Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio N. Nemmi, Valerio Neri, Julinda Stefa.

Speaker
Eugenio N. Nemmi
• Authorship attribution is the task of identifying the author of a given text.
Motivation

Detect real author of a novel

Stephen King
Motivation

Detect author of paper in double blind review submission

Who wrote this paper?

ABSTRACT
Far out in the uncharted backwaters of the unfashionable end of the western spiral arm of the Galaxy lies a small un-regarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green planet whose ape-descended life forms are so amazingly primitive that they still think digital watches are a pretty neat idea. This planet has - or rather had - a problem, which was this: most of the people on it were unhappy for pretty much of the time. Many solutions were suggested for this problem, but most of these were largely concerned with

1 INTRODUCTION
Sadly, however, before she could get to a phone to tell anyone about it, a terribly stupid catastrophe occurred, and the idea was lost forever. This is not her story. But it is the story of that terrible stupid catastrophe and some of its consequences. It is also the story of a book, a book called The Hitch Hiker's Guide to the Galaxy - not an Earth book, never published on Earth, and until the terrible catastrophe occurred, never seen or heard of by any Earthman. Nevertheless, a wholly remarkable book. in fact it was probably the most remarkable book ever to come out of the great publishing houses of Ursa
Motivation

Deanonymize Pseudonyms
Fifotofotofoto donono dodod did ididod odofofof ififododi.

Woooooooa!

Noot noot!
AA Scenarios

Closed-set
Finite set of candidates authors among which there is the real author.

Open-set
The author of a disputed text is not necessarily included in the list of candidates.
Single-Domain vs Cross-Domain
PAN 2019 Authorship Attribution Task

Open-set

Cross-Domain
Main approaches to AA problems

Profile-Based Features

Instance-Based Features

features

features

features

features
Profile-Base features

Profile-Based Features

- Concatenate together texts of the same author.
- Collecting as more information of the user as possible.
- Differences between the training texts by the same author are disregarded.
- Stylometric measures extracted from the concatenated file may be quite different in comparison to each of the original training texts.
Instance-Based Features

- Analyze the texts associated with an author separately.
- Classification algorithms require multiple training instances per class for extracting a reliable model.
- The text samples should be long enough so that the text representation features can represent adequately their style.
Text Pre-Processing

• Pre-processing is a crucial step to prepare the data in almost every NLP problems.

• Text pre-processing usually consists in normalize, sanitize or alter the text to remove noise, error, or completely change the data format.

• We used:

  WordPunctTokenizer   SnowballStemmer   spaCy POS Tagger
Text Distortion


<table>
<thead>
<tr>
<th>Original Text</th>
<th>Text converted with Text Distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>marqué sur la couverture, avant d’avoir un temps d’arrêt. Le dossier se nommait en effet sobrement « Enterrement de vie de garçon ». Plusieurs souvenirs remontèrent. John sourit doucement en se remém</td>
<td><strong><strong><strong>é  ****  ***  ****************  **  ****  <em>'</em></strong></strong></strong>**  **  ***** *'*<strong>é''  ***  *********  ***  *********  ***  ****  **  *********  «  *****************  ***  ****  **  *** é</strong> »  *********  *********  ****<em><strong><strong>é</strong></strong></em>  ***  *********  *********  *********  ***  **  **<em><strong>é</strong></em></td>
</tr>
</tbody>
</table>
Features

- Profile
- Char3-5
- Stem1-3
- Dist3-5
- Stem1&2
- Pos3-5
Model

- Profile -> SVM
- Char3-5 -> SVM
- Stem1-3 -> SVM
- Dist3-5 -> SVM
- Stem1&2 -> SVM
- Pos3-5 -> SVM
Unknown Prediction

$P_i$ $i$-th most probable author for a given text

$Unknown = \begin{cases} 
\text{True, } P_1 - P_2 < 0.1 \land \text{mean } (P_1 - P_2, P_1 - P_3) < 0.7 \\
\text{False, otherwise}
\end{cases}$
## Result on DEV

