The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014

Notebook for PAN at CLEF 2014

Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh

Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
masp1988@hotmail.com, sidorov@cic.ipn.mx, www.gelbukh.com

Abstract. The task of (monolingual) text alignment consists in finding similar text fragments between two given documents. It has applications in plagiarism detection, detection of text reuse, author identification, authoring aid, and information retrieval, to mention only a few. We describe our approach to the text alignment subtask at the plagiarism detection competition of PAN 2014. Our method relies on a sentence similarity measure based on a tf-idf-like weighting scheme that permits us to keep stopwords without increasing the rate of false positives. We introduce a recursive algorithm to extend the matching sentences to maximal length passages. We also introduce a novel filtering method to resolve overlapping plagiarism cases. By the cumulative measure (Plagdet), our approach outperforms the best-performing system of the PAN 2013 competition and resulted in the best-performing system at the PAN 2014 competition. Our system is publicly available in open-source form.

1 Introduction

Plagiarism detection, and more generally text reuse detection, has become a hot research topic given the increasing amount of information being produced as the result of easy access to the Web, large databases and telecommunication in general, and the serious problem it has turned into for publishers, researchers and educational institutions [1]. Plagiarism detection techniques are also useful, for example, in applications such as content authoring systems, which offer fast and simple means for adding and editing content and where avoiding content duplication is desired [2]. Hence, detecting text reuse has become imperative in such contexts.

PAN is a major international competition on uncovering plagiarism, authorship, and social misuse. In 2013 and 2014, the PAN competition consisted of three tasks: plagiarism detection, author verification, and author profiling. The plagiarism detection task was divided in source retrieval and text alignment subtasks. In the text alignment subtask, the systems were required to identify all contiguous maximal-length passages of reused text between a given pair of documents.

In this paper, we present our approach to the text alignment subtask. Our approach outperforms the best-performing system of the PAN 2013 competition on the PAN 2013 evaluation corpus. The official results of the PAN 2014 competition were

Table 1. Main ideas used in the systems participating in PAN 2012 and 2013

Stage	Method	[3]	[4]	[5]	[6]	[7]	[8]	[9]	Our
Preprocessing	Special characters removal	+	–	–	–	–	–	?	+
	Numbers removal	–	–	–	–	+	–	?	–
	Stopwords removal	+	+	–	–	–	–	–	–
	Case conversion	+	+	+	+	+	–	?	+
	Stemming	+	+	–	–	+	–	?	+
Seeding	Bag of words	+	–	–	–	–	+	?	+
	Context n-grams	–	+	+	+	+	–	–	–
	Context skip n-grams	–	+	–	–	–	–	–	–
	Stopword n-grams	–	–	+	+	–	–	–	–
	Named entity n-grams	–	–	–	+	–	–	–	–
Extension	Bilateral Alternating Sorting	+	–	–	–	–	–	–	–
	Distance between seeds	+	+	+	+	–	+	?	+
	Clusters Euclidian distance	–	–	–	–	+	–	–	–
	Extension with multiple features	–	+	–	+	–	–	?	–
Filtering	Passage similarity	+	–	–	–	–	–	?	+
	Small passages removal	–	+	+	–	+	–	+	+
	Overlapping removal	–	–	+	+	–	–	?	+
	Nearby passages join	–	–	–	+	–	–	?	–

announced on two different corpora. Our approach showed the best result (Plagdet 0.87818, precision 0.88168, recall 0.87904, granularity 1.00344 on "corpus-2") out of 11 participating systems. While the announced results include also an evaluation on so-called "corpus-3," this corpus did not correspond to the official training corpus: it did not include summaries and cyclic translations, while the parameters of our system were deliberately optimized for summaries. Even in this unexpected evaluation, our system showed the third best result (Plagdet 0.89197, precision 0.86606, recall 0.91984, granularity 1.00026). Our system is publicly available in open-source form.[1]

2 Related Work

Table 1 summarizes the main ideas employed by the systems participating in PAN 2012 and 2013 [3–9], classified by the four main stages of a typical alignment process suggested in [10]. In some cases (noticeably in case of [9]) we could not find relevant information in the descriptions of the systems; in such cases we used "?" in the table. The last column refers to the system presented in this paper.

