Inter-textual Distances, Three Authors

Thomas Merriam

Presented by Xiaodan Zhu
April 1st, 2004
Outline

- **Background**
  - Labbes’ Method

- **This paper**
  - Data/texts
    - Choice of data
    - Data preprocessing
  - Features used
  - Distance matrix
  - Attribution approaches
  - Experimental results

- **Conclusion**
Outline

- Background
  - Labbes’ Method

- This paper
  - Data/texts
    - Choice of data
    - Data preprocessing
  - Features used
  - Distance matrix
  - Attribution approaches
  - Experimental results

- Conclusion
Labbes’ Method

Inter-textual distance

- Absolute distance
  \[(N_a \cup N_b) - (N_a \cap N_b)\]

- Relative distance
  \[
  \delta_{(a,b)} = \frac{\sum_{V_a} |F_{ia} - F_{ib}| + \sum_{V_b} |F_{ib} - F_{ia}|}{N_a + N_b}
  \] (1)

  \[
  \delta_{(a,b)} = \frac{1}{2} \left( \frac{\sum_{V_a} |F_{ia} - F_{ib}|}{N_a} + \frac{\sum_{V_b} |F_{ib} - F_{ia}|}{N_b} \right)
  \] (2)

Where,

- \(V_a\) and \(V_b\): number of types in A and B
- \(F_{ia}\): frequency of the \(i\)th type in A
- \(F_{ib}\): frequency of the \(i\)th type in B
- \(N_a\) and \(N_b\): number of tokens in A and B with \(N_a = \sum F_{ia}\) and \(N_b = \sum F_{ib}\)
Labbes’ Method

Labbes’ distance

- First, calculate the absolute distance between A and B’

\[ D_{V_{a,b}(u)} = \sum_{V_{a}, V'_{b(E)}} |F_{ia} - E_{ia(u)}| \]

- Then, calculate distance between A and B

\[ D_{(a,b)} = \frac{\sum_{V_{a}, V'_{b(E)}} |F_{ia} - E_{ia(u)}|}{\sum_{V_{a}} F_{ia} + \sum_{E_{ia(u)}}} = \frac{\sum_{V_{a}, V'_{b(E)}} |F_{ia} - E_{ia(u)}|}{N_{a} + N'_{b}} \]

Where,

\[ E_{ia(u)} = F_{ib} \ast U_{(a,b)} \text{ with } U_{(a,b)} = \frac{N_{a}}{N_{b}} \]

\[ N'_{b} = \sum_{V_{b}} E_{ia(u)} \]
Labbes’ Method

Labbes’ distance scale

An author

Different authors

Minimal common nucleus for texts in the same language.

0.65

Different genres, remote topics

0.40

Minimal common nucleus for texts produced by a same author.

Different genres, remote topics

0.30

Similar genre = remote topics
Different genres = close topics

Same genre and topics
Possible authorship attribution

0.25

Same author, genre, topic.

Sure authorship attribution

0.20

0.10
Outline

- **Background**
  - Labbes’ Method

- **This paper**
  - Data/texts
    - Choice of data
    - Data preprocessing
  - Features used
  - Distance matrix
  - Attribution approaches
  - Experimental results

