CS 839 Project Stage 1 Report

Zeping Ren, Stephen Sheen, Shiqi Yang

1 Entity Type

We extracted 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 selected the first 200 documents to be set I and the next 100 documents to be Set J. The leftover 15 documents were 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 suggested models, we chose SVM as it had the best performance. Shown below are 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; we also used the RBF kernel on our SVM model), we obtained the classifier X with the following metrics:

- Precision: 0.99
- Recall: 0.76
- F1: 0.86

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

- Precision: 0.99
- Recall: 0.76
- F1: 0.86