Mono- and Cross-Lingual Paraphrased Text Reuse and Extrinsic Plagiarism Detection

Muhammad Sharjeel

School of Computing and Communications, Lancaster University

Supervisors:

Dr. Paul Rayson
Lancaster University,
Lancaster, United Kingdom
p.rayson@lancaster.ac.uk

Dr. Rao Muhammad Adeel Nawab
COMSATS University Islamabad,
Lahore Campus, Pakistan
adeelnawab@cuilahore.edu.pk

A dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

June 23, 2020

This thesis is dedicated to my mother, and my late father.

Acknowledgements

In the Name of Allah, the Most Gracious, the Most Merciful

First and foremost, I thank the Almighty Allah (SWT), the ultimate source of all knowledge and wisdom in this world, for His countless blessings on me.

Regarding my dissertation, I would like to express my sincere gratitude towards my thesis supervisors, Dr. Paul Rayson and Dr. Rao Muhammad Adeel Nawab. And I wish to "reuse" this sentence in so many ways, to show how grateful I am for their guidance, continuous motivation, and outstanding support that formed an endless "corpus" of wisdom that will be with me, always!

I greatly admire Dr. Paul Rayson for being a kind, accessible, and an amiable supervisor. I am indebted to Dr. Rao Muhammad Adeel Nawab for mentoring my research for the past several years and helping me to develop a strong background in Natural Language Processing and Machine Learning. A thanks also goes to all the anonymous reviewers for their invaluable feedback that has led to significant improvements in my PhD study.

A heartfelt thanks goes to my parents! Words cannot express my feelings, especially towards my mother. I am obliged to her for the countless prayers and efforts for me. After the death of my father, she has played a vital role in my upbringing and the education I received in my early years and finally, achieving this milestone. I extend my thanks to my sister and two brothers for their endless support and encouragement. A special mention for my little niece Yashfa and nephew Munzir, though both of them provided enough distraction in this journey, it could never have been such a joyful ride without them. Finally, I wish to thank my loving and

supportive wife, Dania, who has just come into my life and to our first child which we are expecting to arrive later in the year.

I am fortunate to have the following people around me as colleagues and friends, Mr. Hafiz Rizwan Iqbal, Mr. Jawad Shafi Mian, Mr. Touseef Tahir, Mr. Shahbaz Akhtar Abid, Mr. Abdul Wahab, Mr. Muhammad Anas Masood, Mr. Saqib Mehboob, Mr. Adnan Muzaffar, and Mr. Farrukh Naveed. The help, advice, and leisure they provided to lift my spirits whenever I needed, gave me strength through the difficult times.

Finally, I am thankful to COMSATS University Islamabad, Lahore Campus, Pakistan, for funding this work under the Split-Site PhD program and Lancaster University, United Kingdom, for providing an excellent research environment.

Declaration

I hereby declare that this dissertation or any part thereof has not previously been presented, unless otherwise indicated, in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose. Save for any express acknowledgements and references cited in the work, I confirm that the contents of the work is the result of my own efforts and of no other person.

The right of Muhammad Sharjeel is to be identified as author of this work.

Muhammad Sharjeel

Abstract

Text reuse is the act of borrowing text (either verbatim or paraphrased) from an earlier written text. It could occur within the same language (mono-lingual) or across languages (cross-lingual) where the reused text is in a different language than the original text. Text reuse and its related problem, plagiarism (the unacknowledged reuse of text), are becoming serious issues in many fields and research shows that paraphrased and especially the cross-lingual cases of reuse are much harder to detect. Moreover, the recent rise in readily available multi-lingual content on the Web and social media has increased the problem to an unprecedented scale.

To develop, compare, and evaluate automatic methods for mono- and cross-lingual text reuse and extrinsic (finding portion(s) of text that is reused from the original text) plagiarism detection, standard evaluation resources are of utmost importance. However, previous efforts on developing such resources have mostly focused on English and some other languages. On the other hand, the Urdu language, which is widely spoken and has a large digital footprint, lacks resources in terms of core language processing tools and corpora. With this consideration in mind, this PhD research focuses on developing standard evaluation corpora, methods, and supporting resources to automatically detect mono-lingual (Urdu) and cross-lingual (English-Urdu) cases of text reuse and extrinsic plagiarism.

This thesis contributes a mono-lingual (Urdu) text reuse corpus (COUNTER Corpus) that contains real cases of Urdu text reuse at document-level. Another contribution is the development of a mono-lingual (Urdu) extrinsic plagiarism corpus (UPPC Corpus) that contains simulated cases of Urdu paraphrase plagiarism. Evaluation results, by applying a wide range of state-of-the-art mono-lingual methods on both corpora, shows that it is easier to detect verbatim cases than paraphrased ones. Moreover, the performance of these methods decreases considerably on real cases of

reuse. A couple of supporting resources are also created to assist methods used in the cross-lingual (English-Urdu) text reuse detection. A large-scale multi-domain English-Urdu parallel corpus (EUPC-20) that contains parallel sentences is mined from the Web and several bi-lingual (English-Urdu) dictionaries are compiled using multiple approaches from different sources.

Another major contribution of this study is the development of a large benchmark cross-lingual (English-Urdu) text reuse corpus (TREU Corpus). It contains English to Urdu real cases of text reuse at the document-level. A diversified range of methods are applied on the TREU Corpus to evaluate its usefulness and to show how it can be utilised in the development of automatic methods for measuring cross-lingual (English-Urdu) text reuse. A new cross-lingual method is also proposed that uses bi-lingual word embeddings to estimate the degree of overlap amongst text documents by computing the maximum weighted cosine similarity between word pairs. The overall low evaluation results indicate that it is a challenging task to detect cross-lingual real cases of text reuse, especially when the language pairs have unrelated scripts, i.e., English-Urdu. However, an improvement in the result is observed using a combination of methods used in the experiments.

The research work undertaken in this PhD thesis contributes corpora, methods, and supporting resources for the mono- and cross-lingual text reuse and extrinsic plagiarism for a significantly under-resourced Urdu and English-Urdu language pair. It highlights that paraphrased and cross-lingual cross-script real cases of text reuse are harder to detect and are still an open issue. Moreover, it emphasises the need to develop standard evaluation and supporting resources for under-resourced languages to facilitate research in these languages. The resources that have been developed and methods proposed could serve as a framework for future research in other languages and language pairs.

خلاصہ

متن کا دوبارہ استعمال ایک پہلے سے تحریر شدہ متن سے نقل (لفظ با لفظ یا پیرا فریس) کرنے کا عمل ہے۔ یہ اسی زبان میں (یک لسانی) یا کسی دوسری زبان (بین السانی) میں ہو سکتا ہے جہاں دوبارہ استعمال شدہ متن اصل متن سے مختلف زبان میں ہو۔ متن کا دوبارہ استعمال اور اس سے متعلقہ مسئلہ، سرقہ (متن کا غیر تسلیم شدہ دوبارہ استعمال)، متعدد شعبوں میں سنگین مسائل بن رہے ہیں اور تحقیق سے پتہ چلتا ہے کہ پیرا فریسڈ اور بالخصوص بین السانی دوبارہ استعمال کی مثالوں کا پتہ لگانا بہت مشکل ہے۔ مزید یہ کہ، ویب اور سماجی میڈیا پر آسانی سے دستیاب کثیر السانی مواد میں حالیہ اضافے نے اس مسئلے کو غیر معمولی حد تک بڑھا دیا ہے۔

یک اور بین السانی متن کے دوبارہ استعمال اور خارجی (متن کے وہ حصے تلاش کرنا جو اصل متن سے دوبارہ استعمال ہوئے ہوں) سرقہ کے پتہ لگانے کے خودکار طریقوں کے بنانے، موازنہ کرنے، اور تشخیص کرنے کے لئے، تشخیص کے معیاری وسائل کا ہونا انتہائی اہم ہے۔ تاہم، اس طرح کے وسائل بنانے کے لئے کی گئی گذشتہ کوششوں میں زیادہ تر توجہ انگریزی اور کچھ دوسری زبانوں پر مرکوز رہی ہے۔ دوسری طرف، اردو زبان، جو کہ وسیع پیمانے پر بولی جاتی ہے اور اس کا ایک بہت بڑا ڈیجیٹل ذخیرہ موجود ہے، میں بنیادی لینگویج پروسیسنگ ٹولز اور کورپرا کا فقدان ہے۔ اس بات کو مدنظر رکھتے ہوئے، پی ایچ ڈی کی اس تحقیق کی توجہ معیاری تشخیصی کورپرا، طریقوں، اور معاون وسائل بنانے پر مرکوز ہے جن سے خود بخود یک لسانی (اردو) یا بین السانی (انگریزی اردو) متن کو دوبارہ استعمال کرنے اور خارجی سرقہ کی مثالوں کا پتہ لگایا جاسکے۔

یہ مقالہ ایک یک لسانی (اردو) متن کے دوبارہ استعمال کا کورپس (کاؤنٹر کورپس) پیش کرتا ہے جس میں دستاویز کی سطح پر اردو متن کے دوبارہ استعمال کی حقیقی مثالیں شامل ہیں۔ ایک اور پیش کش یک لسانی (اردو) خارجی سرقہ کورپس (یو پی پی سی کورپس) کا تیار کرنا ہے جس میں اردو پیرا فریس سرقہ کی مصنوعی مثالیں شامل ہیں۔ تشخیصی نتائج، دونوں کورپرا پر ایک وسیع رینج کے جدید ترین یک لسانی طریقوں کا استعمال کرتے ہوئے، ظاہر کرتے ہے کہ لفظ با لفظ مثالوں کی نشاندہی کرنا پیرا فریس سے کہیں زیادہ آسان ہے۔ مزید برآں، دوبارہ استعمال کی حقیقی مثالوں پر ان طریقوں کی کارکردگی کافی حد تک کم ہو جاتی ہے۔ دو معاون وسائل بھی تیار کیے گئے ہیں جو بین السانی (انگریزی اردو) متن کے دوبارہ استعمال کی نشاندہی کے طریقوں میں مددگار ہو سکتے ہیں۔ ایک بڑے پیمانے پر ملٹی ڈومین انگریزی اردو پیرالل کورپس (ای یو پی سی ۲۰) جس میں متوازی جملے شامل ہیں ویب سے اخذ کیا گیا ہے اور متعدد دو لسانی (انگریزی اردو) لغات کئی طریقوں کے ذریعے مختلف ذرائع سے مرتب کی گئی ہیں۔

اس مطالعے کا ایک اور اہم حصہ ایک بڑا معیاری بین السانی (انگریزی اردو) متن کے دوبارہ استعمال کا کورپس (ٹی آر ای یو کورپس) تیار کرنا ہے۔ اس میں دستاویز کی سطح پر انگریزی سے اردو متن کے دوبارہ استعمال کی اصل مثالیں شامل ہیں۔ ٹی آر ای یو کورپس پر ایک متنوع رینج کے طریقوں کا اطلاق کیا گیا ہے تاکہ اس کی افادیت کا اندازہ لگایا اور ظاہر کیا جا سکے کہ یہ بین السانی (انگریزی اردو) متن کے دوبارہ استعمال کو جانچنے کے خود کار طریقے بنانے میں کس طرح استعمال کیا جاسکتا ہے۔ ایک نیا بین السانی طریقہ بھی تجویز کیا گیا ہے جس میں دو لسانی ورڈ امبیڈنگز کا استعمال کرکے متن دستاویزات

کے درمیان اوورلیپ کی حد کا اندازہ لفظوں کے جوڑوں کے مابین زیادہ سے زیادہ ویٹڈ کوسائن مماثلت سے لگایا گیا ہے۔ مجموعی طور پر کم تشخیصی نتائج ظاہر کرتے ہیں کہ بین السانی متن کے دوبارہ استعمال کی اصل مثالوں کا پتہ لگانا ایک انتہائی مشکل کام ہے، خاص طور پر جب زبان کے جوڑے غیر متعلقہ رسم الخط کے ہوں، جیسے کہ، انگریزی اردو۔ تاہم ، تجربات میں استعمال کئے گئے طریقوں کے آپسی امتزاج سے نتیجہ میں بہتری دیکھی گئی ہے۔

اس پی ایچ ڈی مقالہ میں شروع کیا گیا تحقیقی کام انتہائی کم وسائل کی اردو اور انگریزی اردو زبان کی جوڑی کے لئے یک اور بین السانی متن کے دوبارہ استعمال اور خارجی سرقہ کے کورپرا، طریقے، اور معاون وسائل پیش کرتا ہے۔ یہ نمایاں کرتا ہے کہ پیرا فریسڈ اور بین السانی بین رسم الخط متن کے دوبارہ استعمال کی اصل مثالوں کا پتہ لگانا مشکل ہے اور اب بھی اس پر مزید تحقیق کی ضرورت ہے۔ مزید برآں، یہ کم وسائلی زبانوں کے لئے معیاری تشخیصی اور معاون وسائل تیار کرنے کی ضرورت پر زور دیتا ہے تا کہ ان زبانوں میں تحقیق کو فروغ دیا جائے۔ جن وسائل کو تیار کیا گیا ہے اور جو طریقے تجویز کئے گئے ہیں وہ دیگر زبانوں اور زبانوں کے جوڑوں میں مستقبل کی تحقیق میں ایک لائحہ عمل کے طور پر کام کر سکتے ہیں۔

Contents

List of Figures

List of Tables

"To steal ideas from one person is plagiarism; to steal from many is research."

Steven Wright

1 Introduction

Text reuse is the process in which pre-existing text is consciously reused to create a new text [Clough, 2010]. It occurs when information from one website is republished on a different website or when authors derive text for their novels from previously written work. Text reuse often implies different levels of rewriting as it starts from verbatim (or copy-paste), stretches to paraphrasing when the contents are rephrased using different text editing operations, to a case where the re-written text is produced completely independent of its source [Clough et al., 2002, Maurer et al., 2006]. The amount of text that is reused varies from small phrases, sentences, paragraphs, and even up to entire documents. Additionally, reuse is not just limited to text only but programming code, music, images, videos, and even ideas are often subject of reuse [Ganguly et al., 2018, Dittmar et al., 2012, Porter, 2009].

There are two possible scenarios for text reuse: (1) mono-lingual text reuse, and (2) cross-lingual text reuse. In mono-lingual, both the rewritten (also called 'derived', 'reused', or 'suspicious' text) and the source (also called 'original' text) texts share the same language, while in cross-lingual, the rewritten text is in a different language than its source.

In some cases, where a proper citation is provided, text reuse is considered acceptable. In journalism, for example, it is a desirable practice as information generated by the news agencies is edited (and in some cases translated then edited) by newspapers for publishing [Wilks, 2004]. It is also permissible in collaborative authoring, for instance, in Wikipedia, where it is considered fair to generate the contents of an article by reusing the corresponding article's text (even across languages).

Text plagiarism, on the other hand, represents unacknowledged text reuse in which no proper reference about the source is provided [Wood, 2004]. It is defined by the Institute of Electrical and Electronics Engineers (IEEE)[1] as "the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source". Similar to text reuse, text plagiarism can be verbatim, paraphrase, idea, or cross-lingual plagiarism when it crosses language boundaries [Martınez, 2009].

In spite of the fact that text plagiarism has long been considered to be a serious academic offence [Eaton, 2004, Schrimsher et al., 2011], it is not a phenomenon enclosed in a classroom anymore, but has diversified and, more recently, we see a sharp rise in cross-lingual plagiarism cases too [Pupovac et al., 2008, Butakov and Scherbinin, 2009, Osman et al., 2012]. Recently, two journalists of a Portuguese newspaper, a New York Times columnist, and a renowned Time magazine journalist admitted plagiarising their news articles [Sousa-Silva, 2015]. After the famous

[1] http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html

case of former German Defence Minister Guttenberg[2] (2011), two similar and more recent cases where a Romanian Prime Minister, Victor Ponta[3] (2012) and a German Education Minister, Annette Schavan[4] (2013) found guilty of plagiarised PhD dissertations. Furthermore, in media, songs, choreography, lyrics, and stories are reused without citing the corresponding source [Dittmar et al., 2012, Giguere, 2019].

With the increasing volume of reported cases of both mono- and cross-lingual text reuse and plagiarism, the transformation of the Web into a social and multi-lingual hub, expansion of Wikipedia in multiple languages with readily available electronic documents, and widely adopted use of Machine Translation (MT) systems [Logue, 2004, Gipp et al., 2014], the computational study and thorough analysis of text reuse and plagiarism are becoming hot research topics. Consequently, developing reliable systems for their detection has become an interesting intellectual problem and one whose solution promises practical benefits to both individuals and organisations, for example, in academia, where teachers can assess the originality of a student's assignment, in businesses where companies are interested to find breaches of ownership or wish to track the distribution of copyright digital content, detecting infringements of the news monitoring system, and Web search engines wishing to filter duplicate content prior to providing results to the users.

Although the detection of text reuse and plagiarism can be difficult for humans, the best practice is to manually identify the cases. However, it is not practical to keep track of every on-line resource manually. As a result, it is mandatory to have automatic methods that assist humans. The automatic text reuse and plagiarism detection take advantage of state-of-the-art Natural Language Processing (NLP) and Information Retrieval (IR) techniques to determine whether a text (either full or partial) has been reproduced by considering another as its source. No matter

[2] http://www.bbc.co.uk/news/magazine-12613617
[3] http://www.reuters.com/article/2014/12/16/us-romania-ponta-idUSKBN0JU1N520141216
[4] http://www.ithenticate.com/plagiarism-detection-blog/annette-schavan-surrenders-in-fight-to-save-her-phd

what, the final decision, with the help of supportive linguistic evidence, is by a human.

Automatic plagiarism detection is often divided into two subtasks [Stein et al., 2007]. Intrinsic plagiarism detection is the task of checking whether the whole text that contains plagiarism (suspicious text) is documented by a single author. Extrinsic[5] plagiarism detection, on the other hand, is the task of identifying portions of a derived text that are borrowed from the original text(s) (source text(s)). The extrinsic plagiarism detection task is further classified as mono- and cross-lingual, wherein the latter case, the source and derived texts are in different languages.

Text reuse and extrinsic plagiarism, whether mono- or cross-lingual, are comparatively hard to tackle as they can occur at the document, passage, and sentence levels [Martin, 1994]. Moreover, the rewritten text is often obfuscated with different paraphrasing mechanisms and may be borrowed from more than one text document (source text) [Maurer et al., 2006]. In the cross-lingual case, it becomes even more complex, when the source-derived texts are from different languages or belong to different language families (e.g., English-Arabic, English-Urdu, etc.) [Barrón-Cedeño et al., 2010].

To identify a potential case of text reuse or extrinsic plagiarism, the derived text should ideally be compared with all the possible sources. However, in large collections, this is computationally expensive and practically difficult to achieve. Therefore, a small set of sources are first shortlisted (often called candidate text documents[6]) and then a feature-based detailed analysis (also known as the pair-wise comparison) is carried out on each source-derived text document pair. The goal is to pinpoint the source text document(s) used to create the derived text document and further, to identify the portion of text reused from the source text document(s).

[5] In some literature, it is also referred to as external

[6] Mostly IR based approaches are used at this stage, where the derived text document is used as a query to retrieve all the sources from a large set of text documents

1.1 Thesis focus

The main focus of this thesis is to investigate the open issue of mono- and cross-lingual paraphrased text reuse and extrinsic plagiarism detection for the Urdu and English-Urdu language pair. We believe that this is the first work that thoroughly explores this problem in the mono-lingual context for the Urdu language and in the cross-lingual context for the English-Urdu language pair. As far as we know, the majority of the previous efforts were inclined towards English or English-European language pairs [Potthast et al., 2009a, Potthast et al., 2010a, Potthast et al., 2011b, Potthast et al., 2012a, Potthast et al., 2013a, Potthast et al., 2014]. However, there is a large population of the world that speak Indo-Aryan languages (approximately one billion[7]) and there is a clear shortage of corpora and methods proposed for the text reuse and extrinsic plagiarism detection research on these languages.

