Improved Evaluation Framework for Complex Plagiarism Detection

Anton Belyy[1] **Marina Dubova**[2] **Dmitry Nekrasov**[1]
[1]International laboratory “Computer Technologies”, ITMO University, Russia
[2]Saint Petersburg State University, Russia
{anton.belyy, marina.dubova.97, dpokrasko}@gmail.com

Abstract

Plagiarism is a major issue in science and education. It appears in various forms, starting from simple copying and ending with intelligent paraphrasing and summarization. Complex plagiarism, such as plagiarism of ideas, is hard to detect, and therefore it is especially important to track improvement of methods correctly and not to overfit to the structure of particular datasets. In this paper, we study the performance of plagdet, the main measure for Plagiarism Detection Systems evaluation, on manually paraphrased plagiarism datasets (such as PAN Summary). We reveal its fallibility under certain conditions and propose an evaluation framework with normalization of inner terms, which is resilient to the dataset imbalance. We conclude with the experimental justification of the proposed measure. The implementation of the new framework is made publicly available as a Github repository.

1 Introduction

Plagiarism is a problem of primary concern among publishers, scientists, teachers (Maurer et al., 2006). It is not only about text copying with minor revisions but also borrowing of ideas. Plagiarism appears in substantially paraphrased forms and presents conscious and unconscious appropriation of others’ thoughts (Gingerich and Sullivan, 2013). This kind of borrowing has very serious consequences and can not be detected with common Plagiarism Detection Systems (PDS). That is why detection of complex plagiarism cases comes to the fore and becomes a central challenge in the field.

2 Plagiarism Detection

Most of the contributions to the plagiarism text alignment were made during the PAN annual track for plagiarism detection held from 2009 to 2015. The latest winning approach (Sanchez-Perez et al., 2014) achieved good performance on all the plagiarism types except the Summary part. Moreover, this type of plagiarism turned out to be the hardest for all the competitors.

In a brief review, Kraus emphasized (2016) that the main weakness of modern PDS is imprecision in manually paraphrased plagiarism and, as a consequence, the weak ability to deal with real-world problems. Thus, the detection of manually paraphrased plagiarism cases is a focus of recently proposed methods for plagiarism text alignment. In the most successful contributions, scientists applied genetic algorithms (Sanchez-Perez et al., 2018; Vani and Gupta, 2017), topic modeling methods (Le et al., 2016), and word embedding models (Brlek et al., 2016) to manually paraphrased plagiarism text alignment. In all of these works, authors used PAN Summary datasets to develop and evaluate their methods.

3 Task, Dataset, and Evaluation Metrics

3.1 Text Alignment

In this work we deal with an extrinsic text alignment problem. Thus, we are given pairs of suspicious documents and source candidates and try to detect all contiguous passages of borrowed information. For a review of plagiarism detection tasks, see Alzahrani et al. 2012.

3.2 Datasets

PAN corpora of datasets for plagiarism text alignment is the main resource for PDS evaluation. This collection consists of slightly or substantially different datasets used at the PAN competitions

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 157–162
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

since 2009 to 2015. We used the most recent 2013 (Potthast et al., 2013) and 2014 (Potthast et al., 2014) English datasets to develop and evaluate our models and metrics. They consist of copy&paste, random, translation, and summary plagiarism types. We consider only the last part, as it exhibits the problems of plagdet framework to the greatest extent.

3.3 Evaluation Metrics

Standard evaluation framework for text alignment task is **plagdet** (Potthast et al., 2010), which consists of macro- and micro-averaged precision, recall, granularity and the overall plagdet score. In this work, we consider only the macro-averaged metrics, where **recall** can be defined as follows:

$$rec_{macro}(S, R) = \frac{1}{|S|} \sum_{s \in S} \underset{macro}{rec_{single}}(s, R_s), \quad (1)$$

and **precision** can be defined through recall as follows:

$$prec_{macro}(S, R) = rec_{macro}(R, S), \quad (2)$$

where S and R are true plagiarism cases and system's detections, respectively.

Single case recall $\underset{macro}{rec_{single}}(s, R_s)$ is defined as follows:

$$\frac{|s_{plg} \cap (R_s)_{plg}| + |s_{src} \cap (R_s)_{src}|}{|s_{plg}| + |s_{src}|},$$

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

4 Problem Statement

In this section, we explore problems representative to several manual plagiarism datasets (mainly, Summary part of PAN corpora), and show that the plagdet framework can fail to correctly estimate PDS quality on these datasets.

