A Corpus for Analyzing Text Reuse by People of Different Groups

Notebook for PAN at CLEF 2015

Waqas Arshad Cheema, Fahad Najib, Shakil Ahmed, Syed Husnain Bukhari, Abdul Sittar, and Rao Muhammad Adeel Nawab

Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan.
waqascheema06@gmail.com, choudharyfahad@gmail.com, shakil.ahmed@ciitlahore.edu.pk, husnain.syed@live.com, abdulsittar72@gmail.com, adeelnawab@ciitlahore.edu.pk

Abstract Plagiarism; an un-attributed reuse of text, is very significant problem specifically for higher education institutions. Consequently, a number of automated plagiarism detection system have been developed to cater this problem. The comparison of these automated plagiarism detection systems is difficult sue to problem in collecting real cases of plagiarism by students / scholars. This paper describes development of corpus containing simulated cases of plagiarism by the people having different level of writing skills. This corpus will be a very valuable addition in the set of evaluation resources presently available for comparison of plagiarism detection systems.

1 Introduction

The un-acknowledged reuse of information is generally known as plagiarism [9]. Plagiarism is acknowledged as a significant & increasing problem in higher education [7] [11] [20] [12] [5]. Resultantly, plagiarism & its detection has recently received much attention [1] [8] [21] and higher education institutions are now using automated systems to detect plagiarism in students' / scholars' work. Numerous approaches for plagiarism detection are available [2] [19]. However, one of the barriers preventing a comparison among techniques is the lack of a standardised evaluation resource.

This corpus will be a valuable addition to the set of existing corpora for the plagiarism detection task. This corpus, (1) can be used for comparison & evaluation of different techniques for plagiarism detection, (2) will help in further research in the field, (3) will be very helpful in understanding the strategies used by students / scholars when they plagiarise.

The aim of this corpus collection is to investigate how text is reused by students / scholars while writing an article, and to determine whether algorithms can be discovered to detect and quantify such reuse automatically. It is hoped that results will generalise beyond the text reuse & plagiarism in academia and provide broader insights into the nature of text derivation and paraphrase; but the selected scenario provides an ideal initial case study, and one with considerable potential practical application.

2 Related Work

There can be three types of plagiarism in a benchmark corpus; (1) artificially plagiarised documents (automatically generated), (2) simulated (manually created plagiarised documents by humans to simulate plagiarism), and (3) real cases of plagiarism [17]. The construction of a benchmark corpus containing real cases of plagiarism is difficult due to confidentiality issues [4]. The research community constructed corpora containing artificial examples of plagiarism [18], simulated examples of plagiarism [3], and the corpora containing both simulated and artificial cases of plagiarism [17].

A number of corpora have been constructed for evaluation of state of the art techniques for plagiarism detection. An outstanding effort for developing plagiarism corpora is the PAN International Competitions on Plagiarism Detection[1]. A series of evaluation labs have been held on plagiarism detection as part of the CLEF conferences[2]. A number of benchmark corpora generated as an outcome of this series of competitions [18] [17] [14] [16] [15]. Both mono-lingual and cross-lingual examples of plagiarism are present in these corpora, 90% of these are mono-lingual, and remaining 10% are cross-lingual. The distribution of plagiarised and non-plagiarised examples is uniform i.e 50% documents in each corpora are plagiarised, and remaining 50% are non-plagiarised. The plagiarised documents are created using different techniques: (1) artificial (automatically generated documents, which are further categorised into none, low and high), (2) simulated (plagiarised documents were written by humans to simulate plagiarism), (3) cyclic translation (original text in English language was translated into different languages using automated tools and then translated back to English) and (4) summarization (the original text was summarised to create plagiarised text). There is variation in length of plagiarism cases from short passages to very long passages. The mono-lingual plagiarism cases are written in English language.

The Short Answer Corpus [3] contains examples of simulated plagiarism. The Short Answer Corpus was created by asking participants to answer the five questions on different topics from Computer Science domain. In order to create non-plagiarised and plagiarised documents, each participant answered each of the five questions only once. The each answer consists of 200-300 words. All the documents (answers to the questions) in this corpus were manually (simulated) created. This corpus contains total 100 documents, 95 of which are suspicious documents and 5 documents are source Wikipedia[3] articles. Out of 95 suspicious documents, 57 documents are plagiarised with different levels of rewrite (near copy = 19, light revision = 19 and heavy revision = 19) and remaining 38 documents are non-plagiarised.

