CLEU - A Cross-Language English-Urdu Corpus and Benchmark for Text Reuse Experiments

Iqra Muneer
Department of Computer Science, Rachna College of Engineering and Technology, Pakistan

Muhammad Sharjeel
Department of Computer Science, COMSATS Institute of Information Technology, Pakistan and School of Computing and Communications, Lancaster University, UK

Muntaha Iqbal
Department of Computer Science, COMSATS Institute of Information Technology, Pakistan

Rao Muhammad Adeel Nawab
Department of Computer Science, COMSATS Institute of Information Technology, Pakistan

Paul Rayson
School of Computing and Communications, Lancaster University, UK

Text reuse is becoming a serious issue in many fields and research shows that it is much harder to detect when it occurs across languages. The recent rise in multi-lingual content on the Web has increased cross-language text reuse to an unprecedented scale. Although researchers have proposed methods to detect it, one major drawback is the unavailability of large-scale gold standard evaluation resources built on real cases. To overcome this problem, we propose a cross-language sentence/passage level text reuse corpus for the English-Urdu language pair. The Cross-Language English-Urdu Corpus (CLEU) has source text in English whereas the derived text is in Urdu. It contains in total 3,235 sentence/passage pairs manually tagged into three categories that is near copy, paraphrased copy, and independently written. Further, as a second contribution, we evaluate the Translation plus Mono-lingual Analysis method using three sets of experiments on the proposed dataset to highlight its usefulness. Evaluation results ($f_1=0.732$ binary, $f_1=0.552$ ternary classification) indicate that it is harder to detect cross-language real cases of text reuse, especially when the language pairs have unrelated scripts. The corpus is a useful benchmark resource for the future development and assessment of cross-language text reuse detection systems for the English-Urdu language pair.

Introduction

Text reuse, the process of creating new texts using existing ones, has become very common because of free, readily available, and large digital repositories. In addition, state-of-the-art text processing applications have made it very simple to copy-paste text and give it a new identity. Text borrowed from such sources can be reused verbatim (copy-paste) or rewritten (paraphrased). If the rewriting process involves complex editing operations (e.g., lexical substitution, changes in syntax, summarization, synonym replacement, altering word order, or verb or noun nominalization) then the borrowed text transforms into an independently written piece (Clough, Gaizauskas, Piao, & Wilks, 2002; Maurer, Kappe, & Zaka, 2006). Moreover, new text can be created using text from one or more sources and the amount of reused text varies from local text reuse (such as, a single word, small chunks, or sentences) to global text reuse (i.e., an entire document; Mittelbach, Lehmann, Rensing, & Steinmetz, 2010; Seo & Croft, 2008).

Unlike academic plagiarism (the unacknowledged reuse of text), text reuse is a common practice in journalism. Newspapers pay news agencies for their text(s) (here...
termed source text) to generate news stories (termed derived text). The text purchased from a news agency can be reused “verbatim” or “paraphrased” to create the newspaper story. However, at times the newspaper story might also be independently written without using any news agency text (Clough, 2010).

Text reuse can either be mono-lingual (when the source and derived text share the same language) or cross-lingual (when the source text is in one language and the derived text is in another). Mono-lingual text reuse detection has been a subject undergoing intense study for the research community for some time, but recently the focus has shifted towards detecting text reuse across languages (Ceska, Toman, & Jezek, 2008; Franco-Salvador, Gupta, Rosso, & Banchs, 2016; Gupta, Barrón-Cedeño, & Rosso, 2012; Potthast, Barrón-Cedeño, Stein, & Rosso, 2011). A recent study suggested that the scale of cross-language text reuse and plagiarism is increasing (Barrón-Cedeño, Gupta, & Rosso, 2013). This is because of the following reasons: (a) users of under-resourced languages, which are very large in number, commonly use text(s) from resource-rich languages, (b) speakers of one language staying in a country other than their own can consult the text(s) in their native language, and (c) often speakers of one language are keen to write in a foreign language. Likewise, the recent rise in multi-linguality, freely available machine translation systems, and intelligent word processors are contributing to an environment where it is easy to reuse text across languages, but with a perception of being harder to detect such reuse (Somers, Gaspari, & Niño, 2006). Therefore, there is an ever-increasing necessity to develop standard evaluation resources and methods to detect cross-language text reuse for the various language pairs.

To develop, evaluate, and analyse methods for cross-language text reuse (either local or global), gold standard benchmark corpora are needed. These corpora can be generated in three ways: (a) artificial - using an automatic text altering tool, (b) simulated - humans are asked to rewrite source text to create new text, and (c) real - new agency’s text is reused by journalists to create the newspaper story. It seems likely that cross-language text reuse detection methods which are trained on real examples are more likely to give realistic performance that we investigate further in our paper.

This study aims to develop a publicly available large-scale benchmark corpus that contains real examples of cross-language text reuse at sentence/passage level for the English-Urdu language pair. Urdu belongs to the Indo-Aryan family, widely spoken in Pakistan and the northern parts of India (Alam, Mehmood, & Nelson, 2015). Moreover, it has a strong Perso-Arabic influence in its vocabulary and is written in a Perso-Arabic script from right to left. It is also spoken world-wide because of the South Asian Diaspora (with large populations in the Middle East, United States, UK, Norway, and Canada etc.; Daud, Khan, & Che, 2016). Despite that, for the English-Urdu language pair, there are no publicly available cross-language text reuse detection datasets known to us. Moreover, previous research has tended to focus more on European languages.

The corpus developed as an outcome of this study contains 3,235 pairs of real examples of cross-language text reuse at sentence/passage level (the source text is in English whereas derived text is in Urdu). Each sentence/passage pair is categorised as i) Near Copy (NC; 751 pairs), ii) Paraphrased Copy (PC; 1751 pairs), or iii) Independently Written (IW; 733 pairs). The corpus is representative enough to serve as a benchmark for: (a) developing and evaluating techniques for cross-language text reuse detection for the English-Urdu language pair, (b) obtaining an insight into what edit operations are likely used by journalists in reusing text, and (c) to foster text reuse detection research in the English-Urdu language pair.

The remainder of this article is organized as follows. We first review previously developed cross-lingual text reuse or plagiarism detection corpora. Then we present a detailed discussion on the CLEU corpus construction, its statistics, characteristics, linguistic analysis, and example cases. This is followed by the explanation of cross-language text reuse detection experiments that we performed on our corpus to highlight its strengths and its utility for evaluation purposes. Finally, we present the results and their analysis and then conclude the article.

