A new Resource for Analyzing Collaborative Writing Styles and One-Sidedness

Scientific Authorship and Peer Review: Between a Means of Governance and Structural Meaning?

01.–02.09.2022, HU Berlin

Erik Körner
Harry Scells
Benno Stein
Outline

① Multi-Authorship Identification

② Collaborative Writing Styles

③ SMAuC - The Scientific Multi-Authorship Corpus

④ Researching Algorithmic Bias
Multi-Authorship Identification

Introduction

- Multi-Authorship Identification/Analysis an important variant of the vanilla (Single) Authorship Identification problem
  - Single-Author: “Who is the author?” of a letter, an article, or a book
  - Multi-Authorship Identification: questions and issues about documents written by a group of authors

- Authorship in academia ➔ often multiple authors
  - Intentional, collaborative writing
  - Text reuse, plagiarism, …

- Increased attention and application of Multi-Authorship Identification
  - Numerous tasks, datasets and methods over the years
  - PAN, various shared task and datasets

  - However, style of collaborative writing mostly the same
  - Very little (public) academic datasets, affects and hinders comparability of approaches against each other
Multi-Authorship Identification
Author Identification Tasks in Literature

❑ Single-Author: [Thomas Corvin Mendenhall 1887]
❑ Multi-Author: [Glover and Hirst 1996]
- **Task**: Single author or multiple authors?

- Only very few studies that solely address this problem
- Often as consequence or reduction of more complex result, e.g. author count, style changes
- Many datasets with assumption that texts are multi-authored and then just application of more ‘sophisticated’ methods
Multi-Authorship Identification
Author Count Prediction

- **Task**: Number of authors?

- Fundamental multi-author identification task
- Application not limited to human-readable texts, e.g. compiled binary software
Multi-Authorship Identification

Style Change Detection

- **Task**: Identify boundary where style of text changes.

- Sub-tasks that require to first segment the text vs. pre-segmented texts
Multi-Authorship Identification

Multi-Author Attribution

- **Task**: Attribution of text segments

- **Sub-tasks**:
  - Authors: closed-set vs. open-set
  - Document: local vs. global
Multi-Authorship Identification
Multi-Author Attribution Sub-Tasks

Author X
Author Y
Author Z
Author X
Segmented Multi-authored text
Style 1
Style 2
Style 3
Style 1
Unknown text
Known texts/authors
Arbitrary style labels

Erik Körner
At PAN, multi-authorship identification datasets have been constructed so far by combining texts that are written by single authors into a single, multi-authored text. Multi-Author Attribution research often only focuses on the metadata of the text, e.g. author list of journal articles. But Collaborative Writing Styles are not really taken into account when developing methods to address Multi-Author Identification.

What are the different types a text can be written collaboratively? Where does research (currently) happen?
Collaborative Writing Styles
Types of Writing Styles

- Sequential
- Group Single
- Horizontal Division
- Stratified Division
- Reactive

Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice [Lowry et al. 2004]
Collaborative Writing Styles

Sequential Writing Style

Author X
Author Y
Author Z
Author X
Collaborative Writing Styles
Sequential Writing Style

Characteristics
- Each author writes a section of the text, sequentially, independently
- Boundaries of authorial style explicitly defined, co-authors are not allowed to edit outside of their section of text

Examples
- Collaboration of a PhD student and supervisor on a research paper; supervisor writing the introduction and conclusion, student the content in between

Tasks
- Multi-Author Detection, Style Change Detection, Multi-Author Attribution, . . .
Collaborative Writing Styles

Group Single Writing Style

Author X → Author Y → Author Z → Author W

Author W

14 Erik Körner
Collaborative Writing Styles
Group Single Writing Style

Characteristics

❑ Several authors contribution to the ideation phrase of writing
❑ Single author compiles these into a single text
❑ Consistent authorship style, yet involvement of multiple authors in creation

Examples

❑ Grant writing: many principal investigators or collaborators involved in ideation, chief investigator writes proposal document

→ Multi-Authorship Identification methods may not be applicable?
❑ Are there style boundaries for Style Change Detection?
❑ Multi-Author Detection may be possible?
Collaborative Writing Styles
Horizontal Division Writing Style

