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

<table>
<thead>
<tr>
<th>Problem</th>
<th>Number Correct</th>
<th>Total</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>6</td>
<td>100%</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>10</td>
<td>70%</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>8</td>
<td>87.5%</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>17</td>
<td>58.8%</td>
</tr>
<tr>
<td>E</td>
<td>83</td>
<td>90</td>
<td>92.2%</td>
</tr>
<tr>
<td>F</td>
<td>77</td>
<td>80</td>
<td>96.3%</td>
</tr>
<tr>
<td>I</td>
<td>12</td>
<td>14</td>
<td>85.7%</td>
</tr>
<tr>
<td>J</td>
<td>12</td>
<td>16</td>
<td>75.0%</td>
</tr>
<tr>
<td>Total</td>
<td>214</td>
<td>241</td>
<td>88.8%</td>
</tr>
</tbody>
</table>
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 performed 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 perform in identifying them and how many experts it takes to reach a conclusion.