Related Work

In the previous literature, efforts have been made to develop standard evaluation resources for measuring cross-language text reuse (and plagiarism) for different language pairs. For example, PAN authors have developed a series of corpora with artificial and simulated examples of plagiarism at document level (Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010; Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011; Potthast et al., 2012–2014; Stein, Rosso, Stamatatos, Koppel, & Agirre, 2009). The majority (90%) of the text plagiarism cases in these corpora are mono-lingual, however, there exists a small portion (10%) of cross-lingual plagiarism cases too. These cross-language plagiarism cases are for the English-German and English-Spanish language pairs. Most of these cases are artificial (created using automatic MT [Machine Translation] system that is, Google Translate but a small number of them are created manually (i.e., translated by humans). These corpora have been used to evaluate text plagiarism detection methods in the competitions held annually.

The CLITR corpus (Cross-Language Indian Text Reuse) corpus is the first of its kind developed specifically for the analysis of

---

1 The corpus contains both sentences as well as small paragraphs or passages (up to 5 sentences) hence why the term sentence/passage is used when referring to the level of reuse.
cross-language text reuse detection in the Hindi-English language pair at document level (Barrón-Cedeño, Rosso, Devi, Clough, & Stevenson, 2013). The suspicious documents it contains are in Hindi and the source documents in English language. The training set includes 198 suspicious (Hindi) and 5,032 source (English) documents, whereas the test set has 190 suspicious (Hindi) and 5,032 source (English) documents. The CLiTR corpus contains simulated cases of text reuse. The volunteers involved in the study were asked to answer a set of 10 questions, related to the tourism and computer science domains, to create suspicious documents. It contains three types of revisions, categorized by the amount of obfuscation used, namely “Exact” (without any modifications, translation only), “Light” (very few modifications, translation, and manual correction), and “Heavy” (detailed modifications, translation, and manual correction). The corpus also contains “Original” (independently written) documents which were generated without referring to the source documents but using the learning material provided.

Another cross-language corpus of 110 documents (55 source in English and 55 plagiarised in Bangla) that contains simulated plagiarism cases and was built using student’s reports from a university (Arefin, Morimoto, & Sharif, 2013). Two groups of 55 students each, were asked to write a report on a given topic. 50 reports are used as training set whereas the remaining 10 as test set. Plagiarism cases were obfuscated by replacing contents with several plagiarized fragments of different lengths. However, the corpus is not available to download.

Recently, a cross-language (Urdu-English language pair) document level plagiarism detection corpus was submitted for the PAN 2016 shared task (Hanif et al., 2015). The corpus is divided in two sets, 500 source (Urdu) and 500 suspicious (English) documents, and contains only simulated examples of plagiarism. The source documents are Wikipedia excerpts whereas the plagiarized documents were manually created by university students. The students were asked to plagiarize 270 documents on three levels of obfuscation (“Near Copy,” “Light Revision,” and “Heavy Revision”), whereas 230 documents in the corpus are “Nonlabialized.” Moreover, the plagiarism cases inserted in the suspicious documents are of various length that is small (< 50 tokens), medium (50–100 tokens), and large (100–200 tokens). The corpus is the first cross-language (Urdu-English pair) dataset created for plagiarism detection research at the document level.

CLiPA (Cross-Language Plagiarism Analysis) is a publicly available fragment or sentence level corpus containing five source sentences (in English) which were used to generate plagiarized cases (in Spanish and Italian) using both machine translation (artificial) and manual translation (simulated; Barrón-Cedeño, Rosso, Pinto, & Juan, 2008). The machine translation cases were generated using five different services to have variations whereas for manually (human) simulated plagiarism cases, nine volunteers were asked to plagiarize each of the five source fragments. They were further requested to generate the same number of nonplagiarized cases as well. The corpus was used in experiments on text plagiarism detection research in the English-Spanish and English-Italian language pairs.

In summary, the corpora discussed above either contain artificial or simulated examples of cross-language text reuse (or plagiarism). Cross-language text reuse detection methods developed using these non-real types of text reuse are unlikely to perform well on real cases of text reuse that occur in real world scenarios (e.g., academia, journalism; Weber-Wulff, 2010).

Moreover, the simulated cases created in a controlled environment using crowd-sourcing do not represent the strategies used by humans when rewriting text in real life. Because cross-language text reuse is increasing day-by-day, first, there is an urgent need to develop text reuse detection corpora with real examples of text reuse. Second, the available corpora for research are created at document level and there are no corpora available at sentence/passage level for the English-Urdu language pair. Last, the corpora listed above are not large enough to generate robust results. This is not surprising because it takes a lot of manual effort to create corpora with simulated examples of text reuse or plagiarism.

To develop and evaluate cross-language text reuse detection methods for the real-world scenario, we need to create corpora with real examples of text reuse. To fill this gap, our research work proposes a large-scale gold standard benchmark corpus containing real examples to measure cross-language text reuse at sentence/passage level for the English-Urdu language pair. The next section describes the corpus generation process in detail.

Corpus Construction

Extracting Sentence/Passage Pairs

To construct our corpus, we first collected a dump of 1,800 “related” news archives on a wide range of topics from the domain of journalism. Half of them (900) were news agency articles written in English, whereas the remaining half (900) were Urdu news stories published in the popular Urdu newspapers of Pakistan. It is a common practice in the newspaper industry that the news text released by news agencies is directly used by the newspapers verbatim or adapted by rephrasing it (Clough, 2012). Each of the collected news agency articles (English text) had a one-to-one mapping with the newspaper story (Urdu text), but as practiced in journalism, the newspaper story may or may not contain text from the news agency article. We targeted two well-known English news agencies namely, Associated Press of Pakistan (APP) and Independent News Agency (INP) and collected English news articles released by these agencies on a range of topics including sports, politics, business, showbiz, technology, local, and foreign news. The Urdu newspaper stories were extracted from the top five large circulation national dailies that is Daily Jang, Daily Dunya, Express, Daily Aaj, and Nawa-e-Waqt for 6 months from September 2015 to
February 2016. The news text collection was carried out throughout each month excluding the public holidays on which either the newspaper was not published, or the news agency did not provide the service.

