Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

Notebook for PAN at CLEF 2015

Vahid Zarrabi, Javad Rafiei, Khadijeh Khoshnava, Habibollah Asghari, Salar Mohtaj

ICT Research Institute,
Academic Center for Education, Culture and Reseach (ACECR), Iran

{vahid.zarrabi, javad.rafiei, khadijeh.khoshnava,
habib.asghari, salar.mohtaj}@ ictrc.ir

Abstract. This paper addresses the text alignment task of 7th International competition on plagiarism detection; PAN 2015. We investigate five submitted corpora and evaluate them based on their characteristics in two ways: manual and automatic evaluation. The results of evaluation show that the most of plagiarism cases in prepared corporahavea rather high quality in term of "rate of obfuscation" alongside "preserving the concepts".

Keywords: Plagiarism Detection, Text Alignment, Corpora Evaluation, Corpus Construction

1 Introduction

Plagiarism detection in PAN is divided into two separated subtasks: source retrieval and text alignment [1]. Recently, the latter subtask is changed to corpus construction where participants are wanted to provide instances of plagiarism cases with their source documents.

These instances can occur in two forms: real-world samples and generated (artificial or simulated) samples. In this paper, as a participant of corpus construction subtask in PAN 2015, we evaluate other submitted corpora from the point of view of quality and realism of plagiarism cases in a manual way, and also analyze statistical information of corpora. The corpora are in different languages, and even there maybe cross-lingual corpora in which source documents are in different language from suspicious ones.

This paper is organized as follows: Section 2 describes public metadata of corpora. Section 3 analyzes the statistical information extracted from corpus metadata. In section 4, in order to determine the quality of corpora, we manually investigate plagiarism cases and their related original documents for each corpus, based on three factors. In section 5, we describe automatic methods for evaluating the real-world and

summary obfuscation of Kong15 and Palkovskii15 corpora, respectively. Finally in section 6we have discussions and conclusion.

2 Global Metadata

In this section, we describe the global metadata of corpora under evaluation. Table 1 shows the metadata of five corpora. As can be showed in the Table, there is one bi-lingual and four mono-lingual corpora in English and Chinese. The last row shows data resources of documents. It can be seen that in most cases, the resources for source and suspicious documents are the same.

Table 1. - Global information of prepared corpora

	cheema15	**hanif15**	**Kong15**	**Alvi15**	**Palkov-skii15**
Type of Corpus	Mono-Lingual	Bi-Lingual	Mono-Lingual	Mono-Lingual	Mono-Lingual
Source–Suspicious Language	English-English	Urdu-English	Chinese-Chinese	English- English	English-English
Resource Documents	Gutenberg books and Wikipedia	Wikipedia pages	Chinese thesis and http://wenku.baidu.com/ website	"The Complete Grimm's Fairy Tales" book	Internet web pages crawling

3 Corpora Statistical Information

For evaluation of corpora based on statistical information, we categorized the statistical information in three different aspects: The first view describes the numerical information about corpora such as number and length of documents and suspicious cases which has been shown in Table 2. In the second view, the distributions of obfuscation strategies are demonstrated as shown in Table 3. In the third view, we have calculated some ratios for demonstrating a better statistical picture of corpora as shown in Table 4.

Table 2 shows the statistical information of the submitted five corpora in text alignment subtask.We categorized statistical information of corpora in three rows: The first row demonstrates the number of suspicious and source documents. In second row, the length of documents has been determined by Min, Average and Max categories. In the third row, we have shown the information extracted from XML files that provide either one-to-one or one-to-many links between source and suspicious fragments.This information shows the length of plagiarism fragments.

Most of corpora have approximately equal number of source and suspicious documents. Although the corpus of Kong15 has just four suspicious documents, but it

should be noted that it contains real plagiarism casesin suspicious fragments in the corpus.

