Cross-Language Urdu-English (CLUE) Text Alignment Corpus

Notebook for PAN at CLEF 2015

Israr Hanif, Rao Muhammad Adeel Nawab, Affiffa Arbab, Huma Jamshed, Sara Riaz, and Ehsan Ullah Munir

Department of Computer Science, COMSATS Institute of Information Technology (Wah & Lahore Campuses), Pakistan.
aoaisrar@bzu.edu.pk, adeelnawab@ciitlahore.edu.pk, nagrah2012@gmail.com, humaj62@gmail.com, sarariaz15@gmail.com, ehsanmunir@comsats.edu.pk

Abstract Plagiarism is well known problem of the day. Easy access to print and electronic media and ready to use material made it easy to reuse the existing text in new document. The severity of the problem is much reduced in monolingual context by the automated and tailored effort made by the research community but the issue is yet not properly addressed in cross language (CL) text reuse. Any story or article written in any source language like Urdu is simply translated in target language like English and translator claims it as his own. Availability of standard and simulated resource address the issue and act as test bed for analyzing and implementing available plagiarism detection approaches. The research work is aimed at enriching the available cross- language corpus and on the other hand providing a benchmark corpus to Cross Language Plagiarism (CLP) domain.

1 Introduction

Text reuse is the process of developing a new document using the data of existing documents. Plagiarism is a most familiar type of text reuse. In general, plagiarism is considered as reuse of thoughts, procedures, outcomes, or words without clearly showing the original source. The size of text that is reused varies from case to case. In some conditions authors use only phrases, sentences or passages to create new document while in some conditions, word by word document is reused to create a new document. To create a new document data can be collected from different source documents. In some conditions entire document of original text is reused to create new document. Possible ways to detect plagiarism are (1) Intrinsic Plagiarism Detection- indicating whether all passages written by single author and (2) Extrinsic Plagiarism Detection- pointing all sources from where passages are used to create the suspicious document [18].

Plagiarism has crossed the language boundaries now like Urdu to English or any. Translational technologies are giving new ways of plagiarism, known as cross language plagiarism (CLP). In cross language plagiarism, source material is translated from one language to another and then translated data is reused to develop a new document without giving references of the original source. Generally such unattributed text reuse is also labeled as plagiarism [8]. In this type of plagiarism only language change occurs such as from Urdu to English or vice versa. That's why cross language plagiarism is also called translation plagiarism. Barron-Cedeno also defines CLP as a piece of text in one language translated into a target language while keeping the content and semantics same without referring the origin [2].

Availability of ready to use data in different formats and in multiple languages on internet is also boosting the case of CLP. Student assignments, and newspaper stories and articles are hot domains for CLP as education and information has no barriers and boundaries. CLP needs to develop a benchmark corpus having source and target language document pairs to detect any level of plagiarism.

Urdu is a language with more than 100 million native speakers1. Few corpuses are developed for cross-language information retrieval (CLIR) [7] but no serious effort has been made to address CLP problem. English is an official and almost educational language in indo-Pak region. This diversity raised the CLP issues with more potential in this region especially in higher education sector [6]. Therefore developing an Urdu-English corpora for CLP detection is much needed area to be focused.

This research is aimed at generating a standard corpus in Urdu-English language pairs. The corpus will serve as base for CLP detection and analyzing multiple evaluation techniques in context of performance. Three levels of plagiarism (Near Copy, Light Revision, and Heavy Revision) enabled it to detect plagiarism at different levels. Automated and manual effort to generate suspicious document made the corpus more realistic and precise.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Section 3 describes corpus generation process in detail. Analysis about corpus is presented in section 4. Finally, section 5 concludes the paper.

2 Related Work

Generation of corpus using simulated and artificial approach as recommended by Potthast et al. is in practice now [16]. Clough and Stevenson in 2011 created a short answer corpus which contains plagiarized examples generated based on simulated format [4]. Similar effort was made by stein et al. for PAN-PC-09 [17] and by Potthast et al. for PAN-PC-10 corpus [9].