From the literature, it has been observed that the most famous type of obfuscation that people use to rephrase the text is by paraphrasing (or paraphrasing after translation, in the case of cross-lingual settings) [Maurer et al., 2006, Keck, 2006, Osman et al., 2012]. However, the algorithms proposed are limited to detecting mostly verbatim or direct translations of texts (using surface-level matching) as it is a straightforward task [Nawab, 2012]. Previous research has also shown that it is difficult to detect paraphrased text reuse and extrinsic plagiarism especially when it occurs across languages and more specifically, between cross-script language pairs [Barrón-Cedeño et al., 2010]. Furthermore, the experiments conducted on some non-ideographic languages have reported unsatisfactory results [Chen and Vines, 2014, Aljohani and Mohd, 2014]. One of the reasons is that the majority of the methods proposed in the literature rely on the availability of supporting resources (dictionaries, thesaurus, semantic networks, etc.) and these languages lack these supporting resources.

[7] https://en.wikipedia.org/wiki/Indo-Aryan_languages

This thesis presents efforts on developing benchmark evaluation corpora, supporting resources, and methods to detect text reuse and extrinsic plagiarism in Urdu and English-Urdu language pair. Urdu, belonging to the Indo-Aryan language family, is the official language of Pakistan and predominantly spoken in the country. Moreover, it is one of the most popular languages spoken by around 175 million people around the globe [Alam et al., 2015]. 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 [Mukund et al., 2010, Daud et al., 2017]. However, for NLP, it is a low-resource language concerning even the core processing tasks like tokenisation, part-of-speech (POS) tagging, or morphological analysis [Anwar et al., 2006, Jabbar et al., 2018].

This PhD work examines the specific problem of mono- (Urdu) and cross-lingual (English-Urdu) text reuse and extrinsic plagiarism detection (or pairwise text comparison). The aim is to compare a pair of texts, whether written in a same or different language, to determine whether one has reused the other. An exhaustive pairwise comparison of text documents this way is useful in determining the amount of text reused to create the new text. Besides, it could be used to discriminate between different levels of text reuse. The main motivation behind this PhD research work is to foster the text reuse and extrinsic plagiarism detection research in an under-resourced language, i.e., Urdu and discourage the unacknowledged reuse of text.

1.2 Research goals

The main research goals of this thesis are as follows:

- Explore the problem of text reuse and extrinsic plagiarism detection for an under-resourced Urdu language and English-Urdu language pair;
- Develop benchmark gold standard mono-lingual text reuse and extrinsic pla-

giarism corpora for the Urdu language;

- Develop benchmark gold standard cross-lingual text reuse corpora for the English-Urdu language pair;
- Create supporting lexical resources that assist in the detection of cross-lingual (English-Urdu) text reuse cases;
- Evaluate and compare the performance of state-of-the-art mono-lingual text reuse and extrinsic plagiarism detection methods on the Urdu corpora;
- Develop or fine-tune methods for the mono- and cross-lingual text reuse and extrinsic plagiarism detection.

1.3 Contributions

The key contributions of this thesis work are summarised below;

- **Development of benchmark mono-lingual (Urdu) standard evaluation corpora for the Urdu paraphrased text reuse and extrinsic plagiarism detection.**

  Two mono-lingual (Urdu) corpora for the paraphrased text reuse and extrinsic plagiarism detection are developed. (1) The COUNTER Corpus is a benchmark Urdu text reuse corpus that contains real cases of text reuse from the journalism domain. It has 1,200 text documents with three levels of text reuse. (2) The UPPC Corpus contains simulated cases of Urdu paraphrase plagiarism. It has 160 documents divided into paraphrased plagiarised and non-plagiarised types. Both corpora are available as free to download resources to promote NLP research in the Urdu language.

- **Development of a benchmark cross-lingual gold standard text reuse corpus for the English-Urdu language pair.**

A benchmark gold standard cross-lingual (English-Urdu) text reuse corpus is also developed. The TREU corpus has source text documents in the English language and derived documents in the Urdu language. The manually created corpus is considerably large in size and has in total 4,514 text documents, categorised into three types, i.e., Wholly Derived, Partially Derived, and Non Derived. The corpus is saved in a standard XML format and available as a free to download resource.

- **Development of supporting lexical resources for the English-Urdu language pair.**

  A suitably large English-Urdu Parallel Corpus (154,258 parallel sentences) and several bi-lingual dictionaries as supporting lexical resources are also developed. These resources are useful for many NLP applications including (but not limited to) text reuse and extrinsic plagiarism detection, paraphrase identification and generation systems, MT systems, etc.

- **Evaluation and comparison of state-of-the-art mono-lingual methods for text reuse and extrinsic plagiarism detection for the Urdu language.**

  The performance of several state-of-the-art mono-lingual text reuse and extrinsic plagiarism detection methods is evaluated on the proposed benchmark Urdu corpora. A range of methods are used in the experiments performed, i.e., Word n-grams overlap, Vector Space Model, Longest Common Subsequence, Greedy String Tiling, Local alignment, Global alignment, Stop-word n-grams overlap, Sentence ratio, and Token ratio. Evaluation is carried out on real (COUNTER Corpus) as well as simulated (UPPC Corpus) cases of Urdu text reuse and extrinsic plagiarism. The evaluation assisted in depicting a true picture of the performance of these methods as well as helped to identify which method(s) works best for the Urdu language.

- **Evaluation of state-of-the-art and newly proposed methods for the cross-**

lingual (English-Urdu) text reuse detection.

A diversified range of methods are applied for the evaluation of proposed cross-lingual (English-Urdu) text reuse corpus (i.e., TREU Corpus). The methods used broadly fall into three categories, (1) Translation + Mono-lingual Analysis (includes a language normalisation step that first translates source or derived text documents into the same language and then applies mono-lingual methods), (2) Cross-lingual Vector Space Model (uses a bi-lingual dictionary to translate words from the source or derived text documents and then apply Vector Space Model), and (3) Cross-lingual Embeddings (uses cross-lingual word and sentence embeddings to map words into a single embedding space and then calculates similarity). The evaluation is carried out to show the usefulness of the corpus and how it can be utilised in the development and evaluation of cross-lingual (English-Urdu) text reuse detection systems.

- **Newly proposed method for cross-lingual (English-Urdu) text reuse detection.**

  A new method is proposed for the detection of cross-lingual (English-Urdu) text reuse cases at the document level. The proposed method calculates the cosine similarity between word pairs instead of averaging all word vectors in a source or derived text document. Moreover, it only takes into account the weighted maximum similarity which allows for approximate matching. This way the words that are replaced with their synonyms in the derived text document may also be captured.

- **Use of supporting lexical resources for the cross-lingual (English-Urdu) text reuse detection.**

  Several bi-lingual dictionaries, compiled from different sources, are used as supporting resources in the cross-lingual text reuse detection experiments. To see the effect of lexical coverage, the dictionaries are used separately as well as

combined as a single resource. Furthermore, the experiments are performed using 'first word' as well as 'all words' from the dictionaries as translation units.

- **Custom training of multiple word and sentence embeddings models on an Urdu news corpus.**

  A number of word and sentence embeddings models are trained on a large news corpus for the Urdu language (and are made available for free to download). These models could be used not only for the text reuse and extrinsic plagiarism detection but various other NLP tasks.

1.4 Main findings

The key observations as the main findings of this thesis work are as follows.

- **Observation 1:** There is a dire need to develop linguistic resources and tools for under-resourced languages (e.g., Urdu) to foster research in these languages.
- **Observation 2:** It is a lot easier to detect verbatim reuse of text. However, the problem becomes harder when the text is heavily paraphrased or when translated and then paraphrased.
- **Observation 3:** Detecting text reuse across languages is not a trivial task. Moreover, it becomes more challenging when the reuse occurs between non-ideographic languages (e.g., English-Urdu, English-Arabic, etc.).
- **Observation 4:** Simpler methods (e.g., n-grams overlap) perform competitively with the complex methods (e.g., Greedy String Tiling). Moreover, the performance of methods decreases with the increasing length of n-grams.
- **Observation 5:** At the document level, it is easier to differentiate between two levels of text reuse than three levels. Furthermore, it is difficult to discriminate

between paraphrased and independently written text than paraphrased and verbatim or verbatim and independently written.

- **Observation 6:** State-of-the-art text reuse and extrinsic plagiarism detection methods work fairly well on the simulated cases of reuse, however, their performance falls short on real cases of reuse.

- **Observation 7:** In some cases, text pre-processing (removal of stop-words, punctuation masks, foreign characters, and numbers) is helpful in improving the results whereas, in others, it does not. Moreover, text stemming has a positive effect on the performance of the methods than text lemmatisation.

- **Observation 8:** Using the T+MA method for English-Urdu cross-lingual text reuse detection at document level produced reasonably good results. Moreover, combining different methods has proven to be useful in the English-Urdu cross-lingual text reuse detection.

1.5 Thesis structure

The rest of this thesis is organised as follows:

Chapter 2

Literature Review: The second chapter of this thesis starts with a brief history of text plagiarism. It then categorises and reviews in detail the mono- and cross-lingual standard evaluation corpora already developed for the text reuse and extrinsic plagiarism detection. It also classifies and describes the state-of-the-art methods proposed for the tasks. Moreover, the chapter also discusses the evaluation measures commonly used to assess the performance of a text reuse and extrinsic plagiarism detection system.

Chapter 3

Mono- and Cross-lingual Text Reuse and Extrinsic Plagiarism Resources: The third chapter of this thesis presents the efforts in creating mono- and cross-lingual standard evaluation and supporting resources for an under-resourced language, i.e., Urdu. It describes the details of a mono-lingual (Urdu) text reuse corpus (COUNTER Corpus), a mono-lingual (Urdu) extrinsic plagiarism corpus (UPPC Corpus), and one cross-lingual (English-Urdu) text reuse corpus (TREU Corpus) developed as the outcome of this thesis work. Furthermore, it provides the details of a large-scale English-Urdu Parallel Corpus (EUPC-20) and several bi-lingual dictionaries compiled as the supporting resources for the cross-lingual (English-Urdu) text reuse and extrinsic plagiarism detection.

Chapter 4

Mono-lingual (Urdu) Text Reuse and Extrinsic Plagiarism Detection: The fourth chapter of this thesis describes the experiments performed on the proposed mono-lingual (Urdu) text reuse (COUNTER Corpus) and extrinsic plagiarism (UPPC Corpus) corpora. Several state-of-the-art mono-lingual methods are applied on both corpora to evaluate their performance and examine their behaviour on the Urdu text. Results showed that Word n-grams overlap and Greedy String Tiling with smaller values of n performed best on the Urdu text. The results also highlighted the fact that detecting real cases of text reuse is comparatively more difficult than simulated plagiarism cases that are created in a controlled environment.

Chapter 5

Cross-lingual (English-Urdu) Text Reuse Detection: The fifth chapter of this thesis reports the details of the cross-lingual (English-Urdu) text reuse detection experiments carried out on the TREU Corpus. A large set of diverse methods, classified under three categories, are applied on the proposed corpus to

show how it can be used in the development and evaluation of cross-lingual (English-Urdu) text reuse detection systems. Evaluation results indicated that Translation + Mono-lingual Analysis outperformed both cross-lingual Vector Space Model and cross-lingual embeddings and the combination of different methods effectively improves the performance.

Chapter 6

Conclusions and Future Directions: The sixth chapter concludes the thesis by providing a summary of the contributions made and discusses avenues for future work.

1.6 Published work

The following research articles have been published as part of the research work presented in this thesis. Where appropriate, portions of this thesis are based on our contributions to these publications.

Journal

- Iqra Muneer, Muhammad Sharjeel, Muntaha Iqbal, Rao Muhammad Adeel Nawab and Paul Rayson (2019), CLEU-A Cross-Language English-Urdu Corpus and Benchmark for Text Reuse Experiments. Journal of the Association for Information Science and Technology, 70(7), 729-741.
- Sara Sameen, Muhammad Sharjeel, Rao Muhammad Adeel Nawab, Paul Rayson and Iqra Muneer (2017), Measuring Short Text Reuse for the Urdu Language. IEEE Access, 6(1), 7412–7421.
- Muhammad Sharjeel, Rao Muhammad Adeel Nawab and Paul Rayson (2017), COUNTER: COrpus of Urdu News TExt Reuse. Language Resources and Evaluation, 51(3), 777–803.

Conference

- Muhammad Sharjeel, Paul Rayson and Rao Muhammad Adeel Nawab (May 2016), UPPC - Urdu Paraphrase Plagiarism Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA).

"One can steal everything from an artist except their talent."

Marty Rubin

2

Literature Review

In the previous chapter, an introduction to mono- and cross-lingual text reuse and extrinsic plagiarism detection was presented. Moreover, the importance of standard evaluation corpora to estimate the performance of the state-of-the-art methods was also discussed. Furthermore, the chapter highlighted the fact that the majority of the standard evaluation resources are being developed for English and some other languages. Consequently, the state-of-the-art methods proposed and their supporting resources (in the case of cross-lingual) are mostly available for these languages.

This chapter starts with a brief history of the text reuse and extrinsic plagiarism detection in natural language text (Section 2.1). In the next section, the classification of the already developed mono- and cross-lingual corpora based on the reuse examples they contain is presented (Section 2.2). A comprehensive review of an

individual corpus within each classification, concerning how it was constructed, the levels of reuse it contains, and its detailed statistics are provided. The next section categorises state-of-the-art methods already available for the mono- and cross-lingual text reuse and extrinsic plagiarism detection (Section 2.3). A thorough and in-depth survey of the methods proposed under each category is presented and how effective these methods are when evaluated on different corpora is reported. The chapter concludes with the discussion of the evaluation measures used to assess the performance of text reuse and extrinsic plagiarism detection systems (Section 2.4).

2.1 Plagiarism history

Text reuse and extrinsic plagiarism detection have received increasing attention in recent years, however, text plagiarism detection has a rather long history [Pereira et al., 2010, Potthast et al., 2011a, Barrón-Cedeño et al., 2013a, Ferrero et al., 2016]. The research on plagiarism detection for natural languages started around 1990. However, it was at the start of this century that the field gained more attention, new frameworks were proposed and implementations evaluated, and researchers started to highlight the issue of plagiarism detection in written text [Culwin and Lancaster, 2001, Lyon et al., 2001, Vernon et al., 2001, Clough, 2003]. It was further suggested incorporating NLP and Machine Learning (ML) techniques as enhancements to the existing systems.

Initially, the focus was on detecting mono-lingual plagiarism which later shifted towards cross-lingual plagiarism [Ceska et al., 2008, Lee et al., 2008, Pinto et al., 2009]. A survey conducted on textual plagiarism emphasises the issue of detecting extensive paraphrasing [Maurer et al., 2006]. Although the shift of the survey was more towards mono-lingual plagiarism, it concluded that plagiarism detection systems perform poorly when plagiarism crosses language boundaries. The findings of another survey on the existing online commercial plagiarism detection systems suggested that all of the available systems failed against paraphrased and translated

or cross-lingual plagiarism [Köhler and Weber-Wul, 2010]. For the past 10 years, the tests performed to assess the performance of plagiarism detection systems on the direct copy, paraphrased, and cross-lingual plagiarism cases clearly indicate that the available systems can only detect exact copies, and not paraphrased or cross-lingual plagiarism [Weber-Wulff, 2008, Weber-Wulff, 2013]. These reports highlight that the paraphrased and cross-lingual plagiarism cases are hard to detect and are open issues. Moreover, research in cross-lingual plagiarism between non-ideographic languages has been in its infancy and initial experiments have reported unsatisfactory results [Barrón-Cedeño et al., 2010].

2.2 Corpora for text reuse and extrinsic plagiarism

In any NLP task, a standard evaluation corpus is of utmost importance to not only develop, tune, and compare different existing methods under a common setting but it also provides a material basis and a testbed for building new NLP systems. For the majority of NLP tasks, plenty of corpora are readily available for evaluation and comparison purposes but plagiarism involves confidentiality and ethical issues, therefore, it is difficult to compile a corpus that contains real plagiarism examples [Clough, 2003]. However, the research community has made some serious efforts at developing standard evaluation corpora for the mono- and cross-lingual text reuse and extrinsic plagiarism and these efforts have been fruitful.

Benchmark corpora for the detection of text reuse and extrinsic plagiarism, whether in mono- or cross-lingual settings, can be created in three ways: (1) corpora with artificial examples of reuse - these corpora are created using automatic text rewriting software, (2) corpora with simulated examples of reuse - individuals are asked to obfuscate the original text to generate the plagiarised text, and (3) corpora with real examples of reuse - these examples can be obtained from a domain where text reuse is acceptable, e.g., journalism. In the following sections, a survey of already available mono- and cross-lingual text reuse and extrinsic plagiarism corpora

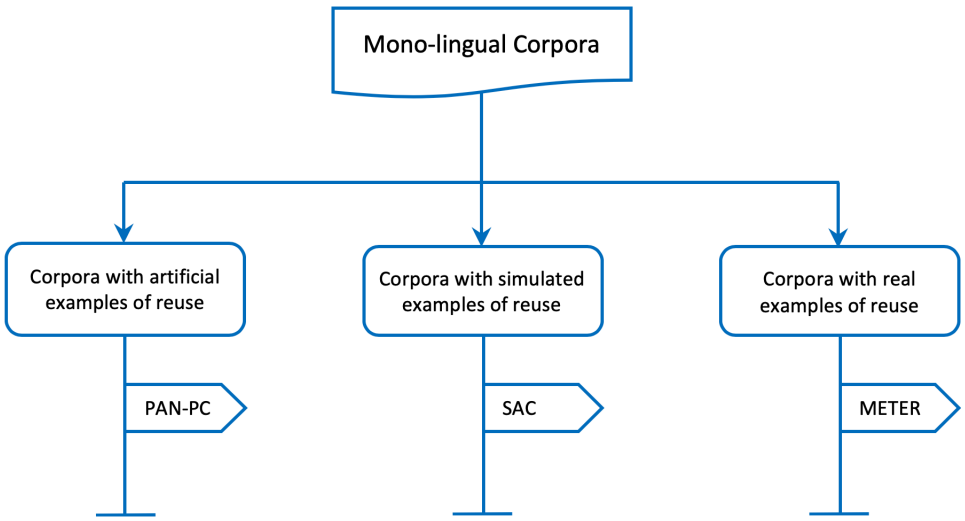

Figure 2.1: Classification of the mono-lingual text reuse and extrinsic plagiarism detection corpora

is presented. Each corpus is first categorised, base on the reuse cases it contains, under one of the three types i.e., (1) artificial, (2) simulated, or (3) real. It is then described at length, emphasising the corpus generation process, its statistics, and the purpose of creating each standard evaluation resource.

2.2.1 Mono-lingual corpora

Over the past decade, researchers have made some notable efforts to develop benchmark mono-lingual text reuse and extrinsic plagiarism corpora. These standard evaluation resources have not only helped in the development of new methods but also comparing the performance of the existing methods for the text reuse and extrinsic plagiarism detection tasks. Figure 2.1 shows the classification of these corpora in terms of the type of reuse cases they contain.

2.2.1.1 Corpora with artificial examples of reuse

The following sections present mono-lingual corpora that contain artificial examples of text reuse and extrinsic plagiarism.

2.2.1.1.1 PAN-PC The PAN[1] Plagiarism Corpora (PAN-PC) are probably the most representative examples of corpora containing artificial examples of reuse. The text documents in the PAN-PC contain plagiarism cases entered automatically (or in some cases manually), in order to allow for the evaluation and the assessment of automatic plagiarism detection systems. These corpora, developed and matured over the years, predominantly include cases of mono-lingual plagiarism but a few cross-lingual (Spanish and German) examples are also part of them. Most of these corpora are based on books (22,135 English, 527 German, and 211 Spanish) from Project Gutenberg[2] and are widely used for research purposes[3]. The following subsections briefly describe the PAN-PC corpora, the nature of the plagiarism cases they contain and summarise the main characteristics of each corpus.