Simulation of actual cases of plagiarism cases requires that the suspicious documents have enough length to embed some plagiarized fragments within their text. As shown in the table, the documents in Kong15's corpus have greatest average length, while the documents in Cheema15's corpus have greatest minimum length. So, in both of them, we can potentially insert more and larger plagiarized fragments in order to construct suspicious documents. Three of corpora have approximately same average length of plagiarism cases; Due to the short length of plagiarism cases in Hanif15's corpus, even with a medium rate of obfuscation, the plagiarism detection will become more difficult. On the other hand, Palkovskii15 corpus has long plagiarism cases and needs to perform more changes in order to build a challenging corpus. These will be discussed later in this paper.

Table 2. – The statistical information of the five corpora

	Cheema15	Hanif15	Kong15	Alvi15	Palkovskii15
Number of Docs					
• Suspicious Docs	248	250	4	90	1175
• Source Docs	248	250	78	70	1950
Length of Docs (in chars)					
• Min Length	2263	361	394	514	519
• Max Length	22471	74083	121829	45222	517925
• Average Length	7239	4382	42839	7718	6512
Length of Plagiarisms Cases (in chars)					
• Min Length	134	78	62	259	157
• Max Length	2439	849	2748	1160	14336
• Average Length	503	361	423	464	782

Extra information also can be extracted from mentioned XML files such as obfuscation strategy. Table 3 demonstrates obfuscation strategies with the number of plagiarism fragments related to these types in the corpora. Some participants have employed one type of obfuscation such as Cheema15 and Hanif15 which applied simulated obfuscation in their corpora. Kong15 corpus includes just real obfuscation strategy of plagiarism without any added fragments to suspicious documents, where each of suspicious documents have passages either are the plagiarism cases or have the potential to be plagiarism.

On the other hand, two participants have multiple obfuscation strategies in their corpora: Alvi15 corpus has employed three types of obfuscation: "retelling-human" is similar to simulated obfuscation; "character-substitution" and "automatic" is similar to artificial obfuscation. "Character-Substitution" obfuscation exchanges vowel sounds with same character glyphs but with different Unicode. Also Palkovskii15

corpus covers four kinds of obfuscation: “None”obfuscationwhich is an exact copy of fragments, “cyclic Translation”, “summary obfuscation” and “random obfuscation”.

Table 3. - Obfuscation strategies employed by participants, PAN 2015

Obfuscation Strategies	Cheema15	Hanif15	Kong15	Alvi15	Palkovskii15
Simulated	123	135	-	-	-
Real	-	-	109	-	-
Automatic	-	-	-	25	-
Retelling-Human	-	-	-	25	-
Character-Substitution	-	-	-	25	-
Translation	-	-	-	-	618
Summary	-	-	-	-	1292
Random	-	-	-	-	626
None	-	-	-	-	624
Sum	**123**	**135**	**109**	**75**	**3160**

Considering the number of suspicious documents from Table 2 and suspicious fragments from Table 3, we can calculate the average number of plagiarism cases per suspicious document as shown in the Table 4. Moreover, the third row in Table 4 demonstrates the following formula

F=AVG [No. of plag. cases in each susp. doc] / AVG [length of susp. docs] * AVG [length of plag. cases] (1)

Among the participants, Kong15 and Palkovskii15 corpora have higher F-measure values in comparison to the others, with 32% and 27% respectively. When the number of plagiarism cases in each suspicious document increase, plagiarism detection would be more difficult. Thus, it seems that plagiarism detection in Kong15 corpus is a challenging matter. We should also mention that it needs detail investigation of corpora for better analysis.

Table 4. – Relative statistical information of corpora

Number/ Percent	Cheema15	Hanif15	Kong15	Alvi15	Palkovskii15
Plagiarism Cases	123	135	109	75	3160
Plagiarism Cases per Suspicious Document	0.49	0.54	27.25	0.83	2.68
Share of plagiarism cases in Suspicious documents	3.4%	4.4%	26.9%	4.9%	32.18%

4 Manual Evaluation of Corpora

In this section we manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus based on the following three measures:

Changes in syntactic structure between source and plagiarized passage (categorized as low, medium and high)

1. Concept preserving from source passage to plagiarized passage
2. Distribution of obfuscation types in suspicious documents
3. These measures are useful for evaluating how much plagiarism cases are near to real ones.

The first measure can be depicted in figure 1 that shows the rate of structural changes based on three categories low, medium and high. The quantified values of these labels are shown in table 5.This table shows the ratio of syntactically alternated sentences to the total sentences in plagiarized passages.