Plagiarism is not acceptable type of text reuse, but there are other forms of text reuse that are acceptable, for example reuse of news agency text by newspapers. The METER corpus[4] [6] is another benchmark corpus, which was mainly built for the study of text reuse in journalism. However, this corpus can also be used for the evaluation of plagiarism detection systems. The METER corpus contains total 1,716 documents,

[1] http://pan.webis.de/ Last visited: 02-06-2015
[2] http://clef2015.clef-initiative.eu/CLEF2015/ Last visited: 02-06-2015
[3] http://www.wikipedia.org/
[4] http://nlp.shef.ac.uk/meter/ Last visited: 18-03-2015

771 documents are Press Association (PA) articles and the remaining 945 documents are news stories published by nine different British newspapers. Each news story (suspicious document) was manually examined to access level of text reuse, and based on the amount of text reused from the PA article (potential source document) classified at document level as: (1) Wholly Derived (301 news stories), (2) Partially Derived (438 news stories), and (3) Non-derived (206 news stories).

All of the above mentioned corpora contains documents with different levels of rewrite. They lack in categorisation of documents on the basis of writer (having certain level of writing skills) of document. To the best of our knowledge, no standard evaluation resource is available for study the variation in text rewritten by groups of people having different writing skills.

3 Corpus Creation Process

3.1 Fragment Generation

Previous studies have shown that detecting paraphrased plagiarism is a difficult task and an open challenge [10] [14]. The proposed corpus aims to collect paraphrased examples of plagiarism from participants i.e. collection contains simulated cases of plagiarism.

In previous studies, simulated examples are generated from university students [3] or by paying workers on Amazon Mechanical Turk [13]. However, none of these contain paraphrased examples of plagiarism generated by different groups of people. This study aims to collect paraphrased examples of plagiarism from different groups. We selected following four groups:

i. **Undergrad in progress:** The students of undergrad program, who have not written final year project report.
ii. **Undergrad:** The people, who have completed undergrad, and they have written report for their final year project. This group also includes the students of Masters program, who have written report for their final year prject of undergrad program but they have not written their Master's thesis.
iii. **Masters:** The group of people, who have completed master degree, and have written masters thesis. This group also includes the students of PhD program, who have written their masters thesis but they have not written their PhD thesis.
iv. **PhD:** The group of people, who have completed their PhD degree.

Another important point is that participants were asked to selected text of their own research area i.e. in which they have sound knowledge and experience. Because to efficiently paraphrase a text one mush have the domain knowledge. The participants were asked to generate paraphrased plagiarism examples with different amount of text because people may have variation in the amount of text reused for plagiarism. The three variants were: small, medium and large.

Documents were collected from domains including (1) Technology, (2) Life Sciences, and (3) Humanities. The abbreviations used in xml annotation files are (1) technology, (2) life_sciences, and (3) humanities respectively for each domain. Total 250 pairs of text fragments collected from all the four groups of people. Table 1 shows detailed statistics of the text fragment pairs.

Fragment Size (Characters)	Groups of people			
	undergrad-in-progress	undergrad	masters	phd
small (<=500)	18	95	6	30
medium (<=1000)	15	51	5	9
large (>1000)	10	7	4	0

Table 1. Statistics of source-suspicious text fragment pairs in the corpus

3.2 Document Collection

After generating source-plagiarised fragment pairs, we collected 500 document pairs from Wikipedia[5], and Project Gutenberg[6] on the same topics that were used in fragment pairs.

3.3 Generating Text Alignment Corpus

The proposed corpus contains total 1000 documents (500 source documents, and 500 suspicious documents). Out of these 500 document pairs, 250 pairs are plagiarised, and remaining 250 document pairs are non-plagiarised. The 250 plagiarised pairs were created by inserting text fragments into documents. Only one source-plagiarised fragment pair was inserted into one source-suspicious document pair. The source-plagiarised fragment pairs belonging to Technology domain were inserted into source-suspicious documents belonging to same domain i.e Technology, and similarly source-plagiarised fragment pairs from other domains were inserted into source-suspicious document pairs which belonged to the same domain. Table 2 presents the domain vise statistics of fragment pairs in the corpus.

Domain	Groups of people			
	undergrad-in-progress	undergrad	masters	phd
Technology	6	126	15	39
Humanities	22	12	0	0
Life Sciences	15	15	0	0

Table 2. Domain vise Statistics of source-suspicious text fragment pairs in the corpus

3.4 Participation / Representatives

To collect corpus while ensuring authenticity and least dependencies on the contributors, three types of contributors were selected for our study: (1) family, (2) colleagues and friends and (3) university students. Note that all the contributors were volunteers, and were not paid for the purpose of data collection.

[5] http://www.wikipedia.org/

[6] http://www.gutenberg.org/

4 Peer Review

In this section we will review the other participant's corpora. We are reviewing only those corpora which are completely in English i.e. both the source and suspicious documents are in English, as it is not possible for us to review the corpora in the other unfamiliar languages. We randomly observed some of the document pairs of the entire corpus and reported our observations. Alvi15's corpus constitute of obfuscation strategies: 'non plagiarized', 'human retelling', 'synonym replacement' and 'character substitution'. In 'non plagiarized' class, we couldn't find any matching pairs at all, as it supposed to be. In 'human retelling' class, the inserted text in the source document has been paraphrased in the suspicious document. In 'synonym replacement' category, most of the words in the inserted text buffer have been replaced by their synonyms. In 'character substitution', the character we found substituted mostly is 'the' has been replaced by 'thy' at some points in the corpus. Mohtaj15's corpus compromise of obfuscation strategies: 'non plagiarized', 'no-obfuscation', 'random obfuscation', and 'simulated obfuscation'. In 'non plagiarized' class, we couldn't find any matching strings. In 'no-obfuscation', string buffers has been inserted into the documents pairs with no obfuscation i.e. the inserted text is exactly same in both the source and suspicious documents. In 'random obfuscation', the text has been randomly obfuscated i.e. the words of the matched string has been reordered randomly and makes no sense grammatically and has no meaning. In 'simulated obfuscation', the text that has been inserted randomly in the documents pairs has been paraphrased. Palkovskii15's corpus consist of obfuscation strategies: 'non plagiarized', 'no-obfuscation', 'random obfuscation', 'translation obfuscation' and 'summary obfuscation'. In 'non plagiarized' class, again we couldn't find any matching text in the observed pairs. Again in 'no-obfuscation', buffers has been inserted into the documents pairs with no obfuscation i.e. no change in text at all. Similarly in 'random obfuscation', the text has been randomly obfuscated i.e. the words of the matched string has been reordered randomly and makes no sense grammatically and has no meaning. In 'summary obfuscation', the source document is basically the short summary of the suspicious document. In "translation obfuscation", the inserted text has been paraphrased in the suspicious document. Overall we find all the three corpora error free and true in realism.

5 Conclusion

This paper explained the construction of a new corpus for text reuse & plagiarism detection research. This corpus contains examples of simulated plagiarism, and has been created manually. Our corpus is available to others for evaluation of techniques developed for plagiarism & text reuse detection. The corpus allows much more deeper analysis of different strategies used by people having different level of education.

In future, we plan to gather more document pairs to increase the size of the corpus. Also we will apply & evaluate different techniques using this corpus for text reuse & plagiarism detection.

Acknowledgements

We thank all the volunteers for their contribution in corpus construction.

References

1. Boisvert, R.F., Irwin, M.J.: Plagiarism on the rise. Communications of the ACM 49(6), 23–24 (2006)
2. Clough, P.: Plagiarism in natural and programming languages: an overview of current tools and technologies. Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK pp. 1–31 (2000)
3. Clough, P., Stevenson, M.: Developing a corpus of plagiarised short answers. Language Resources and Evaluation 45(1), 5–24 (2011)
4. 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)
5. Culwin, F., Lancaster, T.: Plagiarism issues for higher education. Vine 31(2), 36–41 (2001)
6. Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., Piao, S.: The meter corpus: a corpus for analysing journalistic text reuse. In: Proceedings of the Corpus Linguistics 2001 Conference. pp. 214–223 (2001)
7. Judge, G.: Plagiarism: Bringing economics and education together (with a little help from it). Computers in Higher Education Economics Reviews (Virtual edition) 20, 21–26 (2008)
8. Lyon, C., Malcolm, J., Dickerson, B.: Detecting short passages of similar text in large document collections. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. pp. 118–125 (2001)
9. Martin, B.: Plagiarism: a misplaced emphasis. Journal of Information Ethics 3(2), 36–47 (1994)
10. Maurer, H.A., Kappe, F., Zaka, B.: Plagiarism-a survey. J. UCS 12(8), 1050–1084 (2006)
11. McCabe, D.: Research report of the center for academic integrity. http://www.academicintegrity.org (2005)
12. Park, C.: In other (people's) words: Plagiarism by university students–literature and lessons. Assessment & evaluation in higher education 28(5), 471–488 (2003)
13. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd international competition on plagiarism detection. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
14. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 3rd international competition on plagiarism detection. In: Notebook Papers of CLEF 11 Labs and Workshops (2011)
15. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th international competition on plagiarism detection. In: CLEF (Online Working Notes/Labs/Workshop) (2013)
16. Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: CLEF (Online Working Notes/Labs/Workshop) (2013)
17. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Proceedings of the 23rd international conference on computational linguistics: Posters. pp. 997–1005. Association for Computational Linguistics (2010)

18. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
19. White, D.R., Joy, M.S.: Sentence-based natural language plagiarism detection. Journal on Educational Resources in Computing (JERIC) 4(4), 2 (2004)
20. Zobel, J.: Uni cheats racket: A case study in plagiarism investigation. In: Proceedings of the Sixth Australasian Conference on Computing Education-Volume 30. pp. 357–365. Australian Computer Society, Inc. (2004)
21. Zu Eissen, S.M., Stein, B.: Intrinsic plagiarism detection. In: Advances in Information Retrieval, pp. 565–569. Springer (2006)