PAN 2015 Shared Task on Plagiarism Detection: Evaluation of Corpora for Text Alignment *

Notebook for PAN at CLEF 2015

Marc Franco-Salvador[1], Imene Bensalem[2], Enrique Flores[1], Parth Gupta[1], and Paolo Rosso[1]

[1] Universitat Politècnica de València, Spain
mfranco@prhlt.upv.es, {eflores,pgupta,prosso}@dsic.upv.es
[2] Constantine 2 University, Algeria
bens.imene@gmail.com

Abstract. In this paper we describe and evaluate the corpora submitted to the PAN 2015 shared task on plagiarism detection for text alignment. We received mono- and cross-language corpora in the following languages (pairs): English, Persian, Chinese, and Urdu-English, English-Persian. We present an independent section for each submitted corpus including statistics, discussion of the obfuscation techniques employed, and assessment of the corpus quality.

Keywords: Plagiarism detection, Text re-use detection, Cross-language, Evaluation, Corpus construction

1 Introduction

Plagiarism detection [1, 4] refers to automatically identify the plagiarized fragments of a suspicious document in a set of source documents. When the source of plagiarism is in a different language, we refer to cross-language (CL) plagiarism detection [5, 2, 3]. Since 2012, the *Uncovering Plagiarism Authorship and Social Software Misuse*[3] (PAN) CLEF Lab, organized the shared task on plagiarism detection task which is divided in two subtasks: source retrieval and text alignment [6, 7]. Given a suspicious document and a web search API, the source retrieval subtask consists in retrieving all plagiarized sources while minimizing retrieval costs. Given a pair of documents, the text alignment subtask is based on identifying all contiguous maximal-length passages of plagiarized text between them.

The PAN 2015 subtask on text alignment[4] offered a new challenge to participants: the submission of corpora. This new initiative has obtained a considerably high acceptance with a total of six participant teams and eight submissions. They applied different

* This research has been carried out within the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA - Finding Hidden Knowledge in Texts (TIN2012-38603-C02) projects, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

[3] http://pan.webis.de/

[4] http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/plagiarism-detection.html

Table 1. Corpus statistics for 426 documents and 193 plagiarism cases in cheema15's English corpus.

Document statistics					
Document purpose			*Plagiarism per document*		
source documents	50 %		hardly	(5%-20%)	95 %
suspicious documents			medium	(20%-50%)	5 %
- with plagiarism	25 %		much	(50%-80%)	0 %
- w/o plagiarism	25 %		entirely	(>80%)	0 %
Document length					
short	(1-30k characters)	100 %			
medium	(30k-300k characters)	0 %			
long	(300k-3M characters)	0 %			

Plagiarism case statistics		
Type of case		
real plagiarism		100 %
Case length		
short	(50-1k characters)	89 %
medium	(1k-3k characters)	11 %
long	(3k-30k characters)	0 %

obfuscation techniques over text pairs, or collected real plagiarism fragments, in order to generate the plagiarism cases of the corpora. Eight are the corpora that have been submitted: six monolingual -Chinese, Persian and four English- and two CL corpora -Urdu-English and English-Persian. Evaluating whether a submitted corpus is suitable for evaluation purposes requires an in-depth analysis of its content. Therefore, in this paper, we report on our manual assessment of the submitted corpora with regard to quality and realism of the plagiarism cases.

2 Monolingual Text Alignment Corpora

In this first part we study the monolingual submitted corpora. Each subsection title corresponds with the name of the team and the language employed in the plagiarism cases. PAN 2015 shared subtask on text alignment encouraged participants to submit corpora in languages with less resources for plagiarism detection than English. For the analysis of the plagiarism cases, in order to make sure that the topic and structure of the plagiarized fragment and the suspicious document were the same, we employed Google Translate to convert the random[5] selected cases to English.

2.1 cheema15 - English

The corpus statistics are shown in Table 1. We observe that all the corpus has been composed by English paraphrasing cases. PhD, MSc and undergrad students collaborated with authors to manually generate and annotate the cases. Some forced substitutions have been found (e.g. "PC Project" replaced by "computer program"), in addition to minor issues which are not much determinative in order to detect plagiarism, e.g. source and suspicious documents starting from mid-sentence or words. However, the manual study of several random samples provided a positive impression about the plagiarism cases and its usability as corpus for evaluation.