In the next step, from the set of 1,800 related English-Urdu news articles, a total of 3,235 sentence/passage pairs were manually extracted by one of the authors (who is a native speaker of Urdu, with high proficiency in English, and has expertise in the field of text reuse and plagiarism detection). The sentence/passage pairs extracted from the news agency articles (in English) were considered as “source” text whereas the newspaper stories (in Urdu) were “derived” text.

**Annotation Guidelines**

After extracting the related sentence/passage pairs, the next step was to manually classify each pair into one of the three categories that is, (a) Near Copy (NC), (b) Paraphrased Copy (PC), or (c) Independently Written (IW), depending upon the relationship between them. To classify a pair into one of the three categories, the following guidelines were prepared:

**Near Copy:** A sentence/passage pair will be tagged as “NC” if the reused text is almost an exact translation of the source text. However, because of the cross-language setting, small alterations in the reused text will be ignored. Additionally, a small amount of new text may also appear in the reused text because of the structural difference in both languages (see section Examples from the Corpus for NC example).

**Paraphrased Copy:** For a pair to be tagged as “PC,” its contents must be semantically the same, describing the same information. However, the reused text must not be mere translations of the source text. Rather, the source text should be paraphrased using different text editing operations including word/sentence re-ordering, merging/splitting of sentences, insertions/deletions of new text, replacing words/phrases with appropriate synonyms, expansion/compression of text etc. (see section Examples from the Corpus for PC example).

**Independently Written:** To tag a sentence/passage pair as “IW,” the context of the news should be the same or both texts must be describing the same event. However, the reused text must not be borrowed from the news agency text (although there may be individual words that co-occur). Moreover, a lot more new information could be present in the reused text with completely different facts and figures (see section Examples from the Corpus for IW example).

**Annotations**

With these guidelines in mind, each sentence/passage pair was then manually tagged into one of the three classes (i.e., NC, PC, or IW) by three annotators (X, Y, and Z). The annotators were post-graduate NLP (Natural Language Processing) students, native Urdu speakers with a high level of proficiency in English language. Furthermore, they were provided with training about the journalistic text reuse phenomena and with tutorials on different text rewriting operations by a linguist. The purpose of training was to comprehend the concepts, particularly, related to different levels of text reuse. The annotations began with two annotators, annotator X and annotator Y. Both were given a task to annotate a small subset of 100 sentence/passage pairs. The resulting annotations of this subset were discussed to further refine the annotation guidelines and the results were saved. Next, the remaining 3,135 sentence/passage pairs in the corpus were annotated using the final guidelines presented above and inter-annotator agreement was computed on the entire corpus. The conflicting pairs were annotated by annotator Z.

Of the total 3,235 sentence/passage pairs, agreement was reached on 2,925 pairs with unresolved conflicts on 310 pairs. The final inter-annotator agreement (IAA) on the entire corpus is 90.42% and the Kappa Coefficient (Cohen, 1960) is 73.50% (Weighted Kappa 77.40%; Cohen, 1968), which are of “Substantial” level (Landis & Koch, 1977).

As shown in Table 1, most conflicts arose in discriminating between PC and IW sentence/passage pairs (199) because it is hard to distinguish substantially paraphrased text from independently written one. The number of conflicts between NC and PC classes is also a reasonable number (109). Potentially, this is because of the small text fragments and cross-language settings, meaning that it might have become difficult for annotators to distinguish between NC and PC classes. However, there are only two conflicts between NC and IW classes. This means that it was relatively easy for annotators to distinguish between these two classes. The 310 conflicts were resolved by annotator Z, favoring 143 (46%) times annotator A and 165 (53%) times annotator B. Moreover, the two conflicts between NC and IW were resolved as PC.

**Corpus Statistics**

Our benchmark cross-language text reuse detection corpus contains 3,235 source-derived sentence/passage pairs. More than half of the pairs in the corpus belong to the PC category (54.12%). The number of NC (23.21%) and IW (22.65%) pairs are of almost equal proportions. Detailed statistics can be found in the Table 2. The corpus is freely available.

**Table 1. Annotation statistics**

|        | 3,235 |
|--------|-------|
| Total  | 2,925 |
| Conflicted | 310   |
| Conflicts PC and IW | 199   |
| Conflicts PC and NC | 109   |
| Conflicts NC and IW | 2     |
Examples From the Corpus

In this section, we present representative NC, PC, and IW sentence/passage pairs examples from our corpus. The first two examples show a Near Copy sentence and a passage pair from the proposed corpus. As can be noted, the reused text is almost exact translation of the source text. Moreover, in both cases the order of information is also preserved.

NC sentence pair example

Source
The chief minister said he would personally monitor the programme of repair and construction of roads in rural areas and review the pace of progress on fortnightly basis.

Reused
The chief minister said, "I will personally monitor the program of the construction and repair of roads in rural areas and will review the progress after every fifteen days."

NC passage pair example

Source
The Chief Minister directed that the program of Google mapping of the land be extended to the whole province and effective awareness campaign be initiated regarding the importance of the project. He appreciated concerned institutions and authorities over speedy progress on the project and said that it was a public-welfare project and would directly benefit the masses.

Reused
The Chief Minister directed that the scope of Google mapping program of the land be extended to the entire Punjab and effective awareness campaign be launched to highlight the importance of the project. The Chief Minister also praised the relevant institutions and authorities on the rapid progress of the project and said that it is an extremely important project of public interest, which would directly benefit the masses.

The next two examples show cross-language Paraphrased Copy sentence and passage pairs from the corpus. In both examples, the reused text is either compressed or expanded version of the source text. Although the content is similar, both texts convey the same information. It can also be noted that words (or phrases) have been reordered to generate the reused text. In the first example, the source text has only one sentence whereas the derived text has been split into two sentences, whereas in the second example, the source passage is comprised of two sentences whereas the reused text has been reduced to only one sentence. Furthermore, some words have been replaced with appropriate synonyms. For instance, "king" is replaced with "Shah Abdullah," "Medical City" with "Medical Center" in the reused text. The second example also exhibits a temporal change, as in the source text the age of the person mentioned "around 90" is replaced with "91" in the derived text. These transformations highlight the fact that complex rewriting operations have been used by the journalists. However, while creating the derived text from the source text, meanings of the source text have been preserved.

PC sentence pair example

Source
Prime Minister Nawaz Sharif recalled that in start of December he had announced reduction of 2.32 rupees per unit in prices of electricity, but the announcement was deferred because of Peshawar school tragedy.

