d59fe5113b5019091200fb3f2
main.py#L140

Description of the rule: for candidates generated from the same span, only set one candidate to true of all candidates labeled true in the span. And always prefer the candidate that is composed from more words. e.g. if we have Lord Ping Yuan, Ping and Ping Yuan may both set to true, while only Ping Yuan is true.

6 Final Performance

Report the precision, recall, F1 of classifier Y (see description above) on set J. This is the final classifier (plus rule-based post-processing if you have done any).

- Precision = 0.9105
- Recall = 0.8190
- F1 = 0.8623

7 Comment

If you have not reached precision of at least 90% and recall of at least 60%, provide a discussion on why, and what else can you possibly do to improve the accuracy.

- Our performance reaches the requirement most of the time.
- But due to the randomness of decision tree, the precision may below 0.9 occasionally but always above 0.84.