[5] In this paper we employed four reviewers and an average of eight cases per dataset and reviewer. Random cases were independently selected for each reviewer.

Table 2. Corpus statistics for 160 documents and 75 plagiarism cases in alvi15's English corpus.

Document statistics			
Document purpose		*Plagiarism per document*	
source documents	44 %	hardly (5%-20%)	56 %
suspicious documents		medium (20%-50%)	11 %
- with plagiarism	47 %	much (50%-80%)	33 %
- w/o plagiarism	9 %	entirely (>80%)	0 %
Document length			
short	(1-30k characters)	99 %	
medium	(30k-300k characters)	1 %	
long	(300k-3M characters)	0 %	

Plagiarism case statistics	
Type of case	
verbatim copy	33.33 %
artificial obfuscation	33.33 %
real plagiarism	33.33 %
Case length	
short (50-1k characters)	99 %
medium (1k-3k characters)	1 %
long (3k-30k characters)	0 %

2.2 alvi15 - English

The authors of this English corpus employed three types of plagiarism (see Table 2): verbatim, obfuscation and real plagiarism cases. The first type is limited to simply inserting copies of fragments of a source in a suspicious document. The obfuscation cases have automatically replaced different words and nouns by synonyms and pronouns respectively. However, there is a loss of semantic relatedness in some cases, e.g. "already big enough to speak" replaced by "already great adequate to say". Authors used character substitution as well for this type of plagiarism. The real plagiarism cases -extracted from fairy tales- contain a high manual modification level while maintaining the sense. In contrast, some errors have been found in the codification of the XML files of the corpus: wrong case offsets -with starting point at mid-word-, in addition to the attribute "type" established to "real" in all the cases, instead to only the real plagiarism cases and "artificial" for the rest. Despite this errors, the overall opinion about this corpus is positive, especially the real plagiarism cases. The quality of the corpus could be increased in future versions.

2.3 palkovskii15 - English

As it is shown in Table 3, this corpus is composed by English verbatim and automatic obfuscation plagiarism cases of three types: random, translation and summary. The random obfuscation is quantified by degrees to measure the level of automatic obfuscation, by random employed word reordering. Translation obfuscation cases used a chain of translators among ten intermediate languages employing MyMemory[6], Google[7] and Bing[8] translators. Summary obfuscation cases are created by means of an automatic summarization tool. The manual analysis of several cases provided average-negative impressions about the quality of the corpus for its practical usage. It seems that the high level of random obfuscation, the chain of translators and the unspecified summarization

[6] https://mymemory.translated.net/
[7] https://translate.google.com/
[8] https://www.bing.com/translator/

Table 3. Corpus statistics for 3,125 documents and 1,976 plagiarism cases in palkovskii15's English corpus.

Document statistics			
Document purpose		*Plagiarism per document*	
source documents	62 %	hardly (5%-20%)	82 %
suspicious documents		medium (20%-50%)	17 %
- with plagiarism	18 %	much (50%-80%)	0 %
- w/o plagiarism	20 %	entirely (>80%)	0 %
Document length			
short	(1-30k characters)	97 %	
medium	(30k-300k characters)	3 %	
long	(300k-3M characters)	0 %	

Plagiarism case statistics	
Type of case	
artificial obfuscation (summary|random)	69 %
translation-chain	31 %
Case length	
short (50-1k characters)	96 %
medium (1k-3k characters)	4 %
long (3k-30k characters)	0 %

Table 4. Corpus statistics for 2,744 documents and 2,747 plagiarism cases in mohtaj15's English corpus.

Document statistics			
Document purpose		*Plagiarism per document*	
source documents	71.8 %	hardly (5%-20%)	86 %
suspicious documents		medium (20%-50%)	14 %
- with plagiarism	17.6 %	much (50%-80%)	0 %
- w/o plagiarism	10.6 %	entirely (>80%)	0 %
Document length			
short	(1-30k characters)	81 %	
medium	(30k-300k characters)	19 %	
long	(300k-3M characters)	0 %	

Plagiarism case statistics	
Type of case	
verbatim copy	8 %
artificial obfuscation	77 %
manual obfuscation	15 %
Case length	
short (50-1k characters)	99 %
medium (1k-3k characters)	1 %
long (3k-30k characters)	0 %

tool, provided a high number of senseless text fragments and non-related cases. Finally, we found similarities with this corpus and the PAN 2013 text alignment corpus[9], e.g. *suspicious-document00005* and *source-document01090* are present in both corpora.