Reused
Prime Minister said that the price of electricity has been decreased by 2 rupees 32 paisas per unit. In the coming days [we] are trying to make electricity more affordable.
PC passage pair example

Source
The king, whose age and frequent hospitalization have raised concerns about the future leadership of the key oil producer, was admitted to the King Abdulaziz Medical City in Riyadh, a court statement carried by the official SPA news agency said. The king is believed to be around 90 although his exact age is unknown.

Reused
سؤولی، میاپورش، شیعی 91 ساله، مدت طولانی که طبیعت عرب به به این می‌گوید، قبلاً که این می‌گوید، طبیعت عرب به به جای این که این می‌گوید، از دو چندی دیگر به چنین می‌گوید، از دو چندی دیگر به چنین می‌گوید.

Reused-translation
According to Saudi media reports, 91-year-old Shah Abdullah is unwell and has been moved to King Abdulaziz Medical Centre in Riyadh for medical check-up where some important tests will be performed on him.

The final two examples are the cross-language Independently Written sentence and passage pairs. It can be noted that both source and derived texts are describing the same news. However, the facts, figures, or contents are altered, and the way of expression is entirely different. In addition, some new (but related) information is also added in the reused text which is not present in the source text. This shows that reused text is generated independently of the source text and any overlap of words is very low between the text pair.

IW sentence pair example

Source
Prime Minister Muhammad Nawaz Sharif Wednesday announced further reduction in the prices of petroleum products, which will be effective from January 1, 2015.

Reused
وزیر اعظم نواز شریف نے پنج کمیونٹی سٹیشن کی ویژن میں 14 ارب 125 پر بجلی کا اعضا کو 122.25 پر بجلی کا اعضا کو 11.26 رپر کے بدل 6.25 پر ایل کے بدل 7.86 رپر کے بدل 11.26 رپر کے بدل 6.25 پر ایل کے بدل 7.86 رپر کے بدل 11.26 رپر کے بدل 6.25 پر ایل کے بدل 7.86 رپر کے بدل 11.26 رپر.

Reused-translation
Prime Minister Nawaz Sharif has announced reduction in prices of petroleum products up to Rs. 14 per liter. Petrol 6.25, diesel 7.86, kerosene oil 11.26 rupees per liter cheaper [than before].

IW passage pair example

Source
The sources said that two inmates of a house in the precincts of PS Brewery died because of suffocation caused by the gas leakage. The victims were shifted to Bolan Medical College Hospital and later their bodies were handed over to the heirs. Further probe was in progress.

Linguistic Analysis of the Transformations

Further to highlight the breadth of reuse in our corpus, we carried out a linguistic analysis by investigating the most common edit operations (or transformations) that have been used to generate the reused text (in Urdu) from the source text (in English). For our study, we have explored ten different types of transformations (following the taxonomy used by Clough, [2012], which are described, with examples, as follows.

Temporal change (time/date change): This involves changing time or date from the source text to create the new text (e.g., today — yesterday, 1998 — last year, earlier in this year — March etc.). Below is an example from the corpus that shows the use of word “آج” (today) to replace the day/date “Tuesday, 30 Dec.”

Trade bodies announced to close the city markets to mourn the sad incident on Tuesday, Dec 30.

ان کے مکمل ٹیم ٹیم کی تیاری پر اڑا بھی ہے۔

Change of tense: This involves creating the new text by making a shift in the tense. For instance, present tense can be changed to past tense (e.g., he is going — he went). The example that follows shows a source sentence with past tense has been rewritten with present tense.

They were awarded death sentence over insufficient proof.

ان کے مکمل ٹیم ٹیم کی تیاری پر اڑا بھی ہے۔

Use of pronoun for proper noun: While reusing text, proper noun can be changed to pronoun and vice versa (e.g., John Walker — he). The following example from our corpus exhibits the use of a proper noun “نواز شریف” (Nawaz Sharif) in derived text.

This is part of my vision of a developed KPK and a real Naya Pakistan, he said.

ان کے مکمل ٹیم ٹیم کی تیاری پر اڑا بھی ہے۔

Change in phrase order: In this type of transformation, the order of words or phrases in a text are changed when

7 In all the examples, words are underlined to highlight them.
reusing it. The corpus has plenty of such examples and one of them is shown below (because Urdu is written RTL, the underlined phrase in the Urdu text is actually the last phrase in the sentence).

No further information was immediately available, and the cause of the incident was being investigated.

**Addition/deletion of text**: This involves adding new information or deleting text in the source fragment to create the derived one. The following example shows addition of new information "(held from Pakistan)" at the end of derived text.

US transfers five Guantanamo detainees to Kazakhstan.

**Change of voice**: Sometimes the source text is converted from active to passive voice (or vice versa) to create the new text. We show here one such example from our corpus.

Direct to indirect: This type of transformation occurs when a text is converted from direct to indirect speech or vice versa (e.g., Ali said, "I am writing a letter now". — Ali said that he was writing a letter then.). Below is the example from our corpus which shows a sentence changed from direct to indirect speech.

“They must tell the players their requirement from the players,” he said.

**Change in name/title reference**: This type of change occurs when a title/designation is replaced with the actual name or vice versa (e.g., John Smith — director). Below is an example in which “PM” (Prime Minister) is replaced by “نواز شریف” (Nawaz Sharif) in the derived text.

Imran Khan presented wrong facts and figures in his recent news conference addressed at Banigala to avoid embarrassment.

**Synonym substitution**: This type exhibits the replacement of a word or phrase with an appropriate synonym. The following example shows a word “embarrassment” and a phrase “wrong facts and figures” in the source text substituted with “مشع شکار” (humiliation) and “نادری” (erratic), in the derived text, respectively.

Imran Khan presented humiliation and erratic figures in his recent news conference addressed at Banigala to avoid embarrassment.

