CS 839 Project Stage 1 Report

Zeping Ren, Stephen Sheen, Shiqi Yang

1 Entity Type

We extract location names from Winter Olympics news from different news sources (US, UK and Canada). This includes continent names, country names, state names, county names, and city names.

Examples: "Africa", "United States", "Washington", "Jeongseon", and "Sochi".

2 Training and Test Set Information

We labeled 315 documents with 1479 mentions. Then we shuffled the documents and select the first 200 to be set I and then 100 documents to be Set J. The rest 15 documents are not considered.

Set I:

- Number of documents: 200
- Number of mentions: 958

Set J:

- Number of documents: 100
- Number of mentions: 485

3 Model Information

Out of the suggest models, we choose SVM as it has the best performance. Below is the reported values on classifier M before we performed debugging.

- Precision: 0.97
- Recall: 0.67
- F1: 0.79

After debugging (mostly fixed label-tagging errors, and we used RBF kernel on our SVM model), we obtained the classifier X with the following metrics:

- Precision: 0.99
- Recall: 0.64
- F1: 0.78

Since we do not have rule-based post-processing steps. This is the final classifier Y. The precision, recall and F1 values on set J is as follows:

- Precision: 0.99
- Recall: 0.69
- F1: 0.81