PAN-PC-[09-10-11] In 2009, PAN organisers introduced the PAN-PC-09 corpus, an artificially created corpus for evaluating plagiarism detection systems [Potthast et al., 2009b]. It consists of 41,223 documents extracted from books of Project Gutenberg. The corpus has two distinct subsets for both extrinsic and intrinsic plagiarism detection tasks. The test corpus developed for the extrinsic plagiarism detection task contains 7,215 source and 7,214 suspicious documents. To represent paraphrased plagiarised text in the suspicious documents, artificial operations, referred to as obfuscations, are used. These obfuscations are mostly random text operations, which insert, remove, substitute or rearrange words at random. Similarly, words were replaced with their synonyms or antonyms randomly, and word shuffling with an effort to preserve the POS sequences. Out of the total 94,202 plagiarism cases, only 10% are for cross-lingual plagiarism whereas a large number of these are in the English language (90%). These artificially generated passages lack proper

[1] PAN is an acronym for "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection" - http://pan.webis.de
[2] https://www.gutenberg.org
[3] PAN-PC corpora are freely available to download - https://www.uni-weimar.de/en/media/chairs/webis/corpora/

	PAN-PC-09	PAN-PC-10	PAN-PC-11
Document Statistics			
source docs	50%	50%	50%
suspicious docs			
— with plagiarism	25%	25%	25%
— without plagiarism	25%	25%	25%
Document Length			
short(1-10pp.)	50%	50%	50%
medium(10-100pp.)	35%	35%	35%
long(100-1000pp.)	15%	15%	15%
Plagiarism per Document			
hardly (5%-20%)	-	45%	57%
medium (20%-50%)	-	15%	15%
much (50%-80%)	-	25%	18%
entirely (>80%)	-	15%	10%
Obfuscation Statistics			
none	35%	40%	18%
paraphrasing			
— automatic (low)	35%	20%	32%
— automatic (high)	20%	20%	31%
— simulated	-	6%	8%
translation			
— automatic	10%	14%	10%
— manual	-	-	1%
Obfuscation Case Length			
short	-	34%	35%
medium	-	33%	38%
long	-	33%	27%
Cross-Language Sub-corpus Statistics			
Spanish-English			
— source docs	146	187	199
— suspicious docs	110	189	304
German-English			
— source docs	305	414	348
— suspicious docs	251	476	251
translated plagiarism cases			
— automatic	1,685	7,898	5,142
— manual	-	-	433

Table 2.1: Statistics of the PAN-PC-[09-10-11] corpora

semantics due to random operations, but in the absence of genuine cases, the corpus provides a sufficient test base for evaluating plagiarism detection systems.

The PAN-PC-10 corpus, released in 2010, is an enhanced version that comprises 27,073 documents but the specifications are very similar to the previous year [Potthast et al., 2010b]. However, there are two important differences from the previous version, (1) only one corpus was created for both extrinsic (corresponding to about 70%) and intrinsic (corresponding to about 30%) plagiarism detection tasks, (2) for the first time, 6% simulated plagiarism cases were introduced in the corpus. The organisers used Amazon Mechanical Turk[4] workers to generate manually simulated plagiarism cases. The extrinsic plagiarism detection portion of the test corpus contains 11,148 source documents and 15,925 suspicious documents. Out of the total 68,558 plagiarism cases, only 14% contain translated or cross-lingual plagiarism. An additional effort was put to create topical relationships among the source and suspicious documents in the PAN-PC-10 corpus. 20 different clusters were created to group source-suspicious document pairs into either intra-topic (same cluster) or inter-topic (different clusters).

Released in 2011, the PAN-PC-11 corpus is the next incremental version in the series and contains 26,939 documents [Potthast et al., 2010b]. However, there are two exceptions from the previous years, (1) the total number of obfuscated cases increased from 60% to 82% and, (2) manually simulated cases also saw a modest increase from 6% to 8%. The extrinsic plagiarism detection test corpus released for the third year contains 11,094 source and 11,094 suspicious documents. Out of the total 61,064 plagiarism cases, only 11% cases are of translated or cross-lingual plagiarism. One notable difference, however, is the inclusion of cross-lingual simulated plagiarism cases (though only 1%) in the corpus. These were created in a similar fashion to the mono-lingual ones used in the PAN-PC-10 corpus, i.e., workers at Amazon Mechanical Turk were requested to manually paraphrase the

[4] https://www.mturk.com/mturk/welcome

plagiarism cases after translation. As a result, translated plagiarism (or cross-lingual plagiarism) cases are much closer to real plagiarism cases in the corpus.

Table 2.1 summarises the key statistics and cross compares each of the PAN-PC corpora from 2009-2011.

PAN-PC-12 In 2012, the PAN organisers created a new corpus from scratch. The PAN-PC-12 corpus test set contains 3,000 suspicious documents and 3,500 source documents (including the translations of 500 non-English (Spanish and German) source documents) [Potthast et al., 2012b]. The training set comprises 1,804 suspicious documents and 4,210 source documents. Passages from source documents are automatically obfuscated and then inserted into suspicious documents. In PAN-PC-12, the obfuscation strategies used were; exact copy (no obfuscation), artificial plagiarism (low and high), paraphrased plagiarism (manually simulated) and translated (cross-lingual) plagiarism. 500 cases were generated using each of these strategies while 500 non plagiarised cases are also included in the corpus. Additionally, 33 real plagiarism cases of around 75 to 150 words length are also part of the corpus. However, the rather small number of these real cases were not released with the corpus. Table 2.2 displays the case-wise division of documents in the corpus [Potthast et al., 2012b]. In keeping with the tradition, cross-lingual plagiarism cases were also released with the corpus, though, the cross-lingual sub-corpus was completely revised this year. PAN-PC-12 cross-lingual plagiarism cases are based on the multi-lingual Europarl corpus [Koehn, 2005]. To generate plagiarism cases, a passage from a non-English (German or Spanish) source document was selected, then the analogous passage from the English version of the source document was extracted and inserted into a Gutenberg book. Doing so, the Google Translate service was avoided as it was observed that the same strategy was used to detect cross-lingual plagiarism in the competition. Another improvement is that source-suspicious pairs are formed on the basis of similarity of both documents.

Sub-corpus	Number of Cases
real cases	33
simulated	500
translation({de, es} to en)	500
artificial (high)	500
artificial (low)	500
no obfuscation	500
no plagiarism	500
overall	3,033

Table 2.2: Statistics of the PAN-PC-12 corpus

PAN-PC-13 In 2013, the PAN organisers introduced a new evaluation corpus. In contrast to previous practices (Section 2.2.1.1.1 and Section 2.2.1.1.1), the suspicious text documents in the PAN-PC-13 corpus were created manually through crowd-sourcing [Potthast et al., 2013b], while the source text documents were extracted from the Webis-TRC-12 corpus [Potthast et al., 2013c]. The Webis-TRC-12 corpus contains manually written essays by oDesk workers, searching for a topic from ClueWeb09 corpus [Callan et al., 2009]. In the first step, a set of source text documents was compiled from the Web-TRC-12 corpus, for 144 topics, with at least 2 and up to 170 documents. Roughly 50 word long passages were extracted from source text documents, automatically obfuscated and then inserted (concatenated) to create suspicious text documents. Four obfuscation strategies were used to generate plagiarism cases i.e., none, random and (two completely new strategies) cyclic translation and summary obfuscation. In cyclic translation obfuscation, a passage of text was made to undergo a sequence of translations. This exploits the fact that the inherent nature of machine translation systems introduces paraphrasing in the text. For summary obfuscation, already available resources from automatic text summarisation were incorporated. Table 2.3 shows the distribution of documents in the corpus based on plagiarism type. PAN-PC-13 corpus has 4,774 source and 3,653 suspicious documents which contains 6,000 plagiarised cases (i.e., 2,000 for

Sub-Corpus	Number of cases
summary obfuscation	1000
cyclic translation obfuscation	1000
random obfuscation	1000
no obfuscation	1000
no plagiarism	1000
overall / averaged	6000

Table 2.3: Statistics of the PAN-PC-13 corpus

each of the obfuscation strategies), 2,000 containing no obfuscation and 2,000 without plagiarism. Surprisingly, the PAN-PC-13 corpus does not include any cases of cross-lingual plagiarism.

2.2.1.2 Corpora with simulated examples of reuse

The following sections discuss mono-lingual corpora that contain simulated examples of text reuse and extrinsic plagiarism.

2.2.1.2.1 SAC The Short Answer Corpus (SAC) contains simulated plagiarism cases created to imitate plagiarism in academia [Clough and Stevenson, 2011]. The corpus is an outcome to the answers of five different questions on topics related to Computer Science. A group of 19 volunteers manually created the plagiarised and non-plagiarised text documents for the corpus by answering the below mentioned five questions.

1. Explain the inheritance in respect to object oriented programming
2. What is PageRank algorithm which is used by Google search engine?
3. What is Vector Space Model which is formally used for Information Retrieval?
4. Discuss Bayes Theorem from the field of probability theory.
5. Discuss dynamic programming

The approximate length of each text document in the corpus is between 200-300 words. The volunteers were given Wikipedia excerpts related to the questions to help them in writing the answers. They were also briefed on creating plagiarised text with the following different rewrite levels,

Near copy Use the source Wikipedia article to answer the question by using cut-and-paste operations. However, the length of the answer should be between 200-300 words.

Light revision The answer to the question should be based on the source Wikipedia page. The original text should be altered by paraphrasing techniques like synonym replacement and changing the grammatical structure. Moreover, in sentences, the information order should be preserved.

Heavy revision Again the answer should be based on the original Wikipedia page but it should be generated by rephrasing the original text such that same content is expressed using different linguistic expressions. This may include sentence merging and splitting.

Instructions for creating the non-plagiarised answers are as follows,

Non-plagiarism Subjects were instructed to answer the question using their own knowledge and what they have learned from the learning material (lecture notes, relevant sections from textbooks etc.) provided to them. While answering a question they can look at other relevant material but not Wikipedia.

A total of 95 documents were created, 57 plagiarised (near copy = 19, light revision = 19 and heavy revision = 19) and remaining 38 non-plagiarised. The set of non-plagiarised documents is useful in evaluating the ability of a plagiarism detection system to discriminate plagiarised documents from non-plagiarised ones. In total, this corpus contains 100 documents, 95 suspicious documents and five source Wikipedia articles.

2.2.1.3 Corpora with real examples of reuse

The following sections describe mono-lingual corpora that contain real examples of text reuse and extrinsic plagiarism.

2.2.1.3.1 METER The most prominent effort in recent years, for the development of mono-lingual text reuse corpora containing real examples for the English language, is the MEasuring TExt Reuse (METER) corpus [Clough et al., 2002]. It consists of 1,716 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 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 independently 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, however, the corpus has been used in the past to evaluate the performance of extrinsic plagiarism detection systems [Clough, 2003, Barrón-Cedeño et al., 2009].

2.2.2 Cross-lingual corpora

Similar to the mono-lingual corpora, benchmark cross-lingual standard evaluation corpora have also been proposed for the cross-lingual text reuse and extrinsic plagiarism detection. Figure 2.2 shows the classification of these corpora, each categorised under the type of reuse examples they contain.

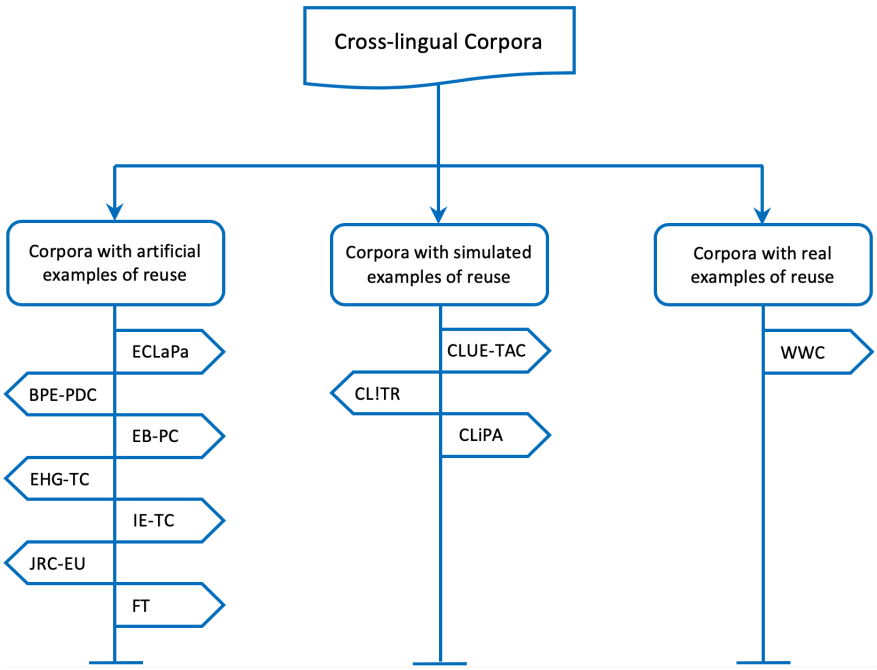

Figure 2.2: Classification of the cross-lingual text reuse and extrinsic plagiarism corpora

2.2.2.1 Corpora with artificial examples of reuse

The sections that follow describe cross-lingual corpora that contain artificial examples of text reuse and extrinsic plagiarism.

2.2.2.1.1 ECLaPa The Europarl Cross-Language Plagiarism analysis (ECLaPa) corpus contains examples of artificial cases of cross-lingual plagiarism [Pereira et al., 2010]. These reuse cases are automatically created using text documents from the Europarl Parallel Corpus[5] [Koehn, 2005]. The ECLaPA corpus contains both monolingual and multi-lingual (French and Portuguese) text plagiarism cases in equal number. To generate artificial plagiarism cases, selected passages from source French or Portuguese text documents are translated and then inserted into suspicious text documents by locating the equivalent English passages. The corpus has been made

[5] The Europarl parallel corpus includes archives of the European Parliament proceedings available in 21 languages.

		English	French	Portuguese
Suspicious documents	300	300	0	0
Source documents	348	0	174	174
Total number of plagiarism cases	2,169			

Table 2.4: Statistics of the ECLaPa corpus

freely available to download[6].

The multi-lingual part of the corpus comprises 348 source text documents and 300 suspicious text documents (Table 2.4). However, 100 (50 French, 50 Portuguese) source text documents are not used in generating suspicious text documents and 100 suspicious text documents are plagiarism free. The corpus has 2,169 cross-lingual plagiarism cases with varying lengths: (1) Short passages (<1,500 characters) used in 30% cases, (2) medium passages (1,501 - 5,000 characters) used in 60% cases and (3) large passages (5,001 - 15,000 characters) used in 10% cases. A suspicious text document may contain up to 15 plagiarised passages from 5 different sources. The purpose of creating the corpus was to detect plagiarised fragments across English-French and English-Portuguese language pairs. The experiment performed achieved F-measure (Section 2.4) score of 0.58.

2.2.2.1.2 BPE-PDC The Bilingual Persian-English Plagiarism Detection Corpus (BPE-PDC) is an artificially created dataset submitted for the PAN 2015 shared task [Asghari et al., 2015]. The corpus was created by following more or less the same approach as PAN corpora, i.e., the source text (from Wikipedia) was obfuscated and inserted back into source documents to create suspicious documents. However, the obfuscation strategy adopted is named as "sentence stitching". To create such obfuscation, sentence pairs from a parallel corpus are extracted and then appended to each other to form aligned passages of text. A total of 11,200 plagiarised passages

[6] http://www.inf.ufrgs.br/~viviane/eclapa.html

created this way were then planted into the source documents. To ensure that the plagiarised passages have at least some similarity to source and suspicious document text, sentence and document clustering was first obtained through IR queries with the Lucene IR engine [Białecki et al., 2012]. The corpus contains 19,973 source and 7,142 suspicious documents. However, half of the suspicious documents contain no plagiarism. In the remaining half, 2,035 documents contain "little", 536 "medium", 642 "much" and 538 "very much" plagiarism.

2.2.2.1.3 EBPC The English Bangla Plagiarism Corpus (EBPC) is a toy corpus of 110 text documents created for the detection of cross-lingual plagiarism in English-Bangla language pair [Arefin et al., 2013]. The corpus was built using students' reports obtained from a public university. Two groups of 55 students each, were asked to write a report on a topic within a specific domain. One group wrote the report in English while the other in Bangla. Out of the 110 reports received, 50 from each group (language) were used as training set while the remaining 10 were used as the test set. Plagiarism cases were obfuscated by replacing contents with several plagiarised contents of different lengths. Not much information about nature or length of plagiarised passages inserted into the suspicious text documents is provided. However, the length of each text document is stated to be almost equal in size. The corpus was used to test a bilingual plagiarism detector, to detect plagiarism in English-Bangla text document pairs but it is not publicly available to download.

2.2.2.1.4 EHG-TC The English Hungarian German Translation Corpus (EHG-TC) was constructed by translating sentences, extracted from English Wikipedia articles to Hungarian and German languages [Pataki, 2012]. A total of 65,000 English sentences were machine translated using Google Translate API. Apart from these, 100 English sentences were hand translated to Hungarian, but these are not part of the final dataset because they are very few in number. The purpose of build-

ing the corpus was to evaluate cross-lingual text reuse and extrinsic plagiarism detection methods on English-Hungarian English-German translated plagiarism cases. However, the corpus is not made available to download.

2.2.2.1.5 IE-TC The Indonesian English Translation Corpus (IE-TC) was created for the experiments on English-Indonesian plagiarism detection [Alfikri and Purwarianti, 2012]. It is composed of a group of suspicious text documents that are plagiarised using a literal translation of some source text documents. The corpus contains 10 documents in English on a topic related to "NLP" or "text processing". The test cases are created in four groups. The following points describe test cases generation:

Test case 1: The whole document is plagiarised using only one source document;

Test case 2: Part of a document (few sentences) are plagiarised using only one source document;

Test case 3: The whole document is plagiarised using multiple source documents;

Test case 4: The whole document is plagiarised using a source document which has similarity with another document in the corpus.

No further details of the corpus are available and it is not provided with an option to download.

2.2.2.1.6 JRC-EU and FTC For investigating cross-lingual plagiarism in the Czech language, two distinct corpora, named JRC-EU and Fairy-tale, were created [Ceska et al., 2008]. Both corpora contain text documents in English and their translations in Czech. JRC-EU corpus is a collection of 400 (half English, half Czech) European Union legislative reports extracted randomly from JRC-Acquis [Steinberger et al., 2006]. Fairy-tale, on the other hand, is a small collection of 54 texts (half English, half Czech) with simplified vocabulary. Not much information about the type of plagiarism used in both the corpora is given and none is made available to

download.

2.2.2.2 Corpora with simulated examples of reuse

The sections that follow present cross-lingual corpora that contain simulated examples of text reuse and extrinsic plagiarism.

2.2.2.2.1 CLUE-TAC The Cross Language Urdu English Text Alignment Corpus (CLUE-TAC) is the first cross-lingual plagiarism detection corpus that contains simulated examples of plagiarism for Urdu-English language pair [Hanif et al., 2015]. Likewise, Bilingual Persian-English Plagiarism Detection Corpus (Section 2.2.2.1.2), it was also created for the PAN 2015 shared task [Potthast et al., 2015]. It contains a collection of 1,000 documents, 500 of which are Urdu source documents and the remaining 500 English suspicious documents. The main source of corpus text is Wikipedia documents related to computer science and some general topics. The simulated cases of plagiarism in the corpus were generated by university students using both manual and semi-automated (using MT with manual editing) approaches. The students were given small (<50 words), medium (50–100 words) and large (100–200 words) text fragments and were requested to obfuscate these fragments on three levels, i.e., Near Copy (CP), Light Revision (LR) and Heavy Revision (HR). These plagiarised fragments were then embedded into the source documents to create suspicious documents. Out of the 500 suspicious documents, only 270 are plagiarised while 230 contains no plagiarism. The CLUE-TAC was submitted for the PAN 2015 shared task to foster research in cross-lingual Urdu-English plagiarism detection.

2.2.2.2.2 CL!TR The Cross-Language Indian Text Reuse (CL!TR) corpus is the first of its kind dataset developed specifically for the analysis of cross-lingual text reuse detection in Hindi language [Barrón-Cedeño et al., 2013b]. The suspicious text documents it contains are in Hindi and source text documents in the English

Training Partition		Test Partition	
Reused	130	Reused	146
— light revision	30	— light revision	69
— heavy revision	55	— heavy revision	43
— exact copy	45	— exact copy	34
Original	68	Original	44
Total	198	Total	190

Table 2.5: Statistics of the CL!TR corpus

language. The training set includes 198 suspicious (Hindi) and 5,032 source (English) text documents, while the test set has 190 suspicious (Hindi) and 5,032 source (English) text documents (Table 2.5 for details).

The source text documents in the corpus are the answers to a set of 10 questions, each related to the tourism and computer science domains, generated from Wikipedia and Incredible !ndia[7]. The creation of 388 potentially reused Hindi text documents is inspired by the approach of [Clough and Stevenson, 2011]. The volunteers were given the task to write short answers (in Hindi) to a set of pre-defined questions, by using either Wikipedia (English version) or from the learning materials such as textbooks, lecture notes, websites etc. (in English) provided as sources. Moreover, for generating reused cases they were free to use automatic MT systems for translating English text to Hindi. An effort was made to manually generate each case of reuse. The corpus includes three levels of rewritten text, i.e., Exact, Light, and Heavy. Exact means word-to-word or translation only copy, Light means text with few revisions or translation plus manual correction and Heavy means text is reused applying substantial paraphrasing or translation plus paraphrasing. The final method, Original, is used to generate answers which were independently written using the learning material. The corpus was used for the PAN@FIRE[8] 2013 competi-

[7] http://incredibleindia.org
[8] FIRE is an acronym for Forum for Information Retrieval Evaluation

Plagiarism Type	Count
Original (English)	5
Human Plagiarised (Spanish)	45
Human Non-Plagiarised (Spanish)	25
Translation Plagiarised (Spanish)	25
Human Plagiarised (Italian)	25
Human Non-Plagiarised (Italian)	25
Related Non-Plagiarised (Spanish)	26
Related Non-Plagiarised (Italian)	25

Table 2.6: Statistics of the CLiPA corpus

tion where six teams participated to detect cross-lingual text reuse in English-Hindi document pair setting [Barrón-Cedeño et al., 2013b]. The best result reported an F-measure score (Section 2.4) of 0.79 [Kothwal and Varma, 2013]. The corpus is available with a free to download option[9].

2.2.2.2.3 CLiPA To evaluate cross-lingual plagiarism offences between Spanish, Italian and English languages, a small automatic plus manually simulated cross-lingual plagiarism analysis corpus named CLiPA (Cross-Language Plagiarism Analysis) corpus was built [Barrón-Cedeño et al., 2008]. It is a toy corpus created at fragment-level. To create the corpus, five text fragments on topic "plagiarism" were plagiarised using both machine translation and by humans to generate cross-lingual plagiarism cases. For machine translation, five different on-line translation services were utilised, in order to have variations in the generated cases whereas for manually (human) simulated plagiarism cases, nine individuals were asked to plagiarise each of the five source fragments. Moreover, the individuals were asked to generate the same number of non-plagiarised cases as well. Table 2.6 shows detailed statistics of the corpus. The corpus was used in the evaluation of cross-lingual plagiarism

[9] http://www.uni-weimar.de/medien/webis/events/panfire-11/panfire11-web/#corpus

detection between English-Spanish and English-Italian language pairs and resulted in F-measure (Section 2.4) score of 0.88. The corpus can be downloaded freely[10].

2.2.2.3 Corpora with real examples of reuse

The sections that follow describe cross-lingual corpora that contain real examples of text reuse and extrinsic plagiarism.

2.2.2.3.1 WWC Debora Weber-Wulff, author of the Copy, Shake, Paste blog[11] has performed several tests over the years on plagiarism detection software using manually created plagiarism cases. The test data set, called Weber Wulff Corpus (WWC), contains short essays between 1 to 1.5 pages, including genuine student plagiarism examples [Weber-Wulff, 2014]. The testing method is repeated almost every year and includes plagiarism cases from English, German, Japanese and Hebrew languages. Some cases contain authentic texts with no plagiarism, others contain machine translation (cross-lingual) plagiarism, and some others, paraphrasing. In one of the test cases, the German politician Guttenberg's doctoral thesis is used (the same thesis has been used in other research work too [Gipp et al., 2011, Gipp et al., 2014]). The detection results, as expected, are pretty poor. The test cases are not publicly available due to their nature.

2.3 Methods for text reuse and extrinsic plagiarism

Computational methods to automatically detect text reuse and extrinsic plagiarism have been around for many years [Diederich, 2006, Clough and Gaizauskas, 2009, Lukashenko et al., 2007, Alzahrani et al., 2012]. In the following sections, a detail description of the existing state-of-the-art methods for both mono- and cross-lingual

[10] http://www.dsic.upv.es/grupos/nle/downloads.html
[11] http://copy-shake-paste.blogspot.co.uk/

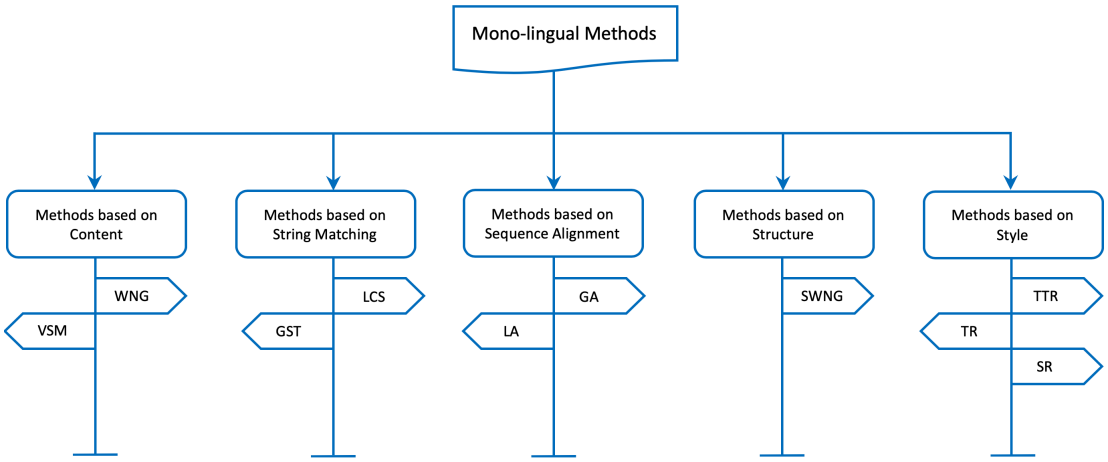

Figure 2.3: Classification of the mono-lingual text reuse and extrinsic plagiarism detection methods

text reuse and extrinsic plagiarism detection is presented. The main aim is to provide an overview of these methods, discuss their development, and how effective they are when evaluated on different corpora.

2.3.1 Mono-lingual methods

In the literature, different mono-lingual text reuse and extrinsic plagiarism detection methods have been proposed that generate similarity scores, by comparing each source-derived text document pair, based on the features which can be extracted from the given texts [Clough et al., 2002, Mihalcea et al., 2006, Bär et al., 2012]. The higher the score, the more similar the contents of the two text documents [Wise, 1992, Brin et al., 1995, Gitchell and Tran, 1999, Lyon et al., 2001].

This section categorises and describes these methods based on five different types i.e., lexical overlap, string matching, sequence alignment, structure, and style of the written text (Figure 2.3).

2.3.1.1 Methods based on lexical overlap

2.3.1.1.1 WNG Word n-grams (WNG) overlap is one of the popular mono-lingual text reuse and extrinsic plagiarism detection method that computes the resemblance

of a text document pair by simply calculating the common n-grams and dividing it by the length of one or both text documents. The method has proven to provide good results for detecting extrinsic plagiarism [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., 2002, Chiu et al., 2010].

Since a derived text document is likely to share more n-grams with the source text document, the n-gram overlap similarity score can be used to distinguish between them. The underlying assumption is that if the similarity score between a source-derived text document pair is higher than a certain threshold value, the source text document was reused to create the rewritten one.

A number of similarity measures, based on set-theoretic principles, have been proposed to quantify the degree of overlap between the two sets of n-grams generated from the text documents [Broder, 1997, Manning and Schütze, 1999]. A similarity measure either falls into the category of asymmetric similarity measure or symmetric similarity measure. In the former case, the length of only one of the sets is employed in the normalisation process whereas in the latter case, normalisation is carried out using the lengths of both sets. Four widely employed similarity measures are, (1) Jaccard (Equation 2.1), (2) Dice (Equation 2.2), (3) Overlap (Equation 2.3), and (4) Containment (Equation 2.4):

$$S_{jaccard}\left(st, dt\right) = \frac{\left|S\left(st, n\right) \bigcap S\left(dt, n\right)\right|}{\left|S\left(st, n\right) \bigcup S\left(dt, n\right)\right|} \tag{2.1}$$

$$S_{dice}\left(st, dt\right) = 2 \times \frac{\left|S\left(st, n\right) \bigcap S\left(dt, n\right)\right|}{\left|S\left(st, n\right)\right| + \left|S\left(dt, n\right)\right|} \tag{2.2}$$

$$S_{overlap}\left(st, dt\right) = \frac{\left|S\left(st, n\right) \bigcap S\left(dt, n\right)\right|}{min\left(\left|S\left(st, n\right)\right|, \left|S\left(dt, n\right)\right|\right)} \tag{2.3}$$

$$S_{containment}\left(st, dt\right) = \frac{\left|S\left(st, n\right) \bigcap S\left(dt, n\right)\right|}{\left|S\left(st, n\right)\right|} \tag{2.4}$$

In the equations 2.1, 2.2, 2.3, and 2.4, $S(st, n)$ and $S(dt, n)$ are the sets of n-grams of length n in source text (st) and derived text (dt), respectively. The similarity score ranges between 0 to 1, where 0 indicates that there are no n-grams in common and 1 indicates that the two texts have all n-grams in common.

2.3.1.1.2 VSM In Vector Space Model (VSM), both source and derived text documents are represented as term (word or phrase) vectors in a high dimensional vector space [Salton et al., 1975]. The number of unique terms in each text document corresponds to a dimension in the vector space. The similarity between two document vectors is calculated by computing the angle between them. The method was originally proposed for IR but has recently been used in the experiments on the detection of text reuse [Clough, 2003, Bendersky and Croft, 2009, Ekbal et al., 2012] and document duplicates [Hoad and Zobel, 2003, Runeson et al., 2007]. Moreover, it was a popular choice for the majority of the participating systems in the PAN competitions [Sanchez-Perez et al., 2014].

To calculate the closeness between source and derived text document vectors (the angle between them), the cosine similarity measure is used (Equation 2.5).

$$S(st, dt) = \frac{\overrightarrow{st} \bullet \overrightarrow{dt}}{|\overrightarrow{st}| \times |\overrightarrow{dt}|} = \frac{\sum_{i=1}^{n} st_i \times dt_i}{\sqrt{\sum_{i=1}^{n}(st_i)^2 \times \sum_{i=1}^{n}(dt_i)^2}} \tag{2.5}$$

where $|\overrightarrow{st}|$ and $|\overrightarrow{dt}|$ represent the lengths of the source and derived text document vectors, respectively.

Before computing the similarity, the popular weighting schemes term frequency *tf*, inverse document frequency *idf*, or their combination *tf-idf* [Jurafsky and Martin, 2009, Baeza-Yates and Ribeiro-Neto, 2011] are applied to assign weights to individual terms in the source and derived text documents (Equation 2.6).

$$tf\text{-}idf_{i,d} = tf_{i,d} \cdot idf_i = \frac{n_{i,d}}{\sum_k n_{k,d}} \cdot log \frac{|D|}{|D_i|} \tag{2.6}$$

2.3.1.2 Methods based on string matching

2.3.1.2.1 LCS Longest Common Subsequence (LCS) is another mono-lingual text reuse and extrinsic plagiarism detection method where the degree of resemblance between a text document pair is calculated by taking into account the total number of changes made when the text was rewritten. Given a piece of text, a subsequence is a contiguous stream of tokens (letters or words) even if some terms are removed from that text. However, LCS is the longest stream of consecutive tokens that are common between the two texts and are in order. Let us assume, *st* and *dt* are two texts (strings) to be compared, if *st* = "123456" and *dt* = "129456", then *456* is a subsequence and *12456* is the longest common subsequence.

The LCS algorithm is order-preserving and can identify the modifications in the text caused by lexical substitutions, word re-ordering and other text altering operations [Cormen et al., 2009]. It has been used in Computer Science for file comparison and compression, detecting duplicate documents [Elhadi and Al-Tobi, 2009], citation-based plagiarism detection [Gipp and Meuschke, 2011], and text reuse and extrinsic plagiarism detection [Chong et al., 2010, Clough and Stevenson, 2011].

2.3.1.2.2 GST Greedy String-Tiling (GST) is another method based on sub-string matching and was proposed for identifying biological sub-sequences and computing similarity between free texts [Wise, 1993, Wise, 1995]. GST can detect *block move* (caused by transposition of tokens), which is missed by LCS (Section 2.3.1.2.1) 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. These tiles are identified in two steps: (1) scan pattern and (2) mark arrays. In the first step, a scan is performed to find the longest possible matches

between a text pair. In the second step, these matches are saved and marked, so they cannot be used again in the next pass. However, to avoid specious matches of very small lengths, a minimum Match Length (mML) value is used.

Given a text document pair and a set of matching tiles of a given length between the two, the normalised GST similarity score can be obtained by taking the ratio of total length of the tiles to the length of one (asymmetric) or both (symmetric) text documents (Section 2.3.1.1.1). This means, to calculate similarity, Jaccard (Equation 2.1), Dice (Equation 2.2), Overlap (Equation 2.3), or Containment (Equation 2.4) can be used. GST has efficiently been used in the past for capturing source code reuse [Wise, 1992, Wise, 1996] and text reuse and plagiarism detection [Clough et al., 2002, Nawab, 2012].

2.3.1.3 Methods based on sequence alignment

2.3.1.3.1 GA Global Alignment (GA) is a sequence alignment method that calculates similarity between two texts by first representing them as sequences of words (tokens) and then identifying similar portions of text (align tokens) between them. The goal is to search for individual terms (words or phrases) that have the same order in both texts. It was proposed by Needleman-Wunsch and works mostly for those sequences that have almost equal lengths [Needleman and Wunsch, 1970]. For GA, the scoring matrix is constructed using Equation 2.7:

$$GA_{score} = S(i,j) = max \begin{cases} S(i\text{-}1,j) - gap \\ S(i,j\text{-}1) - gap \\ S(i\text{-}1,j\text{-}1) + w(a_i,b_j) \end{cases} \tag{2.7}$$

where $w(a_i, b_j)$ value is calculated using match score = 1, mismatch score = -1 and gap value is calculated using gap penalty = 0. For two strings, st = "1212345443636" and dt = "123444336", GA will return the aligned sequence "1–234-443-36".

2.3.1.3.2 LA Local Alignment (LA), is a variation of GA (Section 2.3.1.3.1) which is based on the Smith-Waterman algorithm [Smith and Waterman, 1981]. The algorithm marks similar text portions between two sequences of varying lengths. It uses the same approach as GA (Section 2.3.1.3.1) by first constructing a scoring matrix and then calculating the final score. However, it assigns no penalty to the unaligned portions of sequences (Equation 2.8):

$$LA_{score} = S(i,j) = max \begin{cases} S(i\text{-}1, j) - gap \\ S(i, j\text{-}1) - gap \\ S(i\text{-}1, j\text{-}1) + w(a_i, b_j) \\ 0 \end{cases} \tag{2.8}$$

where $w(a_i, b_j)$ value is calculated using match score = 1, mismatch score = -1 and gap value is calculated using gap penalty = 0. For st = "1212345443636" and dt = "123444336", LA will return the aligned sequence "–1234-44336–".

2.3.1.4 Methods based on structure

2.3.1.4.1 SWNG A method grounded on the syntactic similarity, between source and derived text document pair, is Stop-word n-grams (SWNG) overlap [Stamatatos, 2011]. Stop-words, also known as frequent words, can show useful information for plagiarism analysis as they are often preserved while modifying texts. It has been observed that the plagiarists commonly replace (and rearrange) content words with synonyms, without changing the stop-words.

Given a source and a derived text document, sets of stop-words based n-grams are extracted and compared to find the common chunks. The comparison is performed by counting the number of common n-grams normalised by the length of one (asymmetric) or both (symmetric) sets of n-grams (Section 2.3.1.1.1). The method has performed well to detect text reuse and extrinsic plagiarism in the past [Stamatatos, 2011, Bär et al., 2012, Gupta et al., 2016].

2.3.1.5 Methods based on style

2.3.1.5.1 TTR Type Token Ratio (TTR), as the name suggests, is a ratio between the types (unique words) and the total number of words (tokens) in a corpus [Youmans, 1990]. A higher value of TTR shows that the text is more varied whereas a low TTR value indicates the opposite. However, the ratio is affected by the variation in text length. For longer texts, when the number of tokens (words) increases the number of types comes down. There are methods proposed in the literature to standardised its value [McCarthy and Jarvis, 2010]. Nevertheless, TTR could be used to compare the vocabulary of source and derived texts (Equation 2.9) and it is useful (to some extent) when the texts to be compared for similarity are of equal size.

$$X(st, dt) = \frac{min(TTR(stt), TTR(dtt))}{max(TTR(stt), TTR(dtt))} \tag{2.9}$$

In equation 2.9, *st* is the source text while *dt* is the derived text. Moreover, *stt* represents tokens (words) in a source text while *dtt* represents tokens (words) in a derived text.

2.3.1.5.2 TR Token Ratio (TR) simply calculates the ratio of words (tokens) between the two texts (Equation 2.10). It tries to estimate if the given text pairs have some similarity in terms of writing style.

$$TR(st, dt) = \frac{min(|stt|, |dtt|)}{max(|stt|, |dtt|)} \tag{2.10}$$

In equation 2.10, *st* is the source text while *dt* is the derived text. Furthermore, $|stt|$ and $|dtt|$ represents the number of tokens (words) in a source text and derived text, respectively.

2.3.1.5.3 SR Sentence Ratio (SR) is another stylistic measure used to compute the ratio between the number of sentences in the source and derived texts (Equa-

tion 2.11).

$$SR(st, dt) = \frac{min(|sts|, |dts|)}{max(|sts|, |dts|)} \tag{2.11}$$

In equation 2.11, st is the source text while dt is the derived text. Furthermore, $|sts|$ and $|dts|$ represents the number of sentences in a source text and derived text, respectively.

2.3.2 Cross-lingual methods

In the field of cross-lingual text reuse and extrinsic plagiarism detection, methods have been proposed in the literature for computing similarity between text documents across languages [Ceska et al., 2008, Barrón-Cedeño et al., 2008, Pinto et al., 2009, Pereira et al., 2010, Potthast et al., 2011a, Barrón-Cedeño et al., 2013a]. In the following sections, the existing state-of-the-art methods for cross-lingual text reuse and extrinsic plagiarism detection in natural language text are discussed in detail. The main objective is to describe each method and its effectiveness specifically in a cross-lingual context. Therefore, the development of these methods is presented from the ground up, emerging from the MT based methods to the contemporary and more reliable knowledge-based systems.

The methods have been classified under six categories, based on the knowledge source used by each (Figure 2.4). The following discussion describes in detail each classification category and method(s) that fall under that classification.

2.3.2.1 Methods based on syntax

These methods rely on syntactically similar languages (e.g., English-Spanish) and the presence of foreign words in texts. The idea is that the inherent features of similar terms in languages (cognates) can be exploited when composing small chunks, e.g., character n-grams or prefixes. The methods have proven to be useful in CLIR tasks [Buckley et al., 2000, McNamee and Mayfield, 2004] and similarly to detect text

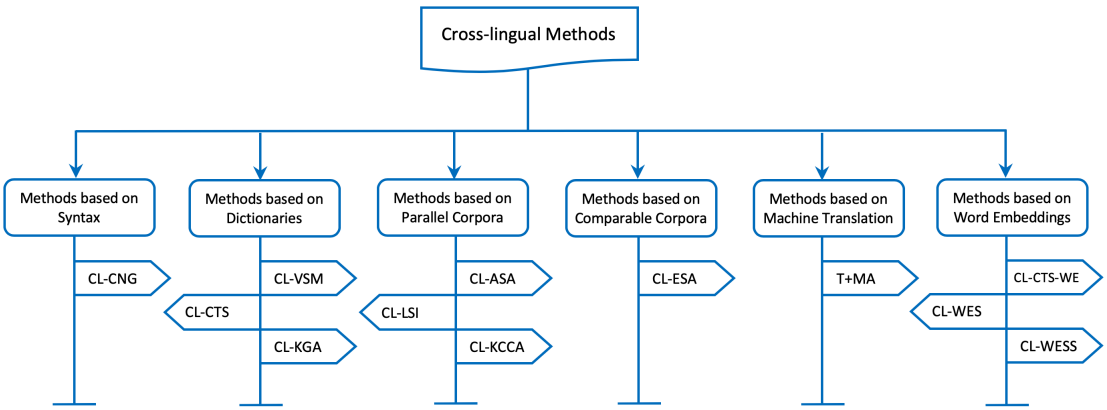

Figure 2.4: Classification of the cross-lingual text reuse and extrinsic plagiarism detection methods

reuse across languages [Potthast et al., 2011a], but have not shown success with distinct language pairs [Barrón-Cedeño et al., 2010]. The key distinction of these methods is that they do not require language-specific information, dictionaries or language translations. On the downside, they work best on languages that belong to the same linguistic family (have a strong influence on each other) and share some elements of the lexicon (e.g., English and related European languages).

Character dot-plot is one of the methods comes under this category, which was originally proposed for bi-texts alignment of parallel corpora[12] [Church, 1993]. The problem of detecting text reuse across languages is considered closely equivalent to bi-texts alignment task, viewed from a different perspective. Another characterisation method originally proposed to identify parallel sentences, is cognateness [Simard et al., 1993]. Cognateness works on prefixes and has been exploited for cross-lingual plagiarism detection. Shingles from the source and derived text documents are extracted according to the defined criteria, to create text document vectors. These vectors are then compared by any standard measure (e.g., cosine between their angles (Section 2.5)). In cross-lingual settings, these methods can only perform bet-

[12] The method has also been used in monolingual plagiarism detection [Grozea et al., 2009].

ter if investigated with languages sharing the Latin alphabet[13]. The most famous method among this category, however, is the Cross-Language Character N-Gram (CL-CNG), discussed in the next section.

2.3.2.1.1 CL-CNG Cross-Language Character N-Gram (CL-CNG), first explored for European languages text retrieval, exploits the overlapping of character n-gram tokens between the source and derived text documents [McNamee and Mayfield, 2004]. For cross-lingual plagiarism detection, the source and derived text documents are represented as vectors by encoding them into character n-grams (optionally, pre-processing and document weighting is applied using some standard scheme, e.g., *tf-idf*). Both vectors are then compared using any standard similarity measure (Equation 2.5).

For a source text document st in one language and a derived text document dt in another, the similarity can be calculated using the Equation 2.12,

$$S(st, dt) = \frac{\overrightarrow{st}.\overrightarrow{dt}}{|st|.|dt|} \tag{2.12}$$

The key difference between CL-CNG and the remaining state-of-the-art methods (discussed below) lies in the fact that it makes possible direct comparison of multi-lingual texts without requiring any supporting resources (parallel or comparable corpora, knowledge-bases, or thesaurus). Nevertheless, it excels only on languages with lexical and syntactic similarities.

The method has been used to detect cross-lingual extrinsic plagiarism with character 3-gram (CL-C3G), *tf-idf* weighting scheme, and cosine similarity (Equation 2.5), on a corpus that has English text documents paired with Spanish, German, Polish, French and Dutch text documents [Potthast et al., 2011a]. It has also been used with character 4-gram (CL-C4G) on English-Spanish text document pairs (Sec-

[13] http://www.omniglot.com/writing/latin.htm

tion 2.2.1.1.1) [Barrón-Cedeño et al., 2013a]. The evaluations results showed CL-CNG outperformed other methods and ranked as best recall (Section 2.4) oriented method.

2.3.2.2 Methods based on dictionaries

These methods overcome the language boundaries by using a multi-lingual dictionary or thesaurus (e.g., Eurovoc [Steinberger et al., 2002] or EuroWordNet [Vossen, 1998]) to translate words or concepts across languages. They may simply be named as cross-language vector space methods, where document vectors are constructed using indexed dictionaries or multi-lingual concept spaces. These methods are known for good retrieval speed but bilingual (or multi-lingual) dictionaries are sparse and the available ones have incompleteness in terms of disambiguation and narrow-domain specific terms [Demner-Fushman and Oard, 2003, Ceska et al., 2008]. Moreover, the translation of one word into many words in cross-lingual settings when dealing with a large vocabulary in a big corpus poses serious issues in respect to ambiguity and computational cost. The methods that fall under this category are Cross-Language Vector Space Method (CL-VSM) and Cross-Language Conceptual Thesaurus based Similarity (CL-CTS).

2.3.2.2.1 CL-VSM Cross-Language Vector Space Method (CL-VSM) follows the traditional VSM approach (Section 2.3.1.1.2), but by constructing vectors of the text documents using bilingual (or multi-lingual) thesaurus, dictionaries, and other concept spaces. Eurovoc [Steinberger et al., 2002], JRC-Acquis Multilingual Parallel Corpus [Steinberger et al., 2006], or similar corpora are used to link multi-lingual pairs of common words. The goal is to transform the two text documents (vectors) into a kind of language-independent form, where they can be compared.

The method was used for experiments with English, German and Hungarian languages (Section 2.2.2.1.4) by using dictionaries to translate lemmas of the extracted keywords [Pataki, 2012]. An ad-hoc metric was used to discard word number vari-

ance as one word in one language stands equivalent to many completely different words in other languages. The experiments, however, were conducted on simple machine translated sentences (no paraphrasing) to find similarity in terms of their translations and achieved 95% probability for German-English and 99% for the Hungarian-English language pair[14].

2.3.2.2.2 CL-CTS Cross-Language Conceptual Thesaurus based Similarity (CL-CTS) is a low computation method based on Eurovoc dictionary [Steinberger et al., 2002]. It utilises the domain-specific mapping of Eurovoc to estimate the conceptual similarity between text documents across languages using the proposed algorithm. The source and derived text documents are first converted into their respective vectors to construct a multidimensional vector space. These vectors are then compared to estimate similarity using the shared entries in the Eurovoc conceptual thesaurus. The method, when tested on English paired with German and Spanish (Section 2.2.1.1.1), offers competitive results and stability across the dataset [Gupta et al., 2012]. Additionally, it outperformed CL-CNG (Section 2.3.2.1.1) and MT based methods (Section 2.3.2.5) during the comparative experiments performed.

2.3.2.2.3 CL-KGA Cross-Language Knowledge Graph Analysis (CL-KGA) makes use of semantic knowledge bases (e.g., BabelNet [Navigli and Ponzetto, 2010], ConceptNet [Havasi et al., 2007], or EuroWordNet [Vossen, 2004]) and graph-based multi-lingual text representation and comparison. In CL-KGA, text documents are first fragmented into paragraphs, then grammatically tagged to create knowledge graphs. These graphs are based on concepts from the text documents themselves as well as concepts from the multi-lingual semantic network. These weighted and labelled graphs are then compared by any standard model to measure similarity [Montes-y Gómez et al., 2001]. Suppose, gs and gd are two graphs of the source

[14] The high percentage of results are because of the experiments were not performed on a cross-lingual corpus but a corpus of mere machine translated sentences.

text document *st* and derived text document *dt*, respectively. The similarity function S(*gs*, *gd*) shown in Equation 2.13 is used for graph comparison,

$$S(gs, gd) = S_c(gs, gd) * (x + y * S_r(gs, gd)) \tag{2.13}$$

where S_c denotes concepts and S_r relations scores between the graphs, x and y are the parameters used to give relevance to concepts and relations using the semantic network used.

The method has been used to detect plagiarism cases between Spanish and German languages (Section 2.2.1.1.1), using BabelNet as knowledge source [Franco-Salvador et al., 2014]. The source and derived texts were fragmented into 5-sentence chunks, lemmatised and POS tagged using TreeTagger[15]. For similarity estimation, the graphs were compared on concepts and their relation scores. The evaluation of the method refines the results of CL-ASA (Section 2.3.2.3.1) and CL-CNG (Section 2.3.2.1.1). In a more recent study, these knowledge graphs were further investigated to see the impact of incorporating Word Sense Disambiguation (WSD) and vocabulary expansion [Franco-Salvador et al., 2016]. Moreover, the weighting scheme used to calculate the similarity was also fine-tuned (Equation 2.13). These improvements further enhanced the results and it performed better than CL-ESA (Section 2.3.2.4.1) due to the high coverage of concepts from the BabelNet. However, there was a notable difference in the evaluation results (a drop from 0.651 to 0.171) when the evaluation was performed on only paraphrased cases of plagiarism (Section 2.2.1.1.1).

2.3.2.3 Methods based on parallel corpora

These methods are trained on parallel corpora, where text documents are sentence aligned based on their translations. They make use of statistical machine trans-

[15] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

lation technology to find related terms across languages or to generate translation units [Littman et al., 1998]. These methods have reported the best performance in high-quality retrieval but are computationally expensive. Methods classified under this category are Cross-Language Alignment based Similarity Analysis (CL-ASA), Cross-Language Latent Semantic Indexing (CL-LSI) and Cross-Language Kernel Canonical Correlation Analysis (CL-KCCA).

2.3.2.3.1 CL-ASA Cross-Language Alignment based Similarity Analysis (CL-ASA) is a statistical method which estimates the probability of derived text document being the translation of a source text document [Barrón-Cedeño et al., 2008]. It is based on methods that are used to align identical sentences from comparable corpora [Munteanu et al., 2004]. CL-ASA is a two-step process based on statistical machine translation technology. The first step is the creation of a bi-lingual statistical dictionary on the basis of the parallel corpus (aligned at word level [Brown et al., 1993, Och and Ney, 2003]). The next step is to estimate the maximum similarity of text document pairs using the Expectation-Maximisation (EM) algorithm concerning the statistical dictionary.

The similarity between a source-derived text document pair (st, dt) can be computed using Equation 2.14,

$$S(st, dt) = l(st, dt) * t(st|dt) \tag{2.14}$$

where $l(st, dt)$ is length normalisation factor which is required when comparing two text documents in different languages with the similar content but different lengths [Pouliquen et al., 2003].

Moreover, Equation 2.14 also requires translation model $t(st|dt)$

$$t(st|dt) = \sum_{a \in st} \sum_{b \in dt} p(a, b) \tag{2.15}$$

In Equation 2.15, $p(a, b)$ defines translation probability when a word a translates

into word b using a bilingual statistical dictionary.

The method was used to detect cross-lingual plagiarism cases between English paired with Spanish, German, Polish, French and Dutch text documents [Potthast et al., 2011a] and English paired with German and Spanish (Section 2.2.1.1.1) [Barrón-Cedeño et al., 2013a]. The experiments used JRC-Acquis Multilingual Parallel Corpus [Steinberger et al., 2006] and three different types of statistical dictionaries ((1)dictionary aligned using IBM M1 model [Brown et al., 1993], (2) inflectional terms dictionary, and (3) stemmed terms dictionary) for the likelihood estimation of words. Evaluation results indicate that the method provides best precision results and works better than CL-CNG (Section 2.3.2.1.1) but not as good as T+MA (Section 2.3.2.5)).

2.3.2.3.2 CL-LSI Cross-Language Latent Semantic Indexing[16] (CL-LSI) is another method that makes use of parallel corpora along with linear algebra, but works without dictionaries, by constructing a multi-lingual semantic space. Latent Semantic Indexing (LSI) is a common approach extensively used mostly by IR systems[17] (e.g., Google). The method is 'latent' as it does not require external knowledge but generates it from within the data itself (based on word co-occurrences). In a cross-lingual scenario, first, a text document pair from a parallel corpus is merged to create a new text document, the co-occurrence of terms in this text document indicates cross-linguistic semantic relatedness. The inherent idea is that the semantically similar terms, even across languages, correspond to similar latent components and are near to each other in the reduced comparison space. Singular Value Decomposition (SVD) (Equation 2.16) of the term-document matrix is used to perform further analysis on the dataset.

[16] Also named as Cross-Language Latent Semantic Analysis (CL-LSA)
[17] The method has also been adapted to use in CLIR [Littman et al., 1998].

$$D = \begin{pmatrix} D_x \\ D_y \end{pmatrix} = U\Sigma V^T \tag{2.16}$$

where D_x and D_y are source and derived documents, respectively.

CL-LSI has been characterised as performance poor method due to the use of SVD algorithm [Potthast et al., 2011a]. Moreover, CL-ESA (Section 2.3.2.4.1) outperforms CL-LSI both in terms of quality and computational performance [Cimiano et al., 2009]. Therefore, different studies have investigated the use of different statistical and linear algebra methods for finer semantic modelling of data over LSI [Vinokourov and Girolami, 2002]. The method has been used for CLIR and received good results, however, it has not been evaluated on any of the cross-lingual corpora [Ballesteros and Croft, 1998].

2.3.2.3.3 CL-KCCA Cross-Language Kernel Canonical Correlation Analysis (CL-KCCA) works by creating a large multi-lingual semantic space which gives the text documents a language-independent representation. It then implements kernel functions to analyse the correspondence of points in two high dimensional spaces, representing a bilingual text document pair, from the parallel corpus. The goal is to measure a set of projections that are maximally correlated. CL-KCCA provides detection of certain semantic similarities, patterns of words that are related in the given pairs of parallel documents.

CL-KCCA has already proved successful in applications on CLIR and text categorisation [Fortuna, 2004]. Moreover, it performs much better than CL-LSI (Section 2.3.2.3.2) for CLIR tasks although it is based on SVD as well [Vinokourov et al., 2002]. However, it has been ranked below CL-CNG (Section 2.3.2.1.1) and CL-ASA (Section 2.3.2.3.1) on cross-lingual extrinsic plagiarism detection task due to the performance issues though the actual results are not shown [Potthast et al., 2011a].

2.3.2.4 Methods based on comparable corpora

These methods are trained on comparable corpora, where text documents in different languages are topically related with a common vocabulary. Comparable corpora are noisier but more flexible than parallel corpora. Hence, these methods do not require language translation, but the mapping of text documents into a multi-lingual vector space. The cross-lingual document similarity can then be measured using high-quality bi-lingual dictionaries or multi-lingual concept spaces. One example method proposed under this category is Cross-Language Explicit Semantic Analysis (CL-ESA).

2.3.2.4.1 CL-ESA Cross-Language Explicit Semantic Analysis (CL-ESA) method is an extension to Explicit Semantic Analysis (ESA), proposed originally for IR, where the similarity between a text document and query is usually determined as a measure of term overlap [Gabrilovich and Markovitch, 2007]. However, ESA aims towards a more semantic dimension and counts semantic relatedness, where the similarity between concepts (derived from comparable corpora) are taken into account. CL-ESA is basically a collection relative method, where the collection documents come from comparable corpora (usually Wikipedia) [Sorg and Cimiano, 2012]. Both source and derived text documents are first represented by the similarities between a term vector representation of text documents and an inverted index of collection documents. Secondly, the resultant vectors are compared using any standard model. Optionally, term weighting schemes such as *tf-idf* are also applied.

CL-ESA has the strength of creating word-level relations among bilingual text documents. These are used to perform comparisons of vocabulary correlation between text documents. The method is 'explicit' as the knowledge is coming from the text documents (concepts) in the comparable corpus [Cimiano et al., 2009]. Hence, the similarity is measured between text document terms and associated concepts derived from the comparable corpus. The strength of CL-ESA lies in the fact that

it does not require a translation step but multi-lingual text document collection written on similar topics.

CL-ESA works best with cross-script cross-lingual text document pairs with unrelated syntax [Barrón-Cedeño et al., 2013c] and even better if the source-derived texts are in a topical relationship [Potthast et al., 2011a].

2.3.2.5 Methods based on machine translation

These methods are widely adopted by researchers for the cross-lingual text reuse and extrinsic plagiarism detection as they simplify the task by translating either source or derived text documents into one language and then tackling it as a monolingual problem [Kent and Salim, 2010, Pereira et al., 2010, Oberreuter et al., 2011]. These methods make use of MT technology and hence are generalised as Translation+Monolingual Analysis (T+MA).

2.3.2.5.1 T+MA In Translation+Monolingual Analysis (T+MA) based methods, the first step is to detect the language of the derived text document, to translate it to the same language as the source text document. In the second step, monolingual analysis is performed using any of the state-of-the-art monolingual methods. Text document preprocessing, i.e., stop-words removal, stemming, lemmatisation, and weighting the document terms with standard schemes e.g., *tf-idf* are also adopted before the similarity measuring step.

The method has been used to detect cross-lingual plagiarism cases between English-Malay [Kent and Salim, 2010]. The Malay text was first translated using Google Translate[18], after preprocessing, the comparison is performed using fingerprinting with three least frequent 4-grams. Another study used English paired with Spanish and German (Section 2.2.1.1.1) [Muhr et al., 2010]. These and other similar approaches provide promising results [Barrón-Cedeño et al., 2013a].

[18] http://translate.google.com

2.3.2.6 Methods based on word embeddings

More recently, word embeddings have shown promising results in all kinds of NLP tasks [Mikolov et al., 2013]. As a result, they have been exploited for the cross-lingual text reuse and extrinsic plagiarism detection tasks too. Using word embeddings, we can extract similar contextual words from the multi-dimensional vector space using cosine similarity between the two word embedding vectors. For cross-lingual text reuse and extrinsic plagiarism detection, word embeddings substitute the use of typical lexical resources (e.g., dictionaries) to compute text similarity across languages. The methods proposed in the literature that make use of word embeddings are Cross-Language Conceptual Thesaurus based Similarity using Word Embeddings (CL-CTS-WE) and it's variant Cross-Language Word Embeddings based Similarity (CL-WES). Moreover, Cross-Language Word Embeddings based Sentence Similarity (CL-WESS) also falls under this category.

2.3.2.6.1 CL-CTS-WE Cross-Language Conceptual Thesaurus based Similarity using Word Embeddings (CL-CTS-WE) works in a similar fashion to CL-CTS (Section 2.3.2.2.2) but benefits from word embeddings instead of the Eurovoc [Steinberger et al., 2002] or BDNary [Sérasset, 2015] as a lexical resource [Ferrero et al., 2017]. In CL-CTS-WE implementation, first, a parallel corpus is used to build the bi-lingual word embeddings vectors. Secondly, top-ranking words (e.g., 10 words) found in the bilingual distributed representations of words are selected to form a Bag-of-Words (BoW) model. This model is further used to compute similarity using any standard similarity estimation measures.

The method when evaluated on English-French corpora [Ferrero et al., 2016] outperformed CL-CTS (Section 2.3.2.2.2) but falls short of CL-CNG (Section 2.3.2.1.1) [Ferrero et al., 2017].

2.3.2.6.2 CL-WES Cross-Language Word Embeddings based Similarity (CL-WES) is another word embeddings based method that tries to directly compare texts in

two languages exploiting sentence vectors [Ferrero et al., 2017]. The source and derived text documents are first split into smaller units (sentences or segments). The sentence vectors are then constructed by computing the summation of the embeddings vectors of each word in the sentence. The similarity between the resulting sentence vectors is calculated using cosine similarity (Equation 2.5).

CL-WES demonstrated lower F_1 scores (Section 2.4)) when compared with three state-of-the art methods i.e., CL-CNG (Section 2.3.2.1.1), CL-CTS (Section 2.3.2.2.2), and CL-ASA (Section 2.3.2.3.1) on English-French corpora [Ferrero et al., 2016].

2.3.2.6.3 CL-WESS Cross-Language Word Embedding based Syntax Similarity (CL-WESS) is an enhanced version of CL-WES (Section 2.3.2.6.2) that integrates POS information along with the embeddings vectors to construct syntactically distributed representation of text [Ferrero et al., 2017]. This addition of grammatical information reduces the word disambiguation problem to some extent. Similar to CL-WES (Section 2.3.2.6.2) implementation, the texts are first split into sentences and assigned normalised POS tags to preserve the syntax structural information. To further improve the efficiency of the results, each POS tag is assigned an optimal weight. The sentence vector V is then constructed by taking the summation of the dot products of each word vector and its weighted POS tag information using Equation 2.17

$$V = \sum_{i=0}^{n} (weight(pos(w_i)).vector(w_i)) \tag{2.17}$$

where w_i is the i^{th} word in a text, *weight*, *pos*, and *vector* are the functions that return POS weight, POS, and word embedding vector of same word, and . is the scalar product (between a number and a vector).

The similarity between two vectors created this way is computed using cosine similarity (Equation 2.5).

CL-WESS when evaluated on English-French corpora [Ferrero et al., 2016] per-

formed notably better than CL-CTS-WE (Section 2.3.2.6.1) and CL-WES (Section 2.3.2.6.2) due to the addition of syntax information into the vectors.

2.4 Evaluation measures

The evaluation of text reuse and extrinsic plagiarism detection methods is commonly performed using standard IR evaluation measures i.e., *precision*, *recall* and *F-measure*. For the specific case of text reuse and extrinsic plagiarism detection, a set of source text document(s) are marked as *relevant* if they were exploited when writing the derived text document. On the opposite side, the set of potential source text document(s) returned are treated as the *retrieved* text documents. Using these sets, precision and recall can be computed using Equations 2.18 and 2.19 respectively [Baeza-Yates and Ribeiro-Neto, 2011].

$$Precision(P) = \frac{|retrieved \cap relevant|}{|retrieved|} \tag{2.18}$$

$$Recall(R) = \frac{|retrieved \cap relevant|}{|relevant|} \tag{2.19}$$

In the above equations, P (Precision) is the percentage of retrieved text documents that are actually relevant, moreover, R (Recall) is the percentage of relevant text documents over the text documents retrieved by the text reuse and extrinsic plagiarism detection system.

High precision means the system has returned all the relevant text documents but recall will be low. A system with high recall is just the opposite. To balance the combination of both precision and recall, the F-measure (Equation 2.20) is used.

$$F_\alpha = \frac{(1 + \alpha^2) \cdot p \cdot r}{\alpha^2 \cdot p + r} \tag{2.20}$$

where α is the weight assigned to precision p or recall r. When both precision and

recall are equally balanced i.e., α= 1, gives the F_1 measure which is the harmonic mean of precision and recall and computed as (Equation 2.21) :

$$F_1 = \frac{2 \cdot p \cdot r}{p + r} \tag{2.21}$$

As the text reuse and extrinsic plagiarism detection evaluation involve the classification of text documents as either derived or non-derived, the above-defined standard-de-facto measures are best suited and useful for the task.

2.5 Chapter summary

This chapter presented a detailed review of the existing standard evaluation corpora and state-of-the-art methods for the mono- and cross-lingual text reuse and extrinsic plagiarism detection tasks.

In the first part of the chapter, benchmark corpora already developed for both the tasks were presented. The corpora were classified into three categories based on the reuse cases they contain, i.e., (1) artificial, (2)simulated, and (3) real. Each corpus was then extensively discussed via four parameters, (1) how it was constructed, (2) nature of text documents used, (3) levels of reuse it contains, and (4) its detailed statistics.

In the second part of the chapter, the state-of-the-art methods already proposed for the text reuse and extrinsic plagiarism detection were categorised and discussed. The mono-lingual methods are classified into five different types based on the characteristics of the written text, i.e., (1) lexical overlap, (2) string matching, (3) sequence alignment, (4) structure, and (5) style. The cross-lingual methods are classified into six categories based on the knowledge source used by them, i.e., (1) syntax, (2) dictionaries, (3) parallel corpora, (4) comparable corpora, (5) machine translation, and (6) word embeddings. For each mono- and cross-lingual method, first the method itself is described in detail and then how effective it is when evaluated on different

corpora.

Finally, standard evaluation measures, most commonly used to evaluate the performance of text reuse and extrinsic plagiarism detection systems, i.e., precision, recall and F measure, are presented and described.

"Behind every plagiarism, there is Google."

Vikash Shrivastava

3

Mono- and Cross-Lingual Text Reuse and Extrinsic Plagiarism Resources

In the previous chapter, a detailed review of the benchmark corpora and state-of-the-art methods for mono- and cross-lingual text reuse and extrinsic plagiarism detection was presented. The chapter highlighted the fact that for the development, analysis, and evaluation of computational methods for both tasks, benchmark corpora are of utmost importance. Additionally, to a large extent, the methods for the cross-lingual task depends on the quality and availability of the supporting language resources as well. It was also stressed the need to develop good quality NLP corpora and tools for the under-resourced languages (e.g., Urdu).

This chapter explains in detail the development of the mono- and cross-lingual

resources for the Urdu and English-Urdu language pair. Mainly two types of resources are developed: (1) standard evaluation resources for both mono- (Urdu) and cross-lingual (English-Urdu) text reuse and mono-lingual (Urdu) extrinsic plagiarism, and (2) supporting resources for cross-lingual (English-Urdu) text reuse detection.

Three benchmark corpora are created for the text reuse and extrinsic plagiarism detection: (1) COUNTER Corpus - an Urdu text reuse corpus containing three levels of reuse examples at the document level (Section 3.1), (2) UPPC Corpus - an Urdu extrinsic plagiarism corpus that contains manually crafted simulated examples of Urdu paraphrased plagiarism at the document level (Section 3.2), and (3) TREU Corpus - an English-Urdu cross-lingual document level text reuse corpus which includes real cases of text reuse from English to Urdu (Section 3.3). Regarding supporting resources for cross-lingual (English-Urdu) text reuse detection, this chapter presents two main contributions: (1) development of a large-scale publicly available English-Urdu Parallel Corpus (EUPC-20) compiled from the Web (Section 3.4), and (2) bi-lingual dictionaries created for the English-Urdu language pair using different methods from online and offline sources (Section 3.5).

3.1 Urdu text reuse corpus

The COUNTER[1] (COrpus of Urdu News TExt Reuse) Corpus [Sharjeel et al., 2017], is a benchmark Urdu text reuse corpus, developed with an approach that is closely related to the METER corpus (Section 2.2.1.3.1) [Clough et al., 2002]. It contains real examples of Urdu text reuse from the field of journalism. There are a total of 1,200 text documents in the corpus, half of them are source text documents and the remaining half derived text documents. The source text documents are produced

[1] 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

by the leading news agencies of Pakistan, whereas the derived text documents are a collection of corresponding newspaper stories published in the major newspapers of Pakistan. The derived collection contains text documents with various degrees of text reuse. Some of the newspaper stories (i.e., derived text documents) are rewritten (either verbatim or paraphrased) from the news agency's text (i.e., source text document) while others have been written by the journalists independently on their own. For the former case, source-derived text 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. For the latter case, they are tagged as Non Derived (ND) as the journalist(s) have not reused anything from the news agency's text but based on their observations and findings, developed and documented the story.

The corpus will serve as a benchmark standard for the evaluation of methods to automatically detect mono-lingual text reuse for the Urdu language. Moreover, it can also be used to develop automatic methods which can be employed in journalism, for measuring the amount of news source copy reused, for taking appropriate actions. Additionally, we believe that the corpus is a considerably good resource to develop or fine-tune methods for the mono-lingual text reuse detection research in languages similar to the Urdu language (e.g., Persian, Arabic etc.).

3.1.1 Corpus generation process

As described in the previous section, the main intention behind the development of such a resource was to evaluate the existing methods available for text reuse detection in general and specifically for the 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. Moreover, 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 that newspaper editors use different paraphrase mechanisms such as lexical or syntactical substitution, inflectional or derivational changes and summarisation to rewrite a newspaper story [Bell, 1991, Fries, 1997, Jing and McKeown, 1999, Clough, 2003]. 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 normalisation.

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[2] available on the Web in electronic form. However, some of the Urdu newspapers publish the text on Web in graphics (images) form. These images were saved and later converted into electronic form (Urdu text) manually.

The corpus consists of news articles released by the 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 were extracted from nine daily published and large circulation national newspapers 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 4 million[3]. News agency texts were provided in electronic form by the news agencies on a daily basis when they released the news. Newspaper stories 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 the data collection. Table 3.1 shows distribution of text documents in the COUNTER Corpus.

[2] All the copyrights of the original news text are owned by the respective newspapers. The data is made available for non-commercial and research purposes only.
[3] https://pakpressfoundation.wordpress.com/2006/05/05/pakistan-press-foundation

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 3.1: Distribution of text documents by the news agencies, newspapers, and their domains in the COUNTER Corpus

3.1.2 Corpus properties and statistics

The corpus is composed of two main document types: (1) source text documents and (2) derived text documents. There is a total of 1,200 text documents in the corpus: 600 are source text documents (news agency articles) and 600 are derived text documents (newspapers stories). The corpus contains in total 275,387 words (tokens[4]), 21,426 unique words (types), and 10,841 sentences. The average length of a source text document is 227 words while for the derived text documents it is 254 words. Table 3.2 shows detailed statistics of the COUNTER Corpus.

	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)	1,377	2,481

Table 3.2: COUNTER Corpus statistics

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

3.1.3 Annotations and inter-rater agreement

The COUNTER Corpus has been annotated at document level 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 annotations were performed to categorise each document pair into one of the three classes of text 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, and (3) conflict resolving. During the training phase, annotators A and B manually annotated 60 text 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 text document (newspaper story) depending on the volume of text rewritten from the source (news agency article) falls into one of the following categories;

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 a 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

text document contain paraphrased text or new facts and figures added by the journalist's 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 text documents), it has completely different facts and figures or is heavily paraphrased from the news agency's copy. In this case, the derived text document is independently written and has a lot more new text.

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

Table 3.3: Classification of text document pairs in the COUNTER Corpus and its comparison with METER corpus [Clough et al., 2002]

After the annotation phase, the inter-annotator agreement was computed. The inter-rater score was calculated to be 85.5% as the annotators had an agreement on 513 of the 600 pairs. The Kappa Coefficient was computed to be 77.28% (Weighted Kappa 81.4%) [Cohen, 1960, Cohen, 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 documents. Table 3.3 lists the classification of documents in the COUNTER Corpus and compares it with the METER corpus (Section 2.2.1.3.1) [Clough et al., 2002]. It highlights the similarities as both the corpora have the majority of the documents in the PD class i.e. 48% (METER) and 46.3% (COUNTER).

3.1.4 Examples of reuse cases from the corpus

This section shows examples of the WD, PD, and ND text document pairs from the COUNTER Corpus[5]. As expected, the derived text document in WD (Section 3.1.4.1) is a word-to-word copy of the source text document. The information described in the derived text is the same as in the text reported by the news agency. In the case of PD (Section 3.1.4.2), 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 news agency source. For ND (Section 3.1.4.3), 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.

3.1.4.1 Example of WD source and derived text documents

Source text document

دبئی میں منعقد ہونے والی بین الاقوامی تجارتی نمائش میں شرکت کے خواہشمند اداروں سے درخواستیں طلب۔ 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 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 اشیاء تیار و برآمد کرنے والے ادارے شرکت کے ذریعے عالمی خریداروں سے کاروباری معاملات طے کرسکیں 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۔

[5] Words common in source and derived text documents are underlined.

Derived text document

دبئی میں عالمی تجارتی نمائش میں شرکت کیلئے درخواستیں طلب۔ ٹیکسٹائل، لیدر اور عام ضروریات کی اشیاء تیار و برآمد کرنے والے ادارے شرکت کریں گے۔ نمائش میں شرکت کیلئے درخواستیں 30 جولائی تک جمع کروائی جاسکتی ہیں۔ ٹریڈ ڈویلپمنٹ 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 اشیاء تیار و برآمد کرنے والے ادارے شرکت کے ذریعے عالمی خریداروں سے کاروباری معاملات طے کرسکیں 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 دیگر تفصیلات اتھارٹی کی ویب سائٹ سے بھی حاصل کی جاسکتی ہیں۔

3.1.4.2 Example of PD source and derived text documents

Source text document

وزیراعظم محمد نوازشریف کا افغان صدر اشرف غنی کو فون۔ صوبہ پکتیکا میں خودکش دھماکے میں 50 سے زائد افراد کی ہلاکت پر گہرے دکھ اور افسوس کا اظہار۔ وزیراعظم محمد نوازشریف نے افغان صوبہ پکتیکا میں خودکش 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 پکتیکا میں خودکش 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 نوازشریف نے خودکش 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۔

Derived text document

دہشت گرد مشترکہ دشمن ہیں، نوازشریف کا افغان صدر غنی کو فون۔ بزدلانہ کاروائیوں سے دونوں ملکوں کے عوام گھبرانے والے نہیں، پکتیکا دھماکے پر افسوس۔ وزیراعظم نوازشریف نے افغان صدر اشرف غنی کو ٹیلی فون کیا اور پکتیکا میں ہونے والے دھماکے پر اظہارے افسوس کیا ہے۔ وزیراعظم ہاؤس کے مطابق وزیراعظم نوازشریف نے بم 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 کاروائیوں 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 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۔

3.1.4.3 Example of ND source and derived text documents

Source text document

تمام سیاسی جماعتیں اختلافات بھلا کر ملک کی بہتری کیلئے ورکنگ ریلیشن شپ کو بہتر کریں، سندھ اس بات کا متحمل نہیں ہوسکتا کہ یہاں قومیتوں کے جھگڑے بڑھیں، خورشید شاہ سید ہیں وہ اپنے نانا کی توہین نہیں کرسکتے، 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۔ 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 ہوسکتا کہ یہاں قومیتوں کے جھگڑے بڑھیں۔ سندھ کی ترقی پاکستان کی ترقی ہے۔ خورشید شاہ سید ہیں وہ اپنے نانا کی توہین نہیں کرسکتے، 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 پگاڑا پیر صبغت اللہ سے ملاقات کے بعد میڈیا سے گفتگو کرتے ہوئے کیا۔ اس موقع پر وفاقی وزیر سمندر پار پاکستانی و انسانی 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 پگاڑا 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 حسینؑ کسی ایک فرقے کہ نہیں تھے اس لئے پوری دنیا میں اچھا پیغام جاتا ہے۔ انہوں نے کہا کہ میں نے پیر صاحب پگاڑا اور ان کے بھائی پیر صدر الدین شاہ راشدی کو منصورہ آنے 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۔

Derived text document

صوبے 4 صوبے چلا کر دکھائیں پھر 20 بنائیں، سراج الحق: پیر پگاڑا سے ملاقات، جمہوری سسٹم کے تحفظ پر اتفاق۔ جماعت اسلامی کے امیر سراج الحق اور پیپلزپارٹی کے رہنما سابق وزیراعظم 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 و انصاف 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 و غصہ 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 چلاکر دکھائیں پھر ،30 چالیس بنا لیں، 4 صوبے چلتے نہیں باقی کیسے چلائینگے۔ انہوں نے کہا کہ جلداز جلد بلدیاتی انتخابات کرائے جائیں، لڑائی سے کچھ ہاتھ نہیں آئیگا، ہو سکتا ہے سیاست اور جمہوریت کا خاتمہ ہو جائے۔ دریں اثناء پیپلزپارٹی کے رہنما سابق وزیراعظم یوسف رضا

گیلانی نے بھی پیر پگاڑا سے انکی رہائش گاہ پر ملاقات کرکے سیاسی صورت حال پر تبادلہ خیال کیا۔ اس موقع پر گفتگو کرتے ہوئے پیر پگاڑا نے کہا کہ کونسی فورس ہے جو کرپشن کرنیوالوں کا احتساب کرے۔ احتساب سے پہلے الیکشن کا کوئی فائدہ نہیں ہو گا۔ انہوں نے کہا کہ وزیراعظم نوازشریف کو صرف پنجاب کی فکر ہے دیگر صوبوں یا وفاق کی نہیں، مڈٹرم 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 آئو۔ انہوں نے کہا کہ میں 70 دن سے دھرنے والوں کے پاس جاتا رہا، دونوں کے رہنمائوں 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۔

3.1.5 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 typologies [Clough, 2003, Barrón-Cedeño et al., 2013c, Vila et al., 2014]. Following the same approach, we also identified the paraphrase mechanisms used (by journalists) to formulate the newspaper story (derived text document), in the COUNTER Corpus.

The paraphrase typology (Table 3.4) followed, to present a linguistic analysis of the COUNTER 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.

The following discussion describes each of the 14 types of paraphrase typology

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

Table 3.4: The paraphrase typology showing 6 classes and 14 types.

with examples[6] from the corpus.

Morphology-based changes

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

پاکستان کے 4 [وکٹوں] پر 261 رنز
پاکستان کے 4 [وکٹ] پر 261 رنز

Derivational changes consist 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

[6] The examples shown are just small fragments extracted from the source/derived text documents. Refer to Section 3.1.4 to see full examples of source/derived text documents. The words/phrases in the focus of discussion are enclosed in square brackets to emphasise them.

[Pakistan-i] (adjective) is changed to [Pakistan] (noun).

امریکی منڈیوں میں [پاکستانی] مصنوعات کیلئے بہتر رسائی کی ضرورت ہے
[پاکستان] کی اشیا کی 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 the example below, abbreviations are changed to their full forms.

پشاور پینتھرز، [این بی پی]، راولپنڈی ریمز، [یو بی ایل] شامل ہیں
پشاور پینتھرز، [نیشنل بینک آف پاکستان]، راولپنڈی ریمز اور [یونائیٹڈ بینک لمیٹڈ] شامل ہیں

Same-polarity substitutions replace 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.

ایبولا وائرس سے [متاثرہ شخص] ہسپتال میں علاج کے دوران جاں بحق ہوگیا
فیصل آباد میں ایبولا کا [مشتبہ مریض] دم توڑ گیا

Synthetic/analytic substitutions involve 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.

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

Opposite-polarity substitutions change the word or phrase to 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.

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

کیوی ٹیم 3-2 گولز سے [فتح یاب]
پاکستان کو نیوزی لینڈ سے [شکست]

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

امریکی یر غمال کے والدین کی داعش سے رحم دلی کی اپیل
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.

ویسٹ انڈیز نے سنیل نارائن کو بھارت کے خلاف [نہ] کھلانے کا فیصلہ کر لیا
مایہ ناز آف سپنر سنیل نارائن دورہ بھارت کے دوران ٹیم کی نمائندگی [نہیں] کریں گے

Discourse-based changes

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

رابن ولیمز کی پراسرار موت سے متعلق تحقیقات شروع ہوگئی ہیں جس میں خودکشی کے پہلوکو بھی مد نظر رکھا جائے گا
ان کی موت بظاہر خودکشی 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 the format of text. Normally these changes do not affect the lexical units. The first part of the following example shows a 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.

کشمیر اور تلنگانہ بھارت کا حصہ نہیں ۔بھارتی رکن پارلیمنٹ
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 with similar content. 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.

اوباما نے عراقی عسکریت پسندوں کے خلاف فضائی حملوں کی منظوری دے دی
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.

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

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

آسٹریلیا نے 49 [گولڈ]،42 [سلور] اور 46 [برانز] [میڈلز] کے ساتھ دوسری پوزیشن حاصل کی
آسٹریلیا 49 [سونے]،42 [چاندی]،46 [کانسی] سمیت137 [تمغوں] کیساتھ داسرے نمبر پر رہا

مچل جونسن اور مچل اسٹارک نے [نصف سنچریاں] جڑیں
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 text document pairs from the corpus[7] and calculated the paraphrase type frequencies for each of the 14 types (Table 3.4).

Table 3.5 shows that 'Same-polarity substitutions' emerges as the most frequent (0.312) paraphrase type present in the subset of the corpus, followed by 'Semantic

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

	$frequencies_{abc}$	$frequencies_{rel}$
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

Table 3.5: Paraphrase type frequencies occurring within the 50 text documents subset corpus. Bold values are the sum of the corresponding types within the main classes.

changes' (0.200) and 'Addition/deletion of information' (0.168) which also contribute to a major extent[8]. This was expected as the corpus text (of derived text 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

[8] It is expected 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.

involve replacing lexical units. Moreover, journalistic writing involves an editor's observations which naturally results in the addition/deletion of information. In conclusion, same polarity substitutions, semantic changes, and addition/deletion of information are the most favourite mechanisms used by journalists as they are relatively easy to apply and preferable by individuals when reusing text.

3.2 Urdu extrinsic plagiarism corpus

The UPPC[9] (Urdu Paraphrase Plagiarism Corpus) [Sharjeel et al., 2016] Corpus is a benchmark standard evaluation resource that contains simulated examples of paraphrase plagiarism for the Urdu language. The corpus contains a total of 160 text documents, with 20 source text documents and 140 suspicious text documents[10]. The source text documents are original Wikipedia articles on well-known personalities while the set of suspicious text documents are either manually paraphrased (plagiarised) versions produced by applying different rewriting techniques or set of independently written (non-plagiarised) documents.

The resource is the first of its kind developed for the Urdu language and we believe that it will be a valuable contribution to the evaluation of Urdu paraphrase plagiarism detection systems. The UPPC Corpus can be used for: (1) the development, analysis and evaluation of automated paraphrase plagiarism detection systems for the Urdu language, (2) identifying which types of obfuscations (paraphrase strategies) are easy or difficult to detect, and (3) would be a valuable resource for the Urdu paraphrase identification task (at document level).

[9] The corpus is freely available to download at http://ucrel.lancs.ac.uk/textreuse/uppc.php and through Lancaster's DOI http://dx.doi.org/10.17635/lancaster/researchdata/67

[10] the term 'suspicious' has been used here instead of 'derived' because these text documents are suspected to contain plagiarism (Section 1)

3.2.1 Corpus generation process

The UPPC Corpus is created to mimic the real world paraphrase plagiarism practised by the students in academia. The text documents in the corpus contain examples of heavily paraphrased texts manually written by the university students. To generate example cases, we decided to use the same strategy followed by [Clough and Stevenson, 2011] since it accurately represents plagiarism approaches followed by the students.

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

Table 3.6: List of Wikipedia articles used for UPPC Corpus generation

To create the UPPC Corpus, a set of twenty Urdu articles were selected from Wikipedia, describing well-known people belonging to a variety of disciplines (Table 3.6). Some of them are famous politicians, others are historical leaders and some notable religious figures. The personalities were chosen carefully, such that the source and learning material (used in creating the text documents) could be easily obtained and the volunteers have general knowledge about them, so they can create good quality text documents for the corpus. A passage of size between $200 - 300$ words was excerpted from each Wikipedia article (source text document). The Wikipedia as chosen as a source since it is a large, reliable and open content on-line repository and hence a favourite source for plagiarists [Martınez, 2009].

The aim was to create a resource that, as accurately as possible, reflects dif-

ferent paraphrasing mechanisms (in the plagiarised documents) to effectively check the behaviour of different paraphrase plagiarism detection algorithms. To generate paraphrased plagiarised and non-plagiarised text documents, five volunteers were asked to manually write essays of length $200 - 300$ words. The volunteers were undergrad students, native Urdu language speakers who had a good understanding of paraphrasing mechanisms. Moreover, the students were given a detailed presentation on how to paraphrase a text and what different techniques are used in the process of rewriting a text. Overall, the intention was to create near realistic plagiarism settings. A formal agreement was signed by the volunteers which enabled us to make the corpus publicly-accessible.

These volunteers wrote paraphrase text documents based on the Wikipedia source articles provided to them. They were told to rephrase text from the source article by replacing words with appropriate synonyms and changing the sentence structure but not the meaning (semantics). There were no hard constraints on how to paraphrase or which paraphrase technique to use. The volunteers were encouraged to use their knowledge of how to paraphrase a piece of text. It could include, but not limited to, synonym replacement, changing in tense or grammatical structure, summarising content, and splitting or combining sentence to make new ones.

For non-plagiarised text document writing task, volunteers were provided with the learning materials in the form of on-line references, essays, and books written on each of the personalities that could be used to generate the text document. They were encouraged to use their knowledge or obtain help from the material provided (or their sources) but strictly required not to use Wikipedia.

Personality	PP	NP
Chaudhry Rehmat Ali	3	3
Muhammad (PBUH)	5	3
Liaquat Ali Khan	4	4
Mirza Ghalib	4	3
Tipu Sultan	4	3
Abdul Qadeer Khan	3	3
Muhammad Ali Jinnah	4	4
Nusrat Fateh Ali Khan	3	3
Benazir Bhutto	4	4
Fatima Jinnah	3	3
Rashid Minhas	3	4
Aafia Siddiqui	3	3
Queen Victoria	4	3
Zaynab bint Ali	4	4
Sher Shah Suri	4	4
Bulleh Shah	4	3
Bill Gates	4	3
Zulfikar Ali Bhutto	4	3
Allama Iqbal	4	3
Umar ibn Al-Khattab	4	2
Total	**75**	**65**

Table 3.7: Number of paraphrased plagiarised (PP) and non-plagiarised (NP) documents in the UPPC Corpus

3.2.2 Corpus properties and statistics

The UPPC Corpus has been saved in standard XML format and made freely available to download[11]. It contains 160 text documents in total, 20 original Wikipedia sources, 75 paraphrased plagiarised text documents and 65 non-plagiarised text documents. Table 3.7 lists the number of text documents in the corpus with respect to the personalities and plagiarism type. It is of reasonable size with 48,387 words

[11] http://ucrel.lancs.ac.uk/textreuse/uppc.php

(tokens) in total[12] and 6,201 unique words (types). Table 3.8 highlights detailed statistics of the UPPC Corpus.

The UPPC Corpus texts include typos (spelling and grammatical errors) written by the volunteers. This emphasises the fact that in the real world scenario when a plagiarist reuses a piece of text, he/she paraphrases it with his/her understanding and knowledge of the language. Moreover, it would be interesting to see the behaviour of plagiarism detection systems on these typographical errors.

Whole Corpus Statistics	
Number of Text Documents	160
Sentence Count	2,711
Word Count	46,729
Word Count (after stop-word removal)	27,076
Unique Word Count	6,201
Plagiarised Text Documents Statistics	
Number of Text Documents	75
Sentence Count	1,134
Word Count	18,247
Word Count (after stop-word removal)	10,647
Non-Plagiarised Text Documents Statistics	
Number of Text Documents	65
Sentence Count	1,341
Word Count	23,978
Word Count (after stop-word removal)	13,676

Table 3.8: UPPC Corpus statistics

3.2.3 Examples of plagiarised and non-plagiarised text documents

This section presents and discusses an example passage from each of the plagiarised and non-plagiarised text document from the UPPC Corpus.

[12] Compound words (or multi-word expressions) are counted as single words

3.2.3.1 Example of paraphrased plagiarised text document

Wikipedia Source Text
مرزا غالب 1797– 1869 اردو زبان کے سب سے بڑے شاعر سمجھے جاتے ہیں۔ ان کی عظمت کا راز صرف ان کی شاعری کے حسن اور بیان کی خوبی ہی میں نہیں ہے۔ ان کا اصل کمال یہ ہے کہ وہ زندگی کے حقائق اور انسانی نفسیات کو گہرائی میں جا کر سمجھتے تھے اور بڑی سادگی سے عام لوگوں کے لیے بیان کر دیتے تھے۔ غالب جس پر آشوب دور میں پیدا ہوئے اس میں انہوں نے مسلمانوں کی ایک عظیم سلطنت کو برباد ہوتے ہوئے اور باہر سے آئی ہوئی انگریز قوم کو ملک کے اقتدار پر چھاتے ہوئے دیکھا۔
Paraphrased Plagiarised Text
اردو زبان کے سب سے بڑے سمجھے جانے والے شاعر کا نام مرزا غالب ہے آپ کی پیدائش 1797 میں اور وفات 1869 میں ہوئی۔ مرزا غالب کی عظمت کا راز صرف اُن کی شاعری کے حسن و بیان کی خوبی ہی میں نہیں ہے۔ اُن کا اصل کمال تو یہ ہے کہ وہ زندگی کے حقائق اورانسانی نفسیات کو گہرائی میں جا کر سمجھتے تھے۔ اور عام زبان میں لوگوں کی سمجھ کے مطابق بیان کر دیتے۔ غالب جس مشکل دور میں پیدا ہوئے آپ نے تب مسلمانوں کی عظیم سلطنت کوزوال پذیر ہوتے اور انگریز قوم کو قابض ہوتے دیکھا۔

3.2.3.2 Example of non-plagiarised text document

Wikipedia Source Text
شادی کے بعد انہوں نے اپنے آبائی وطن کو خیر باد کہہ کر دہلی میں مستقل سکونت اختیار کرلی۔ شادی کے بعد مرزا کے اخراجات بڑھ گئے اور مقروض ہو گئے۔ آخر مالی پریشانیوں سے مجبور ہو کر غالب نے قلعہ کی ملازمت اختیار کرلی اور 1850ء میں بہادر شاہ ظفر نے مرزا غالب کو نجم الدولہ دبیرالملک نظام جنگ کا خطاب عطا فرمایا، اور خاندان تیموری کی تاریخ لکھنے پر مامور کر دیا اور 50 روپے ماہور مرزا کا وظیفہ مقرر ہوا۔
Paraphrased Plagiarised Text

10 اگست 1810ء کو جب غالب کی عمر صرف 13برس تھی اُن کی شادی امراؤ بیگم سے کر دی گئی۔ اُن کی شادی دہلی میں ہوئی تھی اس لیے انہوں نے دہلی میں سکونت اختیار کر لی۔ اس زمانہ میں دہلی میں ملک کے بہترین اہل فضل وکمال جمع تھے شاعری کا شوق ان کو بچپن سے ہی تھا دہلی کی فضاء نے ان کی شاعری کو نکھارنے اور پروان چڑھانے میں بڑی مدد دی۔ مرزا غالب کی مالی حالت کبھی بھی بہتر نہ تھی اُن کی شاہ خرچیوں نے انہیں مقروض کر رکھا تھا۔ اِن کی عمر کا بیش تر حصہ آبائی پنشن بحال کروانے میں گزرا مگر بے سود۔ 1850ء میں آخری مغل بادشاہ بہادر شاہ ظفر نے انہیں آل تیمور کی تاریخ لکھنے پر مامور کیا اور پچاس روپے ماہوار وضیفہ مقرر کر دیا۔

Sections 3.2.3.1 and 3.2.3.2 show example passages from paraphrase plagiarised and non-plagiarised text documents of the corpus (of personality Mirza Ghalib) along with their sources (Wikipedia). From the plagiarised example, it is obvious that a number of obfuscation strategies were employed to paraphrase the source text. The first sentence of plagiarised text example (Section 3.2.3.1) shows a shift in tense. Furthermore, the source sentence is split into two sentences. Similarly, the last two sentences demonstrate synonym replacement and involve complex paraphrasing while a small chunk of the passage is reused verbatim. This also demonstrates that rewriting varies and depends on the volunteer.

For the non-plagiarised example, the rewritten passage is independently constructed of the source (although the same words may still occur in both) and has been extended to include additional information. For example, at the start of the non-plagiarised text example (Section 3.2.3.1), the rewritten text adds new contextual information (i.e. why he got shifted to Delhi). Furthermore, sentences from both passages share the content of the same event but neither of them shares any similarity or have the same meaning.

3.3 English-Urdu text reuse corpus

The TREU (Text Reuse English Urdu) Corpus is developed on the footsteps of COUNTER (Section 3.1) [Sharjeel et al., 2017] and METER (Section 2.2.1.3.1) [Clough et al., 2002] corpora, i.e., compiling data from journalism. The corpus is comprised of cross-lingual English-Urdu real cases of text reuse at the document level. The source text documents in the corpus are in the English language while the derived text documents are in the Urdu language. For source text documents, the English news reports released by the news agencies are used. The derived text documents, on the other hand, are Urdu newspaper stories published in the popular Urdu newspapers of Pakistan. Each of the news agency reports (English text) has a one-to-one mapping with the newspaper story (Urdu text), but as practised in journalism, the newspaper story may or may not contain text from the news agency report.

The TREU Corpus contains a total of 2,257 source-derived text document pairs (total 4,514 text documents). These pairs are divided into three categories, i.e. (1) Wholly Derived (WD), when the derived text document is the mere translation (with small changes due to language structure) of the source text document (verbatim copy), (2) Partially Derived (PD), when the derived text document is the paraphrased version of the translated source text document, and, (3) Non Derived (ND), when the text document is independently written without referring to the source text document.

As far as we are aware, TREU Corpus is the first of its kind cross-script cross-lingual standard evaluation resource developed for the text reuse detection research for the English-Urdu language pair. We believe that the corpus will serve as a benchmark for the evaluation of the state-of-the-art cross-lingual text reuse detection methods in general, and more specifically, for the English-Urdu (or similar) language pair. Moreover, it can also facilitate in developing algorithms that can detect cross-

script cross-lingual text reuse at the document level. The corpus is released as a free to download[13] resource for research purposes with an aim to promote text reuse detection research in the English-Urdu language pair.

3.3.1 Corpus generation process

As with our successful method with the COUNTER Corpus (Section 3.1), to construct the TREU Corpus, we decided to use the journalistic text. The idea was further motivated by the fact that a large amount of journalistic text is freely available and a lot easier to extract in electronic form, especially for the Urdu language. Moreover, borrowing text from the news agency to compose newspaper stories is a well-known practice in journalism. It is a routine drill for the journalists to formulate a news story by using the press report released by the news agency either directly (verbatim) or by rephrasing (paraphrase) it [Clough, 2003, Wilks, 2004]. In addition, it would be interesting to investigate the behaviour of state-of-the-art cross-lingual text reuse detection methods on these real examples of reuse.

The TREU Corpus has two types of text documents: source text documents in the English language and derived text documents in the Urdu language. To create source text documents, the press reports released by two well-known news agencies of Pakistan, i.e., Associated Press of Pakistan (APP) and Independent News Pakistan (INP) were used. The subscription was established with both news agencies to receive the English news reports daily in the email. On the other hand, the derived text documents were hand-picked from the Urdu news stories published in the top four large circulation national dailies of Pakistan, i.e., Nawa-i-Waqt, Daily Express, Daily Pakistan, and Daily Jang. The newspaper stories were collected manually over a period of 12 months (from July 2015 to June 2016). The news text collection was carried out throughout each month excluding the public holidays on which either

[13] https://github.com/muhmmadsharjeel/PhD-Work

the newspaper was not published, or the news agency did not provide the service. To have variation in the data, the news data was collected across National, Foreign, Domestic, Sports, and Business domains. Table 3.9 shows the distribution of text documents in the TREU Corpus.

News Agencies		News Papers		Domains	
APP	2,015	Nawa-e-Waqt	1,525	National	1,127
INN	242	Daily Express	663	Foreign	538
		Daily Jang	57	Domestic	339
		Daily Pakistan	12	Sports	225
			55	Business	28

Table 3.9: Distribution of documents by news agencies, newspapers and domains in the TREU corpus

3.3.2 Corpus properties and statistics

The TREU Corpus contains a total of 4,514 text documents (2,257 source and 2,257 derived text documents). It is substantially large in size and contains in total 1,009,069 (approx. one million) words (tokens), out of which 486,264 are English and 522,805 are Urdu words. The average length of an English source text document is 215 words while for Urdu derived text document it is 231 words. Table 3.10 show detailed statistics of the corpus.

3.3.3 Annotations and inter-rater agreement

Two human annotators performed the annotations of the TREU Corpus with the help of a linguist. Both the annotators were postgraduate NLP students, native speakers of the Urdu language, who studied English as a foreign language and as the language of instruction throughout their academic career. Furthermore, they were provided with training about the journalistic text reuse phenomena and with

	Source	Derived
Total number of documents	2,257	2,257
Total number of words	486,264	522,805
Average number of words per document	215	231
Total number of types	24,105	17,736
Smallest document (by words)	25	26
Largest document (by words)	1,799	2,404
Number of documents $>$ 1000 words	9	33
Number of documents $>$ 500 but $<$ 1000 words	124	139
Number of documents $>$ 100 but $<$ 500 words	1,623	1,564
Number of documents $<$ 100 words	486	512

Table 3.10: TREU Corpus statistics

tutorials on different text rewriting operations by the linguist.

As a first step, an annotation scheme was prepared under the guidance of the linguist. Following is the annotation scheme used to tag a text document pair in one of the three classes (i.e., WD, PD, or ND).

Wholly Derived (WD): A text pair will be tagged as 'WD' if the derived text is almost an exact translation of the source text. However, due to the cross-lingual setting, small changes appearing in the derived text will be ignored. Additionally, a small amount of new text may also appear in the derived text due to the structural difference in both languages.

Partially Derived (PD): For a text pair to be tagged as 'PD', its contents must be semantically the same, describing the same information. However, the derived text must not be mere translations of the source text. Rather, the source text should be paraphrased using different text editing operations including (but not limited to) word or sentence re-ordering, merging or splitting of sentences, insertions or deletions of new text, replacing words or phrases with appropriate synonyms, and expansion or compression of text etc.

Non Derived (ND): To tag a text pair as 'ND', the context of the news should

be the same or both texts must be describing the same event. However, the derived 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 derived text with completely different facts and figures.

Annotations were performed in multiple phases. In the first phase, based on the annotation scheme, a random subset of 50 document pairs were annotated by the two annotators and the linguist. The results of each annotator were compared with the linguist and conflicting pairs were discussed with them individually. Moreover, the annotation scheme was re-examined after the discussion to make a few changes. In the second phase, another subset of 250 document pairs was now annotated by the two human annotators according to the revised scheme. The results were reviewed by the linguist again and it was observed that the rate of conflicts had dropped.

During the third phase, the two annotators manually tagged the remaining 1,957 document pairs and the results were saved. Both annotators agreed on 1,919 and disagreed on 338 document pairs. The final Inter-Annotator Agreement (IAA) score on the entire corpus is 85.02%, and the Cohen's Kappa score was computed to be 0.77% (Unweighted), 0.82% (Linear weighting), 0.87% (Quadratic weighting) [Cohen, 1960, Cohen, 1968]. As can be noted, these scores are of substantial level considering the difficulty of the annotation task. In the last phase, the conflicting 338 pairs were given to the journalist for conflict resolution. The decisions of the third annotator were considered final.

The final gold standard corpus contains 2,257 text document pairs, out of which 672 are WD, 888 are PD, and 697 are ND. Table 3.11 lists the classification of text documents in the TREU Corpus and compares it with the METER (Section 2.2.1.3.1) [Clough et al., 2002] and COUNTER [Sharjeel et al., 2017] corpora.

Classification	TREU	COUNTER	METER
WD	672 (29.7%)	135 (22.5%)	301 (31.8%)
PD	888 (39.3%)	288 (48.0%)	438 (46.3%)
ND	697 (30.8%)	177 (29.5%)	206 (21.7%)

Table 3.11: Classification of text document pairs in the TREU Corpus and its comparison with METER [Clough et al., 2002] and COUNTER corpora [Sharjeel et al., 2017]

3.3.4 Examples of reuse cases from the corpus

This section presents representative WD, PD, and ND examples from the TREU Corpus.