To investigate the usage of the above transformations in our CLEU corpus, we took a random subset of 189 pairs (63 NC, 63 PC, and 63 IW) from the corpus.8 We manually

### Table 3

|                  | NC pairs |            | PC pairs |            | IW pairs |            |
|------------------|----------|------------|----------|------------|----------|------------|
|                  | $f_{abc}$ | $f_{rel}$  | $f_{abc}$ | $f_{rel}$  | $f_{abc}$ | $f_{rel}$  |
| Temporal change  | 4        | 0.210      | 12       | 0.080      | 19       | 0.101      |
| Change of tense  | 1        | 0.052      | 11       | 0.073      | 22       | 0.117      |
| Use of pronoun for proper noun | 3     | 0.157      | 12       | 0.080      | 12       | 0.064      |
| Change in phrase order | 0     | 0.000      | 19       | 0.127      | 17       | 0.090      |
| Addition/deletion of text | 0   | 0.000      | 40       | 0.268      | 72       | 0.385      |
| Change of voice | 2        | 0.105      | 4        | 0.033      | 7        | 0.037      |
| Change in name/title reference | 0     | 0.000      | 3        | 0.020      | 4        | 0.021      |
| Direct to indirect | 4    | 0.210      | 7        | 0.028      | 10       | 0.053      |
| Change in expression | 0    | 0.000      | 11       | 0.046      | 16       | 0.085      |
| Synonym substitution | 5    | 0.263      | 29       | 0.194      | 8        | 0.042      |
| Total            | 19       |            | 149      |            | 187      |            |

8To start, we randomly selected 75 examples per class, but to have the same number of sentences (40) and passages (23) across all three classes, some of the example pairs were removed.
computed the frequencies of the 10 transformations in the subset of 189 pairs. Note that because of large size of the corpus, it was very difficult and time consuming to compute these frequencies for the entire corpus, therefore, a reasonable subset is used.9

Table 3 shows the absolute (frabc) and relative (frrel) frequencies of transformations in the set of 63 each Near Copy, Paraphrased Copy, and Independently Written pairs. For Near Copy pairs, the most common edit operation is “Synonym substitution” (0.263) followed by “Temporal change” (0.210) and “direct to indirect” (0.210). As expected, four of the transformations have (0) frequency. This demonstrates the fact that when sentence/passage pairs have Near Copy relations, they are almost exact copies of each other and very few edit operations are involved.

For Paraphrased Copy pairs, “Addition/deletion of text” is the most common edit operation (0.268), followed by “Synonym substitution” (0.194), and “Change in phrase order” 0.127. This demonstrate that these three are the favorite text altering operations used by the journalists in rewriting a newspaper story using the news agency’s text. Although other transformations have relatively low frequencies, but almost all the edit operations are being used in creating derived text (total = 149). This means that complex rewriting has been used to create the Paraphrased Copy text.

Finally, the absolute and relative frequencies of edit operations for the Independently Written pairs shows “Addition/deletion of text” transformation has the highest frequency (0.385) which is like Paraphrased Copy pairs. However, this relative frequency is much higher than the second best “Change of tense” (0.117). This shows that in Independently Written pairs information is frequently added or deleted. Moreover, it further highlights the fact that reused text is independently written without borrowing text from the news agency’s copy.

Text Reuse Detection Experiments

In this section, we describe the cross-language text reuse detection experiments that we performed on our corpus. A recent study (Barrón-Cedeño, Gupta, & Rosso, 2013) compared different text similarity estimation methods and showed that Translation + Monolingual Analysis (T+MA) outperformed others. Therefore, we also applied the same approach to our corpus.

Translation Plus Monolingual Analysis

In T+MA, the overall approach is to translate the derived text to the language of the source text and then apply state-of-the-art mono-lingual analysis methods. Therefore, to start with, we translated all the derived text sentence/passage pairs from Urdu to English language using an MT system, that is Google Translate. In the second phase, we applied several mono-lingual text similarity estimation methods such as, Word n-gram overlap, Longest Common Subsequence, and Greedy String Tiling, on our corpus.

The Word n-gram overlap method tries to estimate the amount of common n-grams between source and derived texts. It has been observed that overlapping of longer chunks of n-grams indicates a potential text reuse (Adeel Nawab, Stevenson, & Clough, 2012; Clough & Stevenson, 2011). It is one of the simplest method used in text reuse detection that could easily be applied on a large collection of texts because of its low complexity. As our corpus contains sentence/passage level data (smaller units than a document), we computed the scores for lengths of n from [1–3] (i.e., Unigram, Bigram, and Trigram, see the Results and Discussion section) using overlap similarity measure (Manning & Schütze, 1999). Moreover, we merged the set of all three features (i.e., Unigram, Bigram, and Trigram) as combined (Combined, see the Results and Discussion section) input to the classification task.

Some of the other text reuse detection methods based on string matching algorithms are the Longest Common Subsequence (LCS) and Greedy String Tiling (GST). LCS is typically used for comparing files whereas GST is specifically proposed for plagiarism detection in texts (Wise, 1993). The Longest Common Subsequence of two texts, X and Y, is the longest group of elements that are common between the two and in the same order in each text. For example, the sequences “ABCDGH” and “AEDFHR” have an LCS of “ADH.” For each source and derived sentence/passage pair, we computed the normalized Longest Common Subsequence score (LCSnorm, see the Results and Discussion section) by dividing the length of LCS on the length of the shorter text.

Greedy String Tiling identifies the longest rewritten sequence of substrings from the source text and returns the sequence as tiles (for which a minimum match length [mml] value is provided) paired with the derived text. It is a powerful algorithm as it can detect matches even if some of the text is deleted or if additional text has been inserted. We used the well-known Running Karp-Rabin Matching and Greedy String Tiling (RKR-GST; Wise, 1993) implementation and experimented with mml lengths of [1-3] (i.e., GST-mml1, GST-mml2, and GST-mml3, see the Results and Discussion section).

Experimental Setup

In this section, we describe the set of experiments performed on our corpus. To estimate the similarity between source and derived text pairs, we designed three sets of experiments (i.e., Exp1, Exp2, and Exp3). For Exp1, we used the source and derived texts without any pre-processing settings. For Exp2, we excluded all the stop words10

---

9 As the subset of pairs were randomly selected, we believe that the edit operations appearing in the subset will exhibit the same proportions throughout the corpus.