In spite of the fact that research community is addressing the plagiarism issue potentially, it is majorly yet limited to monolingual aspect. The minor effort made in cross lingual aspect of the problem is also limited to few European languages like Spanish and German as source and English as suspicious in source-suspicious language pairs.

Different cross lingual corpora like English-Spanish [19] and English-German corpus [10] [12] [15] [13] [14] and many others have been developed for detection and analysis in this domain. New PAN@FIRE tasks like (CL!NSS) is an effort to trace similar news stories across the languages [5]. In 2009, European Commissions office for official publications (OPOCE) created a corpus for cross language research. Cross-Language Indian Text Reuse Competition corpus is a standard corpus in English-Hindi language pair perspective [1]. Wikipedia articles were selected as source in computer science and tourism with 112 documents as source and 276 suspicious documents for different levels of plagiarized fragments. Along with corpus creation, applying plagiarism detection approaches on newly created and already available corpuses is also in practice. The JRC-Acquis Multilingual Parallel Corpus was used by Potthast et al. to apply CLP detection approaches. 23,564 documents, extracted from legal documents of European Union, incorporate the corpus [11]. Out of 22 languages in legal document collection, only 5 including French, Germen, Polish, Dutch and Spanish was selected to generate source-suspicious language pair with English language as source. Comparable Wikipedia Corpus is another example of experimenting with similar approach. The corpus contains 45,984 documents.

Applying CLP detection approaches on multiple corpora have also been done by Ceska et al. [3]. Two corpuses JRC-EU and Fairy-tale Corpus were used for the purpose. JRC-EU composed of 400 documents randomly extracted from legislation reports of European Union. Out these 400 documents, 200 were in English as source and remaining 200 were in Czech. Fairy-tale Corpus with 54 documents out of which 27 in English and 27 in Czech translated from English, was the part of experiment.

3 Corpus Generation Process

For the PAN 2015 Text Alignment task, we submitted a cross-language corpus (Urdu-English language pair) for evaluating the performance of CLP detection system. The CLUE corpus contains simulated cases of plagiarism (source fragments are in English and suspicious ones in English).

3.1 Generation of Source-Suspicious Fragment Pairs

To generate source-suspicious fragment pairs, we collected source texts from two domains: (1) computer science and (2) general essay topics. All the source fragments were collected from Wikipedia (http://ur.wikipedia.org/wiki/urdu in footnote). It is likely that the amount of text reused for plagiarize may vary from a phrase, sentence, paragraph to entire document. Therefore, the source fragments were divided into three categories: (1) small (less than 50 words), (2) medium (50-100 words) and (3) large (100-200) words. Table 1 shows the distribution of source-suspicious fragment Paris.

To generate simulated cases of plagiarism participants (volunteers), who were university students (undergraduate and postgraduate) were asked to rewrite the source fragment (in Urdu) to generate the plagiarized fragment (in English) using one of the three methods.

i. **Near Copy:** Participants were told to automatically translate the source fragment to generate the plagiarized fragment.

ii. **Light Revision:** Using this approach, the plagiarized fragment was created in two steps. In the first step source fragment (in Urdu) is automatically translated into English. In the second step, the translated fragment is passed through an automatic text rewriting tool called Article Rewriter1 to generate the plagiarized fragment (i.e. light revision of the source fragment).

iii. **Heavy Revision:** Participants were instructed not to use the automatic machine translation tools for generating heavy revision of the source text. Instead, they were asked to manually translate the original source text in such a way that it looks like a paraphrased copy of the source text.

Level of fragments (words) (Approx.)	**Level name**	**No of fragments (270)**	
		CS(180)	**GL(90)**
<=50	Sentence (Small)	100	50
>50 and <=100	**Paragraph (Medium)**	50	25
>=100 and <=200	Essay (Large)	30	15

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

3.2 Document Collection and Corpus Composition

The proposed corpus contains total 1000 documents (500 source documents (in Urdu) and 500 suspicious documents (in English)). All the documents in the corpus are collected from freely available online resources. A document in the corpus belongs to the domain of computer science or general topics. Computer science topics (Total 50) mainly includes: Free software, Open Source, Binary Numbers, Database Normalization, Artificial intelligence, Robotics, Mobile Apps, Yahoo, MSN, Google, Whatsapp, Android, twitter, Facebook, RUBY language, Gmail, Skype, Daily motion, HTML and few others. General domain topics were also same in count and mainly include: Global warming, Muhammad Iqbal, Capitalism, Bookselling, Mosque, Pakistan Air Force, Two-Nation theory, Cricket, Fashion, Capitalism, Lahore Forte, Badshahi Masjid, Globalization and few others. Out of 500 suspicious documents, 270 are plagiarized and remaining 230 are non-plagiarized. Only one source-plagiarized fragment pair was inserted into one source-suspicious document pair. Computer science source-plagiaries fragment pairs were inserted into computer science source-suspicious documents and similarly source-plagiarized fragment pairs on general topics were inserted into source-suspicious document pairs which belonged to the domain of general topics.

Out of 270 source-plagiarized fragment pairs, 180 are from Computer Science domain and 90 from General topics domain.

All the source-plagiarized fragment pairs were randomly inserted into source- suspicious document pairs.

4 Analysis and Discussion

The developed corpus is divided into source and suspicious documents. Although the manual revision is done on each source fragment to generate its NC, LR and HR version but the order of sentences was kept same. Manual revision was done to overcome issues generated by automatic translation tools outcome. Providing Source (Urdu) version to participant for generating its Heavy Revision (HR) made the plagiarized text more realistic.

5 Conclusion

The paper describes the corpus creation process for detection of plagiarism in cross language domain of Urdu-English pairs. The corpus can be used as benchmark or test bed for upcoming tasks of performance evaluation among different plagiarism detection techniques. In future we intend to increase the size of corpus.

6 Peer Review

Following data sets were observed and most of the xml features including length and offset of the fragment inserted in source and suspicious documents were found correct in all data sets. A mismatch was also found in few cases due to newline and some special characters. Dataset wise other findings are described as:

- **cheema15-training-dataset-english**
  Different folders are used to consider cases of plagiarism at undergrad, Master and Ph. D levels. Fragments are inserted at character level at random places. Source to suspicious ratio is on to one as single source fragment is used to make a document suspicious. Obfuscation strategy is almost paraphrasing with good quality

Pair Entry / Example	**Type /Artificial / Simulated**	**Quality of Plagiarism**
suspicious-document0099-source-document0391.xml	Simulated	Well paraphrased
suspicious-document0259-source-document0189.xml	Simulated	Good
suspicious-document0309-source-document0321.xml	Simulated	Well paraphrased
suspicious-document0386-source-document0186.xml	Simulated	Well paraphrased
suspicious-document0485-source-document0447.xml	Simulated	NEAR COPY

- **palkovskii15-training-dataset-english**
  Multilingual features although described but obfuscation is limited to English only. Fragments are inserted at word level at random places in suspicious document. Source and suspicious documents are of large size and from general domain.

Pair Entry / Example	Type /Artificial / Simulated	Quality of Plagiarism
suspicious-document00021-source-document02467.xml	Artificial	NEAR COPY
suspicious-document00067-source-document02563.xml	Artificial	NEAR COPY
suspicious-document00081-source-document03075.xml	Artificial	NEAR COPY
suspicious-document00380-source-document00270.xml	translation-chain	NEAR COPY
suspicious-document00407-source-document02140.xml	translation-chain	NEAR COPY

– **mohtaj15-training-dataset-english**
Multiple fragments are inserted in single document at random places. In most of the cases 3 fragments are inserted at character level. Placement of fragments is at random places in source and suspicious documents. Although in few cases fragments in source and suspicious documents were found irrelevant but dataset is well composed overall. Large sized documents from general domain are used.

