Overview of the Cross-Domain Authorship Verification Task at PAN 2020

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast & Benno Stein

pan@webis.de   https://pan.webis.de

DOI 10.5281/zenodo.3716403
Renewed strategy

- **Long-running task** on relevant problem, but:
  - Lack of diversity in submissions recently
  - Lack of large-scale resources in field in general
  - Lack of task realism (cf. court settings)

- Renewed **3-year strategy**, increasing difficulty, scope and realism
  - Year 1 (2020): increased size
  - Year 2 (2021): increased difficulty
  - Year 3 (2022): “mystery task”
Task

- **Authorship verification** (and not attribution, obfuscation, …)

- Calibration and test set consist of series of “problems”:
  - Given a pair of texts, assign a verification score $[0, 1]$
  - $< 0.5$ (different-author: DA) or $> 0.5$ (same-author: SA)
  - Exactly 0.5: non-response (for “difficult” pairs)

- Only unseen test texts, but “closed” scenario: no new authors. All pairs are cross-fandom.
Benchmark dataset

- **Fanfiction** dataset (from fanfiction.net): non-professional authors expanding “canons” of well-known works and authors (“fandoms”)
  - English-language (but global phenomenon)
  - Huge scale (and no moderation)
  - User-provided metadata
  - **Fandom information** as a proxy for “domain”
- Emphasis: fandom information available to participants (!)
## Dataset size

(Largest resource in verification that we know of)

<table>
<thead>
<tr>
<th></th>
<th>Same-Author Pairs</th>
<th>Different-Author Pairs</th>
<th># fandoms</th>
<th>SA authors</th>
<th>DA authors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train (“large”)</strong></td>
<td>148K</td>
<td>128K</td>
<td>1.6K</td>
<td>41K</td>
<td>250K</td>
</tr>
<tr>
<td><strong>Train (“small”)</strong></td>
<td>28K</td>
<td>25K</td>
<td>1.6K</td>
<td>25K</td>
<td>48.5K</td>
</tr>
<tr>
<td><strong>Test (2020)</strong></td>
<td>10K</td>
<td>6.9K</td>
<td>0.4K</td>
<td>3.5K</td>
<td>12K</td>
</tr>
</tbody>
</table>

Interesting differences: some only used subset of “small”, others enlarged “large” even further
Evaluation framework

• Varied set of 4 metrics, sensitive to different aspects:
  • AUC: conventional area-under-the-curve score
  • c@1: variant of F1, rewarding systems that leave difficult problems unanswered
  • F1: classic metric, but *not* taking into account non-answers
  • F0.5u: new measure, emphasis on deciding same-author cases correctly

• Combined score for final ranking
Two baselines

Straightforward but competitive

Calibrated on “small” set only (give “large” systems edge):

1. **Cosine similarity** between TF-IDF BOW of 4-grams (with naive “hack” to shift scores)

2. **Text compression method**, based on cross-entropy for “text2” using Prediction by Partial Matching

[All code available from Github (https://github.com/pan-webis-de/pan-code/tree/master/clef20); all data from Zenodo (https://zenodo.org/record/3716403#.X2neLpMzZ25)]
Submissions

• 13 submissions from 10 teams

• Novelty: no calibration on Tira (only testing/deployment) for more flexibility

• 3 teams submitted “small” and “large” versions
  • Others used “small” (or subset!) apart from ordonez20

• Much more diverse array of methods, including use of e.g. siamese nets and fandom info
Results

