UPPC - Urdu Paraphrase Plagiarism Corpus

Muhammad Sharjeel†, Paul Rayson*, Rao Muhammad Adeel Nawab†

*School of Computing and Communications, Lancaster University, UK
†Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan

Email: {s.sharjeel6.p.rayson}@lancaster.ac.uk, adeel.nawab@ciitlahore.edu.pk

Abstract

Paraphrase plagiarism is a significant and widespread problem and research shows it is hard to detect. Several methods and automatic systems have been proposed to deal with it. However, evaluation and comparison of such solutions is not possible because of the unavailability of benchmark corpora with manual examples of paraphrase plagiarism. To deal with this issue, we present the novel development of a paraphrase plagiarism corpus containing simulated (manually created) examples in the Urdu language - a language widely spoken around the world. This resource is the first of its kind developed for the Urdu language and we believe that it will be a valuable contribution to the evaluation of paraphrase plagiarism detection systems.

Keywords: Paraphrase Plagiarism, Corpus Generation, Urdu Plagiarism Detection, Natural Language Processing

1. Introduction

Plagiarism is the use of material without the specification of its source. The freely available nature of on-line text has made unacknowledged reuse much more prevalent. Plagiarism has long been considered to be a serious academic offence (Ali et al., 2011). Surveys conducted in the past reveal that more than 90% of students are involved in plagiarism (Butakov and Scherbina, 2009) while 40% committed plagiarism during assignment submission (Osman et al., 2012). These figures are alarming high. Consequently, detection of plagiarism has attracted attention of the research community (McCabe et al., 2006; Chong et al., 2010; Potthast et al., 2010b; Sanchez-Perez et al., 2014) and higher education institutions are now regularly using automatic plagiarism detection software(s) to check student’s work for plagiarism.

The task of identifying plagiarism from a suspicious document (i.e. one that is suspected to contain plagiarised text) can be accomplished using: (1) intrinsic plagiarism detection; whether the entire document is written by one single author or not and (2) extrinsic plagiarism detection; identification of the original document(s) which were used in producing the plagiarised one. Detection of plagiarism source(s) is difficult because there can be different levels of plagiarism: (1) word-to-word plagiarism, where the text is reused verbatim from the source, (2) paraphrase plagiarism, when the content of the text is obfuscated using different linguistic techniques and (3) plagiarism of idea, when the main idea of the original document is reused independently of the words in the original text (Martin, 1994). Research indicates that the current plagiarism detection systems can only detect verbatim copies and paraphrase plagiarism cases are hard to detect and therefore, are an open challenge (Maurer et al., 2006; Potthast et al., 2010a; Potthast et al., 2011; Weber-Wulff et al., 2013).

To develop and evaluate paraphrase plagiarism detection systems, we need benchmark corpora with examples that imitate real life cases. Benchmark datasets have been developed previously (Potthast et al., 2010b; Clough and Stevenson, 2011) but majority of the available resources (see Section 2) are for the English language. We see a dearth of benchmark corpora being developed for South Asian languages (Baker and McEnery, 1999) especially Urdu. Therefore, to foster plagiarism detection research in the Urdu language, we need to develop standard corpus resources for it.

As part of that endeavour, our study contributes a benchmark corpus with simulated examples of paraphrase plagiarism for one of the widely spoken language in South Asia, i.e. Urdu language. The corpus contains in total 160 documents, with 20 source documents and 140 suspicious ones. The source documents are original Wikipedia articles on well-known personalities while the set of suspicious documents are either manually paraphrased (plagiarised) versions produced by applying different rewriting techniques or set of independently written (non-plagiarised) documents. The resource is the first of its kind developed for the Urdu language and we believe that it will be a valuable contribution to the evaluation of paraphrase plagiarism detection systems. The corpus can be used for: (1) the development, analysis and evaluation of automated paraphrase plagiarism detection systems for Urdu language, (2) identifying which types of obfuscations (paraphrase strategies) are easy or difficult to detect and (3) would be a valuable resource for Urdu paraphrase identification task (at document level).