Author X
Author Y
Author Z

Author W
Collaborative Writing Styles
Horizontal Division Writing Style

Characteristics

- Several authors contribute ‘sub-documents’
- Single authors compiles these into a single text
- Compiled text may contain authorship styles of co-authors, depending on the amount of editing applied

Examples

- Academic book: several academics write different chapters, an editor combines them into a cohesive manuscript
- Text Reuse

Notes

- Mainly targets *Style Change Detection* task
- Easiest and most obvious way to create artificial datasets
Collaborative Writing Styles

Stratified Division Writing Style

- Author
- Editor
- Reviewer
Collaborative Writing Styles
Stratified Division Writing Style

Characteristics

- Similar to *Horizontal Division*
- Each co-author plays a certain role in the creation of a text, e.g. author, editor, reviewer

Examples

- Scholarly article: one author writes majority of text, another author edits the text, an independent reviewer provides critical feedback that feeds back into the creation process
Collaborative Writing Styles

Reactive Writing Style

Author X

Author W

Author Y

Author Z
Collaborative Writing Styles

Reactive Writing Style

Characteristics

- Authors write synchronously on the same text while adjusting the writing of others

Examples

- Several undergraduate students in a group assignment writing a report together
- Collaborative writing platforms, e.g. Overleaf, Etherpad, Google Docs

Notes

- Blurred authorial style boundaries
- Most complex in terms of developing Multi-Author Identification methods
Collaborative Writing Styles

Observations

- Different writing styles may be *easier* or *harder* to apply Multi-Author Identification methods to.

- Boundaries are more clearly defined in *Horizontal Division* compared to *Reactive Writing Style*.

- Some multi-authorship approaches are impossible to apply, e.g. *Style Change Detection* to *Reactive Writing Style*.

- In literature most datasets for Multi-Author Identification are created using *Horizontal Division*!
Collaborative Writing Styles

Observations

- Different writing styles may be *easier* or *harder* to apply Multi-Author Identification methods to.
- Boundaries are more clearly defined in *Horizontal Division* compared to *Reactive Writing Style*.
- Some multi-authorship approaches are impossible to apply, e.g. *Style Change Detection* to *Reactive Writing Style*.
- In literature most datasets for Multi-Author Identification are created using *Horizontal Division*!

→ Existing methods may not be robust against different Collaborative Writing Styles.