Pair Entry / Example	Type /Artificial / Simulated	Quality of Plagiarism
suspicious-document110926-source-document307308.xml	Artificial	NEAR COPY
suspicious-document179883-source-document517709.xml	Artificial	NEAR COPY
suspicious-document235057-source-document534046.xml	Artificial	NEAR COPY
suspicious-document102450-source-document106487.xml	Artificial	Poor
suspicious-document405184-source-document26685.xml	Simulated	Good
suspicious-document105415-source-document149775.xml	Artificial	Good
suspicious-document157936-source-document198805.xml	Simulated	Good

– **kong15-training-dataset-chinese**
Same text is used to suspect many documents. Small sized dataset with only 4 suspicious and 78 source documents. Suspicious text is inserted at consecutive locations probably at character level. Both source and suspicious documents are in Chinese but documents also have large English text in few cases. Quality of plagiarism cannot be judged.

- **khoshnava15-training-dataset-persian**
  A data set with 720 suspicious and 802 source documents. Almost one-to-one source to suspicious ratio is there. Artificial type of plagiarism cases with no obfuscation strategy mostly. Both source and suspicious documents are in Persian therefore quality of plagiarism cannot be judged.
- **Asghari15-training-dataset-english-persian**
  Large data set with 15959 source and 5470 suspicious documents. Most of the Plagiarism cases are artificially generated. Due to English to Persian nature quality of plagiarism cannot be judged properly. Formation of dataset is fine.
- **alvi15-training-dataset-english**
  A data set with 70 source and 90 suspicious documents. Three types of obfuscation strategies are used: character substitution, synonym replacement and human retelling. One source fragment is used in different obfuscation strategies to suspect the suspicious document. Insertion is at sentence level and almost near copy or exact copy of source fragment is used in suspicious documents. There is some difference in the source length, source offset, suspicious length and suspicious offset because of new line character.

Pair Entry / Example	**Type /Artificial / Simulated**	**Quality of Plagiarism**
suspicious-document00003.txt source-document00002.txt	Retelling	Good
suspicious-document00043- source-document00018.xml	Retelling	Good
suspicious-document00102- source-document00040.xml	Retelling	Good
suspicious-document00128- source-document00078.xml	Automatic	Well paraphrased
suspicious-document00039- source-document00010.xml	character-substitution	Good
suspicious-document00078- source-document00020.xml	character-substitution	Well paraphrased
suspicious-document00099- source-document00025.xml	character-substitution	Well paraphrased

Acknowledgements

We are thankful to all volunteers for their valuable contribution in construction of this corpus.

References

1. Barrón-Cedeno, A., Rosso, P., Devi, S.L., Clough, P., Stevenson, M.: Pan@ fire: Overview of the cross-language! ndian text re-use detection competition. In: Multilingual Information Access in South Asian Languages, pp. 59–70. Springer (2013)
2. Barrón-Cedeno, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: PAN (2008)
3. Ceska, Z., Toman, M., Jezek, K.: Multilingual plagiarism detection. In: Artificial Intelligence: Methodology, Systems, and Applications, pp. 83–92. Springer (2008)
4. Clough, P., Stevenson, M.: Developing a corpus of plagiarised short answers. Language Resources and Evaluation 45(1), 5–24 (2011)
5. Gupta, P., Clough, P., Rosso, P., Stevenson, M.: Pan@ fire: Overview of the cross-language! ndian news story search (cl! nss) track. In: Forum for Information Retrieval Evaluation, ISI, Kolkata, India (2012)
6. 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)
7. Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Cross-language information retrieval, pp. 51–62. Springer (1998)
8. Martin, B.: Plagiarism: a misplaced emphasis. Journal of Information Ethics 3(2), 36–47 (1994)
9. 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)
10. 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)
11. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Language Resources and Evaluation 45(1), 45–62 (2011)
12. Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., OberlÂĺander, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P.: Overview of the 4th international competition on plagiarism detection. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
13. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of panâĂŹs shared tasks. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pp. 268–299. Springer (2014)
14. 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)
15. 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)
16. 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)
17. 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`

18. Stein, B., zu Eissen, S.M., Potthast, M.: Strategies for retrieving plagiarized documents. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 825–826. ACM (2007)
19. Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E.: 3rd pan workshop on uncovering plagiarism, authorship and social software misuse. In: 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN). pp. 1–77 (2009)