- **Conclusion**
## Choice of Data

<table>
<thead>
<tr>
<th>Text</th>
<th>Code</th>
<th>Author</th>
<th>Date</th>
<th>Genre</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Pooh Perplex</em> 3–30</td>
<td>C63a</td>
<td>Crews</td>
<td>1963</td>
<td>Parodies</td>
<td></td>
</tr>
<tr>
<td><em>The Pooh Perplex</em> 87–99; 100–12</td>
<td>C63b</td>
<td>Crews</td>
<td>1963</td>
<td>Parodies</td>
<td></td>
</tr>
<tr>
<td><em>Postmodern Pooh</em> 4–17; 20–31</td>
<td>C01a</td>
<td>Crews</td>
<td>2001</td>
<td>Parodies</td>
<td></td>
</tr>
<tr>
<td><em>Postmodern Pooh</em> 147–61; 163–70</td>
<td>C01b</td>
<td>Crews</td>
<td>2001</td>
<td>Parodies</td>
<td></td>
</tr>
<tr>
<td><em>Shakespeare: Text, Stage, and Canon</em> 1–21</td>
<td>P01a</td>
<td>Proudfoot</td>
<td>2001</td>
<td>Prose Exposition</td>
<td>Text</td>
</tr>
<tr>
<td><em>Shakespeare: Text, Stage, and Canon</em> 32–56</td>
<td>P01b</td>
<td>Proudfoot</td>
<td>2001</td>
<td>Prose Exposition</td>
<td>Stage</td>
</tr>
<tr>
<td><em>Shakespeare: Text, Stage, and Canon</em> 61–82</td>
<td>P01c</td>
<td>Proudfoot</td>
<td>2001</td>
<td>Prose Exposition</td>
<td>Canon</td>
</tr>
<tr>
<td>‘New Conservatism and the Theatrical Text’ 127–37</td>
<td>P02</td>
<td>Proudfoot</td>
<td>2002</td>
<td>Prose Exposition</td>
<td>Text</td>
</tr>
<tr>
<td><em>Shakespeare: a Dramatic Life</em> 22–37</td>
<td>W94b</td>
<td>Wells</td>
<td>1994</td>
<td>Prose Exposition</td>
<td>The Actor</td>
</tr>
<tr>
<td><em>Shakespeare: a Dramatic Life</em> 214–31</td>
<td>W94c</td>
<td>Wells</td>
<td>1994</td>
<td>Prose Exposition</td>
<td>Summaries</td>
</tr>
</tbody>
</table>
Data Preprocessing

- Input the texts
  - No quantified estimation

- Trim text length
  - By tokens (5,355 tokens per text)

- What is counted as a word
  - An English word
  - A Roman number
  - A Arabic number
  - A hyphenated words
Features Used

- Features for authorship attribution
  - Lexicon (*This paper use lexical features*)
  - N-gram
  - Sentence length
  - Redundancy information
  - Others
Distance Matrix

Distance between text A and B

\[ \delta_{(a,b)} = \frac{\sum_{i \in V_{(a+b)}} |F_{ia} - F_{ib}|}{2N_a} \]

Where,

- \( V_{(a+b)} \), number of types in A and B = 10,710 words
- \( F_{ia} \): frequency of the \( i \)th type in A
- \( F_{ib} \): frequency of the \( i \)th type in B
- \( N_a \) and \( N_b \): number of tokens in A and B with \( N_a = \sum F_{ia} = 5,355 \) and \( N_b = \sum F_{ib} = 5,355 \).
## Distance Matrix

Inter-textual distances between texts

<table>
<thead>
<tr>
<th></th>
<th>C63a</th>
<th>C63b</th>
<th>C01a</th>
<th>C01b</th>
<th>P01a</th>
<th>P01b</th>
<th>P01c</th>
<th>P02</th>
<th>W67</th>
<th>W72</th>
<th>W86</th>
<th>W94a</th>
<th>W94b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C63b</td>
<td>0.410</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C01a</td>
<td>0.457</td>
<td>0.458</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C01b</td>
<td>0.467</td>
<td>0.465</td>
<td>0.469</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01a</td>
<td>0.532</td>
<td>0.545</td>
<td>0.561</td>
<td>0.558</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01b</td>
<td>0.507</td>
<td>0.523</td>
<td>0.536</td>
<td>0.530</td>
<td>0.428</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01c</td>
<td>0.527</td>
<td>0.538</td>
<td>0.555</td>
<td>0.544</td>
<td>0.406</td>
<td>0.415</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P02</td>
<td>0.531</td>
<td>0.544</td>
<td>0.567</td>
<td>0.559</td>
<td>0.344</td>
<td>0.412</td>
<td>0.420</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W67</td>
<td>0.480</td>
<td>0.497</td>
<td>0.529</td>
<td>0.542</td>
<td>0.496</td>
<td>0.456</td>
<td>0.481</td>
<td>0.496</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W72</td>
<td>0.483</td>
<td>0.489</td>
<td>0.527</td>
<td>0.533</td>
<td>0.507</td>
<td>0.472</td>
<td>0.479</td>
<td>0.501</td>
<td>0.397</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W86</td>
<td>0.547</td>
<td>0.545</td>
<td>0.550</td>
<td>0.549</td>
<td>0.480</td>
<td>0.449</td>
<td>0.441</td>
<td>0.493</td>
<td>0.463</td>
<td>0.460</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W94a</td>
<td>0.492</td>
<td>0.483</td>
<td>0.496</td>
<td>0.490</td>
<td>0.470</td>
<td>0.453</td>
<td>0.443</td>
<td>0.488</td>
<td>0.456</td>
<td>0.444</td>
<td>0.336</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W94b</td>
<td>0.488</td>
<td>0.496</td>
<td>0.512</td>
<td>0.516</td>
<td>0.455</td>
<td>0.421</td>
<td>0.429</td>
<td>0.459</td>
<td>0.434</td>
<td>0.431</td>
<td>0.379</td>
<td>0.376</td>
<td></td>
</tr>
<tr>
<td>W94c</td>
<td>0.496</td>
<td>0.509</td>
<td>0.522</td>
<td>0.546</td>
<td>0.511</td>
<td>0.485</td>
<td>0.497</td>
<td>0.508</td>
<td>0.430</td>
<td>0.405</td>
<td>0.486</td>
<td>0.460</td>
<td>0.450</td>
</tr>
</tbody>
</table>
Typical Authorship Attribution methods