→ The way in which multi-authored texts are created is fundamental to which tasks are applicable and to the difficulty in applying methods to those tasks.
## Collaborative Writing Styles
### Overview over Datasets and Methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset Task</th>
<th>Dataset Source</th>
<th>Collaborative Writing Style</th>
<th>#Docs &amp; Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MULTI-AUTHOR DETECTION (MAD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Glover et al. 1996]</td>
<td>AV</td>
<td>film summaries</td>
<td>HD</td>
<td>20 <strong>self</strong></td>
</tr>
<tr>
<td><strong>MULTI-AUTHOR ATTRIBUTION (MAA)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Tschuggnall et al. 2014]</td>
<td>MAD</td>
<td>Gutenberg/FED</td>
<td>HD</td>
<td>75 <strong>self</strong></td>
</tr>
<tr>
<td>[Payer et al. 2014]</td>
<td>AA</td>
<td>conference papers</td>
<td>S</td>
<td>3,516/-/378 <strong>self</strong></td>
</tr>
<tr>
<td>[Dauber et al. 2017]</td>
<td>AA</td>
<td>Wookiepedia</td>
<td>R</td>
<td>- <strong>self</strong></td>
</tr>
<tr>
<td>[Sarwar et al. 2018]</td>
<td>MAA</td>
<td>Gutenberg/arXiv</td>
<td>HD, S</td>
<td>6,173 <strong>self</strong> +1</td>
</tr>
<tr>
<td>MLPA-400 [Boumber et al. 2018]</td>
<td>MAA</td>
<td>ML papers</td>
<td>S</td>
<td>400 <strong>self</strong></td>
</tr>
<tr>
<td>[Brian Yu 2019]</td>
<td>MAA</td>
<td>Gutenberg</td>
<td>HD</td>
<td>- <strong>self</strong></td>
</tr>
<tr>
<td><strong>AUTHOR COUNT PREDICTION (ACP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Rexha et al. 2016]</td>
<td>ACP</td>
<td>PubMed</td>
<td>S</td>
<td>6,144 <strong>self</strong></td>
</tr>
<tr>
<td>[Alrabaee et al. 2019]</td>
<td>ACP</td>
<td>open-source code</td>
<td>HD</td>
<td>31,150 <strong>self</strong></td>
</tr>
<tr>
<td>PAN19 SCD [Zangerle et al. 2019]</td>
<td>ACP</td>
<td>StackExchange</td>
<td>HD</td>
<td>2,546/1,272/1,210 PAN: 2</td>
</tr>
<tr>
<td><strong>STYLE-CHANGE DETECTION (SCD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Graham et al. 2005]</td>
<td>SCD</td>
<td>Usenet</td>
<td>HD</td>
<td>- <strong>self</strong></td>
</tr>
<tr>
<td>[Akiva et al. 2013]</td>
<td>SCD/AC</td>
<td>Biblical/Blogs/ NYT</td>
<td>HD</td>
<td>- <strong>self</strong> +2</td>
</tr>
<tr>
<td>PAN16 AD [E. Stamatatos 2016]</td>
<td>AD</td>
<td>Webis-TRC-12</td>
<td>HD</td>
<td>174/-/8 PAN: 2</td>
</tr>
<tr>
<td>PAN18 SCD [Kestemont et al. 2018]</td>
<td>MAD</td>
<td>StackExchange</td>
<td>HD</td>
<td>2,980/1,492/1,352 PAN: 5</td>
</tr>
<tr>
<td>PAN21 SCD [Zangerle et al. 2021]</td>
<td>SCD</td>
<td>StackExchange</td>
<td>HD</td>
<td>11,200/2,400/2,400 PAN: 5</td>
</tr>
</tbody>
</table>

**HD**: Horizontal Division (randomly combining text fragments from different authors), **R**: Reactive, **S**: scientific papers (combination of Group-single, Stratified Division, Reactive; no stylistic ‘editing’ by dataset creators)
Motivation

- *Scientific writing* as a new and interesting domain for authorship analysis, especially for *Multi-Authorship Analysis*

- Most datasets lack material from science domain or required metadata

- Research often only with small *unpublished* datasets using arXiv preprints, PubMed articles or journal papers
  ➔ Reproduction and comparability difficult due to varying approaches for data preprocessing and dataset curation

- Very few publication that publish their research, e.g. MLPA-400 [Boumber et al. 2018]

➔ Requirement for large, openly accessible dataset of scientific works
SMAuC - The Scientific Multi-Authorship Corpus

Motivation

- *Scientific writing* as a new and interesting domain for authorship analysis, especially for *Multi-Authorship Analysis*

- Most datasets lack material from science domain or required metadata

- Research often only with small *unpublished* datasets using arXiv preprints, PubMed articles or journal papers
  ➔ Reproduction and comparability difficult due to varying approaches for data preprocessing and dataset curation

- Very few publication that publish their research, e.g. MLPA-400 [Boumber et al. 2018]

➔ Requirement for large, openly accessible dataset of scientific works
SMAuC - The Scientific Multi-Authorship Corpus

Dataset Sources

- **CORE database** [Knoth et al. 2011] [Knoth and Zdrahal 2012]
  - Collection of metadata and full texts of open access scientific publications
  - Dump from 2018-03-01\(^1\)
  - 123M metadata items, 85.6M items w/ abstracts, 9.8M items w/ **full texts**

- **Microsoft Open Academic Graph (OAG)** [Sinha et al. 2015]
  - Openly accessible heterogeneous knowledge graph based on scientific articles, authors, and institutions
  - Source for identifying and disambiguating authors and fields of study
  - Version 2 of the OAG [Hu et al. 2020]\(^2\)
  - 179M nodes, 2B edges

---
\(^1\)[https://core.ac.uk/services/dataset](https://core.ac.uk/services/dataset)
## SMAuC - The Scientific Multi-Authorship Corpus
### Dataset Curation Process