Table 5. – Quantifying of predifiend labels

Label	**The ratio of syntactical alterations**
Low	<10%
Medium	>10% and <50%
High	> 50%

High degree shows high obfuscation rate, so in this case, the plagiarism detection would be more difficult. Hanif15 corpus has more short length plagiarism cases. Moreover, in this corpus, most plagiarism cases have “high structure changes”, which labeled as “high” as depicted in Figure 1. So, plagiarism detection can be more difficult in Hanif15 corpus.

In Palkovskii15 corpus, the plagiarism cases have highest “low structure changes” and also most of plagiarism cases have long length.As a result, detector tools can find most of plagiarism cases from the corpus.

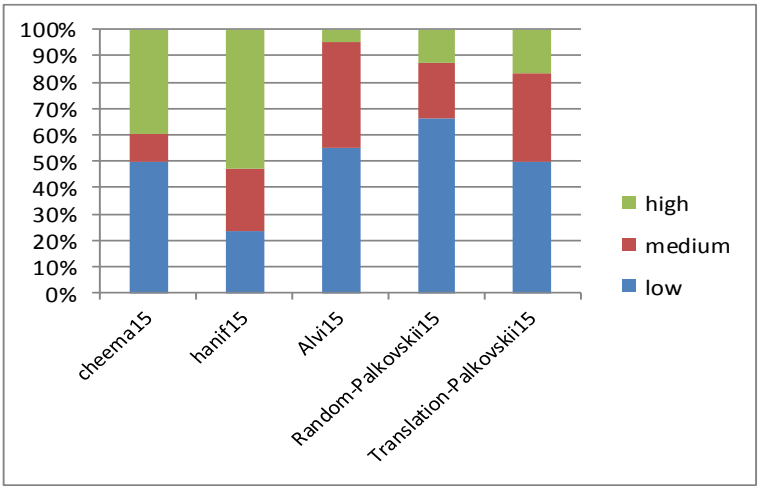

Fig. 1. - Changes in syntactic structure

The second measure determines how many plagiarism cases preserve the concept of the original one. For each plagiarism case, "preserving the concept" is considered as the ratio of maintained keywords to the total number of keywords in the plagiarized passage. Table 6 shows the quantified values of low, medium and high labels.

Table 6. – Quantifying of predifiend labels

Labels	**Ratio of maintained keywords to the total ones**
Low	< 20%
Medium	>20% and <65%
High	>65%

It is better that a corpus can preserve the concept of the original content. As figure 2 shows,the number of "high" label forall corpora is more than 50%. So plagiarism cases preserve the concept of the contentin all corpora quite well.

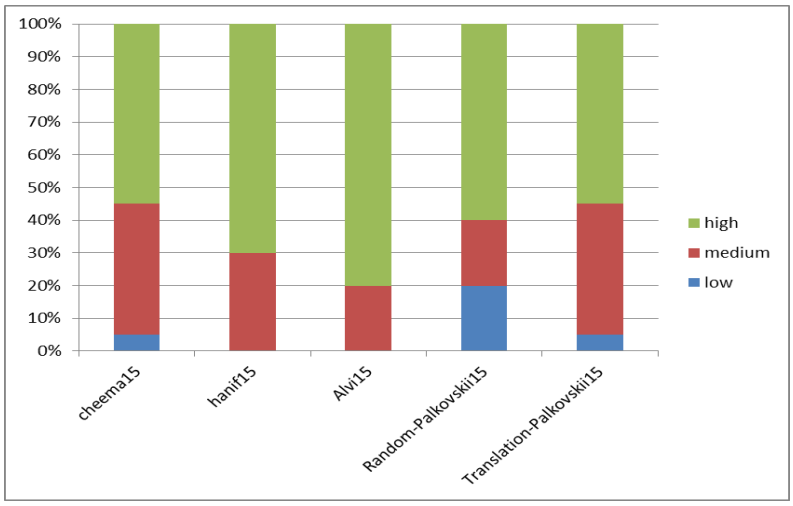

Fig. 2. - Rate of concept maintaining

It can firmly be said that the third measure has a main role to determine the quality of the plagiarism cases and thus quality ofa whole typical corpus.We have considered four types of obfuscation: 'Add', 'delete', 'replacement' and 'reorder'. These types are extracted from [2] and computed manually for each plagiarizedfragment. This measure discusses about how these four types of obfuscation contribute to build plagiarism cases.