Pair-wise differences mostly significant: indicative of diversity

<table>
<thead>
<tr>
<th>Submission</th>
<th>AUC</th>
<th>c@1</th>
<th>F0.5u</th>
<th>F1</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>boenninghoff20-large</td>
<td>0.969</td>
<td>0.928</td>
<td>0.907</td>
<td>0.936</td>
<td>0.935</td>
</tr>
<tr>
<td>weerasinghe20-large</td>
<td>0.953</td>
<td>0.880</td>
<td>0.882</td>
<td>0.891</td>
<td>0.902</td>
</tr>
<tr>
<td>boenninghoff20-small</td>
<td>0.940</td>
<td>0.889</td>
<td>0.853</td>
<td>0.906</td>
<td>0.897</td>
</tr>
<tr>
<td>weerasinghe20-small</td>
<td>0.939</td>
<td>0.833</td>
<td>0.817</td>
<td>0.860</td>
<td>0.862</td>
</tr>
<tr>
<td>halvani20-small</td>
<td>0.878</td>
<td>0.796</td>
<td>0.819</td>
<td>0.807</td>
<td>0.825</td>
</tr>
<tr>
<td>kipnis20-small</td>
<td>0.866</td>
<td>0.801</td>
<td>0.815</td>
<td>0.809</td>
<td>0.823</td>
</tr>
<tr>
<td>araujo20-small</td>
<td>0.874</td>
<td>0.770</td>
<td>0.762</td>
<td>0.811</td>
<td>0.804</td>
</tr>
<tr>
<td>niven20-small</td>
<td>0.795</td>
<td>0.786</td>
<td>0.842</td>
<td>0.778</td>
<td>0.800</td>
</tr>
<tr>
<td>gagala20-small</td>
<td>0.786</td>
<td>0.786</td>
<td>0.809</td>
<td>0.800</td>
<td>0.796</td>
</tr>
<tr>
<td>araujo20-large</td>
<td>0.859</td>
<td>0.751</td>
<td>0.745</td>
<td>0.800</td>
<td>0.789</td>
</tr>
<tr>
<td>baseline (naive)</td>
<td>0.780</td>
<td>0.723</td>
<td>0.716</td>
<td>0.767</td>
<td>0.747</td>
</tr>
<tr>
<td>baseline (compression)</td>
<td>0.778</td>
<td>0.719</td>
<td>0.703</td>
<td>0.770</td>
<td>0.742</td>
</tr>
<tr>
<td>ordonez20-large</td>
<td>0.696</td>
<td>0.640</td>
<td>0.655</td>
<td>0.748</td>
<td>0.685</td>
</tr>
<tr>
<td>ikae20-small</td>
<td>0.840</td>
<td>0.544</td>
<td>0.704</td>
<td>0.598</td>
<td>0.672</td>
</tr>
<tr>
<td>faber20-small</td>
<td>0.293</td>
<td>0.331</td>
<td>0.314</td>
<td>0.262</td>
<td>0.300</td>
</tr>
</tbody>
</table>
Analysis (1): “small” distributions

Number heaping but strong metaclassifier

[Last year, metaclassifier did not outperform strongest participant...]

![Graph showing precision and recall for different submissions and the mean ensemble.](image-url)
Analysis (2): Non-answers
boeninghoff20 surprisingly solid non-response without compromising score

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>c@1</th>
<th>F1</th>
<th>F0.5u</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>*boeninghoff20-large</td>
<td>0.974</td>
<td>0.930</td>
<td>0.936</td>
<td>0.934</td>
<td>0.943</td>
</tr>
<tr>
<td>*weerasinghe20-large</td>
<td>0.957</td>
<td>0.886</td>
<td>0.897</td>
<td>0.888</td>
<td>0.907</td>
</tr>
</tbody>
</table>

Number of non-answers vs. c@1 score.
Analysis (3): topic model

NMF on TFIDF | 150 dims | top 5K tokens
Analysis (4): Topic effect is real

All score: “small” meta-classifier on test set
Conclusions

• Higher diversity in submissions lead to interesting edition

• Reliably established that scale and size matter

• Promising new neural approaches (but brittle, cf. ordonez20)

• Closing in on solution for in-domain authorship attribution

• Topic-author orthogonality remains holy grail (next year!)

• Next year:
  • same training data
  • but much (!) more challenging test dataset
Many thanks to the AV@PAN team, but especially the participants!

Check out the task overview paper for more info and see you next year!