Mixture of Experts Authorship Attribution

Michael Ryan, John Noecker Jr
Evaluating Variations in Language Lab
Duquesne University
mryan, jnoecker @ jgaap.com

Tools

- JGAAP (Java Graphical Authorship Attribution Program) - a modular test bed for authorship attribution methods.
  - All methods used are either available in JGAAP or were extensions of it
  - Source code for the methods used in this experiment is available at jgaap.com

Mixture of Experts

- Combined three Authorship Attribution techniques
- Each technique assigns a vote on the author of the document
- If there is not majority author assume the author was not in the sample group

Centroid L1

- Break documents into feature vectors of character 3-grams using relative frequencies of 3-grams
- Build Centroids for the known authors
  - Take the average of that authors feature vectors
- Measure the L1 Distance between the authors' centroids and the unknown's feature vector
- Assign your vote to the author whose centroid had the smallest L1 Distance

WEKA SMO

- Break documents into feature vectors of character 3-grams using relative frequencies of 3-grams
- Train WEKA's Sequential Minimal Optimization Support Vector Machines (SMO) using the known authors' feature vectors
- SMO will rate authors similarity
- Assign a vote to the most similar author

Repeated Microdocument Analysis

- Break all documents into 3,000 character chunks
- Reduce all contiguous whitespace to single spaces and all character to lower case
- Break chunks into feature vectors of character 11-grams using relative frequencies of 11-grams
- Generate Centroids for the known authors
  - Take the average of the author's feature vectors
- Measure the Intersection Distance between the author centroids and chunks, assigning the closest centroid's author to each chunk
- Vote on the author who receives a majority of the chunks

Author Diarization Method

- Break documents into paragraphs
- Extract named entities from paragraphs
- Group paragraphs with named entities in common
- Assume each group is an author
- Use the grouped paragraphs as known chunks with Repeated Microdocument Analysis and ungrouped paragraphs as unknowns
- Add the ungrouped paragraph that is closest to a group to that group and re-run the analysis until all paragraphs are grouped

Results

Problem	Number Correct	Total	Accuracy
A	6	6	100%
B	7	10	70%
C	7	8	87.5%
D	10	17	58.8%
E	83	90	92.2%
F	77	80	96.3%
I	12	14	85.7%
J	12	16	75.0%
Total	214	241	88.8%

Conclusions

- These methods show promise with document accuracy of 88.8% and mean accuracy of 83.2%, respectively first and third in the competition.
- The method used preformed poorly on open-class problems because they were developed with only closed class in mind, removing the open-class portions changes our accuracies to 91.6% and 88.5%

Future Work

- Refine analysis of open-class problems by examining how different experts preform in identifying them and how many experts it takes to reach a conclusion.