The measure 3 expresses the ratio of alternated words (based on 4 types of obfuscations) to total number of the source fragment's words, based on three labels: low, medium and high. Table 7 shows the quantified values of low, medium and high labels.

Table 7. – Quantifying of predifiend labels

Labels	**The ratio of alternated words to the total number**
Low	<20%
Medium	>20% and <40%
High	>40%

As can be seen in figure 3, corpora have different percent of labels. Cheema15 corpus has the largest number of 'high' label, which has a great difference compared to other corpora. As a result, plagiarism cases mostly have a great degree of obfuscation and thus plagiarism can be hard to find. Other corpora mainly have more 'medium' label than 'low' and 'high' label. However, the number of 'low' label is few and it can be concluded that most corpora have enough degree of obfuscations in their plagiarism cases and this can cause challenges in plagiarism detection process.

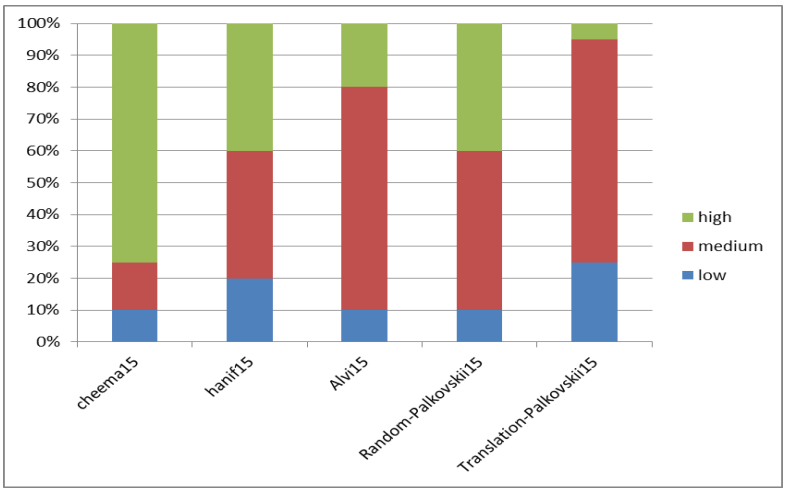

Fig. 3. - Rate of obfuscation

In addition to rate of obfuscation, we discusses about what type of obfuscation is used for corpora construction. Table 8 shows statistical information about the "obfuscation types" in plagiarism cases. 'Delete' and 'Replacement' have the greatest impact on obfuscation degree. As shown in Table 8, Hanif15 and Cheema15 corpora are most consumers of these two operations. As a result, plagiarism detection can be a challenging problem in both corpora.

According to our study, 'character-substitution' obfuscation is used in Alvi15 and Palkovskii15 corpora can be simply solved by exchange vowel sounds with their original characters, so we didn't consider them in our evaluation process.

Table 8. – The perecent of the obfuscation type ineach corpus

Types of Obfuscation	Cheema15	Hanif15	Alvi15	Random-Palkovskii15	Translation-Palkovskii15
Add	10%	0%	5%	0%	0%
Delete	20%	35.40%	0%	36.40%	20%
Replacement	70%	64.60%	95%	27.25%	50%
Reorder	0%	0%	0%	36.35%	30%

5 Automatic Evaluation of Corpora

In this section, we separately evaluate two remained obfuscation scenarios: real obfuscation from Kong15 corpus and summary obfuscation from Palkovskii15 corpus.

5.1 Automatic Evaluation of Kong15 corpus

For Kong15 corpus, all source and correspond suspicious fragments are extracted, and the total number of similar "characters n-grams" between source and suspicious plagiarized passagesare calculated for n in range of one to four [3]. In the next step, the normalized total numbers (in percent) are clustered using k-means clustering algorithm [4]. The similarity numbers are classified into three clusters. Table 9 shows the clusters of similarity numbers as ordered pairs (cluster centroid, number of cluster nodes in percent) for n=1, 2, 3, 4. In the last column, the total average of similarity numbers of all n-grams is calculated.