<table>
<thead>
<tr>
<th>Conditions applied</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORE</td>
<td>123,988,821 (100.00%)</td>
</tr>
<tr>
<td>↔ full texts</td>
<td>9,835,064 (7.93%)</td>
</tr>
<tr>
<td>↔ text language filtering</td>
<td>6,531,442 (5.27%)</td>
</tr>
<tr>
<td>↔ OAG matching</td>
<td>3,508,509 (2.82%)</td>
</tr>
<tr>
<td>↔ text quality assurance</td>
<td>3,356,686 (2.70%)</td>
</tr>
</tbody>
</table>

- **High requirements on data quality**
  - Multi-step language filtering with fastText
  - Improved mapping of full texts and OAG metadata using DOIs and titles
  - Manual mapping of heterogenous OAG *field of study* → *DFG Classification of Scientific Disciplines and Research Areas* [DFG 2016]
  - Removal of markup, non-ASCII characters; lowercasing, collapsing whitespaces
  - Additional (heuristical) filtering for text quality, e.g. text length, language
SMAuC - The Scientific Multi-Authorship Corpus
Counts for all types of documents and their total

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single author w/o multi author</td>
<td>711,471</td>
</tr>
<tr>
<td>Single author w/ multi author</td>
<td>261,629</td>
</tr>
<tr>
<td>Multi author w/o single author</td>
<td>1,481,106</td>
</tr>
<tr>
<td>Multi author w/ single author</td>
<td>894,945</td>
</tr>
<tr>
<td>No author information</td>
<td>7,535</td>
</tr>
<tr>
<td>Total</td>
<td>3,356,686</td>
</tr>
</tbody>
</table>
## SMAuC - The Scientific Multi-Authorship Corpus

Number of documents in the corpus by text length in characters and document type with percentage per row

<table>
<thead>
<tr>
<th>Length</th>
<th>Total</th>
<th>Single author (%)</th>
<th>Multi author (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 3,000</td>
<td>39,300</td>
<td>13,680 (1.41%)</td>
<td>25,567 (1.07%)</td>
</tr>
<tr>
<td>– 5,000</td>
<td>96,067</td>
<td>32,059 (3.29%)</td>
<td>63,832 (2.69%)</td>
</tr>
<tr>
<td>– 50,000</td>
<td>2,273,246</td>
<td>467,844 (48.07%)</td>
<td>1,799,435 (75.73%)</td>
</tr>
<tr>
<td>– 250,000</td>
<td>771,756</td>
<td>301,975 (31.03%)</td>
<td>468,473 (19.72%)</td>
</tr>
<tr>
<td>&gt; 250,000</td>
<td>176,317</td>
<td>157,542 (16.19%)</td>
<td>18,744 (0.79%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3,356,686</td>
<td>973,100 (100.00%)</td>
<td>2,376,051 (100.00%)</td>
</tr>
</tbody>
</table>
# SMAuC - The Scientific Multi-Authorship Corpus

Document counts by research area [DFG 2016]

<table>
<thead>
<tr>
<th>Research Area</th>
<th>SA</th>
<th>MA</th>
<th>A</th>
<th>TL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering Sciences</td>
<td>55,015</td>
<td>375,206</td>
<td>3</td>
<td>28,467</td>
</tr>
<tr>
<td>Humanities</td>
<td>58,317</td>
<td>199,926</td>
<td>3</td>
<td>37,224</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>48,723</td>
<td>715,218</td>
<td>5</td>
<td>32,616</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>147,024</td>
<td>651,076</td>
<td>3</td>
<td>26,103</td>
</tr>
</tbody>
</table>

Single author documents (SA), multi author documents (MA), median authors per document (A) and median text length (TL).
**SMAuC - The Scientific Multi-Authorship Corpus**

Total author count over the number of single-author and multi-author publications per author

