IMPROVED EVALUATION FRAMEWORK FOR COMPLEX PLAGIARISM DETECTION

ANTON BELY, MARINA DUBOVA, DMITRY NEKRASOV
{ANTON.BELY, MARINA.DUBOVA.97, DPOKRASKO}@GMAIL.COM

PLAGIARISM DETECTION

- Plagiarism is a major issue in science and education. Complex plagiarism is hard to detect ⇒ important to track improvement of methods.
- Plagiarism and source parts of complex PD datasets are often imbalanced as a result of paraphrasing or summarization.
- The main PD evaluation framework is Plagdet. We study its performance on PAN Summary datasets and show that it fails to distinguish good PD systems from bad ones under certain conditions.
- We propose normalized version of Plagdet which is resilient to dataset imbalance.

DATASET IMBALANCE EXAMPLE

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Plagiarism (plg)</th>
<th>Source (src)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>626 ± 43</td>
<td>5109 ± 2431</td>
</tr>
<tr>
<td>Test-1</td>
<td>639 ± 40</td>
<td>3874 ± 1427</td>
</tr>
<tr>
<td>Test-2</td>
<td>627 ± 42</td>
<td>5318 ± 3310</td>
</tr>
</tbody>
</table>

The average plagiarism case is much shorter than the source case in PAN 2013 Summary datasets.

COMPARISON OF METRICS

We constructed two adversarial models, M1 and M2, that exploit dataset imbalance in their prediction to achieve high plagdet on PAN Summary datasets, but significantly lower normalized plagdet.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Year</th>
<th>Plagdet</th>
<th>Normalplagdet</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAN 2013 Test-1</td>
<td>Sanchez-Perez et al.</td>
<td>2014</td>
<td>0.6703</td>
<td>0.7965</td>
</tr>
<tr>
<td></td>
<td>Brik et al.</td>
<td>2016</td>
<td>0.8180</td>
<td>0.8783</td>
</tr>
<tr>
<td></td>
<td>Sanchez-Perez et al.</td>
<td>2018</td>
<td>0.8841</td>
<td>0.9319</td>
</tr>
<tr>
<td></td>
<td>Adversarial M1</td>
<td>2018</td>
<td>0.8320</td>
<td>0.2614</td>
</tr>
<tr>
<td></td>
<td>Adversarial M2</td>
<td>2018</td>
<td>0.4739</td>
<td>0.1700</td>
</tr>
<tr>
<td>PAN 2013 Test-2</td>
<td>Sanchez-Perez et al.</td>
<td>2014</td>
<td>0.5638</td>
<td>0.7470</td>
</tr>
<tr>
<td></td>
<td>Brik et al.</td>
<td>2016</td>
<td>0.7072</td>
<td>0.8107</td>
</tr>
<tr>
<td></td>
<td>Sanchez-Perez et al.</td>
<td>2018</td>
<td>0.8125</td>
<td>0.8859</td>
</tr>
<tr>
<td></td>
<td>Adversarial M1</td>
<td>2018</td>
<td>0.8789</td>
<td>0.2869</td>
</tr>
<tr>
<td></td>
<td>Adversarial M2</td>
<td>2018</td>
<td>0.4848</td>
<td>0.1559</td>
</tr>
</tbody>
</table>

NORMPLAGDET: PROPOSED EVALUATION FRAMEWORK

- Plagdet framework consists of precision, recall, and their weighted harmonic mean:

\[
\text{prec}(S, R) = \sum_{r \in R} \frac{|\{s \in S | d_{plg} \cap d_{src} \} |}{|R|}, \quad \text{rec}(S, R) = \sum_{s \in S} \frac{|\{r \in R | d_{plg} \cap d_{src} \} |}{|S|}, \quad \text{gran}(S, R) = \frac{1}{|S|} \sum_{s \in S} |R_s|,
\]

\[
\text{plagdet}(S, R) = \frac{\text{F}_\alpha(\text{prec}(S, R), \text{rec}(S, R))}{\log_2 (1 + \text{gran}(S, R))}.
\]

- Let us rewrite recall using the notion of single-case recall:

\[
\text{rec}(S, R) = \sum_{s \in S} \frac{|\{r \in R_s | d_{plg} \cap d_{src} \} |}{|S|} = \sum_{s \in S} \frac{|\{r \in R_s | d_{plg} \cap d_{src} \} |}{|S|} + \frac{|\{r \in R_s | d_{plg} \cap d_{src} \} |}{|S|},
\]

where \( R_s \) is the union of all detections of a given case \( s \).

- Then we apply normalization to the inner term in Eq. 1 to obtain normalized single-case recall:

\[
\text{normrec}(S, R) = \sum_{s \in S} \frac{w_{plg}(|\{s \in S | d_{plg} \cap d_{src} \} |)}{\log_2 (1 + \text{gran}(S, R))} + \frac{w_{rec}(|\{s \in S | d_{plg} \cap d_{src} \} |)}{\log_2 (1 + \text{gran}(S, R))},
\]

where \( w_i(x) = (x - a_i) \frac{b_i}{b_i - a_i}, i \in \{plg, src\} \), and \( a_i / b_i \) is a minimum / maximum possible size of the case \( s \) intersecting all of its detections: \( s_i \cap (R_s) \).

- Finally, we see that \( \text{prec}(S, R) = \text{rec}(S, R) \) and therefore we define normalized plagdet as

\[
\text{normalplagdet}(S, R) = \frac{\text{F}_\alpha(\text{normrec}(S, R), \text{normrec}(S, R))}{\log_2 (1 + \text{gran}(S, R))}.
\]

LESSONS LEARNED

- Plagdet, standard evaluation metric for PD, does not reflect the performance correctly and can be misused on datasets for manual plagiarism detection to achieve higher scores.
- Normalization of inner terms in single-case precision and recall prevents misusage of dataset imbalance on text alignment tasks.
- When introducing new dataset, the evaluation metric should be checked to match its properties.

ACKNOWLEDGEMENTS

This work was financially supported by Government of Russian Federation (Grant 08-08).

Our implementation is freely available at https://github.com/AVBelyy/norm plagdet