Table 9. – The clusters of the n-grams similarities for real obfuscation

	Pair of (centroid, 1-gram)	Pair of (centroid , 2-gram)	Pair of (centroid , 3-gram)	Pair of (centroid , 4-gram)	Pair of (centroid, Average of n-grams)
Non-Relevant	(0.29, 0.27%)	(0.07, 79.6%)	(0.02, 62.4%)	(0.02, 71.57%)	(0.12, 60.56%)
Medium to High	(0.48, 67.61%)	(0.37, 21.1%)	(0.27, 19.26%)	(0.33, 14.67%)	(0.39, 21.1%)
Low to Medium	(0.87, 32.11%)	(0.80, 20.18%)	(0.76, 18.34%)	(0.81, 13.76%)	(0.81, 18.34%)

According to the last column, the centroid value of first cluster is small, that means the source fragment has different topic against the suspicious fragment, or maybe a little sub-fragment are shared between them, for example:

Suspicious:
避免的要把表现与业务逻辑代码混合在一起，都给前期开发
与后期维护带来巨大的复杂度。为了摆脱上述的约束与局限，把业务逻辑代码从表现层中清晰的分离出来，2000**年**，Craig McClanahan

Source:
如表41**所示。本章完成了系**统数据库的数据需求分析的过程，说明
了数据库由概念结构设计转换成逻辑结构设计的过程，并把各个物**理数据模型**结合起来形成了一个整体的关系数据库模型，为系统详细设计作好了充足的准备工作

Here, the topic of suspicious fragment is about "business logic" and the topic of source fragment is about "database system". The centroid value of second cluster shows that the source fragment and the suspicious fragment are similar; either in terms of "medium to high obfuscation" or a large sub-fragment are shared, for example:

Suspicious:
如：状态管理服务、持续性服务、分布式共享数据对象的

缓冲服务等，它对开发人员来说是很重要的，这样开发人员可以集中精力在如何创建业务逻辑上，相应地缩短了开发时间。**并**发用户的访问而急剧下降，另外系统也同时具备了很好的可扩展性和伸缩性，即在请求并发量增大或减少时，可根据实际情况增加或减少应用服务器数量，以便保证性能的前提下，合理利用硬件资源。**任**务由应用服务器…

Source:
当请求并发量巨大时，数据库性能下降很快。针对这一不足，
基于J2EE**架构的**处理方式是：业务逻辑分布到应用服务器上，数据库上不**再具有**业务逻辑处理单元，而只负责基础业务数据的管理，主要的计算任务由应用服务**器完成**，**从而充分利用了**应用服务器在并发处理和逻辑计算方面的优势。另外，应用服**供水**调度应急、预警信息平台的设计与实现

Here, suspicious fragment has medium to high obfuscation in comparison with the source fragment. The last cluster has high centroid value, which means the source fragment has same topic in comparison with the suspicious fragment with "low to medium obfuscation" or maybe exact copy, for example:

Suspicious:
层开发任务**交**给中间件供应商去完成，而这些复杂的系统级功能
是常规应用开发中难度最大、开发成本最高的一部分工作。高级中间件供应商提供复杂的中间件服务，如：状态管理服务、持续性服务、分布式共享数据对象的缓冲服务等，它对开发人员来说是很重要的，这样开发人员可以集中精力在如何创建业务逻辑上，相应地缩短了开发时间。**并**发用户的访问而急剧下降，另外系统也同时具备了很好的可扩展性和伸缩性，即在请求并发量增大或减少时，可根据实际情况增加或减少应用服务器数量，以便保证性能的前提

Source:
提供复杂的中间件服务，如：状态管理服务、持续性服务、**分布式共**
享数据对象的缓冲服务等，它对开发人员来说是很重要的，这样开发人员可以集中精力在如何创建业务逻辑上，相应地缩短了开发时间。**性和伸**缩**性**，**即在**请求并发量增大或减少时，可根据实际情况增加或减少应用服务器数量，**以便保**证性能的前提下，合理利用硬件

Here, source and suspicious fragments have same topics, while the number of "low to medium obfuscation" is higher in suspicious document with respect to source document. Now by using statistical information in the last column and the above examples, clusters are labeled with "Non-Relevant", "Medium to High" and "Low to Medium" labels that are shown on first column of Table 9.