2.4 mohtaj15 - English

This English corpus (see Table 4) contains plagiarism cases of three types: verbatim, random and manual obfuscation. Random obfuscation is performed at two levels (low and high), with higher word reordering and synonym substitution for the second. We observed that there exist, especially with the high level, senseless and semantically unrelated cases of this type. The manual obfuscation cases suffered manual paraphrasing and are in general suitable for plagiarism detection evaluation. Random obfuscation should be improved in order to have a representative corpus for evaluation.

Table 5. Corpus statistics for 82 documents and 109 plagiarism cases in kong15's Chinese corpus.

Document statistics					Plagiarism case statistics	
Document purpose		*Plagiarism per document*			*Type of case*	
source documents	95 %	hardly	(5%-20%)	0 %	real plagiarism	100 %
suspicious documents		medium	(20%-50%)	100 %		
- with plagiarism	5 %	much	(50%-80%)	0 %	*Case length*	
- w/o plagiarism	0 %	entirely	(>80%)	0 %	short (50-1k characters)	92 %
Document length					medium (1k-3k characters)	6 %
short	(1-30k characters)	35 %			long (3k-30k characters)	2 %
medium	(30k-300k characters)	65 %				
long	(300k-3M characters)	0 %				

Table 6. Corpus statistics for 1,522 documents and 411 plagiarism cases in khoshnava15's Persian corpus.

Document statistics					Plagiarism case statistics	
Document purpose		*Plagiarism per document*			*Type of case*	
source documents	53 %	hardly	(5%-20%)	47 %	verbatim copy	31 %
suspicious documents		medium	(20%-50%)	53 %	artificial obfuscation	69 %
- with plagiarism	21 %	much	(50%-80%)	0 %	*Case length*	
- w/o plagiarism	26 %	entirely	(>80%)	0 %	short (50-1k characters)	42 %
Document length					medium (1k-3k characters)	58 %
short	(1-30k characters)	99 %			long (3k-30k characters)	0 %
medium	(30k-300k characters)	1 %				
long	(300k-3M characters)	0 %				

2.5 kong15 - Chinese

The corpus of Table 5 is formed by real plagiarism cases in Chinese. Unfortunately, XML files do not contain information about the type of strategy employed. Therefore, it is impossible to determine how the real cases were created. In addition, the manual analysis of several cases proved that there is not topic and structural relatedness between annotated cases. It is possible that some error with offsets tagging have been produced. Note also the low number of suspicious documents, which may produce non-significant results when using this corpus during evaluation.

2.6 khoshnava15 - Persian

The corpus of the Table 6 is formed by Persian verbatim and random obfuscation cases. Despite the low information about how the corpus was created, we note the high quality

[9] `http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-web/plagiarism-detection.html`

Table 7. Corpus statistics for 21,429 documents and 5,606 plagiarism cases in asghari15's English-Persian corpus.

Document statistics				**Plagiarism case statistics**	
Document purpose		*Plagiarism per document*		*Type of case*	
source documents	74 %	hardly (5%-20%)	88 %	translated (English to Persian)	100 %
suspicious documents		medium (20%-50%)	12 %		
- with plagiarism	13 %	much (50%-80%)	0 %	*Case length*	
- w/o plagiarism	13 %	entirely (>80%)	0 %	short (50-1k characters)	100 %
Document length				medium (1k-3k characters)	0 %
short (1-30k characters)	85 %			long (3k-30k characters)	0 %
medium (30k-300k characters)	15 %				
long (300k-3M characters)	0 %				

of the cases. Random selected and revised samples of both types of cases are well annotated, semantic and structurally related. Therefore, also by its large size, we consider this corpus has a good quality to be used for Persian plagiarism detection.

3 Cross-language Text Alignment Corpora

In this section we study the cross-language submitted corpora. Each subsection title corresponds with the name of the team and the source-suspicious document language-pairs employed. As for the monolingual plagiarism cases not in English, also in the following CL- text alignment corpora we used Google Translate in order to validate the topic and structural relatedness.