<table>
<thead>
<tr>
<th>Multi-author docs. per author</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>20627</td>
<td>3990</td>
<td>1399</td>
<td>667</td>
<td>344</td>
<td>208</td>
<td>137</td>
<td>106</td>
<td>56</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>11222</td>
<td>2491</td>
<td>947</td>
<td>465</td>
<td>251</td>
<td>168</td>
<td>99</td>
<td>80</td>
<td>45</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>7711</td>
<td>1863</td>
<td>759</td>
<td>319</td>
<td>181</td>
<td>122</td>
<td>83</td>
<td>53</td>
<td>32</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>5742</td>
<td>1420</td>
<td>589</td>
<td>308</td>
<td>176</td>
<td>116</td>
<td>59</td>
<td>52</td>
<td>48</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>4371</td>
<td>1167</td>
<td>519</td>
<td>242</td>
<td>154</td>
<td>94</td>
<td>57</td>
<td>41</td>
<td>30</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>3603</td>
<td>1022</td>
<td>460</td>
<td>249</td>
<td>131</td>
<td>79</td>
<td>58</td>
<td>37</td>
<td>31</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>2862</td>
<td>833</td>
<td>372</td>
<td>192</td>
<td>119</td>
<td>74</td>
<td>46</td>
<td>36</td>
<td>22</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>2426</td>
<td>677</td>
<td>298</td>
<td>172</td>
<td>112</td>
<td>61</td>
<td>41</td>
<td>35</td>
<td>15</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>2076</td>
<td>613</td>
<td>287</td>
<td>166</td>
<td>77</td>
<td>53</td>
<td>44</td>
<td>19</td>
<td>22</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1815</td>
<td>541</td>
<td>238</td>
<td>142</td>
<td>84</td>
<td>50</td>
<td>36</td>
<td>27</td>
<td>19</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single-author docs. per author</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>22624</td>
<td>677</td>
<td>315</td>
<td>158</td>
<td>74</td>
<td>37</td>
<td>19</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1815</td>
<td>541</td>
<td>238</td>
<td>142</td>
<td>84</td>
<td>50</td>
<td>36</td>
<td>27</td>
<td>19</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1576</td>
<td>507</td>
<td>237</td>
<td>141</td>
<td>83</td>
<td>51</td>
<td>36</td>
<td>27</td>
<td>19</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1318</td>
<td>453</td>
<td>207</td>
<td>128</td>
<td>69</td>
<td>39</td>
<td>21</td>
<td>11</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

32 Erik Körner
SMAuC - The Scientific Multi-Authorship Corpus
Publication and Access

- Features
  - Full-text extracts, annotated with author metadata
  - Publications from different scientific domains, stylistically diverse texts
  - Monographs and multi-authored documents

- Paper currently under review
  SMAuC - The Scientific Multi-Authorship Corpus

- Dataset will be made accessible via Zenodo, restricted to academia

- Ongoing experiments in context of multi-authorship and algorithmic bias
Researching Algorithmic Bias

Motivation

Background

- Increasing reliance on *machine learning* processes in various domains, esp.
  - Plagiarism Detection,
  - Authorship Attribution of scientific research,
  - Digital Text Forensics.

Problem

- Detection of Plagiarism or Authorship Attribution may perform worse or fail for
  (a) one *gender* compared to another, or (b) *non-native speakers* compared to
  native speakers e.g. in court decisions, job assessment, etc.

→ Unfair advantages, faulty predictions, monetary loss, etc. due to ML model bias
Researching Algorithmic Bias

Focus
- Scientific domain / academia
- Algorithmic bias

Types
- Native Speakers (English)
- Gender

Data
- SMAuC - The Scientific Multi-Authorship Corpus
Researching Algorithmic Bias

Work in Progress
- Manually annotating *gender* and *native language* for authors in SMAuC
- Prototype using *Generalized Unmasking* [Koppel and Schler 2004] [Bevendorff et al. 2019]

Future Plans
- Creating experiment framework to easily substitute different algorithms and datasets/authorship tasks
Researching Algorithmic Bias

Work in Progress

- Manually annotating gender and native language for authors in SMAuC
- Prototype using Generalized Unmasking [Koppel and Schler 2004] [Bevendorff et al. 2019]

Future Plans

- Creating experiment framework to easily substitute different algorithms and datasets/autorship tasks

Thank you for your attention!