4.1 Dataset Imbalance

The PAN Summary datasets turn out to be highly imbalanced.

- Source part of each plagiarism case takes up the whole source document:

$$\forall s \in S \; \exists d_{src} \in D_{src} : s_{src} = d_{src}. \quad (3)$$

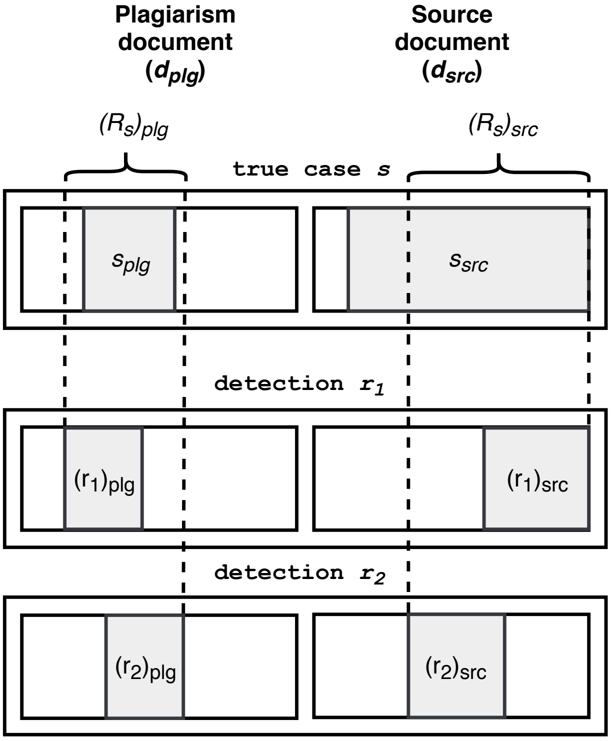

Figure 1: Single case recall computation for text alignment task. Note the imbalance in this case: plagiarism part s_{plg} is much shorter than source part s_{src}.

- For any given case, its plagiarism part is much shorter than its source part[1]:

$$\forall s \in S : |s_{plg}| << |s_{src}|. \quad (4)$$

As these datasets are publicly available, anyone can figure out these details and, therefore, construct an algorithm where statements 3 and 4 are true for detections R as well.

Let us now consider a true case s, its detections R_s and its source document d_{src}. Then single case recall for PAN Summary document will be equal to:

$$\frac{|s_{plg} \cap (R_s)_{plg}| + |d_{src}|}{|s_{plg}| + |d_{src}|} \quad (5)$$

(here we used that and $s_{src} = (R_s)_{src} = d_{src}$).

Since plagiarism part s_{plg} of the case s is much shorter than source document d_{src}, the term $|d_{src}|$ dominates numerator and denominator in eq. 5, which results in **inadequately high document-level precision and recall** on PAN Summary datasets.

[1] For exact lengths, see table 3

Other datasets for manual plagiarism detection display the similar properties, however, not to the PAN Summary extent. Examples include: Palkovskii15, Mashhadirajab et al. 2016, and Sochenkov et al. 2017.

Discussion

The important question is whether such dataset imbalance reflects the real-world plagiarizers' behavior.

There is an evidence that performing length unrestricted plagiarism task people tend to make texts shorter, however, not to the PAN Summary extent (Barrón-Cedeño et al., 2013). Moreover, we can find some supporting theoretical reasons. Firstly, summarization and paraphrasing are the only techniques students are taught to use for the transformation of texts. Hence, they can use summarization to plagiarize intellectually. Secondly, in the cases of inadvertent plagiarism and the plagiarism of ideas details of source texts are usually omitted or forgotten. This should also lead to smaller plagiarized texts. Though we can find some reasons, such huge imbalance does not seem to be supported enough and may be considered as a bias.

