Improved Evaluation Framework for Complex Plagiarism Detection (ACL '18)

Anton Belyy[1] Marina Dubova[2] Dmitry Nekrasov[1]

[1]ITMO University

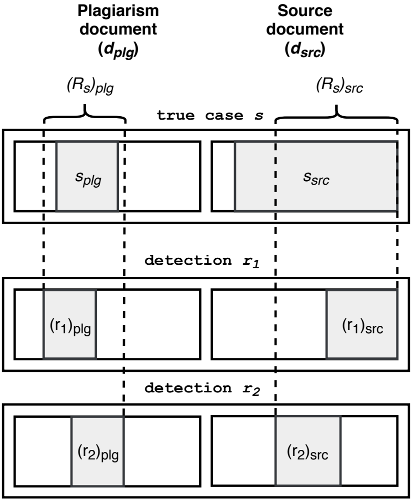
[2]Saint Petersburg State University

NLP Seminar
01.12.2018, St Petersburg

Executive Summary

- Plagiarism is a major issue in science and education. Complex plagiarism is **hard to detect** $\Rightarrow$ important to track improvement of methods.
- Plagiarism and source parts of complex PD **datasets are often imbalanced** as a result of paraphrazing 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.

Text Alignment Problem



- **Given two documents** d_{plg} and d_{src},
- **Detect all pairs of passages** $r \in R$, such that $r_{plg} \in d_{plg}$ is a "plagiariasm" of $r_{src} \in d_{src}$,
- **Calculate their intersection** with golden-set of true cases $s \in S$ as a quality measure.

Dataset Imbalance Example

Dataset	**Plagiarism (*plg*)**	**Source (*src*)**
Train	626 ± 45	5109 ± 2431
Test-1	639 ± 40	3874 ± 1427
Test-2	627 ± 42	5318 ± 3310

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

Plagdet Framework

Plagdet framework consists of precision, recall, granularity and their weighted harmonic mean[1]:

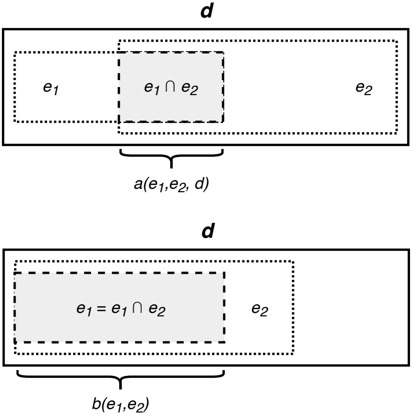
- $prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S} (s \sqcap r)|}{|r|}$,
- $rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R} (s \sqcap r)|}{|s|}$,
- $gran(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|$,
- $plagdet(S, R) = \frac{F_\alpha(prec(S,R), rec(S,R))}{\log_2 (1 + gran(S,R))}$.

However, this works poorly on imbalanced datasets. **Why?**

[1] Here we consider *macro-averaged* precision and recall; the results hold for *micro-averaged* case as well, but they are harder to explain in a limited space.

Degenerate Intersection Lemma

The size of the interesection of two sets, s and r, which are both subsets of d, is bound by: $a(s,r,d) \leqslant |s \cap r| \leqslant b(s,r,d)$.



- $a(s,r,d) = \max(0, |s| + |r| - |d|)$
- $b(s,r,d) = \min(0, |s|, |r|)$
- In *extreme* cases (when $|s| = |d|$) this interval becomes **degenerate**, i.e. $\forall r : a(s,r,d) = b(s,r,d) = |r|$
- W.r.t. Plagdet it means that an *adversary* can achieve arbitrary high score by increasing $|r|$.

Let's make Plagdet great again

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

$$\begin{aligned} rec(S, R) &= \frac{1}{|S|} \sum_{s \in S} rec_{single}(s, R_s) \\ &= \frac{1}{|S|} \sum_{s \in S} \frac{|s_{plg} \cap (R_s)_{plg}| + |s_{src} \cap (R_s)_{src}|}{|s_{plg}| + |s_{src}|}, \end{aligned}$$

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

Note that $prec(S, R) = rec(R, S)$.

Let's make Plagdet great again [2]

Then we apply normalization to the inner term in previous formula to obtain **normalized single-case recall**:

$$nrec(S,R) = \frac{1}{|S|}\sum_{s\in S} nrec_{single}(s,R_s)$$
$$= \frac{1}{|S|}\sum_{s\in S} \frac{\mathbf{w_{plg}}(|s_{plg} \cap (R_s)_{plg}|) + \mathbf{w_{src}}(|s_{src} \cap (R_s)_{src}|)}{\mathbf{w_{plg}}(|s_{plg}|) + \mathbf{w_{src}}(|s_{src}|)},$$

where, for $i \in \{plg, src\}$,

- $w_i(x) = (x - a_i)\frac{b_i - a_i}{|d_i|}$, is a *normalization function*,
- $a_i = a(s_i, (R_s)_i, d_i)$ and $b_i = b(s_i, (R_s)_i, d_i)$ are derived from Degenerate Intersection lemma.

Let's make Plagdet great again [3]

Finally, we define **normalized plagdet** as

$$normplagdet(S, R) = \frac{F_\alpha(nprec(S, R), nrec(S, R))}{\log_2(1 + gran(S, R))}.$$

Comparisons of Metrics

We constructed two adversarial models, **M1** and **M2**, that exploit dataset imbalance to achieve high **plagdet** on PAN 2013 Summary datasets, but significantly lower **normalized plagdet**.

Dataset	Model	Year	Plagdet	Normplagdet
Test-1	Sanchez-Perez et al.	2014	0.6703	0.7965
	Brlek et al.	2016	0.8180	0.8783
	Sanchez-Perez et al.	2018	**0.8841**	**0.9319**
	Adversarial M1	2018	0.8320	0.2614
	Adversarial M2	2018	0.4739	0.1700
Test-2	Sanchez-Perez et al.	2014	0.5638	0.7470
	Brlek et al.	2016	0.7072	0.8107
	Sanchez-Perez et al.	2018	0.8125	**0.8859**
	Adversarial M1	2018	**0.8789**	0.2869
	Adversarial M2	2018	0.4848	0.1559

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.

Thank you!

Improved Evaluation Framework
for Complex Plagiarism Detection

- **Anton Belyy**: anton.belyy@gmail.com
- Marina Dubova: marina.dubova.97@gmail.com
- Dmitry Nekrasov: dpokrasko@gmail.com

The implementation is available online at:
`https://github.com/AVBelyy/normplagdet`