10 We used the English stop words list from NLTK http://www.nltk.org
### TABLE 4. Weighted average $p$, $r$, and $f_1$ results for binary and ternary classifications (Exp1)

|       | Binary | Ternary |
|-------|--------|---------|
|       | $p$    | $r$    | $f_1$ | $p$    | $r$    | $f_1$ |
| Unigram | 0.668  | 0.762  | 0.684 | 0.522  | 0.550  | **0.493** |
| Bigram  | 0.598  | 0.773  | 0.674 | 0.433  | 0.555  | 0.444  |
| Trigram | 0.598  | 0.773  | 0.675 | 0.422  | 0.551  | 0.434  |
| Combined | 0.697  | 0.743  | **0.711** | 0.481  | 0.502  | 0.486  |
| LCS     | 0.721  | 0.774  | 0.695 | 0.518  | 0.555  | 0.486  |
| GST-mml1 | 0.687  | 0.765  | 0.691 | 0.521  | 0.551  | 0.483  |
| GST-mml2 | 0.684  | 0.772  | 0.679 | 0.521  | 0.557  | 0.460  |
| GST-mml3 | 0.636  | 0.772  | 0.675 | 0.462  | 0.553  | 0.446  |

### TABLE 5. Weighted average $p$, $r$, and $f_1$ results for binary and ternary classifications (Exp2)

|       | Binary | Ternary |
|-------|--------|---------|
|       | $p$    | $r$    | $f_1$ | $p$    | $r$    | $f_1$ |
| Unigram | 0.730  | 0.776  | 0.715 | 0.559  | 0.571  | 0.528  |
| Bigram  | 0.599  | 0.774  | 0.674 | 0.426  | 0.551  | 0.432  |
| Trigram | 0.598  | 0.773  | 0.674 | 0.418  | 0.546  | 0.415  |
| Combined | 0.698  | 0.753  | 0.709 | 0.507  | 0.524  | 0.511  |
| LCS     | 0.740  | 0.779  | 0.709 | 0.550  | 0.568  | 0.517  |
| GST-mml1 | 0.746  | 0.783  | **0.735** | 0.567  | 0.580  | **0.549** |
| GST-mml2 | 0.723  | 0.774  | 0.682 | 0.540  | 0.555  | 0.450  |
| GST-mml3 | 0.598  | 0.772  | 0.674 | 0.431  | 0.551  | 0.421  |

### TABLE 6. Weighted average $p$, $r$, and $f_1$ results for binary and ternary classifications (Exp3)

|       | Binary | Ternary |
|-------|--------|---------|
|       | $p$    | $r$    | $f_1$ | $p$    | $r$    | $f_1$ |
| Unigram | 0.731  | 0.773  | **0.732** | 0.558  | 0.573  | 0.538  |
| Bigram  | 0.599  | 0.773  | 0.675 | 0.422  | 0.551  | 0.438  |
| Trigram | 0.598  | 0.772  | 0.674 | 0.424  | 0.548  | 0.420  |
| Combined | 0.701  | 0.743  | 0.714 | 0.516  | 0.534  | 0.518  |
| LCS     | 0.723  | 0.773  | 0.714 | 0.555  | 0.570  | 0.524  |
| GST-mml1 | 0.723  | 0.770  | 0.725 | 0.565  | 0.579  | **0.552** |
| GST-mml2 | 0.735  | 0.774  | 0.680 | 0.518  | 0.554  | 0.450  |
| GST-mml3 | 0.598  | 0.772  | 0.674 | 0.447  | 0.556  | 0.430  |

---

FIG. 1. Weighted average $f_1$ results plots for binary classification
from both source and derived texts, then ran the experiment. For Exp3, after removing stop words, we further stem the source and derived texts using the Snowball stemmer (Porter, 2001). The reason for including Exp2 and Exp3 was to analyze the effect of preprocessing on text reuse detection experiments. For all three experiments, the whole corpus (i.e., 751 NC, 1,751 PC, and 733 IW pairs) was used. Moreover, we investigated both binary and ternary classifications settings for each experiment. For the set of experiments using binary classification, NC and PC sentence/passage pairs were combined to make derived (D) class whereas IW pairs were used as nonderived (ND) class.

For both (binary and ternary) supervised classification tasks, the Random Forest classifier,\(^{11}\) with 10-fold cross-validation, implemented in the WEKA\(^{12}\) machine learning platform is used. To evaluate the performance, results are reported using weighted average precision (\(p\)), recall (\(r\)), and the harmonic mean of both that is the F1-measure (\(f_1\)).

**Results and Discussion**

In this section, we discuss the results obtained as an outcome of the experiments performed on the whole corpus. The aim is to distinguish between multiple levels of text reuse at sentence/passage level. The evaluation results obtained using Exp1, Exp2, and Exp3 are shown in Tables 4, 5, and 6, respectively.

Consistently and following expectations across all three experiments, the results are higher using binary classification settings. This indicates that in text reuse detection experiments it is comparatively easier to differentiate between two classes than three (see Figure 1 and Figure 2). Overall, Unigram and GST-mml1 performed better than LCS for Exp2 and Exp3, and Combined and Unigram performed better for Exp1. This is consistent with the METER study (Clough, Gaizauskas, Piao, & Wilks, 2002) and our recent experiments on Urdu text reuse detection (Sharjeel, Nawab, & Rayson, 2017), which also highlights that Word \(n\)-gram overlap and GST are the most appropriate methods for the problem of text reuse detection. Furthermore, it can also be noticed that the results for higher values of \(n\) (in Word \(n\)-gram overlap) and mml (in GST) are low. This shows that when the text is heavily derived from the source, using smaller lengths of lexical units (words) are more effective in detection experiments. Moreover, the effect of preprocessing (stop words removal and stemming) on text reuse detection is visible in Exp2 and Exp3, where results are slightly better than Exp1 (see Figure 1 and Figure 2). This supplements the previous studies on text normalization and its effects on text reuse and plagiarism detection experiments (Ceska & Fox, 2009; Chong & Specia, 2011). Last, the high error rate in low quality MT for the Urdu-English language pair has resulted in relatively lower precision results in all three experiments.

For Exp1 (see Table 4), Word \(n\)-gram overlap and the combination of Unigram, Bigram, and Trigram that is Combined performed better than the more complex LCS and GST methods. For binary classification, the best results are obtained using Combined approach (\(f_1=0.711\)), whereas for ternary classification it is by using Unigram (\(f_1=0.493\)). This highlights that \(n\)-gram overlap with smaller values of \(n\) can capture the overlap of text between source and derived sentence/passage pairs much better. Moreover, the classifier is better suited for the binary classification problem using a combination of features than one and has produced significantly better results.

For Exp2 (see Table 5), the GST method reported comparatively better results. For both classifications, highest...
results are obtained using mml1 (binary $f_1=0.735$, ternary $f_1=0.549$). As the stop words are removed from the text, GST can find longer chunks in the paraphrased text which results in its high performance. Word $n$-gram overlap performed second best (binary $f_1=0.715$, ternary $f_1=0.528$) for smallest value of $n$, but its performance dropped with the increasing length of $n$. This shows that even if stop words are removed, small values of $n$ produce best results in text reuse detection experiments on our corpus. On the other hand, LCS reported lower scores as it suffers from the block move problem (Wise, 1993).

Interestingly, for Exp3 (see Table 6), Word $n$-gram overlap scored highest in binary classification ($f_1=0.732$) whereas GST is best in ternary classification ($f_1=0.552$). Again, the best results are with the smallest values of $n$ (in Word $n$-gram overlap) and mml1 (in GST). This further supports our statement that in the text reuse detection problem, even after applying preprocessing, selecting a smaller length of tokens is more effective. However, preprocessing of text (stop words removal and stemming) has an impact on the detection of text reuse as can be seen in the results in Table 6 which are significantly better than Table 4. This improvement is statistically significant as tested with Wilcoxon signed-rank test ($p < .05$; Wilcoxon, Katti, & Wilcox, 1973).

Overall, the cross language text reuse detection results obtained on our corpus (best $f_1=0.552$) are comparatively lower than for the METER corpus (best $f_1=0.664$), which is a gold standard mono-lingual text reuse detection corpus (Clough, Gaizauskas, Piao, & Wilks, 2002). Moreover, the results are also low when compared with CL!TR corpus (best $f_1=0.600$), which contains cross language text reuse cases for the English-Hindi language pair (Barrón-Cedeño, Rosso, Devi, Clough, & Stevenson, 2013). The rationale being that our corpus contains real examples of paraphrased text whereas CL!TR is a simulated corpus, manually created by volunteers in a controlled environment, who were allowed to use online automatic tools to translate the text (from English to Hindi language) and then modify it. On the other hand, METER is a mono-lingual corpus (both source and suspicious texts are in the English language) but the problem becomes much harder when the source and derived texts do not share the same language (Barrón-Cedeño, Rosso, Agirre, & Labaka, 2010). Our study indicates that it is challenging to detect real cases of text reuse, especially, when they occur across languages. Furthermore, it becomes much more challenging when the languages involved are non-ideographic (e.g., English and Urdu), thus finding similarity between two sentences that have disparate structure is much harder. Unlike English, the Urdu language has a very complicated morphological structure where a word could have up to sixty different forms (Rizvi & Hussain, 2005). Therefore, when the text is reused across cross-script languages, it becomes hard to detect even the exact copy pairs because of the unique characteristics, features, and structure of different languages.

Both the corpora discussed above are annotated at document level and T+MA suffers on short cases of reuse (Barrón-Cedeño, Gupta, & Rosso, 2013), which has resulted in its low performance on our corpus. One of the key phases in the T+MA method is the language normalization step. During this phase, it was observed that the available online MT systems are not of a good quality, moreover, very few of them support Urdu-English machine translation. We noticed that even the translation of verb-adj or exact copy pairs were far from perfect. This inherited flaw has also been the reason for low performance of T+MA method.

To further support this observation, we randomly selected a sample of 300 sentence/passage pairs (100 from each class) and performed both machine and manual translations of the subset. Table 7 shows the comparison of T+MA results for both machine and manual translated examples. It can be noticed that results of the manually translated sentence/passage pairs are better than machine translated ones in both binary and ternary classification tasks and across all experiments performed. This shows that T+MA method could perform comparatively better provided it incorporates a good MT system.

Table 8 shows the confusion matrix (columns=predicted, rows=actual instances) for the Exp3 ternary classification experiment using GST mml1 method that reported best result overall (see Table 6). The confusion matrix shows the misclassification of 432 NC and 510 IW as PC. This makes PC

---

**TABLE 7.** Comparison of Weighted average $f_1$ results for both machine and manually translated sample subset

| Machine Translation | Manual Translation |
|---------------------|--------------------|
|                      | Binary Classification |                      |
|                      | $f_1$               | $f_1$               |
| Exp1                | Unigram            | 0.638               | Bigram            | 0.756               |
| Exp2                | GST-mml1           | 0.560               | GST-mml1          | 0.793               |
| Exp3                | Combined           | 0.562               | Combined          | 0.781               |
|                      | Ternary Classification |                      |
|                      | $f_1$               | $f_1$               |
| Exp1                | LCS                 | 0.524               | GST-mml2          | 0.613               |
| Exp2                | Unigram            | 0.482               | GST-mml1          | 0.635               |
| Exp3                | GST-mml1           | 0.488               | Combined          | 0.578               |

---

**TABLE 8.** Confusion matrix produced for Exp3 ternary classification experiment using GST mml1 (columns=predicted, rows=actual instances)

|                   | NC  | PC  | IW |
|-------------------|-----|-----|----|
| predicted         |     |     |    |
| actual            |     |     |    |
| NC                | 280 | 432 | 39 |
| PC                | 188 | 1,401 | 162 |
| IW                | 31  | 510 | 192 |

---

13 METER contains three and CL!TR contains four levels of rewrite.

14 We only reported the best $f_1$ results on the subset. The detailed results can be accessed at http://ucrel.lancs.ac.uk/textreuse/cleu.php
the most problematic class for the three-class problem and indicates that it is difficult to discriminate between NC-PC and PC-IW pairs. Therefore, for ternary classification, the overall results were low.

Conclusion and Future Work

Cross-language text reuse detection has recently gained a lot of interest because of the rapid increase in multilingual content being readily available online. However, for the development of state-of-the-art cross-language text reuse detection methods, a major bottleneck is the unavailability of standard evaluation resources containing real cases of text reuse, especially when one of the languages (e.g., Urdu) is highly under-resourced in general. In this paper, we address this issue by contributing the first cross-language text reuse corpus for the English-Urdu language pair with real examples of reuse. The corpus contains sentence/passages collected from the news domain and manually tagged into one of the three classes (i.e., NC, PC, and IW) depending on the amount of adaptation. The source text is from the news agency articles in English language whereas the derived text extracted from the newspapers is in the Urdu language. We described the linguistic properties of such text reuse for the first time and compared related frequencies in a taxonomy of ten transformation types. In addition, we applied three sets of experiments using the popular T+MA cross-language text reuse detection method on the developed dataset. Evaluation results indicated that lower values of \( f_1 \) (for Word n-gram overlap, best \( f_1 = 0.732 \) binary classification) and \( mmr/l \) (for GST, best \( f_1 = 0.552 \) ternary classification) are the most fruitful in text reuse detection experiments on our corpus. Our results (best \( f_1 = 0.552 \)) also showed that the T+MA method applied cross-lingually is less accurate than in the mono-lingual case (METER corpus best \( f_1 = 0.664 \)) and less accurate with real reuse compared to simulated reuse (CLEFT corpus best \( f_1 = 0.600 \)). Furthermore, our experiments also revealed that text preprocessing operations that is, stop words removal and stemming have a positive effect on cross-language text reuse detection.