5.2 Automatic Evaluation of Palkovskii15 Corpus: Summary Obfuscation

For evaluation of summary obfuscation from the point of "concept preserving" measure, we have extracted 10% of top words from source fragments based on tf.idf weight. We used PAN2011 corpus for idf calculation. Figure 4 shows the percent of "concept preserving" of top words for suspicious fragments. We evaluate 108 fragment pairs in the diagram.

Using k-means clustering algorithm [4], the suspicious fragments are classified into three clusters with low, medium and high labels. Now we can calculate the number of fragments in each cluster as low, medium and high percent. Following is a list of this statistical information as ordered pairs (cluster centroid, number of cluster nodes as a percent):

- Low percent: (0.25, 28.8%)
- Medium percent: (0.42, 40.7%)
- High percent: (0.56, 30.5%)

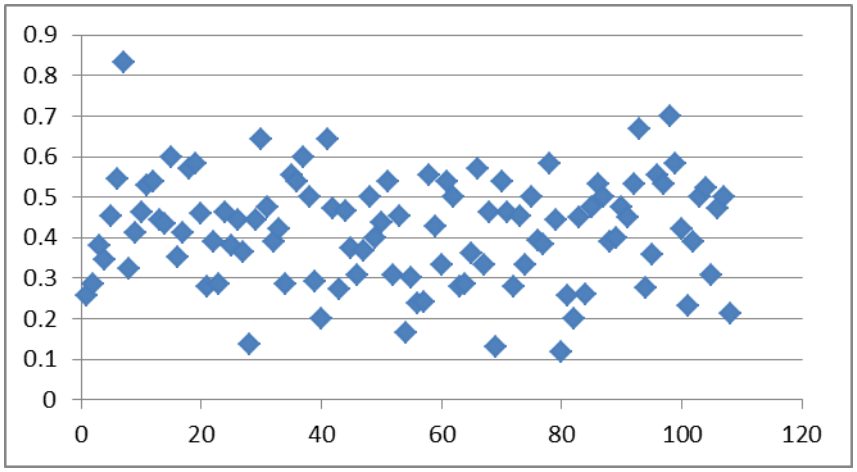

Fig. 4. – Maintaining the key words in summarization process (Palkovskii15 Corpus)

In this figure, the horizontal axis shows fragment id for source-suspicious pairs and vertical axis demonstrate the "concept preserving" percent of key words in summarization process.

6 Conclusion

In this paper, we have evaluated five text reuse corpora that are submitted to text alignment task of 7th international competition on plagiarism detection. At first, the statistical information of the corpora was analyzed. Then the plagiarism cases were manually investigated based on three measures. Finally we used automatic methods for evaluation of real and summary type of obfuscations. The result of evaluation shows that the quality of plagiarism cases in submitted corpora is rather high. However, there are some possibilities of enhancement for each of corpora from view point of quality and quantity.

Acknowledgement

This work has been accomplished in ICT research Institute, ACECR, under the support of Vice Presidency for Science and Technology of Iran - grant No. 1164331. The authors gratefully acknowledge the support of aforementioned organizations. Special thanks go to the members of ITBM research group for their valuable collaboration.

References

1. Potthast, Martin, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. "Overview of the 4th international competition on plagiarism detection." In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, Working Notes Papers of the CLEF 2012 Evaluation Labs, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

2. Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. "Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection." Computational Linguistics 39, no. 4 (2013): 917-947.

3. Clough, Paul, and Mark Stevenson. "Developing a corpus of plagiarised short answers." Language Resources and Evaluation 45, no. 1 (2011): 5-24.

4. Frank, Eibe, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H. Witten. "Data mining in bioinformatics using Weka." Bioinformatics 20, no. 15 (2004): 2479-2481.