3 Methodology

Our system is organized in the four stages identified in [10]: preprocessing, seeding, extension, and filtering. At the **pre-processing** stage, applied sentence splitting and tokenizing, removed all tokens (in what follows, we refer to tokens as words) that did

[1] http://www.gelbukh.com/plagiarism-detection/PAN-2014

not start from a letter or digit, reducing all letters to lowercase, applied stemming, and joined each small sentence ($minSentLength = 3$ words or shorter) with the next one (if the joint "sentence" was still "small," we again joined it with the next one, etc.). In the following sections, we describe our processes of seeding, extension, and filtering.

3.1 Seeding

Given a suspicious document and a source document, the task of the seeding stage is to construct a large set S of small candidate plagiarism cases called seeds. Each such plagiarism case is a pair that consists of a small fragment of the suspicious document and a small fragment of the source document that are in some sense similar.

In our case, the units to form the pairs were sentences (maybe joined; see preprocessing above).

To measure the similarity between two sentences, we represented individual sentences with a tf-idf vector space model (VSM), as if each sentence were, in terminology of VSM, a separate "document" and all sentences in the pair of original document formed a "document collection." The idf measure calculated in this way is called *isf measure* (inverse sentence frequency) to emphasize that it is calculated over sentences as units and not documents:

$$tf(t,s) = f(t,s), \tag{1}$$

$$isf(t,D) = \log \frac{|D|}{|\{s \in D : t \in s\}|}, \tag{2}$$

$$w(t,s) = tf(t,s) \times isf(t,D), \tag{3}$$

where for term frequency $tf(t,s)$ we simply used the number of occurrences $f(t,s)$ of the term t in the sentence s; D is the set of all sentences in both given documents, and $w(t,s)$ is the t-th coordinate of the sentence s in our VSM representation.

A pair of sentences $susp_i$ from the suspicious document and src_j from the source document was included in S if

$$\cos(susp_i, src_j) = \frac{susp_i \cdot src_j}{|susp_i||src_j|} \geq th1 \tag{4}$$

$$Dice(susp_i, src_j) = \frac{2|\delta(susp_i) \cdot \delta(src_j)|}{|\delta(susp_i)|^2 + |\delta(src_j)|^2} \geq th2 \tag{5}$$

where the two sentences are represented as vectors, *cos* is the cosine measure, *Dice* is the Dice coefficient, $|\cdot|$ is the Euclidean length, $\delta(x) = 1$ if $x \neq 0$ and 0 otherwise, and $th1$ and $th2$ are some thresholds.

3.2 Extension

Given the seed set S of pairs (i, j) of small similar text fragments (single sentences in our case), the task of the extension stage is to form larger text fragments that are similar between two documents. For this, the fragments i are joint into maximal contigu-

ous fragments of the suspicious document and fragments *j* into maximal contiguous fragments of the source document, so that those large fragments be still similar.

In our implementation, we measured the similarity *similarity* (F_1, F_2) between two sets of sentences by adding together the vectors corresponding to all sentences of F_1, all sentences of F_2, and computing the cosine between these two vectors: $similarity\left(F_1, F_2\right) = cos\left(\sum_{x \in F_1} x, \sum_{y \in F_2} y\right)$.

We say that a sentence *s* is covered by S if it belongs to at least one pair from S, i.e., $s = i$ or $s = j$ for some $(i, j) \in S$. We say that a contiguous fragment (range of sentences) $F = \{s_l, ..., s_m\}$ of a document is covered by S if every sentence of F is covered by S, except possible gaps up to $maxGap$ sentences long. In other words, F is covered by S if its first and last sentences, s_l and s_m, are covered by S, and of each $maxGap + 1$ consecutive sentences from F, at least one sentence is covered by S.

We denote by $S \cap F$ the set of pairs from S that contain a sentence from F. Sometimes the same sentence *s* belongs to more than one pair from S, then $|S \cap \{s\}| > 1$.