4.2 Degenerate Intersection

Lemma 4.1. *For any sets* $e_1 \subseteq d$ *and* $e_2 \subseteq d$*, their intersection length* $|e_1 \cap e_2|$ *is bounded by:*

$$a(e_1, e_2, d) \leqslant |e_1 \cap e_2| \leqslant b(e_1, e_2),$$

where:

$$a(e_1, e_2, d) = max(0, |e_1| + |e_2| - |d|),$$
$$b(e_1, e_2) = min(|e_1|, |e_2|).$$

Let us take a fresh look at a source part of $rec_{\underset{macro}{single}}$. We assume that $\frac{|s_{src} \cap (R_s)_{src}|}{|s_{src}|} \in [0; 1]$, and this is actually the case if:

$$0 \leqslant |s_{src} \cap (R_s)_{src}| \leqslant |s_{src}|.$$

But, according to lemma 4.1, we see that:

$$0 \leqslant a_{src} \leqslant |s_{src} \cap (R_s)_{src}| \leqslant b_{src} \leqslant |s_{src}|,$$

where:

$$a_{src} = a(s_{src}, (R_s)_{src}, d_{src}),$$
$$b_{src} = b(s_{src}, (R_s)_{src}).$$

.

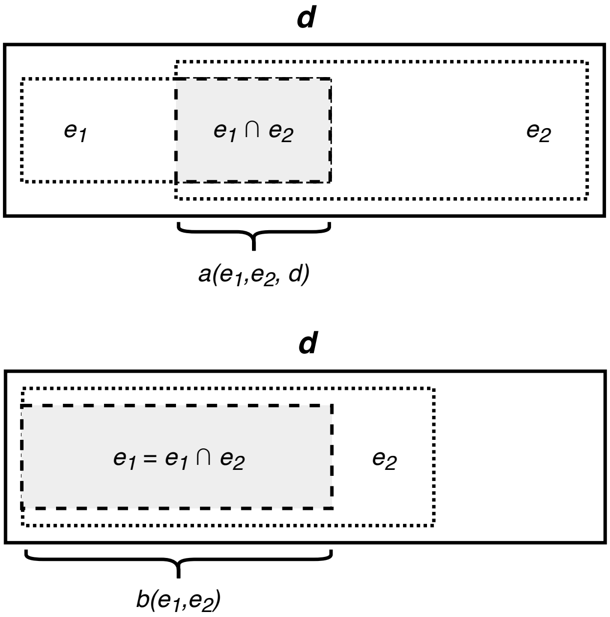

Figure 2: Degenerate intersection lemma. Intuitively, lower bound (a) is achieved when e_1 and e_2 are "farthest" away from each other in d, and upper bound (b) is achieved when $e_1 \subseteq e_2$ (or $e_2 \subseteq e_1$).

This results in a **smaller possible value range** of intersection length and, therefore, range **of precision and recall** values. Because of (3), on PAN Summary this leads to the extreme case of $a_{src} = b_{src} = |d_{src}|$, which causes precision and recall to take constant values on the source part of the dataset.

5 Proposed Metrics

5.1 Normalized Single Case Recall

To address issues of dataset imbalance (section 4.1) and degenerate intersection (section 4.2), we propose the following **normalized** version of **single case recall** $nrec_{\underset{macro}{single}}(s, R_s)$ for macro-averaged case:

$$\frac{w_{plg}(|s_{plg} \cap (R_s)_{plg}|) + w_{src}(|s_{src} \cap (R_s)_{src}|)}{w_{plg}(|s_{plg}|) + w_{src}(|s_{src}|)},$$

where:

$$w_i(x) = \frac{(x - a_i)(b_i - a_i)}{|d_i|},$$
$$a_i = a(s_i, (R_s)_i, d_i),$$
$$b_i = b(s_i, (R_s)_i),$$
$$i \in \{plg, src\}.$$

5.2 Normalized recall, precision and plagdet

The result of (1) where every $rec_{\substack{single\\macro}}(s, R_s)$ term is replaced for $nrec_{\substack{single\\macro}}(s, R_s)$ is defined as **normalized recall** $nrec_{macro}(S, R)$. **Normalized precision** $nprec_{macro}(S, R)$ can be obtained from normalized recall using eq. 2.

Normalized macro-averaged plagdet, or **normplagdet**, is defined as follows:

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

where F_α is the weighted harmonic mean of $nprec_{macro}(S, R)$ and $nrec_{macro}(S, R)$, i.e. the F_α-measure, and $gran(S, R)$ is defined as in Potthast et al. 2010.

6 Adversarial models

To justify the proposed evaluation metrics, we construct two models, **M1** and **M2**, which achieve inadequately high macro-averaged precision and recall.

6.1 Preprocessing

We represent each plagiarism document d_{plg} as a sequence of sentences, where each sentence $sent_{d_{plg},i} \in d_{plg}$ is a set of tokens. Each source document d_{src} will be represented as a set of its tokens.

For each sentence $sent_{d_{plg},i}$ we also define a measure of similarity $sim_{d_{plg},d_{src},i}$ with respect to the source document as:

$$sim_{d_{plg},d_{src},i} = \frac{|sent_{d_{plg},i} \cap d_{src}|}{|sent_{d_{plg},i}|}.$$

6.2 Models

Our models are rule-based classifiers, which proceed in three steps for each pair of documents d_{plg}, d_{src}:

1. Form a candidate set according to similarity score: $cand = \left\{i | sim_{d_{plg},d_{src},i} > \frac{3}{4}\right\}$.
2. Find the candidate with highest similarity score (if it exists): $best = \arg\max_i \left\{sim_{d_{plg},d_{src},i} | i \in cand\right\}$.
3. **(M1)** Output sentence $best$ as a detection (if it exists).
   (M2) Output sentences $\left\{i | i \neq best\right\}$ as a detection (or all sentences if $best$ does not exist).

7 Results and Discussion

We evaluated our adversarial models as well as several state-of-the-art algorithms, whose source code was available to us, using plagdet and normplagdet scores on all PAN Summary datasets available to date.

In plagdet score comparison (Table 1) we included additional state-of-the-art algorithms' results (marked by $*$), borrowed from respective papers. Proposed models M1 and M2 outperform all algorithms by macro-averaged plagdet and recall measures on almost every dataset. Despite their simplicity, they show rather good results.

On the contrary, while measuring normplagdet score (Table 2), M1 and M2 exhibit poor results, while tested state-of-the-art systems evenly achieve better recall and normplagdet scores. These experimental results back up our claim that normplagdet is more resilient to dataset imbalance and degenerate intersection attacks and show that tested state-of-the-art algorithms do not exploit these properties of PAN Summary datasets.

The code for calculating normplagdet metrics, both macro- and micro-averaged, is made available as a Github repository[2]. We preserved the command line interface of plagdet framework to allow easy adaptation for existing systems.

8 Conclusion

Our paper shows that the standard evaluation framework with plagdet measure can be misused to achieve high scores on datasets for manual plagiarism detection. We constructed two primitive models that achieve state-of-the-art results for detecting plagiarism of ideas by exploiting flaws of standard plagdet. Finally, we proposed a new framework, normplagdet, that normalizes single case scores to prevent misuse of datasets such as PAN Summary, and proved its correctness experimentally. The proposed evaluation framework seems beneficial not only for plagiarism detection but for any other text alignment task with imbalance or degenerate intersection dataset properties.

Acknowledgments

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

[2] https://github.com/AVBelyy/normplagdet

Table 1: Results of Summary Plagiarism Detection using Plagdet

Dataset	Model	Year	Precision	Recall	Plagdet
PAN 2013 Train	Sanchez-Perez et al.	2014	**0.9942**	0.4235	0.5761
	Brlek et al.	2016	0.9154	0.6033	0.7046
	Le et al. $*$	2016	0.8015	0.7722	0.7866
	Sanchez-Perez et al.	2018	0.9662	0.7407	0.8386
	Adversarial M1	2018	0.9676	0.7892	**0.8693**
	Adversarial M2	2018	0.5247	**0.8704**	0.4816
PAN 2013 Test-1	Sanchez-Perez et al.	2014	**1.0000**	0.5317	0.6703
	Brlek et al.	2016	0.9832	0.7003	0.8180
	Vani and Gupta $*$	2017	0.9998	0.7622	0.8149
	Sanchez-Perez et al.	2018	0.9742	0.8093	**0.8841**
	Adversarial M1	2018	0.9130	0.7641	0.8320
	Adversarial M2	2018	0.4678	**0.8925**	0.4739
PAN 2013 Test-2	Sanchez-Perez et al.	2014	**0.9991**	0.4158	0.5638
	Brlek et al.	2016	0.9055	0.6144	0.7072
	Le et al. $*$	2016	0.8344	0.7701	0.8010
	Vani and Gupta $*$	2017	0.9987	0.7212	0.8081
	Sanchez-Perez et al.	2018	0.9417	0.7226	0.8125
	Adversarial M1	2018	0.9594	0.8109	**0.8789**
	Adversarial M2	2018	0.5184	**0.8938**	0.4848