3.1 asghari15 - English-Persian

This is a considerably large corpus for CL English-Persian plagiarism detection (see the Table 7 caption). It is formed by documents with encyclopedic knowledge. Authors generated all the plagiarism cases using obfuscation -we assume that by means of translation-, and divide the level of obfuscation on three types: low, medium and high. No further details have been provided about how this obfuscation and translation have been performed. However, the manual analysis of several random samples showed that the topic and structural relatedness have been maintained in the CL plagiarism cases and their quality is high enough to consider this corpus for benchmarking English-Persian plagiarism detection.

3.2 hanif15 - Urdu-English

In Table 8 we can see the statistics of this Urdu-English plagiarism detection corpus. The corpus has been created using three types of obfuscation by means of manual Urdu-English translation. Unfortunately, the tags employed in the XML annotation files do

Table 8. Corpus statistics for 500 documents and 135 plagiarism cases in hanif15's Urdu-English corpus.

Document statistics			
Document purpose		*Plagiarism per document*	
source documents	50 %	hardly (5%-20%)	90 %
suspicious documents		medium (20%-50%)	9 %
- with plagiarism	27 %	much (50%-80%)	1 %
- w/o plagiarism	23 %	entirely (>80%)	0 %
Document length			
short	(1-30k characters)	99 %	
medium	(30k-300k characters)	1 %	
long	(300k-3M characters)	0 %	

Plagiarism case statistics		
Type of case		
translated (Urdu to English)		100 %
Case length		
short	(50-1k characters)	100 %
medium	(1k-3k characters)	0 %
long	(3k-30k characters)	0 %

not allow to understand which is the real difference between these types. Manual analysis of several random cases offered an average impression about the corpus. There are semantically unrelated cases but the number of correct instances is higher. However, we found also some minor typos in the English writing, in addition to some cases which start at mid-word or in the last word of a sentence. A future revision of the corpus fixing these errors could provide an interesting corpus for benchmarking Urdu-English plagiarism detection.

4 Conclusions

In this paper we evaluated the quality of the corpora submitted at the PAN 2015 shared task on text alignment. Among the eight evaluated corpora, seven used some obfuscation strategy to generate their plagiarism cases, five used also verbatim cases, and three contained real plagiarism cases too. The preferred obfuscation method has been the random obfuscation, followed by the synonym substitution. Most of the used documents and plagiarism cases has been short. Documents and cases with average lengths have been present in a small amount and corpora authors discarded the use of long ones. In general, suspicious documents were hardly formed by plagiarism cases, followed by documents with an average amount of them. Only two corpora contained a percentage of documents with much plagiarism. Despite English has been the most used language (in six corpora), the contributions in other languages have been highly appreciated and some cases denote a remarkable effort to create high quality corpus to evaluate these languages. It is encouraging to see the high acceptance of this new initiative of allowing the participants to submit new corpora for text alignment. Future editions will require a short summary of the strategies and methodology employed to create the plagiarism cases in order to ease the evaluation of the corpora. We will work also to include statistics about the approximate number of errors per reviewed corpus.

References

1. Clough, P., et al.: Old and new challenges in automatic plagiarism detection. In: National Plagiarism Advisory Service, 2003; http://ir. shef. ac. uk/cloughie/index. html. Citeseer (2003)
2. Franco-Salvador, M., Gupta, P., Rosso, P.: Cross-language plagiarism detection using a multilingual semantic network. In: Proc. of the 35th European Conference on Information Retrieval (ECIR'13). pp. 710–713. LNCS(7814), Springer-Verlag (2013)
3. Franco-Salvador, M., Gupta, P., Rosso, P.: Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In: Ferro, N. (ed.) Bridging Between Information Retrieval and Databases, Lecture Notes in Computer Science, vol. 8173, pp. 227–236. Springer Berlin Heidelberg (2014), http://dx.doi.org/10.1007/978-3-642-54798-0_12
4. Maurer, H.A., Kappe, F., Zaka, B.: Plagiarism-a survey. J. UCS 12(8), 1050–1084 (2006)
5. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Language Resources and Evaluation 45(1), 45–62 (2011)
6. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th international competition on plagiarism detection. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 845–876 (2014)
7. Potthast, M., Hagen, M., Göring, S., Rosso, P., Stein, B.: Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes