Lang Resources & Evaluation (2017) 51:777–803
DOI 10.1007/s10579-016-9367-2

ORIGINAL PAPER

COUNTER: corpus of Urdu news text reuse

Muhammad Sharjeel[1,2] · **Rao Muhammad Adeel Nawab**[2] · **Paul Rayson**[1]

Published online: 10 September 2016

Abstract Text reuse is the act of borrowing text from existing documents to create new texts. Freely available and easily accessible large online repositories are not only making reuse of text more common in society but also harder to detect. A major hindrance in the development and evaluation of existing/new mono-lingual text reuse detection methods, especially for South Asian languages, is the unavailability of standardized benchmark corpora. Amongst other things, a gold standard corpus enables researchers to directly compare existing state-of-the-art methods. In our study, we address this gap by developing a benchmark corpus for one of the widely spoken but under resourced languages i.e. Urdu. The COrpus of Urdu News TExt Reuse (COUNTER) corpus contains 1200 documents with real examples of text reuse from the field of journalism. It has been manually annotated at document level with three levels of reuse: wholly derived, partially derived and non derived. We also apply a number of similarity estimation methods on our corpus to show how it can be used for the development, evaluation and comparison of text reuse detection systems for the Urdu language. The corpus is a vital resource for the development and evaluation of text reuse detection systems in general and specifically for Urdu language.

✉ Muhammad Sharjeel
s.muhammad6@lancaster.ac.uk

Rao Muhammad Adeel Nawab
adeelnawab@ciitlahore.edu.pk

Paul Rayson
p.rayson@lancaster.ac.uk

[1] School of Computing and Communications, Lancaster University, Bailrigg, UK

[2] Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan

Keywords Mono-lingual text reuse · Urdu news corpus · Urdu text reuse detection · Corpus generation

1 Introduction

Text reuse occurs when pre-existing text(s) (source(s)) are reused to create a new text (derived). It is the process of reusing someone else's work by changing its form. Text reuse has become a common phenomenon in recent years due to the large amount of readily available text on the Web. It can vary from literal word-by-word reuse or paraphrasing the content using substitutions, insertions, deletions and re-orderings (Clough et al. 2002a; Maurer et al. 2006), or reuse of facts, concepts and even style. In general, reuse is not limited to text only but ideas, software source code, images and music, are often subjects of reuse, however, our focus is on text reuse only.

As the amount of text that is reused varies, text reuse is commonly classified as either local or global. When small phrases, sentences or paragraphs are borrowed from the source, it is considered *local text reuse* whereas when the text from the entire source document(s) is considered to create new document, we name it as *global text reuse* (Seo and Croft 2008; Mittelbach et al. 2010).

Text reuse can be mono-lingual or cross-lingual. In *mono-lingual*, source-derived text pair is in the same language while in the case of *cross-lingual*, the derived text is in a different language than the source text. In journalism, text reuse is known to be a standard practice. Plagiarism, on the other hand, represents unacknowledged text reuse in which no proper reference to the source is provided.

In recent years, due to the exponential growth of World Wide Web with vast amounts of information easily accessible, exposure to social media and collaborative content authoring systems, the reuse of text is on the rise (Butakov and Scherbinin 2009; Osman et al. 2012; Sousa-Silva 2014). Consequently, it has become a serious issue for educational institutions, online publishers and researchers worldwide (Maurer et al. 2006). To address this challenge, text reuse detection has become vitally important. Moreover, detecting text reuse has a number of key applications in different fields such as automatic plagiarism detection (Hoad and Zobel 2003; Sánchez-Vega et al. 2013), paraphrase identification (Thenmozhi and Aravindan 2015; Tsatsaronis et al. 2010), detecting breach of copyright (Aplin 2010) and news monitoring systems (Clough et al. 2002a).

Automatic text reuse detection is the task of determining whether a text, either full or partial, has been produced by exploiting another as its source. However, in both cases the task depends heavily on the underlying algorithm. The task is much simpler in the case of global text reuse detection whereas in local text reuse detection, the algorithm requires not only to find all the source(s) from where a small part of the document may have been borrowed but also the location of the borrowed fragment within the derived document (Seo and Croft 2008).

One key bottleneck in the development and evaluation of computational methods for automatic text reuse detection, is the lack of benchmark corpora which contain

various levels of reuse, e.g. exact copy, minor paraphrasing, extensive paraphrasing and so on. Although in the past, the research community has developed benchmark datasets but the majority (see Sect. 2) are for English language and we see much less focus been devoted on South Asian languages (Becker and Riaz 2002). The research on these languages is still in its infancy (Anwar et al. 2006) and we are not aware of any sizeable corpora with real examples of text reuse cases. However, the Natural Language Processing (NLP) community seems highly desirous in research of South Asian languages (McEnery et al. 2000), and a review by Baker and McEnery (1999) showed that there is a deficiency of work on these under resourced Indic (or Indo-Aryan[1]) languages. Hence, there is a need to develop standard evaluation resources to foster research in these languages.

In this paper, we present research on developing a benchmark Urdu text reuse corpus. Urdu, belonging to the Indo-Aryan language family, is the official language of Pakistan and one of the most popular languages spoken by around 175 million people around the globe. In contrast to English, Urdu is conventionally written right-to-left in Nastaliq style and relies heavily on Arabic and Persian sources for literary and technical vocabulary. However, for NLP it is a low-resource language with respect to even the core processing tasks like part-of-speech (POS) tagging or morphological analysis. Our corpus, named COrpus of Urdu News TExt Reuse[2] (COUNTER) is developed with an approach that is closely related to the METER corpus (Gaizauskas et al. 2001). It contains real examples of Urdu text reuse from the field of journalism. There are a total of 1200 documents in the corpus, half of them are source documents and the remaining half, derived documents. The source documents are produced by leading news agencies of Pakistan, whereas the derived documents are a collection of corresponding newspapers stories published in the major newspapers of Pakistan. The derived collection contains documents with various degrees of text reuse. Some of the newspaper stories (derived documents) are rewritten (either verbatim or paraphrased) from the new agency's text (source document) while others have been written by the journalists independently on their own. For the former case, source-derived document pairs are either tagged as Wholly Derived (WD) or Partially Derived (PD) depending on the volume of text reused from the news agency's text for creating the newspaper article while for the latter case, they are tagged as Non Derived (ND) as the journalists have not reused anything from the news agency's text but based on their own observations and findings, developed and documented the story.

The need for such a corpus is clear from the above discussion, and for us, it represents the first stage in a larger project. First, we intend to use this corpus to inform the design of an Urdu text reuse detection system. Second, the corpus will serve as a benchmark standard for evaluation of the proposed methods to automatically detect mono-lingual text reuse for Urdu language. Third, it can be used to develop automatic techniques which can be employed in journalism, for measuring the amount of news source copy reused, for taking appropriate actions.

[1] http://en.wikipedia.org/wiki/Indo-Aryan_languages—Last visited: 16-06-2016.

[2] The corpus is freely available to download at http://ucrel.lancs.ac.uk/textreuse/counter.php and through Lancaster's DOI: http://dx.doi.org/10.17635/lancaster/researchdata/96.

The rest of the paper is organized as follows: Section 2 describes existing corpora developed for the text reuse detection. Section 3 introduces the COUNTER corpus, explaining in detail the corpus generation process, its statistics and annotations, sample documents from the corpus and an analysis on the linguistic properties of the corpus. Section 4 explains the similarity estimation methods that we applied on our corpus to show how it can be useful in the development and evaluation of text reuse detection systems for Urdu language. Section 5 presents the experimental setup. In Sect. 6, we report and discuss the experimental results and Sect. 7 concludes the paper.

2 Related work

To develop large scale freely available resources to investigate the problem of text reuse detection is not a trivial task. However, there has been a number of efforts in the recent past, to develop standard evaluation datasets for text reuse detection, although mostly for the English language. The outcome of these efforts are the METER corpus (Clough et al. 2002a) and the Lancaster Newsbooks corpus (McEnery et al. 2010). There are a few others, the Reuters-21578 news corpus (Lewis et al. 2004) and the Text REtrieval Conference (TREC)[3] collections, that contain repeated news stories released by news-wire services. While these have not been designed to study text reuse, some researchers have used them for this purpose (Chowdhury et al. 2002; Metzler et al. 2005).

The most prominent effort in the recent years, for the development of monolingual text reuse corpora for English language, is the METER corpus (Gaizauskas et al. 2001). It consists of 1716 documents with over 500,000 words. The corpus contains 771 Press Association (PA) articles as source documents. The remaining 945 documents are news stories published in nine British newspapers (five tabloids and four broadsheets) that are derived from some of the source(s) documents. These derived documents are categorised as (1) Wholly Derived (WD); where the newspaper text is entirely based on the source document, (2) Partially Derived (PD); where the newspaper text is partly based on the source document and (3) Non Derived (ND); the situation in which the news story is written completely independent of the source document. The corpus includes documents from two domains: court and law (769 documents) and show-business (176 documents). From the 945 derived documents, 301 are tagged as WD, 438 as PD and 206 as ND. Although, in journalism, text reuse is acceptable, but as suggested by Clough (2003) the corpus has been used in the past to evaluate the performance of extrinsic plagiarism detection systems (Barrón-Cedeño et al. 2009).

The Lancaster Newsbooks corpus (McEnery et al. 2010) is a compilation of news stories texts from newsbooks published in the 17th century (especially foreign and political news). Journalists of that time used more or less the same paraphrasing mechanisms we use today for reproducing the source text about similar events in generating the newsbooks. To develop the corpus, the text was extracted from

[3] http://trec.nist.gov/—Last visited: 16-06-2016.

newsbooks between December 1653 and May 1654 and comprised of approximately 800,000 words. The authors used a sentence alignment algorithm (Piao et al. 2003) to determine the extent of similarity between two newsbook stories. However, the corpus has rarely been used for the development and evaluation of text reuse detection systems.

There are similar efforts for building datasets that contains artificial as well as simulated (manual) examples of plagiarism (a superficial type of text reuse). We discuss two such datasets, (1) the Short Answer Corpus (Clough and Stevenson 2011) (simulated plagiarism), and (2) the PAN-PC Corpora (Stein et al. 2009; Potthast et al. 2010b, 2011, 2012, 2013, 2014) (simulated and artificial plagiarism). The Short Answer corpus consists of 100 documents of length between 200 and 300 words. The documents are manually created with four levels of reuse i.e. Near copy, Light revision, Heavy revision and Non-plagiarism. The corpus has five source documents which are used to create 57 plagiarised and 38 non-plagiarised documents. The PAN-PC corpora (Stein et al. 2009; Potthast et al. 2010a, 2011, 2012, 2013, 2014) have been developed and matured over the years, and contain documents from Project Gutenberg.[4] In these corpora, the plagiarised documents contain either artificial, simulated or both cases of plagiarism. The majority of plagiarism cases are mono-lingual (in English language). A number of modification strategies were applied to create different levels of obfuscation. PAN-PC corpora provides an opportunity for NLP researchers to evaluate plagiarism detection systems using common resources and evaluation criteria, in a competition held annually.[5]

Although this research is aimed at developing a mono-lingual text reuse corpus for Urdu language, a recently released cross-lingual plagiarism corpus for Urdu-English language pair (CLUE) is worth mentioning here. The CLUE Text Alignment Corpus (Hanif et al. 2015) contains 1000 documents (500 Urdu source and 500 English suspicious documents). 270 of the suspicious documents are plagiarised while the remaining 230 are non-plagiarised. The documents of the corpus are collected from on-line sources (mainly Wikipedia[6]) and belong to two domains i.e. computer science and general topics. Volunteers (University students) were asked to generate (by manual and semi automated means) plagiarism cases (fragments) of lengths i.e. small (<50 words), medium (50–100 words) and large (100–200 words) and three levels of obfuscation i.e. Near Copy (CP), Light Revision (LR) and Heavy Revision (HR). These fragments were then inserted into the suspicious documents. The basic purpose of the corpus is to facilitate research in cross-language (Urdu–English) plagiarism detection.

Table 1 summarizes the corpora and their properties discussed above. It can been seen that the mono-lingual corpora are available only for English language and contain artificial and simulated cases of reuse (plagiarism) only. In order to stimulate research in Urdu, there is a need to develop standard evaluation resources

[4] https://www.gutenberg.org/—Last visited: 16-06-2016.

[5] http://pan.webis.de/—Last visited: 16-06-2016.

[6] https://en.wikipedia.org/wiki/Main_Page—Last visited: 16-06-2016.

Table 1 Summary of the available text reuse (and plagiarism) corpora (English)

Corpus	Source docs	Derived docs	Levels of rewrite	Domain
METER	771	945	WD, PD, ND	Journalism
Lancaster Newsbooks	N/A	N/A	N/A	Journalism
Short Answer	5	95	NC, LR, HR, NP	Wikipedia
PAN-PC[a]	11,094	11,094	P, NP	Literature

[a] Statistics of the PAN-PC-11 corpus which contains both artificial and simulated cases of plagiarism

for this language as well. As far as we are aware, no Urdu language text reuse corpus with real cases of text reuse has been previously developed.

3 Corpus

3.1 Corpus generation process

Our main intention was to develop a standard benchmark resource for the evaluation of existing systems available for text reuse detection in general and specifically for Urdu language. To generate a corpus with realistic examples, we opted for the field of journalism. In journalism, the same news story is published in different newspapers in different forms. It is a standard practice followed by all the newspapers (reporters and editors) to reuse (verbatim or modified) a news story released by the news agency.

It has been observed (Bell 1991; Fries 1987; Jing and McKeown 1999) that newspaper editors use different paraphrase mechanisms such as lexical or syntactical substitution, inflectional or derivational changes and summarisation to rewrite a newspaper story. Mostly these operations include deletion due to redundancy, making syntactic changes, use of appropriate synonyms, word re-ordering, splitting or merging sentences, tense and voice changes, use of abbreviation and verb/noun nominalisation. The choice of data collection from the press was further motivated by the fact that it is straightforward to collect news stories data with the majority of it readily and freely available on the Web in electronic form. However, some of the Urdu newspapers publish text on Web in graphics (images) form. These images were saved and later converted into electronic form (Urdu text) manually.

The COUNTER corpus consists of news articles (source documents) released by five news agencies in Pakistan i.e. Associated Press of Pakistan (APP), International News Network (INN), Independent News Pakistan (INP), News Network International (NNI) and South Asian News Agency (SANA). The corresponding news stories (derived documents) were extracted from nine daily published and large circulation national news papers of the All Pakistan Newspapers Society (APNS), who are subscribed to these news agencies. These include Nawa-e-Waqt, Daily

Dunya, Express, Jang, Daily Waqt, Daily Insaf, Daily Aaj, Daily Islam and Daily Pakistan. All of them are part of the mainstream national press, long established dailies with total circulation figures of over four million.[7] News agency texts (source documents) were provided (in electronic form) by the news agencies on a daily basis when they released the news. Newspaper stories (derived documents) were collected by three volunteers over a period of six months (from July to December 2014). National, Foreign, Business, Sports and Showbiz were the domains targeted for data collection. Table 2 shows distribution of documents in the proposed COUNTER corpus.

3.2 Corpus properties and analysis

The corpus is composed of two main document types: (1) source documents and (2) derived documents. There are total 1200 documents in the corpus: 600 are news agency articles (source documents) and 600 are newspapers stories (derived documents). The corpus contains in total 275,387 words (tokens[8]), 21,426 unique words and 10,841 sentences. The average length of a source document is 227 words while for derived documents it is 254 words. Table 3 shows detailed statistics of the proposed COUNTER corpus.

3.3 Annotations and inter-rater agreement

The annotations were performed by three annotators (A, B and C), who were native Urdu language speakers and experts of paraphrasing mechanisms. All three were graduates, experienced in text annotations and having an advanced Urdu level. The corpus has been annotated at the document level with three classes of reuse i.e. Wholly Derived (WD), Partially Derived (PD) and Non Derived (ND). The annotations were carried out in three phases: (1) training phase, (2) annotations, (3) conflict resolving. During the training phase, annotators A and B manually annotated 60 document pairs, following a preliminary version of the annotation guidelines. A detailed meeting was carried out afterwards, discussing the problems and disagreements. It was observed that the highest number of disagreements were between PD and ND cases, as both found it difficult to distinguish between these two classes. The reason being that adjusting the threshold where a text is heavily paraphrased or new information added to it that it becomes independently written (ND). Following the discussion, the annotation guidelines were slightly revised, and the first 60 annotations results were saved. In the annotation phase, the remaining 540 document pairs were manually examined by the two annotators (A and B). Both were asked to judge, and classify (at document level) whether a document (newspaper story) depending on the volume of text rewritten from the source (news agency article) falls into one of the following categories:

[7] https://pakpressfoundation.wordpress.com/2006/05/05/pakistan-press-foundation—Last visited: 16-06-2016.

[8] Compound words in Urdu were treated as single words during tokenisation.

Table 2 Distribution of documents by news agencies, newspapers and domains

News agencies		News papers		Domains	
APP	543	Nawa-e-Waqt	145	Sports	222
INN	39	Daily Dunya	132	National	181
NNI	8	Express	115	Foreign	121
SANA	6	Daily Waqt	89	Showbiz	49
INP	4	Daily Insaf	55	Business	27
		Daily Islam	36		
		Jang	21		
		Daily Aaj	6		
		Daily Pakistan	1		

Table 3 Corpus statistics

	Source	Derived
Total number of documents	600	600
Average no of words per document	227	254
Average no of sentences per document	9	8
Smallest document (by words)	52	43
Largest document (by words)	1377	2481

Wholly Derived (WD) The News agency text is the only source for the reused newspaper text, which means it is a verbatim copy of the source. In this case, most of the reused text is word-to-word copy of the source text.
Partially Derived (PD) The Newspaper text has been either derived from more than one news agency or most of the text is paraphrased by the editor when rewriting from news agency text source. In this case, most parts of the derived document contain paraphrased text or new facts and figures added by the journalist's own findings.
Non Derived (ND) The News agency text has not been used in the production of the newspaper text (though words may still co-occur in both documents), it has completely different facts and figures or is heavily paraphrased from the news agency's copy. In this case, the derived document is independently written and has a lot more new text.

After the annotation phase, the inter-annotator agreement was computed. The inter-rater score was calculated to be 85.5 % as the annotators had agreement on 513 of the 600 pairs. The Kappa Coefficient was computed to be 77.28 % (Weighted Kappa 81.4 %) (Cohen 1960, 1968). The inter-rater agreement score of 85.5 % is good, considering three levels of classification involved in the difficulty of the rating task. In the third and last phase, the conflicting 87 pairs were given to the third annotator (C) for conflict resolution. The decision of the third annotator was considered final. Out of the 600 document pairs, the final gold standard annotated dataset contains 135 (22.5 %) WD, 288 (48 %) PD and 177 (29.5 %) ND

Table 4 Classification of document pairs in the COUNTER corpus and its comparison with METER corpus (Gaizauskas et al. 2001)

Classification	COUNTER	METER
WD	135 (22.5 %)	301 (31.8 %)
PD	288 (48.0 %)	438 (46.3 %)
ND	177 (29.5 %)	206 (21.7 %)

documents. Table 4 lists the classification of documents in the COUNTER corpus and compares it with the METER corpus (Gaizauskas et al. 2001). It highlights the similarity of our corpus with METER as both corpora have majority of the documents in the PD class i.e. 48 % (METER) and 46.3 % (COUNTER).

3.4 Examples of text reuse cases from the corpus

This section shows examples of the WD, PD and ND document pairs from the corpus. As expected, the derived document in WD (see Fig. 1) is word-to-word copy of the source document.[9] The information described in the derived text is the same as in the text reported by the news agency. In case of PD (see Fig. 2), source text has been rephrased by changing the passages with different paraphrasing techniques. Also, in some cases, the derived text contains additional events not reported by the new agency source. For ND (see Fig. 3), a lot more new information has been added in the derived document independently without using the source. For standardisation purposes, the documents in the corpus have been saved as standard XML documents. Details of the XML tags and DTD can be found in the README file available with the corpus.

3.5 Linguistic analysis of the corpus

There are numerous ways to rewrite texts and in the previous studies, researchers have classified the 'edit operations' (paraphrase mechanisms) into different types, in different corpora, to form paraphrase topologies (Clough 2003; Barrón-Cedeño et al. 2013; Vila et al. 2014). Following the same approach, we also identified the paraphrase mechanisms used (by journalists) to formulate the newspaper story (derived document), in our corpus.

The typology (see Table 5) we followed, to present a linguistic analysis of our corpus, consists of a concise but concrete list of linguistic phenomena underlying paraphrasing. It is a two level typology, with 6 classes and 14 paraphrasing types. At the first level, each class describes the nature of paraphrase phenomenon while a second more fine-grained level lists the actual paraphrase mechanism used.

[9] Words common in both documents are underlined.

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<COUNTER_document classification="" domain="foreign" filename="0472.xml" newsdate="19.07.14" newspaper="APP" noofwordswithSWR="96" totalnoofsentences="4" totalnoofwords="130">
<headline>
دبئی میں منعقد ہونے والی بین الاقوامی تجارتی نمائش میں شرکت کے خواہشمند اداروں سے درخواستیں طلب
</headline>
<body>
ٹریڈ ڈویلپمنٹ اتھارٹی آف پاکستان (ٹی ڈی اے پی) نے دبئی متحدہ عرب امارات میں منعقد ہونے والی تین روزہ بین الاقوامی تجارتی نمائش "میڈان ایشیائی" میں شرکت کے خواہشمند اداروں سے درخواستیں طلب کی ہیں۔ اتھارٹی کی جاری تفصیلات کے مطابق عالمی تجارتی میلے میں ٹیکسٹائل، لیدر اور عام ضروریات کی اشیاء تیار و برآمد کرنے والے ادارے شرکت کے ذریعے عالمی خریداروں سے کاروباری معاملات طے کر سکیں گے جس سے ملکی برآمدات کو فروغ حاصل ہوگا۔ ٹی ڈی اے پی نے نمائش میں شرکت کے خواہشمند اداروں سے 30 جولائی 2014ء تک درخواستیں طلب کی ہیں جبکہ عالمی تجارتی میلہ 12 تا 14 اکتوبر 2014ء کے دوران دبئی میں منعقد ہوگا۔
</body>
</COUNTER_document>

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<COUNTER_document classification="WD" domain="foreign" filename="0472p.xml" newsdate="20.07.14" newspaper="daily_dunya" noofwordswithSWR="120" totalnoofsentences="6" totalnoofwords="169">
<headline>
دبئی میں عالمی تجارتی نمائش میں شرکت کیلئے درخواستیں طلب
</headline>
<body>
ٹیکسٹائل، لیدر اور عام ضروریات کی اشیا تیار و برآمد کرنے والے ادارے شرکت کریں گے نمائش میں شرکت کیلئے درخواستیں 30 جولائی تک جمع کروائی جاسکتی ہیں۔ ٹریڈ ڈیولپمنٹ اتھارٹی آف پاکستان (ٹی ڈی اے پی) نے دبئی متحدہ عرب امارات میں منعقد ہونے والی 3 روزہ بین الاقوامی تجارتی نمائش "میڈان ایشیا" میں شرکت کے خواہشمند اداروں سے درخواستیں طلب کی ہیں۔ اتھارٹی کی جاری تفصیلات کے مطابق عالمی تجارتی میلے میں ٹیکسٹائل، لیدر اور عام ضروریات کی اشیا تیار و برآمد کرنے والے ادارے شرکت کے ذریعے عالمی خریداروں سے کاروباری معاملات طے کر سکیں گے جس سے ملکی برآمدات کو فروغ حاصل ہوگا۔ ٹی ڈی اے پی نے نمائش میں شرکت کے خواہشمند اداروں سے 30 جولائی 2014ء تک درخواستیں طلب کی ہیں جبکہ عالمی تجارتی میلہ 12 تا 14 اکتوبر 2014ء کے دوران دبئی میں منعقد ہوگا۔ میلے میں شرکت کیلئے دگر تفصیلات اتھارٹی کی ویب سائٹ سے بھی حاصل کی جاسکتی ہیں۔
</body>
</COUNTER_document >

Fig. 1 Example of a WD document pair

In the following discussion, we describe each of the 14 types of our typology with examples[10] from our corpus.

Morphology-based changes

Inflectional changes often involves changing a grammatical category (e.g. from singular to plural or vice versa) with a prefix/suffix. In the example below, word [wickets] is transformed into [wicket] to produce the change.

S: پاکستان کے 4 [وکٹوں] پر 261 رنز

D: پاکستان کے 4 [وکٹ] پر 261 رنز

Derivational changes consists of word alteration that forms a new word by adding an affix to the root form of the word. In the example below, the word [Pakistan-i] (adjective) is changed to [Pakistan] (noun).

[10] The examples shown here are just small fragments extracted from the source/derived documents. Refer to Sect. 3.4 to see full examples of source/derived documents. The words/phrases in focus of discussion are enclosed in square brackets to emphasize them.

<?xml version="1.0" encoding="UTF-8"?>
<COUNTER_document totalnoofwords="181" totalnoofsentences="5" noofwordswithSWR="118"
newspaper="APP" newsdate="24.11.14" filename="0318.xml" domain="national" classification="">
<headline>
وزیراعظم محمد نواز شریف کاافغان صدر اشرف غنی کو فون صوبہ پکتیا میں خود کش دھماکے میں 50 سے زائد افراد کی ہلاکت پر گہرے دکھ اور افسوس کا
اظہار
</headline>
<body>
وزیراعظم محمد نواز شریف نے افغان صوبہ پکتیا میں خود کش حملے کے نتیجے میں 50 سے زائد افراد کی ہلاکت پر گہرے دکھ اور افسوس کا اظہار کرتے ہوئے
کہا ہے کہ پاکستان اور افغانستان مشترکہ جدوجہد کے ذریعے دہشت گردی کا خاتمہ کرنے میں کامیاب ہونگے۔ پیر کو وزیراعظم نے افغانستان کے صدر
اشرف غنی کو ٹیلیفون کیا جس میں انہوں نے گزشتہ روز افغان صوبہ پکتیا میں خود کش دھماکے کے نتیجے میں 50 سے زائد افراد کی ہلاکت پر گہرے دکھ اور
افسوس کا اظہار کیا۔ افغان صدر سے بات چیت کرتے ہوئے وزیراعظم محمد نواز شریف نے خود کش حملے کی مذمت کی اور اسے بزدلانہ اقدام قرار دیا۔
وزیراعظم نے افغان بھائیوں کے ساتھ یکجہتی کا اظہار کرتے ہوئے اس یقین کا اظہار کیا کہ پاکستان اور افغانستان باہم مل کر دہشت گردی کا خاتمہ کرنے
میں کامیاب ہونگے۔
</body>
</COUNTER_document>

<?xml version="1.0" encoding="UTF-8"?>
<COUNTER_document totalnoofwords="161" totalnoofsentences="8" noofwordswithSWR="101"
newspaper="daily_dunya" newsdate="25.11.14" filename="0318p.xml" domain="national"
classification="PD">
<headline>
دہشتگرد مشترکہ دشمن ہیں، نواز شریف کاافغان صدر غنی کو فون بزدلانہ کارروائیوں سے دونوں ملکوں کے عوام گھبرانے والے نہیں پکتیکا دھماکے پر
افسوس
</headline>
<body>
وزیراعظم نواز شریف نے افغان صدر اشرف غنی کو ٹیلی فون کیا اور پکتیکا میں ہونے والے دھماکے پر اظہار افسوس کیا ہے۔ وزیراعظم ہاؤس کے
مطابق وزیراعظم نواز شریف نے ہر دھماکے میں انسانی جانوں کے ضیاع پر افسوس کا اظہار کیا۔ وزیراعظم کا کہنا تھا کہ دہشت گرد پاکستان اور افغان عوام
کو کبھی شکست نہیں دے سکتے، بزدلانہ کارروائیوں سے دونوں ملکوں کے عوام گھبرانے والے نہیں ہیں۔ انہوں نے دہشت گرد دونوں ممالک کے
مشترکہ دشمن ہیں اور دونوں ممالک مشترکہ طور پر اس پر لعنت کا خاتمہ کر رہے ہیں۔ اس موقع پر افغان صدر نے وزیراعظم کا شکریہ ادا کیا۔ ذرائع کے
مطابق وزیراعظم نے افغان صدر سے خطے کی صورتحال پر بھی تبادلہ خیال کیا۔ دونوں رہنماؤں کے درمیان پانچ منٹ تک گفتگو ہوتی رہی۔
</body>
</COUNTER_document>

Fig. 2 Example of a PD document pair

S: امریکی منڈیوں میں [پاکستانی] مصنوعات کیلئے بہتر رسائی کی ضرورت ہے

D: [پاکستان] کی اشیاکی امریکی منڈیوں تک رسائی ضروری ہے

Lexicon-based changes

Spelling and format changes are lexical changes that occur in the spellings and representation of the text (e.g. abbreviations, or digit/letter alternations). In example below, abbreviations are changed to their full forms.

S: پشاور پینتھرز، [این بی پی]، راولپنڈی ریمز، [یو بی ایل] شامل ہیں

D: پشاور پینتھرز، [نیشنل بینک آف پاکستان]، ... ، راولپنڈی ریمز اور [یونایٹڈ بینک لمیٹڈ] شامل ہیں

Same-polarity substitutions comprises of replacing the appropriate word or phrase with similar meaning (synonym). The corpus text has many such examples, the sentence below shows a word in the source text [victim] substituted with [suspected case] in the derived text.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<COUNTER_document classification="" domain="national" filename="0581.xml" newsdate="28.10.14"
newspaper="APP" noofwordswithSWR="295" totalnoofsentences="15" totalnoofwords="511">
<headline>
تمام سیاسی جماعتیں اختلافات بھلا کر ملک کی بہتری کیلئے ورکنگ ریلیشن شپ کو بہتر کریں، سندھ اس بات کا متحمل نہیں ہو سکتا کہ یہاں قومیتوں کے
جھگڑے بڑھیں، خورشید شاہ سید ہیں وہ اپنے نانا کی توہین نہیں کر سکتے، مسلم لیگ فنکشنل اس مسئلہ کے حل کے لئے اپنا کردار ادا کرے امیر جماعت
اسلامی سراج الحق کی پیر صاحب پگاڑا سے ملاقات کے بعد میڈیا سے گفتگو
</headline>
<body>
امیر جماعت اسلامی سراج الحق نے کہا ہے کہ تمام سیاسی جماعتوں کو چاہئے کہ وہ سیاسی اختلافات بھلا کر ملک کی بہتری کیلئے آپس میں ورکنگ ریلیشن شپ
کو بہتر کریں، سندھ اس بات کا متحمل نہیں ہو سکتا کہ یہاں قومیتوں کے جھگڑے بڑھیں سندھ کی ترقی پاکستان کی ترقی ہے۔ خورشید شاہ سید ہیں وہ اپنے نانا
کی توہین نہیں کر سکتے، مسلم لیگ فنکشنل اس مسئلہ کے حل کے لئے اپنا کردار ادا کرے۔ ان خیالات کا اظہار انہوں نے منگل کو راجہ ہاؤس کراچی میں پیر
صاحب پگارا پیر صبغت اللہ سے ملاقات کے بعد میڈیا سے گفتگو کرتے ہوئے کیا۔ اس موقع پر وفاقی وزیر سمندر پار پاکستانی وانسانی ترقی پیر صدر الدین شاہ
راشدی، شہریار مہر، جام مدد علی، نصرت سحر عباسی، کامران ٹیسوری، خضر حیات منگریو اور دیگر بھی موجود تھے۔ سراج الحق نے کہا کہ پیر صاحب پگارا سے
ملاقات کی کافی عرصے سے خواہش تھی، آج ان سے اور ان کے بھائی پیر صدر الدین راشدی سے تفصیلی ملاقات ہوئی ہے۔ انہوں نے کہا کہ ملک میں عام
آدمی کے مسائل ہیں لیکن اس جانب کوئی توجہ نہیں دے رہا ہم چاہتے ہیں تمام سیاسی قیادت آپس کے اختلافات بھلا کر بیٹھیں اور عام آدمی جو محنت
مزدوری کر کے اپنے بچوں کو کھلاتا ہے وہ مشکلات کا شکار ہے۔ انہوں نے کہا کہ موجودہ الیکشن کمیشن کے تحت اگر الیکشن ہوئے تو پھر ملک کا وہی حال ہوگا
اس لئے الیکشن ریفارمز لائی جائیں اور الیکشن سسٹم کو بہتر کیا جائے۔ انہوں نے یہ ٹھیک ہے ایک پارٹی نے اسلام آباد میں دھرنے کے خاتمے کا اعلان کیا
ہے لیکن ملک کا سیاسی بحران ختم کرنے کیلئے حکومت اور عمران خان کو مزاکرات کی ٹیبل پر بیٹھنا ہوگا۔ سراج الحق نے کہا کہ محرم الحرام کا مہینہ ہے اس
لئے میں اپیل کرتا ہوں کہ آپس کے اتفاق کو بڑھایا جائے حضرت امام حسینؓ کسی ایک فرقے کہ نہیں تھے اس لئے پوری دنیا میں اچھا پیغام جاتا ہے۔ انہوں
نے کہا کہ میں نے پیر صاحب پگارا اور ان کے بھائی پیر صدر الدین شاہ راشدی کو منصورہ کرنے کی دعوت دی ہے جو انہوں نے قبول کی ہے۔ اس موقع پر
پاکستان لیگ فنکشنل سندھ کے صدر پیر صدر الدین شاہ راشدی نے کہا کہ میں امیر جماعت اسلامی اور ان کے ساتھ راجہ ہاؤس آمد پر شکریہ ادا کرتا ہوں۔
انہوں نے کہا کہ ہمارے اور ان کے بزرگوں کے تعلقات تھے۔ انہوں نے کہا کہ ہماری ملاقات میں فیڈریشن کو بچانے کے لئے مل جل کر کوششوں پر
اتفاق ہوا ہے۔ انہوں نے کہا کہ اس ملاقات کے اچھے نتائج نکلیں گے۔
</body>
</COUNTER_document>

----------

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<COUNTER_document classification="ND" domain="national" filename="0581p.xml"
newsdate="29.10.14" newspaper="nawa-e-waqt" noofwordswithSWR="281" totalnoofsentences="17"
totalnoofwords="470">
<headline>
4 صوبے چلا کر دکھائیں پھر 20 بنائیں، سراج الحق: پیر پگاڑا سے ملاقات، جمہوری سسٹم کے تحفظ پر اتفاق
</headline>
<body>
جماعت اسلامی کے امیر سراج الحق اور پیپلز پارٹی کے رہنما سابق وزیر اعظم یوسف رضا گیلانی نے گزشتہ روز کراچی میں فنکشنل لیگ کے سربراہ پیر پگاڑا
صبغت اللہ راشدی سے انکی رہائش گاہ پر علیحدہ علیحدہ ملاقاتیں کیں۔ جن میں ملک کی موجودہ صورتحال پر تبادلہ خیال کیا گیا۔ سراج الحق نے صحافیوں سے
گفتگو کرتے ہوئے کہا کہ جب تک ملک میں عدل وانصاف کا موثر نظام قائم نہیں ہوگا مسائل حل نہیں ہوں گے۔ پیر پگاڑا کے ساتھ ملاقات بہت اچھی
رہی۔ ہم نے اتفاق کیا ہے کہ ملک کی سلامتی، بقاء اور جمہوری نظام کے تحفظ کے لئے ہم آپس میں تعاون کریں گے اور مشاورت کا یہ عمل جاری رکھیں
گے۔ انہوں نے کہا کہ عوام کے اندر مایوسی، غم وغصہ ہے اگر سیاسی قائدین نے مل کر مسائل کا حل نہ نکالا تو عوام کی یہ خاموشی اور اضطراب کسی بڑے
طوفان کا پیش خیمہ ثابت ہو سکتی ہے۔ سراج الحق نے بتایا کہ ملاقات میں اس بات پر اتفاق کیا گیا کہ موجودہ الیکشن سسٹم کے اندر ریفارمز ضروری ہے اور
ان ریفارمز کے لئے سب کو اعتماد میں لیا جائے، موجودہ سسٹم کے تحت ہونیوالے انتخابات سے تنازعات کا خاتمہ نہیں ہو سکے گا۔ سراج الحق نے کہا پہلے
4 صوبے چلا کر دکھائیں پھر 30، چالیس بنالیں، 4 صوبے چلتے نہیں باقی کیسے چلا لینگے۔ انہوں نے کہا کہ جلد از جلد بلدیاتی انتخابات کرائے جائیں، لڑائی
سے کچھ ہاتھ نہیں آئیگا، ہو سکتا ہے سیاست اور جمہوریت کا خاتمہ ہو جائے۔ دریں اثناء پیپلز پارٹی کے رہنما سابق وزیر اعظم یوسف رضا گیلانی نے بھی
پیر پگاڑا سے انکی رہائش گاہ پر ملاقات کر کے سیاسی صورت حال پر تبادلہ خیال کیا۔ اس موقع پر گفتگو کرتے ہوئے پیر پگاڑا نے کہا کہ کو نسی فورس ہے جو
کرپشن کرنیوالوں کا احتساب کرے۔ احتساب سے پہلے الیکشن کا کوئی فائدہ نہیں ہوگا۔ انہوں نے کہا کہ وزیر اعظم نواز شریف کو صرف پنجاب کی فکر ہے
دیگر صوبوں یا وفاق کی نہیں، مڈ ٹرم الیکشن ملک کیلئے خطرناک ہونگے، موجودہ دور میں کرپشن کے سارے ریکارڈ ٹوٹ گئے۔ دوسری طرف کراچی میں
تقریب سے خطاب کرتے ہوئے سراج الحق نے کہا کہ پاکستان کو بیرونی عناصر سے زیادہ خطرہ اشرافیہ سے ہے۔ کراچی والوں نے 20 سال ایم کیو ایم کو
ووٹ دیئے مگر انہیں حقوق نہیں ملے۔ انہوں نے کہا کہ ایک ڈاکٹر ناکام ہو تو دوسرے کے پاس علاج کیلئے جایا جاتا ہے کراچی والو ہماری طرف پلٹ آؤ۔
انہوں نے کہا کہ میں 70 دن سے دھرنے والوں کے پاس جاتا رہا دونوں کے رہنماؤں سے مہنگائی اور غربت کیخلاف ملکر جہاد کرنے کو کہا سب نے
جواب دیا یہ ہمارا کام نہیں ہم آپس میں لڑینگے میں نے بھی کہا لڑتے لڑتے ہو گئی گم ایک کی چونچ ایک کی دم۔
</body>
</COUNTER_document>
```

Fig. 3 Example of a ND document pair

Table 5 The paraphrase typology showing 6 classes and 14 types

Class	Type
Morphology-based changes	Inflectional changes
	Derivational changes
Lexicon-based changes	Spelling and format changes
	Same-polarity substitutions
	Synthetic–analytic substitutions
	Opposite-polarity substitutions
Syntax-based changes	Diathesis alterations
	Negation switching
Discourse-based changes	Punctuation and format changes
	Direct/Indirect style alterations
Semantics-based changes	Semantic changes
Miscellaneous changes	Change of order
	Addition/deletion of information
	English to Urdu translation changes

S: ایبولا وائرس سے [متاثرہ شخص] ہسپتال میں علاج کے دوران جاں بحق ہوگیا

D: فیصل آباد میں ابیولا کا [مشتبہ مریض] دم توڑ گیا

Synthetic/analytic substitutions involves addition/deletion of single to multiple lexical terms that do not affect the meaning of the word. The example that follows shows specifier deletions in the derived text.

S: اس فلم میں شاہد کپور کے والد [معمر اداکار پنکج کپور] اور [سوتیلی] بہن ثناء نے بھی کام کیا ہے

D: اس فلم میں میرے والد اور بہن ثناء نے بھی کام کیا ہے

Opposite-polarity substitutions contains change in the word or phrase with its antonym. However, to preserve the meaning, either double polarity change or inverse argument is needed. In the first example text from our corpus, [lose] is replaced with [success] and another substitution [win] is added in the derived text.

S: نیوزی لینڈ نے پاکستان کو دلچسپ مقابلے کے بعد 3-2 گول سے [ہرا] دیا

D: نیوزی لینڈ نے میچ 3-2سے [جیت] کر ٹورنامنٹ میں پہلی [کامیابی] حاصل کر لی

The second example again shows an antonym substitution, but to preserve the meaning, the order of the subject (country name i.e. New Zealand) is shuffled.

S: کیوی ٹیم 3-2 گولز سے [فتح یاب]

D: پاکستان کو نیوزی لینڈ سے [شکست]

Diathesis alternations are changes that occur when a participating verb can be used in its various diathesis frames.

S: امریکی یرغمال کے والدین کی داعش سے رحم دلی کی اپیل

D: رحم کریں اور ہمارے بیٹے کو چھوڑ دیں، امریکی والدین کی داعش سے اپیل

Syntax-based changes

Negation switching in a text occurs when swapping a 'negation term' occurrence. The below example depicts one such occurrence in our corpus.

S: ویسٹ انڈیز نے سنیل نارائن کو بھارت کے خلاف [نہ] کھلانے کا فیصلہ کر لیا

D: مایہ ناز آف سپنر سنیل نارائن دورہ بھارت کے دوران ٹیم کی نمائندگی [نہیں] کریں گے

Discourse-based changes

Direct/indirect style alternations changes employ active to passive style changing and vice versa. In the example below, the statement is expressed in direct and indirect style.

S: رابن ولیمز کی پراسرار موت سے متعلق تحقیقات شروع ہوگئی ہیں جس میں خودکشی کے پہلوکوبھی مد نظر رکھا جائے گا

D: ان کی موت بظاہر خود کشی کا نتیجہ معلوم ہوتی ہے لیکن ابھی تحقیقات جاری ہے

Punctuation and format changes often include changes that appear due to placement of punctuation marks or change in format of text. Normally these changes do not effect the lexical units. The first part of the following example shows punctuation mark (,) added in the derived text. Further, the sentence delimiter (.) is replaced with a comma to add a new clause in the derived sentence.

S: کشمیر اور تلنگانہ بھارت کا حصہ نہیں ۔بھارتی رکن پارلیمنٹ

D: کشمیر، تلنگانہ بھارت کا حصہ نہیں ،سرحد بدلی جائے، رکن لوک سبھا

Semantics-based changes

Semantic changes consist of rephrasing lexical units in the derived text by adding new words or word patterns but of the similar contents. The COUNTER corpus has plentiful examples of such cases. The one case shown in the example below highlights the words [Iraqi militants] replaced with [ISIS] and [approved] rephrased as [declared] in the derived sentence.

S: اوباما نے [عراقی عسکریت پسندوں] کے خلاف فضائی حملوں کی [منظوری دے دی]

D: امریکا نے [داعش] کے خلاف فضائی حملوں کا [اعلان کر دیا]

Miscellaneous changes

Add/delete information often implies compression or expansion of the source text. The lexical and functional units are added to or deleted from the source text to recompose it.

S: تیراہ میں امن لشکر پرخودکش حملہ ، 5 افراد جاں بحق ، 7زخمی ہوگئے

D: وادی تیراہ میں امن لشکر پر خود کش حملہ،7 افراد جاں بحق

Change of order includes any type of change of order from the word level to the sentence level. In the example, a word [noun: Nawaz Sharif] and a phrase [verb: do not care] changed their position in the derived text.

S: وزیراعظم [نواز شریف] نے کہا ہے کہ عوام تبدیلی اور انقلاب والوں کی [پرواہ نہ کریں] وہ ملک کا کچھ نہیں بگاڑ سکتے

D: عوام [پروا نہ کریں]، تبدیلی اور انقلاب والے ملکی ترقی نہیں روک سکتے: [نوازشریف]

English to Urdu translation changes consists of changes that occur when an English word written using Urdu script can be rewritten by translating it into Urdu language word. Our corpus is rich with such examples, some of which are added below.

S: آسٹریلیا نے 49 [گولڈ]،42 [سلور] اور 46 [برانز] [میڈلز] کے ساتھ دوسری پوزیشن حاصل کی

D: آسٹریلیا 49 [سونے]،42 [چاندی]،46 [کانسی] سمیت137 [تمغوں] کیساتھ دوسرے نمبر پر رہا

S: مچل جونسن اور مچل اسٹارک نے [نصف سنچریاں] جڑیں

D: مچل جونسن اور اسٹارک [ففٹیز] بنانے میں سرخرو رہے

To show which paraphrase mechanisms are most frequently used (by journalists) to constitute the newspaper stories, we took a subset of first 50 documents from the corpus[11] and calculated the paraphrase type frequencies for each of the 14 types (see Table 5).

Table 6 shows that 'Same-polarity substitutions' emerges as the most frequent (0.312) paraphrase type present in the subset of the corpus, followed by 'Semantic changes' (0.200) and 'Addition/deletion of information' (0.168) which also contribute to a major extent.[12] This was expected as the corpus text (of derived documents) is reformulated by journalists and in the process they have opted for the most simple paraphrase mechanism i.e. substituting words with others of more or less the same meaning. Closely related to this, and in general, are the semantic changes which involve replacing lexical units. Moreover, journalistic writing involves an editor's own observations which naturally results in the addition/deletion of information. We conclude that same polarity substitutions, semantic changes and addition/deletion of information are the most favourite mechanism used by journalists as they are relatively easy to apply and preferable by individuals when reusing text.

[11] This sub-corpus is also available to download with the main corpus.

[12] We expect that the paraphrase types occurring most frequently in the subset of the corpus will be reflected with similar proportions in the whole corpus since this subset is a substantial representative sample of the whole corpus.

Table 6 Paraphrase type frequencies occurring within the 50 document subset corpus. Bold values are the sum of the corresponding types within the main classes

	$\text{Frequencies}_{\text{abc}}$	$\text{Frequencies}_{\text{rel}}$
Morphology-based changes	**17**	**0.030**
Inflectional changes	8	0.014
Derivational changes	9	0.016
Lexicon-based changes	**212**	**0.379**
Spelling and format changes	6	0.011
Same-polarity substitutions	174	0.312
Synthetic/analytic substitutions	24	0.043
Opposite-polarity substitutions	8	0.014
Syntax-based changes	**18**	**0.032**
Diathesis alternations	11	0.019
Negation switching	7	0.012
Discourse-based changes	**47**	**0.084**
Punctuation and format changes	18	0.032
Direct/indirect style alternations	29	0.052
Semantics-based changes	**112**	**0.200**
Semantic changes	112	0.200
Miscellaneous changes	**152**	**0.272**
Change of order	32	0.057
Addition/deletion of information	94	0.168
English to Urdu translation changes	26	0.046

4 Text reuse similarity estimation methods

In the past, different text similarity estimation methods have been proposed based on syntactic or semantic features (Clough et al. 2002a; Mihalcea et al. 2006; Daniel et al. 2012). This section describes a few popular text similarity estimation methods that we choose to apply on the corpus in order to show how it can be used in the evaluation of state-of-the-art methods for text reuse detection. These methods generate similarity scores, by comparing each source-derived document pair, based on features which can be derived from the given texts. The higher the score the more similar the contents of the two documents (Wise 1992; Brin et al. 1995; Gitchell and Tran 1999; Lyon et al. 2001).

We choose to apply a range of methods, based on three different characteristics i.e. content, structure or style of the given text (Daniel et al. 2012). For content based methods, we chose Word *n*-grams overlap (see Sect. 4.1), Vector Space Model (VSM; see Sect. 4.3), Longest Common Subsequence (LCS; see Sect. 4.4) and Greedy String-Tiling (GST; see Sect. 4.5). For structural similarity we opted for Stop-words based *n*-grams overlap (see Sect. 4.2) and for stylistic features extraction, we applied sentence/token ratio (see Sect. 4.6).

4.1 Word *n*-grams overlap

One of the popular methods, word *n*-grams overlap, computes the resemblance of a document pair by simply calculating the common *n*-grams and dividing it by the length of one or both documents. The method has already proven to provide good results for detecting plagiarism (on mono-lingual English corpora) (Lane et al. 2006; Barrón-Cedeño et al. 2009; Clough and Stevenson 2011), detection of near duplicates (Shivakumar and Garcia-Molina 1995) and measuring text reuse (Clough et al. 2002a; Chiu et al. 2010). In our experiments, we used the Containment similarity co-efficient measure[13] (Broder 1997) to compute similarity between document pairs (see Eq. 1).

$$C_n(X,Y) = \frac{|S(X,n) \bigcap S(Y,n)|}{|S(X,n)|} \tag{1}$$

In the above equation, $S(X, n)$ and $S(Y, n)$ represents the number of unique word *n*-grams (tokens) of size n in documents X and Y, respectively. The method computes how much content (word *n*-grams) of the document X is shared by Y. Further, it generates a similarity score between 0 and 1. A similarity score of 0 means that the two documents have no common word *n*-grams whereas 1 means that all the word *n*-grams are common. The scores are reported for sets of *n*-grams of length [1–5], to indicate the degree of similarity between source-derived document pairs for various lengths of n. Moreover, we experiment both with and without text preprocessing. During text preprocessing, all punctuation marks, illegal characters[14] (if any) and stop-words were removed.

4.2 Stop-words based *n*-grams overlap

Another method, however grounded on the syntactic similarity, between source and derived document pair, is stop-words based *n*-grams overlap (Stamatatos 2011). The method works with a list of stop-words (also known as very frequent words) and the fact that these words are often preserved while modifying texts where the editor commonly replaces or rearranges content words (with synonyms). In our experiments, we first extracted all the stop-words[15] from a source-derived document pair. Secondly, all the stop-words based *n*-grams of both documents were then compared using the same Eq. 1 i.e. Containment measure.

The similarity scores between source-derived document pairs are computed for sets of stop-words based *n*-grams of length [1–5].

[13] We also applied Jaccard, Dice and Overlap similarity coefficients but the results were low when compared to Containment similarity measure. Therefore, we only reported results with Containment measure in this study.

[14] The characters that are not part of the standard Urdu language character set.

[15] The stop-words list that we used is available with the corpus download.

4.3 Vector space model

Vector Space Model (VSM) or its variants (Salton et al. 1975), originally proposed for IR, have recently been used in the experiments on text reuse (Clough 2003; Bendersky and Croft 2009) and detecting document duplicates (Hoad and Zobel 2003; Runeson et al. 2007). Moreover, it was a popular choice for majority of the participating systems in the PAN Competitions (Sanchez-Perez et al. 2014).

In VSM, both source and derived documents are represented as term (word or phrase) vectors. The number of unique terms in each document corresponds to a dimension in the vector space. The similarity between both (source-derived document pair) vectors is measured by the cosine similarity measure (the angle between them), calculated as:

$$sim(d_{SOU}, d_{DER}) = \frac{\overrightarrow{d_{DER}} \cdot \overrightarrow{d_{SOU}}}{|\overrightarrow{d_{DER}}| \times |\overrightarrow{d_{SOU}}|} = \frac{\sum_{i=1}^{n} d_{DERi} \times d_{SOUi}}{\sqrt{\sum_{i=1}^{n} (d_{DERi})^2 \times \sum_{i=1}^{n} (d_{SOUi})^2}} \quad (2)$$

where $|\overrightarrow{d_{DER}}|$ and $|\overrightarrow{d_{SOU}}|$ represent the lengths of the derived and source document vectors respectively. Before computing the similarity, we applied the popular *tf.idf* (see Eq. 3) weighting scheme (Jurafsky et al. 2000) to weight individual terms in the source and derived documents.

$$tfidf_{i,d} = tf_{i,d} \cdot idf_i = \frac{n_{i,d}}{\sum_k n_{k,d}} \cdot log \frac{|D|}{|D_i|} \quad (3)$$

Using the VSM method, we also investigated the effect of stop-words removal.

4.4 Longest common subsequence

Longest Common Subsequence (LCS) is another similarity estimation method used in our experiments. In LCS, the degree of resemblance between a document pair is calculated by taking into account the total number of changes made when the text was rewritten. In the first step, both documents are represented as sequences of tokens (words or phrases). Given a piece of text (called sub-string), a subsequence is a contiguous stream of tokens even if some terms are removed from that sub-string. Let us assume, X and Y are two strings (texts) to be compared, then LCS is the longest subsequence *common* between them. For example, if X = "123456" and Y = "129456", then 456 is a subsequence and 12,456 is the longest common subsequence.

A normalised similarity score (LCS_{norm}) (see Eq. 4), is computed by dividing the length of LCS ($|LCS\ (X, Y)|$) with the length of shorter string.

$$LCS_{norm}(X, Y) = \frac{|LCS(X, Y)|}{min(|X|, |Y|)} \quad (4)$$

Moreover, the LCS algorithm is order preserving. The length of LCS_{norm} shows the modifications in the text caused by lexical substitutions, word re-ordering and other

text altering operations. Again, similar to other methods, the effect of pre-processing was explored for this method as well.

4.5 Greedy string-tiling

The Greedy String-Tiling (GST) algorithm is based on sub-string matching and was proposed for identifying biological sub-sequences and computing similarity between free texts (Wise 1992). GST can detect *block move* (caused by transposition of tokens), which are missed by LCS (Longest Common Subsequence, see Sect. 4.4) method. GST method tries to find a 1:1 match of tokens between two texts, such that one sequence of tokens is covered with maximum length (called tiles) sub-strings from the other. However, to avoid specious matches of very small lengths, a minimum Match Length (mML) value is used.

In our experiments, we were interested to know how much derived text (words) is overlapped with source text. So, given source a document X, a derived document Y and a set of matching tiles of a given length between the two documents, the similarity, gst-sim(X,Y), is obtained using Eq. 5

$$gst - sim(X, Y) = \frac{\sum_{i \in tiles} length_i}{|Y|} \tag{5}$$

The GST experiments are conducted on the corpus, both with and without text preprocessing.

4.6 Sentence/token ratio

Based on the fact that rewritten texts are, to a certain degree, similar in terms of stylistic features, we also experiment with statistical properties of texts to estimate similarity among them. We applied two simple methods, sentence ratio and token ratio (Yule 1939) to compute average number of sentences and tokens respectively. As the corpus contains news stories, documents are mostly structured as single paragraph essays. Therefore, we computed the number of sentences per document and the average number of tokens per sentence.[16] Further, for sentence ratio we computed the ratio of sentences whereas for token ratio we compared the average token length between the reused text and the source text.

5 Experimental set-up

5.1 Dataset

For the set of experiments carried out in this study, the entire COUNTER Corpus is used (see Sect. 3). There are total 600 document pairs in the corpus (WD = 135, PD = 288 and ND = 177).

[16] For sentence boundary detection, we used potential sentence termination markers such as '؟', '_' and '!'.

5.2 Evaluation methodology

In the experiments performed, to distinguish between multiple levels of Urdu text reuse at document level, the problem is tackled as a supervised classification task. We used both binary and ternary classifications of the task. In the former, the target is to differentiate between two classes [(i.e. Derived (D) and Non Derived (ND)] while in the latter case, the target is to differentiate between three classes [(i.e. Wholly Derived (WD), Partially Derived (PD) and Non Derived (ND)]. For the binary classification task, the documents categorised as Wholly Derived and Partially Derived are coupled to make the "Derived" class while the documents categorised as Non Derived are part of the "Non Derived" class. Due to the adequate number of examples (600) present in the corpus, and to better evaluate the performance of the similarity estimation methods used, we applied 10-fold cross-validation. The WEKA[17] (Hall et al. 2009; Witten et al. 2011) implementations of the Bayes theorem based Naïve Bayes classifier, with its default parameter settings, is used for the classification task. Naïve Bayes is appropriate for these kind of experiments as it can handle the numeric features generated by the similarity estimation methods applied on the corpus (see Sect. 4). The similarity scores for each source-derived document pair are used as features for the classifier. Weighted average F_1 results are computed and reported for both binary and ternary classification tasks.

6 Results and analysis

Table 7 presents Naïve Bayes classifier reported F_1 results on the COUNTER corpus for the binary and ternary classifications tasks using Word *n*-grams overlap, Vector Space Model, Longest Common Subsequence, Greedy String Tiling, Stop-words based *n*-grams overlap and Sentence/Token ratio methods. *Uni-gram* means that the results are obtained using word 1-g as a single feature for the classifications task. Similarly, *Bi-gram*, *Tri-gram*, *Four-gram* and *Five-gram* means that the results are obtained using word 2-, 3-, 4- and 5-g respectively as a single feature. *Combined* means that results are obtained by similarity scores of word unigram, bigrams, trigrams, fourgrams and fivegrams as a set of features (5 features) for the classification task. *SWR* after each method means that the similarity score is computed for the method after removing stop-words. Likewise, *Stop-words Uni-gram* means that the results are reported using stop-words based 1-g, *Stop-words Bi-gram* means stop-words based 2-g, *Stop-words Tri-gram* means stop-words based 3-g, *Stop-words Four-gram* means stop-words based 4-g, *Stop-words Five-gram* means stop-words based 5-g and *Stop-words Combined* means that similarity scores of stop-words based *n*-grams of length 1–5 are used as a set of features (5 features) for the classification tasks. *VSM* means results obtained using Vector Space Model, *LCS* means results obtained using Longest Common Subsequence and *GST* means results obtained using Greedy String Tiling methods. For GST, *mML1* to *mML10*

[17] http://www.cs.waikato.ac.nz/ml/weka/—Last visited: 16-06-2016.

means results with minimum match lengths of tiles from 1 to 10, respectively. Again, *SWR* means results computed after stop-words removal. In the last part of the table, "All features combined" means that the results are reported by combining features of all the methods used in this study. The best results obtained overall are presented as bold letters whereas best resulted obtained category-wise are Italics in the table.

From Table 7, as expected, overall, results are lower for the ternary classification task (best $F_1 = 0.73$) compared to the binary classification task (best $F_1 = 0.81$). For both classifications, the same pattern of differences in the results can be seen across all the methods used in the study. This demonstrates that, in text reuse problem, it is easier to distinguish between two levels of reuse than three. For binary classification problem, best F_1 score is obtained using GST *mML1* ($F_1 = 0.81$), nearly matching the result with Word Uni-gram overlap ($F_1 = 0.80$). It can also be noticed that both of these results didn't improve after removal of stop-words. For ternary classification task, the highest F_1 score of 0.73 is obtained for both GST *mML1 + SWR* and Word *n*-grams overlap *Uni-gram* and we can see a small effect of stop-words removal on both methods (improvement of 0.01 in GST while decline of 0.01 in Word *n*-grams overlap). These results show that GST and Word *n*-grams overlap are the most appropriate methods for Urdu text reuse detection on the COUNTER Corpus. It also highlights that, in text reuse detection, a smaller length of blocks (tokens; $n = 1$ or $mML = 1$) is more effective especially when the text has been heavily modified or rephrased (as majority of examples in our corpus are rewritten).

GST outperformed all other methods for binary classification task and its performance for ternary classification task is same as Uni-gram method. Word *n*-grams overlap was the second best. This shows that GST is able to deal better with paraphrased text, identifying individually longest sub-strings in the rearrangements of tokens (lexical units) of the rephrased text. For both classification tasks, decline in performance was observed as the length of tokens/chunks increases ($n > 1$ or $mML > 1$). The possible reason for this is that the derived text is rewritten in PD and ND documents, which makes it difficult to find matching chunks of longer lengths ($n = 2$–5 or $mML = 2$–10). Consequently, that makes it difficult to discriminate different levels of text reuse. Note that these observations are consistent with the METER study (Clough et al. 2002b), which also showed that best results are obtained using word unigrams and an mML of 1, and further an increase in the length of *n* or mML effects performance.

As expected, performance using the LCS method ($F_1 = 0.77$) is lower compared to the GST because it is not able to deal with *block move* problem. Furthermore, the removal of stop-words did not show any improvement in LCS results for the binary classification task, however, there is a slight improvement of 0.01 for ternary classification task.

The results using the VSM method, for both binary ($F_1 = 0.66$) and ternary classifications ($F_1 = 0.54$) are lowest compared to all the other content based methods (Word *n*-grams overlap, LCS, GST). This is likely to happen because VSM aims to identify topical similarity among document pairs for Information Retrieval

Table 7 Weighted average F_1 results for binary and ternary classification tasks using different text reuse detection methods

	Binary	Ternary
Content based measures		
Word n-grams overlap		
Uni-gram	**0.80**	**0.73**
Uni-gram + SWR	**0.80**	0.72
Bi-gram	0.66	0.64
Bi-gram + SWR	0.70	0.68
Tri-gram	0.57	0.56
Tri-gram + SWR	0.60	0.64
Four-gram	0.52	0.52
Four-gram + SWR	0.55	0.57
Five-gram	0.49	0.52
Five-gram + SWR	0.50	0.53
Combined	0.56	0.54
Combined + SWR	0.57	0.57
Vector space model		
VSM	0.66	0.54
VSM + SWR	0.64	0.53
Longest common subsequence		
LCS	0.77	0.70
LCS + SWR	0.77	0.71
Greedy string tiling		
mML1	**0.81**	0.72
mML1 + SWR	**0.81**	**0.73**
mML2	0.77	0.71
mML2 + SWR	0.74	0.67
mML3	0.70	0.65
mML3 + SWR	0.63	0.60
mML4	0.63	0.60
mML4 + SWR	0.60	0.57
mML5	0.58	0.59
mML5 + SWR	0.55	0.53
mML6	0.56	0.53
mML6 + SWR	0.53	0.51
mML7	0.54	0.52
mML7 + SWR	0.48	0.50
mML8	0.51	0.50
mML8 + SWR	0.46	0.50
mML9	0.47	0.49
mML9 + SWR	0.44	0.47

Table 7 continued

	Binary	Ternary
mML10	0.46	0.49
mML10 + SWR	0.43	0.45
Structure based measures		
Stop-words based n-grams overlap		
Stop-words uni-gram	0.58	0.40
Stop-words bi-gram	**0.63**	0.42
Stop-words tri-gram	0.47	0.44
Stop-words Four-gram	0.41	**0.46**
Stop-words five-gram	0.35	0.34
Stop-words Combined	0.40	0.37
Style based measures		
Sentence/token ratio		
Sentence Ratio	**0.58**	0.32
Token Ratio	**0.68**	0.45
Combination of features		
All features combined	0.70	0.68

(IR) task, whereas in text reuse detection task, aim is to identify overlap between document pairs.

The performance of the structure-based and stylistic-based methods i.e. Stop-words based n-grams overlap ($F_1 = 0.63$ (Bi-gram) for binary classification; $F_1 = 0.46$ (Four-gram) for ternary classification) and Sentence/Token ratio ($F_1 = 0.58$ and 0.68 for binary classification), is low overall and they demonstrated poor results in both classification tasks. This shows that structure-based as well as stylistic-based methods are comparatively not suitable for the Urdu text reuse detection task.

The results for the combination of features, using Word n-gram overlap feature "Combined" and Stop-words based n-gram overlap feature "Stop-words Combined", does not improve performance. For both classification tasks, from all the methods used in this study, Word n-grams overlap performed consistency better for $n > 1$ and above, after the removal of stop-words from the text. This improvement is statistically significant as tested with Wilcoxon signed-rank test ($p < 0.05$) (Wilcoxon et al. 1970). LCS also demonstrated slightly better results, for ternary classification task, on pre-processed text with stop-words removed. However, results using VSM and GST methods does not show improvement after the removal of stop-words. This highlights the fact that this pre-processing is useful in some cases for text reuse detection on the Urdu text.

We also conducted experiments by combining all the features from all the methods (*All features combined* method) used in this study i.e. similarity scores reported by 12 features of *Word n-grams overlap*, 20 features of *GST*, 6 features of *Stop-words based n-gram overlap* and 2 features of each *VSM*, *LCS* and *Sentence/*

Table 8 Confusion matrix for ternary classification using GST mML1

	WD	PD	ND
WD	91	43	1
PD	16	232	40
ND	2	68	107

Token Ratio methods were combined and best feature selection method applied on the combination of all features. We applied the attribute selected classifier from Weka (again, the highest results were reported by Naïve Bayes' classifier). However, the *All features combined* method does not improve performance.

Table 8 shows the confusion matrix for the GST "mML1" method (it produced best results for both classification problems, see Table 7). The columns and rows of the matrix represents the instances in the predicted and actual classes respectively.

Among all the three classes shown in the confusion matrix, it can be noted that it is easier to discriminate between WD and ND, however, difficult in the cases of WD–PD and PD–ND pairs. Furthermore, many WD instances are misclassified as PD (43) and similarly ND ones are also misclassified as PD (68), highlighting PD as the most problematic class for the classification problem. As a consequence, for ternary classification, the overall performance decreases.

7 Conclusion

Text reuse detection has attracted the attention of researchers for more than a decade now and it has gained increasing attention recently. For any language, the lack of large scale standardized evaluation resources with real examples of text reuse is a major problem in the analyses and development of text reuse detection systems. This paper presented our novel contribution in terms of the development of the first mono-lingual text reuse corpus for the Urdu language. The new corpus is modelled on the original English METER corpus and contains source and derived documents extracted from the news domain. The source documents contain news articles released by the news agencies whereas the derived documents are the news stories published in newspapers rewritten by journalists using the news agencies text as source. The corpus has been manually annotated by three annotators at document level with three classes of rewrite i.e. Wholly Derived, Partially Derived and Non Derived, and we have made it freely available online. A detailed set of twenty-four similarity estimation methods (content, structure, and style based measures) were used to conduct experiments on the corpus to show how such a resource can be useful in the development and evaluation of mono-lingual text reuse detection systems. Results showed that GST with mML1 feature is the most effective in text reuse detection on our corpus.

In the future, we plan to use character *n*-grams which is capable of capturing both stylistic and content information based on the selected value of *n*. Furthermore, the

corpus will be evaluated on other state-of-the-art semantic similarity estimation methods, after customisation, if necessary, for the Urdu language.

Acknowledgments This work has been supported by the COMSATS Institute of Information Technology, Pakistan and Lancaster University, UK under the Split-Site Ph.D. programme.

References

Anwar, W., Wang, X., & Wang, X.-L. (2006). A survey of automatic Urdu language processing. In *2006 International conference on machine learning and cybernetics, IEEE* (pp. 4489–4494). IEEE.

Aplin, T. (2010). Reflections on measuring text re-use from a copyright law perspective. *Copyright and Piracy: An Interdisciplinary Critique, 13*, 260–268.

Baker, P., & McEnery, A. (1999). Needs of language-engineering communities; corpus building and translation resources. In *Technical report, MILLE working paper 7*, Lancaster University.

Barrón-Cedeño, A., Rosso, P., & Benedi, J.M. (2009). Reducing the plagiarism detection search space on the basis of the Kullback–Leibler distance. In *Proceedings of 10th international conference on computational linguistics and intelligent text processing* (pp. 523–534). Springer.

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

Becker, D., & Riaz, K. (2002). A study in urdu corpus construction. In *Proceedings of the 3rd workshop on Asian language resources and international standardization-volume 12, Association for Computational Linguistics* (pp. 1–5). Association for Computational Linguistics.

Bell, A. (1991). *The language of news media*. Oxford: Blackwell.

Bendersky, M., & Croft, W. B. (2009). Finding text reuse on the web. *Proceedings of the second ACM international conference on web search and data mining, ACM* (pp. 262–271). ACM.

Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. In *Proceedings of the 1995 ACM SIGMOD international conference on management of data, ACM* (pp. 398–409).

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and complexity of sequences 1997. Proceedings, IEEE* (pp. 21–29). IEEE.

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

Chiu, S., Uysal, I. & Croft, W. B. (2010). Evaluating text reuse discovery on the web. In *Proceeding of the third symposium on information interaction in context, ACM* (pp. 299–304). ACM.

Chowdhury, A., Frieder, O., Grossman, D., & McCabe, M. C. (2002). Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, *20*(2), 171–191.

Clough, P. (2003). Measuring text reuse. Ph.D. Dissertation, University of Sheffield, UK.

Clough, P., Gaizauskas, R., Piao, S. & Wilks, Y. (2002a). Measuring text reuse. In *Proceedings of the 40th annual meeting of the association of computational linguistics* (pp. 152–159).

Clough, P., Gaizauskas, R., Piao, S.S. & Wilks, Y. (2002b). Meter: Measuring text reuse. In *Proceedings of the 40th annual meeting on association for computational linguistics, association for computational linguistics* (pp. 152–159). Association for Computational Linguistics.

Clough, P., & 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.

Cohen, J., et al. (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.

Daniel, B., Zesch, T., & Gurevych, I. (2012). Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING* (Vol. 1, pp. 167–184).

Fries, U. (1987). Summaries in newspapers: A textlinguistic investigation. *The Structure of Texts, 3*, 47–63.

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., & Piao, S. (2001). The METER corpus: A corpus for analysing journalistic text reuse. In *Proceedings of the conference on corpus linguistics* (pp. 214–223).

Gitchell, D., & Tran, N. (1999). Sim: A utility for detecting similarity in computer programs. *ACM SIGCSE Bulletin*, *31*(1), 266–270.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

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

Hoad, T. C., & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology*, *54*, 203–215.

Jing, H., & McKeown, K. R. (1999). The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 129–136). ACM.

Jurafsky, D., Martin, J. H., Kehler, A., Linden, K. V., & Ward, N. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 2). New Jersey: Prentice Hall.

Lane, P. C. R., Lyon, C., & Malcolm, J. A. (2006). Demonstration of the ferret plagiarism detector. In *Proceedings of the 2nd international plagiarism conference*.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, *5*, 361–397.

Lyon, C., Malcolm, J., & Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 118–125).

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism—A survey. *Journal of Universal Computer Science*, *12*(8), 1050–1084.

McEnery, T., Baker, P., & Burnard, L. (2000). Corpus resources and minority language engineering. In *Proceedings of the Second International Conference on Language Resources and Evaluation, (LREC), 31 May–2 June, 2000, Athens, Greece*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2000/pdf/187.pdf.

McEnery, T., Hardie, A., & Piao, S. (2010). A corpus-based approach to text reuse in the newsbooks of the commonwealth. In B. M. Dooley (Ed.), *The dissemination of news and the emergence of contemporaneity in early modern Europe* (pp. 251–286). Farnham: Ashgate.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In A. Cohn (Ed.), *AAAI* (Vol. 6, pp. 775–780). Boston, MA: AAAI Press.

Mittelbach, A., Lehmann, L., Rensing, C., & Steinmetz, R. (2010). Automatic detection of local reuse. In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Sustaining TEL: From innovation to learning and practice* (pp. 229–244). Berlin: Springer.

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

Piao, S.S., Rayson, P., Archer, D., Wilson, A., & McEnery, T. (2003). Extracting multiword expressions with a semantic tagger, In *Proceedings of the ACL 2003 workshop on multiword expressions: analysis, acquisition and treatment-Volume 18, Association for Computational Linguistics* (pp. 49–56). Association for Computational Linguistics.

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

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd international competition on plagiarism detection. In *Notebook papers of CLEF 11 labs and workshops*.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., et al. (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., Efstathios, S., & 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., & Stein, B. (2014). Overview of the 6th international competition on plagiarism detection. In *CLEF (Online Working Notes/Labs/Workshop)*.

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

Runeson, P., Alexandersson, M., & Nyholm, O. (2007). Detection of duplicate defect reports using natural language processing. In *Proceedings of the 29th international conference on software engineering, IEEE Computer Society* (pp. 499–510). IEEE Computer Society.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Sanchez-Perez, M. A., Sidorov, G.,&, Gelbukh, A. (2014). A Winning approach to text alignment for text reuse detection at PAN 2014. In *CLEF (Working Notes)* (pp. 1004–1011).

Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gomez, M., Villasenor-Pineda, L., & Rosso, P. (2013). Determining and characterizing the reused text for plagiarism detection. *Expert Systems with Applications*, *40*(5), 1804–1813.

Seo, J., & Croft, W. B. (2008). Local text reuse detection. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 571–578). ACM.

Shivakumar, N., & Garcia-Molina, H. (1995). SCAM: A copy detection mechanism for digital documents. In *Proceedings of the 2nd annual conference on the theory and practice of digital libraries*. Texas.

Sousa-Silva, R. (2014). Detecting translingual plagiarism and the backlash against translation plagiarists. *Language and Law/Linguagem e Direito*, *1*(1), 70–94.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, *62*(12), 2512–2527.

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., & 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)* (pp. 1–77).

Thenmozhi, D., & Aravindan, C. (2015). Paraphrase identification by using clause-based similarity features and machine translation metrics. *The Computer Journal*. doi:10.1093/comjnl/bxv083.

Tsatsaronis, G., Varlamis, I., & Vazirgiannis, M. (2010). Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research*, *37*(1), 1–40.

Vila, M., Martí, M. A., Rodríguez, H., et al. (2014). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, *4*(01), 205.

Wilcoxon, F., Katti, S., & Wilcox, R. A. (1970). 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. (1992). Detection of similarities in student programs: Yap'ing may be preferable to plague'ing. *ACM SIGCSE Bulletin*, *24*(1), 268–271.

Witten, I. H., Hall, M. A., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, *30*(3/4), 363–390.