Table 2: Results of Summary Plagiarism Detection using NormPlagdet

Dataset	Model	Year	Precision	Recall	Plagdet
PAN 2013 Train	Sanchez-Perez et al.	2014	**0.9917**	0.6408	0.7551
	Brlek et al.	2016	0.8807	0.7889	0.8064
	Sanchez-Perez et al.	2018	0.8929	**0.9238**	**0.9081**
	Adversarial M1	2018	0.9673	0.1617	0.2770
	Adversarial M2	2018	0.1769	0.2984	0.1634
PAN 2013 Test-1	Sanchez-Perez et al.	2014	**0.9997**	0.7020	0.7965
	Brlek et al.	2016	0.9384	0.8254	0.8783
	Sanchez-Perez et al.	2018	0.9180	**0.9463**	**0.9319**
	Adversarial M1	2018	0.9130	0.1525	0.2614
	Adversarial M2	2018	0.1488	0.4237	0.1700
PAN 2013 Test-2	Sanchez-Perez et al.	2014	**0.9977**	0.6377	0.7470
	Brlek et al.	2016	0.8701	0.8104	0.8107
	Sanchez-Perez et al.	2018	0.8771	**0.9067**	**0.8859**
	Adversarial M1	2018	0.9585	0.1687	0.2869
	Adversarial M2	2018	0.1552	0.3299	0.1559

Table 3: Average Length of Plagiarism and Source Cases in Summary Datasets

Dataset	Plagiarism (plg)	Source (src)
PAN 2013 Train	626 ± 45	5109 ± 2431
PAN 2013 Test-1	639 ± 40	3874 ± 1427
PAN 2013 Test-2	627 ± 42	5318 ± 3310

References

Salha M Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2):133–149.

Alberto Barrón-Cedeño, Marta Vila, M Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947.

Arijana Brlek, Petra Franjic, and Nino Uzelac. 2016. Plagiarism detection using word2vec model. *Text Analysis and Retrieval 2016 Course Project Reports*, pages 4–7.

Amanda C Gingerich and Meaghan C Sullivan. 2013. Claiming hidden memories as one's own: a review of inadvertent plagiarism. *Journal of Cognitive Psychology*, 25(8):903–916.

Christina Kraus. 2016. Plagiarism detection-state-of-the-art systems (2016) and evaluation methods. *arXiv preprint arXiv:1603.03014*.

Huong T Le, Lam N Pham, Duy D Nguyen, Son V Nguyen, and An N Nguyen. 2016. Semantic text alignment based on topic modeling. In *Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on*, pages 67–72. IEEE.

Fatemeh Mashhadirajab, Mehrnoush Shamsfard, Razieh Adelkhah, Fatemeh Shafiee, and Chakaveh Saedi. 2016. A text alignment corpus for persian plagiarism detection. In *FIRE (Working Notes)*, pages 184–189.

Hermann A Maurer, Frank Kappe, and Bilal Zaka. 2006. Plagiarism-a survey. *J. UCS*, 12(8):1050–1084.

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th international competition on plagiarism detection. In *CEUR Workshop Proceedings*, volume 1180, pages 845–876. CEUR Workshop Proceedings.

Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 301–331. CELCT.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 997–1005. Association for Computational Linguistics.

Miguel A Sanchez-Perez, Alexander Gelbukh, Grigori Sidorov, and Helena Gómez-Adorno. 2018. Plagiarism detection with genetic-based parameter tuning. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(01):1860006.

Miguel A Sanchez-Perez, Grigori Sidorov, and Alexander F Gelbukh. 2014. A winning approach to text alignment for text reuse detection at pan 2014. In *CLEF (Working Notes)*, pages 1004–1011.

Ilya Sochenkov, Denis Zubarev, and Ivan Smirnov. 2017. The paraplag: russian dataset for paraphrased plagiarism detection. In *Proceedings of the Annual International Conference "Dialogue" 2017 (1)*, pages 284–297.

K Vani and Deepa Gupta. 2017. Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. *Expert Systems with Applications*, 73:11–26.