<table>
<thead>
<tr>
<th>Problem</th>
<th>Baseline-SVM</th>
<th>Baseline-Comp</th>
<th>Ensemble</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>69.5</td>
<td>68.2</td>
<td>82.2</td>
<td>12.7</td>
</tr>
<tr>
<td>02</td>
<td>44.7</td>
<td>33.6</td>
<td>56.2</td>
<td>11.5</td>
</tr>
<tr>
<td>03</td>
<td>49.3</td>
<td>50.1</td>
<td>73.0</td>
<td>23.7</td>
</tr>
<tr>
<td>04</td>
<td>33.1</td>
<td>49.0</td>
<td>51.1</td>
<td>18.0</td>
</tr>
<tr>
<td>05</td>
<td>47.1</td>
<td>34.0</td>
<td>56.2</td>
<td>9.1</td>
</tr>
<tr>
<td>06</td>
<td>70.2</td>
<td>69.1</td>
<td>65.6</td>
<td>-4.6</td>
</tr>
<tr>
<td>07</td>
<td>49.9</td>
<td>54.2</td>
<td>63.8</td>
<td>13.9</td>
</tr>
<tr>
<td>08</td>
<td>50.6</td>
<td>49.2</td>
<td>65.6</td>
<td>15.0</td>
</tr>
<tr>
<td>09</td>
<td>59.9</td>
<td>60.8</td>
<td>73.8</td>
<td>13.9</td>
</tr>
<tr>
<td>10</td>
<td>44.2</td>
<td>50.1</td>
<td>57.3</td>
<td>13.1</td>
</tr>
<tr>
<td>11</td>
<td>65.1</td>
<td>59.5</td>
<td>73.7</td>
<td>8.6</td>
</tr>
<tr>
<td>12</td>
<td>59.4</td>
<td>50.8</td>
<td>71.0</td>
<td>11.6</td>
</tr>
<tr>
<td>13</td>
<td>68.7</td>
<td>73.1</td>
<td>74.3</td>
<td>5.6</td>
</tr>
<tr>
<td>14</td>
<td>59.8</td>
<td>78.0</td>
<td>83.3</td>
<td>23.5</td>
</tr>
<tr>
<td>15</td>
<td>74.5</td>
<td>71.2</td>
<td>82.1</td>
<td>7.6</td>
</tr>
<tr>
<td>16</td>
<td>76.8</td>
<td>70.5</td>
<td>88.3</td>
<td>11.5</td>
</tr>
<tr>
<td>17</td>
<td>58.4</td>
<td>62.3</td>
<td>81.7</td>
<td>23.3</td>
</tr>
<tr>
<td>18</td>
<td>70.3</td>
<td>65.9</td>
<td>87.8</td>
<td>17.5</td>
</tr>
<tr>
<td>19</td>
<td>55.6</td>
<td>40.3</td>
<td>71.0</td>
<td>15.4</td>
</tr>
<tr>
<td>20</td>
<td>51.3</td>
<td>22.3</td>
<td>54.1</td>
<td>2.8</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>57.9</strong></td>
<td><strong>55.6</strong></td>
<td><strong>70.5</strong></td>
<td><strong>12.6</strong></td>
</tr>
</tbody>
</table>
Further analysis

- Closed-Set scenario accuracy of 87% on a total of 2,646 documents.

- Closed-Set scenario with Unknowns detector achieve as overall result an accuracy of 78.7%

- Difference in results of 8.7%
TIRA
<table>
<thead>
<tr>
<th>User</th>
<th>Software</th>
<th>Run</th>
<th>Input run</th>
<th>mean macro-f1</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>muttenthaler19</td>
<td>software1</td>
<td>2019-05-12-22-41-24</td>
<td>2019-05-12-21-58-10</td>
<td>0.69</td>
<td>00:33:16</td>
</tr>
<tr>
<td>ner19</td>
<td>software1</td>
<td>2019-05-11-17-41-45</td>
<td>2019-05-11-16-30-11</td>
<td>0.68</td>
<td>01:06:08</td>
</tr>
<tr>
<td>devries19</td>
<td>software3</td>
<td>2019-05-11-08-11-38</td>
<td>2019-05-10-16-46-09</td>
<td>0.644</td>
<td>11:19:32</td>
</tr>
<tr>
<td>delcamporodriguez19</td>
<td>software5</td>
<td>2019-05-12-10-42-54</td>
<td>2019-05-12-08-39-19</td>
<td>0.642</td>
<td>01:59:17</td>
</tr>
<tr>
<td>isbister19</td>
<td>software1</td>
<td>2019-05-11-14-51-16</td>
<td>2019-05-10-11-00-34</td>
<td>0.622</td>
<td>01:05:32</td>
</tr>
<tr>
<td>johansson19</td>
<td>software1</td>
<td>2019-05-07-10-52-58</td>
<td>2019-05-07-08-53-03</td>
<td>0.616</td>
<td>01:05:30</td>
</tr>
<tr>
<td>basile19</td>
<td>software1</td>
<td>2019-05-16-16-25-40</td>
<td>2019-05-16-16-02-32</td>
<td>0.613</td>
<td>00:17:08</td>
</tr>
<tr>
<td>vanhalteren19</td>
<td>software1</td>
<td>2019-05-16-12-08-13</td>
<td>2019-05-14-15-13-20</td>
<td>0.598</td>
<td>37:05:47</td>
</tr>
<tr>
<td>rahgouy19</td>
<td>software1</td>
<td>2019-05-08-17-27-16</td>
<td>2019-05-08-13-56-28</td>
<td>0.58</td>
<td>02:52:03</td>
</tr>
</tbody>
</table>
Conclusion

• Ensemble model with a classifier for each feature.

• We combine Profile-Based and Instance-Based features together.

• We introduced a method that takes into account the three most similar author for the disputed text, instead of only the first two.

• We outperform the baseline in almost every problems.
Future Work

- Although our methodology to detect the unknown authors performs slightly better than the baseline, further improvements are needed.

- In one problem we reach a score lower than the baseline. It could be useful to understand the reason of it.

- Neural Networks approach could be tested.
Gracias
Arigato
Shukura
Thank you
Bolzín
Merci
Question?