The rest of this paper is organised as follows. Section 2 covers related corpora and their properties, Section 3 describes corpus creation process and its analysis while Section 4 concludes the paper.

2. Background

The standard evaluation resources to investigate the performance of plagiarism detection systems, can contain either artificial, manual, or real cases of plagiarism. Nevertheless, it is difficult to compile a corpus with real examples of plagiarism due to the issue of confidentiality (Clough, 2003). To develop freely available and manually created resources to investigate paraphrase plagiarism detection is a labour-intensive and time-consuming task. Therefore, the
corpora developed in the past by the research community for paraphrase plagiarism detection task are either small in size or artificially created. Moreover the corpora are available mostly for English language.

The Short Answer corpus (Clough and Stevenson, 2011) contains manually created English language plagiarised and non-plagiarised texts of length between 200-300 words. The paraphrased plagiarised texts are created with either ‘light revision’ or ‘heavy revision’ of the source. The corpus contains 100 files, 57 plagiarised (19 near copy, light revision and heavy revision each), 38 non-plagiarised and 5 original texts. The PAN-PC corpora (Stein et al., 2009; Potthast et al., 2010a; Potthast et al., 2011; Potthast et al., 2012; Potthast et al., 2013; Potthast et al., 2014) are based on Project Gutenberg books and largely contain automatically generated artificial plagiarism cases. However, the later versions contain sufficient number of manually paraphrased cases, though in English language only (PAN-PC-10 (Potthast et al., 2010a) and PAN-PC-11 (Potthast et al., 2011) corpus contains 3,671 and 4,609 cases respectively). The PAN-PC (Potthast et al., 2013) was built using examples of simulated plagiarism passages found in the PAN-PC-10 Corpus (Potthast et al., 2010b). It contains 847 paraphrase sentence pairs in English language of length 50 words or less.

The above mentioned resources are for English language only and contain artificial and simulated examples of paraphrase plagiarism. In order to stimulate research in Urdu and other languages, there is a dire need to develop benchmark paraphrase plagiarism corpora in those languages. To the best of our knowledge, no paraphrase plagiarism corpus for Urdu language has been developed previously.

3. Corpus creation process

Our main purpose behind the creation of such a resource is that it could be helpful in the evaluation and comparison of state-of-the-art mono-lingual paraphrase plagiarism detection systems for Urdu language. Our Urdu Paraphrase Plagiarism Corpus (UPPC) is created to mimic the real world paraphrase plagiarism practised by students in academia. To generate example cases, we decided to use the same strategy followed by Clough and Stevenson (2011) since it accurately represents plagiarism approaches followed by students. The documents in our corpus contain examples of heavily paraphrased texts manually written by university students.

We selected a set of twenty articles, from Wikipedia, written in Urdu language, describing well-known people belonging to a variety of disciplines (see Table 1). Some of them are famous politicians, others are historical leaders and some notable religious figures. The personalities were chosen carefully, such that the source and learning material (used for creating non-plagiarised documents) could be easily obtained and the volunteers have general knowledge about them, so they can create good quality documents for the corpus. A passage of size between 200 - 300 words was excerpted from each source Wikipedia article. We chose to use Wikipedia as a source since it is a large, reliable and open content on-line repository and hence a favourite source for plagiarists (Martínez, 2009). For each of these articles, a set of documents (some plagiarised, other non-plagiarised) were generated using different rewriting approaches. Our aim was to create a resource that as accurately as possible reflects different paraphrasing mechanisms (in the plagiarised documents) to effectively check the behaviour of different paraphrase plagiarism detection algorithms. To generate paraphrased plagiarised and non-plagiarised documents, five volunteers were asked to manually write essays of length 200 - 300 words. The volunteers were undergrad students, native Urdu language speakers and had good understanding of paraphrasing mechanisms. Moreover, the students were given a detailed presentation on how to paraphrase a text and what different techniques are used in the process of rewriting a text. Overall, we tried to create near realistic plagiarism settings. A formal agreement was signed by the volunteers which enable us to make the corpus publicly-accessible.

These volunteers wrote paraphrase documents based on the Wikipedia source articles provided to them. They were told to rephrase text from the source article by replacing words with appropriate synonyms and changing sentence structure but not the meaning (semantics). There were no hard constraints on how to paraphrase or which paraphrase technique to use. The volunteers were encouraged to use their own knowledge of how to paraphrase a piece of text. It could include, but not limited to, synonym replacement, changing in tense or grammatical structure, summarising content, splitting or combining sentence to make new ones. For non-plagiarised document writing task, volunteers were provided with the learning materials in the form of online references, essays and books written on each of the personalities that could be used to generate the document. They were encouraged to use their own knowledge or obtain help from the material provided (or their own sources) but strictly required not to use Wikipedia.

3.1. Corpus properties and analysis

The corpus is saved in standard XML format and made freely available to download[1]. It contains 160 documents in total, 20 original Wikipedia sources, 75 heavily paraphrased and 75 non-plagiarised documents. The sizes of these documents vary from 200-300 words to 1000-1500 words. The corpus publicly-accessible.

Table 1: List of Wikipedia articles used for Urdu Paraphrase Plagiarism Corpus (UPPC) generation

| Number | Wikipedia article in English language |
|--------|--------------------------------------|
| 1      | Chaudhry Rehmat Ali                   |
| 2      | Liaquat Ali Khan                      |
| 3      | Tipu Sultan                           |
| 4      | Muhammad Ali Jinnah                  |
| 5      | Benazir Bhutto                        |
| 6      | Rashid Minhas                         |
| 7      | Queen Victoria                        |
| 8      | Sher Shah Suri                       |
| 9      | Bill Gates                           |
| 10     | Allama Iqbal                          |
| 11     | Muhammad (PBUH)                      |
| 12     | Mirza Ghalib                         |
| 13     | Abdul Qadeer Khan                    |
| 14     | Nusrat Fateh Khan                    |
| 15     | Fatima Jinnah                        |
| 16     | Aafia Siddiqui                       |
| 17     | Zaynab bint Ali                      |
| 18     | Bulleh Shah                          |
| 19     | Zulfiqar Ali Bhutto                   |
| 20     | Umar ibn Al-Khattab                   |

[1] http://ucrel.lancs.ac.uk/textreuse/uppc.php and via the DOI 10.17635/lancaster/researchdata/67

http://www.gutenberg.org/
Table 2: Number of Paraphrased Plagiarised (PP) and Non-Plagiarised (NP) documents in the corpus

| Personality                  | PP | NP |
|------------------------------|----|----|
| Chaudhry Rehmat Ali          | 3  | 3  |
| Muhammad (PBUH)              | 5  | 3  |
| Liaquat Ali Khan             | 4  | 4  |
| Mirza Ghalib                 | 4  | 3  |
| Tipu Sultan                  | 4  | 3  |
| Abdul Qadeer Khan            | 3  | 3  |
| Muhammad Ali Jinnah          | 4  | 4  |
| Nusrat Fateh Ali Khan        | 3  | 3  |
| Benazir Bhutto               | 4  | 4  |
| Fatima Jinnah                | 3  | 3  |
| Rashid Minhas                | 3  | 4  |
| Aafia Siddiqui               | 3  | 3  |
| Queen Victoria               | 4  | 3  |
| Zaynab bint Ali              | 4  | 4  |
| Sher Shah Suri              | 4  | 4  |
| Bulleh Shah                  | 4  | 3  |
| Bill Gates                   | 4  | 3  |
| Zulfikar Ali Bhattu          | 4  | 3  |
| Allama Iqbal                 | 4  | 3  |
| Umar ibn Al-Khatab           | 4  | 2  |

Total 75 65

Table 3: Corpus statistics

| Whole Corpus Statistics       |        |
|-------------------------------|--------|
| No. of Documents              | 160    |
| Sentence Count                | 2,711  |
| Word Count                    | 46,729 |
| Word Count (after stop-word removal) | 27,076 |
| Unique Word Count             | 6,201  |

Plagiarised Documents Statistics

| No. of Documents | 75    |
| Sentence Count   | 1,134 |
| Word Count       | 18,247|
| Word Count (after stop-word removal) | 10,647 |

Non-Plagiarised Documents Statistics

| No. of Documents | 65    |
| Sentence Count   | 1,341 |
| Word Count       | 23,978|
| Word Count (after stop-word removal) | 13,676 |

For the non-plagiarised example, the rewritten passage is independently constructed of the source (although the same words may still occur in both) and has been extended to include additional information. For example, at the start of the non-plagiarised text example (See Figure 3), the rewritten text adds new contextual information (i.e. why he got shifted to Delhi). Furthermore, sentences from both passages share content of the same events (his marriage and job) but neither of them share any similarity or have same meaning.

4. Conclusion and future work

The paper presented the construction of a manually generated and freely available paraphrase plagiarism corpus for Urdu language created to evaluate and compare Urdu plagiarism detection systems. The corpus as realistically as possible represents the strategies used by plagiarists when paraphrasing a text. Volunteers belonging to one of our educational institutions manually created paraphrased plagiarised and non-plagiarised documents on 20 renowned personalities using their own paraphrasing skills. Although the size of corpus is small, it is the first of its kind manually constructed for the Urdu language. In future, we will include further examples of paraphrased plagiarised and non-plagiarised texts and apply state-of-the-art plagiarism detection techniques and report their performance on our corpus.

5. Acknowledgements

This research is supported by the split-site PhD programme between COMSATS Institute of Information Technology and Lancaster University.

6. References

Ali, A. M. E. T., Abdulla, H. M. D., and Snasel, V. (2011). Survey of plagiarism detection methods. In 2011 Fifth Asia Modelling Symposium, pages 39–42. IEEE.
Baker, P. and McEnery, A. (1999). Needs of language-engineering communities; corpus building and translation resources. Technical report, MILLE working paper 7, Lancaster University.

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

Butakov, S. and Scherbinin, V. (2009). The toolbox for local and global plagiarism detection. *Computers & Education*, 52(4):781–788.

Chong, M., Specia, L., and Mitkov, R. (2010). Using Natural Language Processing for Automatic Detection of Plagiarism. In *Proceedings of the 4th International Plagiarism Conference (IPC-2010)*.

Clough, P. and Stevenson, M. (2011). Developing a Corpus of Plagiarised Short Answers. *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis*, 45(1):5–24.

Clough, P. (2003). Old and New Challenges in Automatic Plagiarism Detection: National Plagiarism Advisory Service.

Martin, B. (1994). Plagiarism: A Misplaced Emphasis. *Journal of Information Ethics*, 3(2):36–47.

Martínez, I. (2009). Wikipedia usage by mexican students. the constant usage of copy and paste. *Wikimania 2009*.

Maurer, H., Kappe, F., and Zak, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084.

McCabe, D., Butterfield, K., and Trevino, L. (2006). Academic Dishonesty in Graduate Business Programs: Prevalence, Causes, and Proposed Action. *Academy of Management Learning and Education*, 5(3):1–294.

Osman, A. H., Salim, N., and Abuobieda, A. (2012). Survey of text plagiarism detection. *Computer Engineering and Applications Journal (ComEngApp)*, 1(1):37–45.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010a). Overview of the 2nd international competition on plagiarism detection. In *CLEF (Notebook Papers/LABS/Workshops)*.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010b). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 997–1005. Association for Computational Linguistics.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In *Notebook Pa-
Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeno, A., Gupta, P., and Rosso, P. (2012). Overview of the 4th international competition on plagiarism detection. In CLEF (Online Working Notes/Labs/Workshop).

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Paolo, R., Efthathios, S., and Stein, B. (2013). Overview of the 5th international competition on plagiarism detection. In CLEF (Online Working Notes/Labs/Workshop).

Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., and Stein, B. (2014). Overview of the 6th international competition on plagiarism detection. In CLEF (Online Working Notes/Labs/Workshop).

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

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E. (2009). 3rd PAN Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. In 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN), pages 1–77.

Weber-Wulff, D., Möller, C., Touras, J., and Zincke, E. (2013). Plagiarism detection software test 2013. [http://plagiat.htw-berlin.de/software-en/test2013/report-2013/] [Online; accessed 25-Feb-2016].