- **Lexical based methods**
  - PCA *(This paper use it)*
  - Cusum Analysis

- **Non-lexical based methods**
  - N-gram
  - Zipping
Results

- Texts distribution based on first two principal components
## Results

**Euclidian distances between texts, using first two principal components**

<table>
<thead>
<tr>
<th></th>
<th>C63a</th>
<th>C63b</th>
<th>C01a</th>
<th>C01b</th>
<th>P01a</th>
<th>P01b</th>
<th>P01c</th>
<th>P02</th>
<th>W67</th>
<th>W72</th>
<th>W86</th>
<th>W94a</th>
<th>W94b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C63b</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C01a</td>
<td>0.26</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C01b</td>
<td>0.49</td>
<td>0.52</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01a</td>
<td>2.63</td>
<td>2.73</td>
<td>2.69</td>
<td>2.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01b</td>
<td>2.20</td>
<td>2.31</td>
<td>2.34</td>
<td>2.24</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01c</td>
<td>2.38</td>
<td>2.48</td>
<td>2.51</td>
<td>2.39</td>
<td>0.74</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P02</td>
<td>2.62</td>
<td>2.73</td>
<td>2.68</td>
<td>2.49</td>
<td>0.06</td>
<td>0.92</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W67</td>
<td>1.99</td>
<td>2.04</td>
<td>2.25</td>
<td>2.39</td>
<td>2.74</td>
<td>1.87</td>
<td>2.05</td>
<td>2.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W72</td>
<td>2.11</td>
<td>2.16</td>
<td>2.37</td>
<td>2.54</td>
<td>2.99</td>
<td>2.13</td>
<td>2.31</td>
<td>3.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W86</td>
<td>2.43</td>
<td>2.51</td>
<td>2.67</td>
<td>2.74</td>
<td>2.35</td>
<td>1.48</td>
<td>1.62</td>
<td>2.40</td>
<td>0.75</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W94a</td>
<td>2.15</td>
<td>2.21</td>
<td>2.40</td>
<td>2.50</td>
<td>2.51</td>
<td>1.63</td>
<td>1.79</td>
<td>2.55</td>
<td>0.35</td>
<td>0.56</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W94b</td>
<td>2.30</td>
<td>2.37</td>
<td>2.54</td>
<td>2.62</td>
<td>2.33</td>
<td>1.45</td>
<td>1.60</td>
<td>2.37</td>
<td>0.65</td>
<td>0.84</td>
<td>0.14</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>W94c</td>
<td>1.86</td>
<td>1.91</td>
<td>2.13</td>
<td>2.29</td>
<td>2.80</td>
<td>1.95</td>
<td>2.14</td>
<td>2.83</td>
<td>0.18</td>
<td>0.26</td>
<td>0.93</td>
<td>0.52</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Results

- Results analysis
  - Distance and similarity between P01a and P02
  - Distance and similarity between C63a and C63b
    - Whether an author can change his underlying style while consciously imitating that of others?
  - Distance and similarity between w86 and w94B
Conclusions

- Conclusions drawn by this paper
  - The distance used discriminates the three authors
  - Chronology does not feature strongly
  - This is no marked difference in the parodies in the books by Crews

- Contribution of this paper
Thanks