Now, our extension algorithm is as follows:

Algorithm 1. Seeds integrator

1. For each fragment F in the suspicious document covered by S
2. $S' = S \cap F$
3. If $|S' \cap F| \geq minSize$
4. For each fragment F′ in the source document covered by S′
5. $S'' = S' \cap F'$
6. If $|S'' \cap F'| \geq minSize$
7. For each fragment F″ in the suspicious document covered by S″
8. If *similarity* (F″, F′) ≥ *th*3
9. add the pair (F″, F′) to the output
10. Else
11. If $maxGap > maxGapLeast$
12. recursively apply this algorithm using S″ instead of S and $maxGap - 1$ instead of $maxGap$

Here, the thresholds *minSize*, *maxGap*, *maxGapLeast*, and *th*3 are parameters of the algorithm; see Section 3.4 for a discussion of their values. Note that at the last step of the algorithm, the algorithm is recursively applied to the two fragments F″ and F′ as if they were the suspicious and the source document, their seed set being S″.

3.3 Filtering

Given the set {(F″, F′)} of plagiarism cases, the task of the filtering stage is to improve precision (at the expense of recall) by removing some "bad" plagiarism cases. We did the filtering in two stages: first, we resolved overlapping fragments; then, we removed too short fragments (in the sequel we only refer to fragments that represent plagiarism cases, not to arbitrary fragments of the documents).

Resolving overlapping cases We call two plagiarism cases (F_1'', F_1') and (F_2'', F_2') overlapping if the fragments F_1'' and F_2'' share (in the suspicious document) at least one sentence. We assume that the same source fragment can be used several times in a suspicious document, but not vice versa: each sentence can be plagiarized from only one source and thus can only belong to one plagiarism case. To simplify things, instead of re-assigning only the overlapping parts, we simply discarded whole cases that overlapped with other cases. Specifically, we used the following algorithm:

1. While exists a case P ("pivot") that overlaps with some other case
2. Denote $\mathcal{O}(P)$ be the set of cases $O \neq P$ overlapping with P
3. For each $O \in \mathcal{O}(P)$, compute the quality $q_O(P)$ and $q_P(O)$ (see below)
4. Find the maximum value among all obtained $q_y(x)$
5. Discard all cases in $\mathcal{O}(P) \cup \{P\}$ except the found x

In our implementation, at the first step we always used the first case from the beginning of the suspicious document.

We compute the quality function $q_y(x)$ of the case x with respect to an overlapping case y as follows. The overlapping cases $x = (X'', X')$ and $y = (Y'', Y')$ are pairs of corresponding fragments. Let $O = X'' \cap Y''$ be the overlap and $N = X'' \setminus O$ be the non-overlapping part. Then the quality

$$q_y(x) = sim_{X'}(O) + \big(1 - sim_{X'}(O)\big) \times sim_{X'}(N), \tag{6}$$

where *sim* is a non-symmetric similarity of a fragment F (in the suspicious document) to a reference fragment R (in the source document):

$$sim_R(F) = \frac{1}{|F|} \sum_{s \in F} \max_{r \in R}(\cos(s, r)) \tag{7}$$

The formula (6) combines the similarity of the overlapping part and of the non-overlapping part of suspicious fragment to the source counterpart.

Removing small cases We also discard the plagiarism that relate too small fragments: if either suspicions or source fragment of a case has the length in characters less than $minPlagLength$, then the case is discarded.

3.4 Adaptive behavior

At PAN competition, the methods are evaluated on four different corpora: no obfuscation, random obfuscation, translation obfuscation, and summary obfuscation, the final result being averaged over those four corpora. We observed that the optimal parameters of our method are different for such different types of plagiarism. Therefore, we introduce adaptive selection of parameters: we detect which type of plagiarism case we are likely dealing with in each specific document pair, and adjust the parameters to the optimal set for this specific type.