In the future, we plan to apply other state-of-the-art cross-language text reuse detection methods which work on the semantics of the text rather than the surface level to better capture the details of real cases of cross-language text reuse.

Acknowledgments

This work has been supported by the COMSATS Institute of Information Technology, Pakistan and Lancaster University, UK under the Split-Site PhD program.

References

Adeel Nawab, R. M., Stevenson, M., & Clough, P. (2012). Detecting text reuse with modified and weighted n-grams. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 54–58). Montreal, Canada.

Alam, S., Mehmood, F. u. d. B., & Nelson, M. L. (2015). Improving accessibility of archived raster dictionaries of complex script languages. In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 47–56). doi: 10.1145/2756406.2756926

Arefin, M. S., Morimoto, Y., & Sharif, M. A. (2013). BAENPD: A bilingual plagiarism detector. Journal of Computers, 8(5), 1145–1156.

Barrón-Cedeño, A., Gupta, P., & Rosso, P. (2013). Methods for cross-language plagiarism detection. Knowledge-Based Systems, 50, 211–217.

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. (2010). Plagiarism detection across distant language pairs. In Proceedings of the 23rd international conference on computational linguistics (pp. 37–45). Beijing, China.

Barrón-Cedeño, A., Rosso, P., Devi, S. L., Clough, P., & Stevenson, M. (2013). PAN@FIRE: Overview of the cross-language! ndian text re-use detection competition. In Multilingual information access in south asian languages (pp. 59–70). Bombay, India: Springer.

Barrón-Cedeño, A., Rosso, P., Pinto, D., & Juan, A. (2008). On cross-lingual plagiarism analysis using a statistical model. In ECAI Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (pp. 9–13). CEUR-WS.org.

Ceska, Z., & Fox, C. (2009). The influence of text pre-processing on plagiarism detection. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) (pp. 55–59).

Ceska, Z., Toman, M., & Jezek, K. (2008). Multilingual plagiarism detection. Artificial Intelligence: Methodology, Systems, and Applications, 83–92.

Chong, M., & Specia, L. (2011). Lexical generalisation for word-level matching in plagiarism detection. In Proceedings of the Recent Advances in Natural Language Processing (RANLP) (pp. 704–709). Hissar, Bulgaria: Association for Computational Linguistics.

Clough, P. (2010). Measuring text reuse in the news industry. In L. Bently, J. Davis, & J. C. Ginsburg (Eds.), Copyright and piracy (pp. 247–259). Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511761577.013 (Cambridge Books Online).

Clough, P. (2012). Mono-lingual paraphrased text reuse and plagiarism detection (PhD Dissertation). University of Sheffield, UK.

Clough, P., Gaizauskas, R., Piao, S., & Wilks, Y. (2002). Measuring text reuse. In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (pp. 152–159).

Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. Language Resources and Evaluation, 45(1), 5–24.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213.

Daud, A., Khan, W., & Che, D. (2016). Urdu language processing: a survey. Artificial Intelligence Review, 1–33.

Franco-Salvador, M., Gupta, P., Rosso, P., & Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. Knowledge-Based Systems, 111, 87–99.

Gupta, P., Barrón-Cedeño, A., & Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In Information access evaluation, multilinguality, multimodality, and visual analytics (pp. 67–75). Springer.

Hanif, I., Nawab, R. M. A., Arbab, A., Jamshed, H., Riaz, S., & Munir, E. U. (2015). Cross-language Urdu-English (chu) text alignment corpus. In Working notes papers of the clef.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. doi: 10.2307/2529310

Manning, C., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.
Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - a survey. Journal of Universal Computer Science, 12(8), 1050-1084.
Mittelbach, A., Lehmann, L., Rensing, C., & Steinmetz, R. (2010). Automatic detection of local reuse. In Sustaining tel: From innovation to learning and practice (pp. 229–244). Berlin, Heidelberg: Springer-Verlag.
Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html
Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd international competition on plagiarism detection. In CLEF (notebook papers/labs/workshops).
Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. Language Resources and Evaluation, 45(1), 45–62.
Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In CLEF (online working notes/labs/workshop).
Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., & Stein, B. (2014). Overview of the 6th International Competition on Plagiarism Detection. In Clef (online working notes/labs/workshop).
Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., ... Stein, B. (2013). Overview of the 5th International Competition on Plagiarism Detection. In CLEF (online working notes/labs/workshop).
Rizvi, S. J., & Hussain, M. (2005). Analysis, design and implementation of Urdu morphological analyzer. In SCONEST Student Conference on Engineering Sciences and Technology (pp. 1–7). Karachi, Pakistan.
Seo, J., & Croft, W. (2008). Local text reuse detection. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 571–578).
Sharjeel, M., Nawab, R. M. A., & Rayson, P. (2017, September 01). COUNTER: corpus of Urdu news text reuse. Language Resources and Evaluation, 51(3), 777–803. doi: 10.1007/s10579-016-9367-2
Somers, H., Gaspari, F., & Niño, A. (2006). Detecting inappropriate use of free online machine translation by language students—a special case of plagiarism detection. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (pp. 41–48). Oslo, Norway.
Stein, B., Rosso, P., Stamatatos, E., Koppel, M., & Agirre, E. (2009). 3rd PAN workshop on uncovering plagiarism, uthorship and social software misuse. In 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN) (pp. 1–77).
Weber-Wulff, D. (2010). Test cases for plagiarism detection software. Retrieved 2016-12-22, from https://copy-shake-paste.blogspot.co.uk
Wilcoxon, F., Katti, S., & Wilcox, R. (1973). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. Selected Tables in Mathematical Statistics, 1, 171–259.
Wise, M. J. (1993). Running karp-rabin matching and greedy string tiling. Basser Department of Computer Science, University of Sydney.