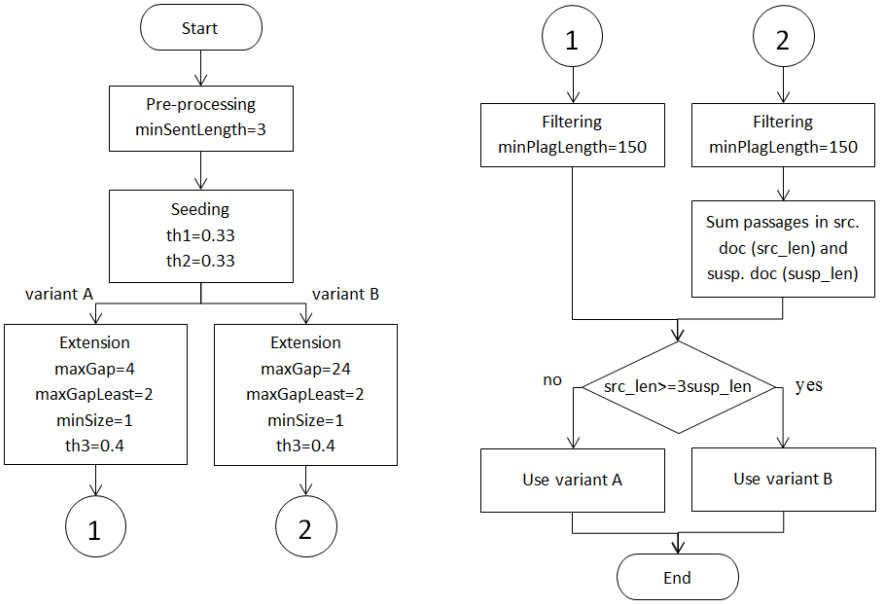

Fig. 1. Parameters and adaptive behavior

Our implementation of this approach is shown in Figure 1. After initial pre-processing and seeding, we applied the same processes twice, with different $maxGap$ values: one value that we found to be best for the summary obfuscation sub-corpus (variant B) and one that was best for the other three corpora (variant A). After we obtain the plagiarism cases using these two different settings, we decide whether those cases are likely to represent summary obfuscation or not, judging by the relative length of the suggested suspicious fragments with respect to the source fragments, and depending on this, choose to output the results of one of the two variants.

Specifically, the decision is made based on the variables *src_len* and *susp_len*, which correspond to the total length of all passages, in characters, in the source document and the suspicious document, respectively: when *susp_len* is much smaller than *src_len*, then we are likely dealing with summary obfuscation.

4 Experimental Results

We trained our system using the corpus provided for PAN 2014 competition (pan13-text-alignment-training-corpus-2013-01-21) [13]. We also evaluated our model on the test corpus of PAN 2013 (pan13-text-alignment-test-corpus2-2013-01-21) in order to compare our approach with existing approaches. Table 2 shows our results on the training corpus of PAN 2014, which was the same as training corpus of PAN 2013, and on the test corpus of PAN 2013. Table 3 compares our results (the cumulative Plagdet measure) with those of the systems submitted to PAN 2013.

We experimented with each one of our improvements separately and verified that they do boost the cumulative Plagdet measure. Both the use of the tf-isf measure and

Table 2. Our results on PAN 2013 training corpus

Obfus-cation	PAN 2013 training corpus				PAN 2013 test corpus			
	Plagdet	**Recall**	**Precision**	**Granul.**	**Plagdet**	**Recall**	**Precision**	**Granul.**
None	0.89381	0.97823	0.82280	1.00000	0.90032	0.97853	0.83369	1.00000
Random	0.88864	0.85819	0.92134	1.00000	0.88417	0.86067	0.91015	1.00086
Translation	0.88394	0.89026	0.87770	1.00000	0.88659	0.88959	0.88465	1.00081
Summary	0.57727	0.42472	0.99418	1.04348	0.56070	0.41274	0.99910	1.05882
Entire	0.87735	0.87995	0.87745	1.00213	0.87818	0.87904	0.88168	1.00344

Table 3. Comparative results according to the Plagdet measure. Performance of the systems, except our system, was tested using TIRA [11] and published in [10].

Team	**Year**	**None**	**Random**	**Translation**	**Summary**	**Entire corpus**
Sanchez-Perez	-	0.90032	0.88417	0.88659	0.56070	0.87818
Torrejón	2013	0.92586	0.74711	0.85113	0.34131	0.8222
Kong	2013	0.8274	0.82281	0.85181	0.43399	0.81896
Suchomel	2013	0.81761	0.75276	0.67544	0.61011	0.74482
Saremi	2013	0.84963	0.65668	0.70903	0.11116	0.69913
Shrestha	2013	0.89369	0.66714	0.62719	0.1186	0.69551
Palkovskii	2013	0.82431	0.49959	0.60694	0.09943	0.61523
Nourian	2013	0.90136	0.35076	0.43864	0.11535	0.57716
Baseline	2013	0.93404	0.07123	0.1063	0.04462	0.42191
Gillam	2013	0.85884	0.04191	0.01224	0.00218	0.40059
Jayapal	2013	0.3878	0.18148	0.18181	0.0594	0.27081

our recursive extension algorithm considerably improved recall without a noticeable detriment of precision. On the other hand, resolution of overlapping cases improved precision without considerably affecting recall. Finally, the dynamic adjustment of the gap size improved Plagdet on summary corpus by 35%, without considerably affecting other corpora.

5 Conclusions and Future Work

We have described our approach to the task of text alignment in the context of PAN 2014 competition. With this approach, our system showed the best result of all 11 participating systems of PAN 2014 (on "corpus-2"). Even in an unexpected evaluation on so-called "corpus-3" whose parameters differed significantly from the official training corpus, our system showed the third best result. Also on the test corpus of PAN 2013, our approach outperforms the state-of-art systems according to the results published by PAN 2013 organizers [10]. Our system is publicly available in the form of open-source software.[1]

Our main contributions are: (1) the use of tf-isf (inverse sentence frequency) measure for "soft" removal of stopwords instead of using a predefined stopword list; (2) a recursive extension algorithm, which allows for dynamically adjusting the tolerance of the algorithm to gaps in the fragments that constitute plagiarism cases; (3) a novel

algorithm for resolution of overlapping plagiarism cases, based on comparison of competing plagiarism cases; (4) dynamic adjustment of parameters according to the type of plagiarism case (summary vs. other types). Each one of these contributions separately improves the performance of the system.

In our future work, we plan to use linguistically motivated methods to address possible paraphrase obfuscation. We also plan to build a meta-classifier that would guess which type of plagiarism case we deal with at each moment and dynamically adjust the set of parameters as adequate for each specific type.

Acknowledgements. Work done under partial support of FP7-PEOPLE-2010-IRSES: Web Information Quality – Evaluation Initiative (WIQ-EI) European Commission project 269180, Government of Mexico (SNI, CONACYT), and Instituto Politécnico Nacional, Mexico (projects SIP 20144274 and 20144534, PIFI, COFAA).

References

1. Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science*, 12 (8), 1050–1084.
2. Bär, D., Zesch, T., & Gurevych, I. (2012). Text Reuse Detection Using a Composition of Text Similarity Measures. *Proceedings of COLING*, Mumbai, India, pp. 167–184.
3. Leilei, K., Haoliang, Q., Cuixia, D., Mingxing, W., & Zhongyuan, H. (2013). Approaches for Source Retrieval and Text Alignment of Plagiarism Detection. *Notebook for PAN at CLEF 2013.* In [12].
4. Rodríguez Torrejón, D. A., & Martín Ramos, J. M. (2013). Text Alignment Module in CoReMo 2.1 Plagiarism Detector. *Notebook for PAN at CLEF 2013.* In [12].
5. Suchomel, Š., Kasprzak, J., & Brandejs, M. (2013). Diverse Queries and Feature Type Selection for Plagiarism Discovery. *Notebook for PAN at CLEF 2013.* In [12].
6. Shrestha, P., & Solorio, T. (2013). Using Variety of n-Grams for the Detection of Different Kinds of Plagiarism. *Notebook for PAN at CLEF 2013.* In [12].
7. Palkovskii, Y., & Belov, A. (2013). Using Hybrid Similarity Methods for Plagiarism Detection. *Notebook for PAN at CLEF 2013.* In [12].
8. Küppers, R., & Conrad, S. (2012). A Set-Based Approach to Plagiarism Detection. *Notebook for PAN at CLEF 2012.*
9. Gillam, L. (2013). Guess again and see if they line up: Surrey's runs at plagiarism detection. *Notebook for PAN at CLEF 2013.* In [12].
10. Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., & Stein, B. (2013). Overview of the 5th International Competition on Plagiarism Detection. *CLEF 2013 Evaluation Labs and Workshop.* Valencia, Spain.
11. Gollub, T., Potthast, M., Beyer, A., Busse, A., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. (2013). Recent Trends in Digital Text Forensics and its Evaluation. In [12].
12. Forner, P., Müller, H., Paredes, R., Rosso, P., & Stein, B. (eds). *Information Access Evaluation Meets Multilinguality, Multimodality, and Visualization*. 4th International Conference of the CLEF Initiative (CLEF 2013), September 2013. Springer.
13. Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, & Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In 23rd International Conference on Computational Linguistics (COLING 10), August 2010. Association for Computational Linguistics.