3.3.4.1 Example of WD source and derived text documents

The following example shows a WD text document pair from the corpus. It can be noted that the reused text is almost the exact translation of the source text. Moreover, the order of information is also preserved. However, a very small amount of information is added/removed in the reused text document due to the language structural change. Furthermore, the source text document has one sentence (*Principal Staff Officers and a large number of Airmen attended the ceremony*) that is not present (reused) in the derived text document.

Source text document

PAF observes Martyrs' Day across the country. Pakistan Air Force (PAF) on Sunday observed September 07 as Martyrs' Day at all PAF Bases throughout the country. A ceremony was also held here at Air Headquarters, in which Chief of the Air Staff, Pakistan Air Force (PAF) Air Chief Marshal Tahir Rafique Butt laid floral wreath and offered "Fateha" at the Martyrs' Monument. Principal Staff Officers and a large number of Airmen attended the ceremony. The day started with special Du'aa and Quran Khawani for the

Shuhada of 1965 and 1971 wars and those who laid down their lives in action since creation of Pakistan. A similar ceremony was held at Karachi, where a PAF contingent led by Air Vice Marshal Azhar Hasan Rizvi, Air Officer Commanding, Southern Air Command offered "Fateha" and laid floral wreath at the grave of Pilot Officer Rashid Minhas Shaheed (Nishan-e- Haider) on behalf of Chief of the Air Staff, Pakistan Air Force. Wreaths were also laid on the graves of PAF martyrs throughout the country.

Derived text document

ملک بھر میں پاک فضائیہ نے گذشتہ روز اپنے تمام فضائی اڈوں پر 7 ستمبر یوم شہداء کے طور پر منایا۔ ائر ہیڈکوارٹرز اسلام آباد میں ایک سادہ مگر پروقار تقریب منعقد کی گئی جس میں پاک فضائیہ کے سربراہ ائر چیف مارشل طاہر رفیق بٹ نے شہداء کی یادگار پر پھولوں کی چادر چڑھائی اور فاتحہ خوانی کی۔ دن کا آغاز 1965 اور 1971 کی جنگوں کے پاک فضائیہ کے شہداء اور ان تمام افراد کے لئے خصوصی دعاؤں اور قرآن خوانی سے کیا گیا جنہوں نے پاکستان کے معرض وجود میں آنے سے لے کر اب تک پاکستان کے لئے اپنی جانیں قربان کیں۔ اسی طرح فضائیہ کے ایک دستے نے ائر وائس مارشل اظہر حسن رضوی، ائر آفیسر کمانڈنگ، سدرن ائر کمانڈ کی سرکردگی میں پاک فضائیہ کے سربراہ ائر چیف مارشل طاہر رفیق بٹ کی جانب سے پائلٹ آفیسر راشد منہاس شہید (نشانِ حیدر) کی قبر پر فاتحہ خوانی کی اور پھولوں کی چادر چڑھائی۔ ملک بھر میں پاک فضائیہ کے شہداء کی قبروں پر بھی پھولوں کی چادریں چڑھائی گئیں۔

Derived text document (translation)

Yesterday Pak Air Force observed September 7th as Martyrs Day at all its airports throughout the country. A simple but inauspicious ceremony was held at the Air Headquarters Islamabad in which Air Chief Marshal Tahir Rafiq Butt laid a floral wreath and offered "Fateha'' at the Martyrs' Monument. The day began with special prayers and Quranic recitation for the Pak Air Force martyrs of 1965 and 1971 and all those who have sacrificed their lives since the creation of Pakistan. Similarly, an Air Force team led by Air Vice Marshal Azhar Hassan Rizvi, Air Officer Commanding, Southern Air Command, offered "Fateha" and laid floral

wreath at the grave of Pilot Officer Rashid Minhas Shaheed (Nishan-e-Haider) on behalf of Chief of the Air Staff, Pakistan Air Force. Wreaths were also laid on the graves of PAF martyrs throughout the country.

3.3.4.2 Example of PD source and derived text documents

The example below shows a cross-lingual PD text reuse example from the corpus. It is worth noting that sentences (or phrases) have been reordered to generate the derived text. The information at the start in the source text document is added (after paraphrasing) at the end of the derived text document. Moreover, some extra details have been added in the derived text (which may be based on the journalist's observations), i.e., the name of the person who offered the Namaz-e-Janaza. The source text document has general information (*representatives of MQM*) whereas the derived text document has more detailed and specific information, i.e., actual names of the representatives. Furthermore, some words have been replaced with appropriate synonyms. These changes highlight the fact that different editing operations have been used by the journalists in formulating the newspaper story. However, while creating the derived text, meanings of the source text have been preserved.

Source text document
Allama Akbar Kumaili laid to rest. Allama Ali Akbar Kumaili was laid to rest at Ali Bagh graveyard, Lyari here on Sunday. Earlier, his Namaz-e-Janaza was offered at Imambargah Shah-e-Khorasan, Naumaish Chowrangi after Zohar prayer. Large number of people, religious scholars, representatives of MQM attended the funeral prayers. Ali Akbar Kumaili, son of Allama Abbas Kumaili and his security guard were shot dead by unknown armed men in Azizabad area on Saturday evening.
Derived text document

کراچی کے علاقے عزیزآباد ہنگوریہ گوٹھ میں گذشتہ شام موٹرسائیکل سواروں کی فائرنگ سے جاں بحق ہونے والے معروف عالم دین علامہ عباس کمیلی کے بیٹے علامہ علی اکبر کمیلی کو نماز جنازہ کی ادائیگی کے بعد باغ علی لیاری کے قبرستان میں سپرد خاک کر دیا گیا۔ نماز جنازہ نمائش چورنگی ایم اے جناح روڈ پر ادا کی گئی جس میں ایم کیو ایم کے رہنما ڈاکٹر صغیر احمد، فیصل سبزواری، حیدر عباس رضوی، بابر غوری، رؤف صدیقی، پیپلزپارٹی کے وقار مہدی، سہیل عابدی سمیت سینکڑوں افراد نے شرکت کی۔ نمازجنازہ علامہ رضی جعفر نے پڑھائی۔ اس موقع پر سکیورٹی کے سخت انتظامات کئے گئے تھے۔

Derived text document (translation)

Allama Ali Akbar Kameli, son of prominent scholar Allama Abbas Kameli, who was killed by motorcyclists in Azizabad Hangoria Goth area of Karachi, last evening, was buried in Bagh Ali Lyari cemetery after the funeral prayer. Hundreds of people, including MQM leaders Dr Saghir Ahmed, Faisal Sabzwari, Haider Abbas Rizvi, Babar Ghauri, Rauf Siddiqui, Peoples Party's Waqar Mehdi, Sohail Abidi attended the funeral prayer offered at the Chorangi MA Jinnah Road. Allama Raza Jaffer led the funeral prayer. Strict security arrangements were made on the occasion.

3.3.4.3 Example of ND source and derived text documents

The example that follows shows an ND text document pair from the corpus. Both source and derived texts are describing the same news event, i.e., proceedings of a Senate meeting and the walkout of members from the meeting. However, the explanation of the event and the way of expressing it is entirely different in the source and derived text documents. In the source text document, two members (*Haji Adeel and Zahid Khan*) are requesting the Deputy Chairman to adjourn the proceedings whereas the derived text states it was requested collectively by the opposition members. Furthermore, in the source text, it is mentioned that the meeting was adjourned for half an hour while the derived text details that it was

restarted after half an hour but postponed again until Friday. In addition to this, the information is very compressed in the source text whereas the event has been reported in greater depth in the derived text document. This shows that reused text is generated independently of the source text and any overlap of words (phrases) is very low (mainly stop-words are common) between the text pair.

Source text document

Opposition stages walkout from Senate. The united opposition on Wednesday staged walkout from the Senate to protest what they said the absence of treasury members especially ministers. Speaking in Senate on Point of Order Haji Adeel of Awami National Party (ANP) pointed out that the ministers do not give importance to the Upper House. He said ANP is with the government for the sake of democracy. Senator Zahid Khan of ANP pointed the lack of quorum. Deputy Chairman adjourned the proceedings of the for half an hour.

Derived text document

سینٹ میں وزیراعظم نوازشریف اور مسلم لیگ (ن) کی حکومت کے خلاف شیم شیم کے نعرے بلند کئے گئے۔ یہ صورتحال اس وقت پیدا ہوئی جب سینٹ کا اجلاس سوا گھنٹے کی تاخیر سے بمشکل شروع ہو سکا۔ اس کے باوجود سرکاری بنچوں پر ایک بھی رکن موجود نہیں تھا۔ اپوزیشن نے سرکاری ارکان کی عدم موجودگی کے خلاف ایوان سے واک آوٹ کیا جبکہ اس دوران کورم بھی پورا نہیں تھا اور اجلاس کی کارروائی آدھ گھنٹے کیلئے ملتوی کر دی گئی۔ آدھ گھنٹے کے بعد بھی کورم پورا نہیں ہو سکا تو اجلاس جمعہ کے روز تک ملتوی کر دیا گیا۔ حکومت کی اتحادی جے یو آئی (ف) نے بھی واک آوٹ میں حصہ لیا اور ڈپٹی چیئرمین سے کہا کہ آپ یہاں کیوں بیٹھے ہیں آپ بھی باہر نکلیں۔ پارلیمانی امور کے وزیر مملکت آفتاب شیخ اس دوران ایوان میں آئے اور انہوں نے کہا کہ ہمیں اپنی ذمہ داری پوری کرنی چاہئے۔ ڈپٹی چیئرمین صابر بلوچ نے کہا کہ ہم وزراء کی عدم موجودگی کا رونا روتے تھے لیکن اب ایک بھی سرکاری رکن یہاں موجود نہیں ہے۔ یہ صورتحال باعث افسوس ہے۔ حکومتیں اس طرح نہیں چلا کرتیں۔

Derived text document (translation)

In the Senate, the slogans of 'sheim sheim' were raised against Prime Minister Nawaz Sharif and the PML-N government. This situation arose when the meeting of the Senate could barely begin with an hour and a half delay. Yet, there was not a single member on the government benches. The opposition walked out of the House against the absence of government members, moreover, during which even the quorum was not full, and the meeting was adjourned for half an hour. After half an hour, the quorum was still not full, so the meeting was adjourned till Friday. Government ally, the JUI-F also took part in the walkout and told the deputy chairman why you are sitting here, you should also walkout. During this, Minister of State for the Parliamentary Affairs Aftab Sheikh came to the House and said that we should fulfil our responsibility. Deputy Chairman Sabir Baloch said that we used to weep over the absence of ministers but now no government member is here. This situation is worrisome. Governments do not work this way.

3.4 English-Urdu parallel corpus

A parallel corpus is defined as a large collection of texts aligned at the sentence level, in two or more languages, that are exact translations of each other [Resnik and Smith, 2003]. It is an essential resource for bi-lingual and multi-lingual NLP research and especially for training automatic MT systems [Hutchins, 2005, Steinberger et al., 2014, Fantinuoli and Zanettin, 2015]. More specifically, for Statistical Machine Translation (SMT) methods, a large-scale good quality parallel corpus is essential to get useful results [Callison-Burch et al., 2004, Koehn, 2005, Eisele and Chen, 2010]. Such a resource is also valuable for contrastive linguistics [Ebeling, 1998] and has applications in CLIR [Nie et al., 1999], bi-lingual lexicon induction [Caseli et al., 2006, Apidianaki, 2008], Word Sense Disambiguation (WSD) [Kazakov

and Shahid, 2013], and cross-lingual text reuse and extrinsic plagiarism detection [Barrón-Cedeño et al., 2008].

Despite applications in a large number of NLP tasks, these corpus resources are very scarce for low-resource language pairs such as English-Urdu. The available English-Urdu parallel corpora are very small in size, domain-specific, not freely available, or are of not so good quality [Baker et al., 2002, Jawaid and Zeman, 2011, Post et al., 2012].

A good quality English-Urdu parallel corpus was desired to automatically extract translation pairs for a bi-lingual dictionary (Section 3.5) which was further used in the cross-lingual text reuse detection experiments (Section 5.1.2). Accordingly, the already available English-Urdu parallel data resources were reviewed as well as new data collected for a large-scale multi-domain English-Urdu parallel corpus. The subsequent sections highlight some of the existing and newly developed English-Urdu parallel data resources.

3.4.1 Existing English-Urdu parallel corpora

Several English-Urdu parallel corpora are publicly available on the Web. However, most of them contain noisy data or have alignment issues. A sentence alignment tool[14], developed in Java as part of this thesis work, was used to correct these issues. The following sections list each of the already available English-Urdu parallel corpora.

OS-18: The Open Subtitles parallel corpus[15] [Lison and Tiedemann, 2016] is available from the OPUS[16] (open source parallel corpus) project. It contains translated subtitles of movies in 62 languages, including Urdu. However, the text is translated using automatic MT systems and is of poor quality, not properly

[14] The tool is available at https://github.com/muhmmadsharjeel/PhD-Work
[15] http://www.opensubtitles.org
[16] http://opus.nlpl.eu/index.php

aligned at the sentence level, and contains a lot of noise.

Tatoeba: Tatoeba[17] (Japanese word which means "for example") website is a collaborative platform that offers a freely available collection of sentences and their translation in many languages. The translations are done by volunteers and contain grammatical errors.

CLE-PC: Centre for Research in Urdu Language Processing (CRULP) (now Center for Language Engineering CLE[18]) has developed an English-Urdu-Nepali parallel corpus[19]. They use the English text from a subset of the Penn Treebank (PTB) project which contains stories from the Wall Street Journal (WSJ) collection [Marcus et al., 1993]. It was then manually translated into Urdu and Nepali by a team from the CRULP. However, due to the licensing issues (licensed under Linguistic Data Consortium (LDC)), there is only a subset of the corpus freely available to download.

IPC: The Indic Parallel Corpus[20] is a collection of Wikipedia documents of six Indian low resourced languages (i.e., Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu) translated into the English language through crowdsourcing via the Amazon Mechanical Turk[21] platform [Post et al., 2012]. For each language, the 100 most visited Hindi Wikipedia pages on different topics were manually translated by four MTurk workers. The translation pairs are of mixed quality as they are created by amateurs.

EURPC: The English-Urdu Religious Parallel Corpus[22] contains publicly available English and Urdu translations of verses from the Holy Quran and Bible [Jawaid

[17] https://tatoeba.org/eng
[18] http://www.cle.org.pk
[19] http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm
[20] https://github.com/joshua-decoder/indian-parallel-corpora
[21] https://www.mturk.com/
[22] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2582

and Zeman, 2011]. Although the corpus texts are domain-specific, they are manually aligned and properly tokenised parallel sentences.

Table 3.12 shows the number of parallel sentences present in each of the corpora discussed above.

Source	Sentences
OS-18	27,540
Tatoeba	1,274
CLE-PC	3,863
IPC	41,244
EURPC	14,371
Total	88,292

Table 3.12: Number of parallel sentences in existing English-Urdu parallel corpora

3.4.2 Newly Proposed English-Urdu parallel corpus

The proposed English-Urdu Parallel Corpus (EUPC-20) consists of sentences extracted from parallel documents collected from different online sources. It is a first considerably large and publicly available collection that comes from several domains such as religion, technology, politics, literature, and Wikipedia. Although the corpus data comes from the Web, the translation pairs are of excellent quality as they are manually translated mostly by experts. The texts in the corpus have been pre-processed, normalised, correctly aligned at sentence-level, and saved in a standard format. The proposed resource improves on the already available ones (Section 3.4.1) as it is large, good quality, multi-domain, and available to download for NLP research.

3.4.2.1 Corpus generation process

To construct EUPC-20, the Web was exploited as a potential resource. The Web by far is the largest multi-lingual resource with plenty of parallel texts readily available.

Different websites, blogs, and social media pages were searched to find translated text documents in both English and Urdu languages. With a list of web links in hand, the data were either copied manually or the parallel web documents were scraped using a script written in Python[23].

The following discussion gives details of all the online sources from where the parallel data was collected for the EUPC-20.

Bible translation: The Bible has been translated into many languages and these translations are easily available on the Web. WordProject[24] is one of the most authentic online sources that provide the translations of the Bible in more than 60 international languages. The English-Urdu translation of the Bible was downloaded from the WordProject website.

HRCP reports: The Human Rights Commission of Pakistan[25] (HRCP) website publishes, apart from English, Urdu translations of some of their press releases, reports, and articles. A web scraper was used to download only those reports which were available in both English and Urdu languages.

MIT-TRP magazine: The Massachusetts Institute of Technology (MIT) Technology Review Pakistan[26] (TRP) is an online magazine that produces Urdu translations of the English articles originally published by the MIT Technology Review[27] website. The articles are related to the latest trends in the science and technology domain and are manually translated into the Urdu language by a team at Information Technology University (ITU), Pakistan[28]. All the articles were scraped from the website for which both English and Urdu versions were available.

[23] The script is available at https://github.com/muhmmadsharjeel/PhD-Work
[24] https://wordproject.org/bibles/ur/
[25] http://hrcp-web.org/hrcpweb/
[26] http://www.technologyreview.pk
[27] https://www.technologyreview.com
[28] http://itu.edu.pk

Tanqeed blog: Tanqeed[29] is an electronic blog related to Pakistani politics and culture. It posts stories and essays highlighting social issues from Pakistan and South Asia in both English and Urdu languages. The manually translated Urdu stories are provided by a team of authors of the blog. The English-Urdu stories were manually extracted from the blog to be used in the corpus.

TED talks transcripts: The famous TED talks[30] website provides translated subtitles of videos in 100+ languages. However, very few translated transcripts are available in the Urdu language. Using Ted2Srt.org[31], all the available English-Urdu translations of the TED transcripts were extracted.

Daniel Pipes stories: Daniel Pipes, the president of 'Middle East Forum'[32], posts articles on American foreign policy and the Middle East on his website[33]. He publishes his stories in 38 international languages including Urdu. The website is very popular and has around 69 million page views. English stories and their Urdu translations were manually copied from the website.

General sentences: Urdu2eng.com[34] is an English language learning website that contains Urdu to English translations of dialogues, interviews, debates, idioms, and general sentences of everyday use. These short sentences and their translations are publicly available and are of excellent quality. All the English and their corresponding Urdu sentences were scraped from the website. Moreover, some of the Urdu Wikipedia articles, on some famous personalities, were manually translated (through two graduate students) into the English language.

[29] http://www.tanqeed.org
[30] http://www.ted.com
[31] https://ted2srt.org
[32] https://www.meforum.org
[33] http://www.danielpipes.org
[34] http://www.urdu2eng.com

Novels: Saadat Hassan Manto was a renowned Urdu poet and writer who is acknowledged as one of the finest writers in Urdu literature. He bravely highlighted the bitter aspects of society in his writings. Manto's famous short stories are available in both English[35] and Urdu[36] languages. Pir-e-Kamil by Umera Ahmad is a famous novel, originally published in the Urdu language[37] and later translated by the author into the English language[38]. Both of these authors' works were obtained in English and Urdu languages to be included in the corpus.

3.4.2.2 Corpus properties and statistics

Table 3.13 shows the total number of parallel sentences compiled from each of the online sources described above.

Source	Sentences
Bible translations	31,037
HRCP reports	2,101
MIT TRP magazine	6,011
Tanqeed blog	3,363
TED talks transcripts	10,861
Daniel Pipes stories	8,096
General sentences	1,758
Novels	17,204
Total	80,431

Table 3.13: Number of parallel sentences collected from online sources

As the data were scraped or manually extracted from different online sources, they contain a lot of noise (especially the Urdu text) and had alignment issues.

[35] https://zjeddy.wordpress.com/
[36] https://mantonama.wordpress.com
[37] http://bit.ly/peer-e-kamil-urdu
[38] http://bit.ly/peer-e-kamil-english

Moreover, some of the data was not properly encoded in Unicode (UTF-8) format. Therefore, the downloaded text was first pre-processed to remove HTML tags and garbage characters. The text was further normalised to remove Urdu diacritics and standardised to Unicode (UTF-8) characters. Then, automatic sentence alignment was performed using a script written in Java[39].

During automatic alignment, a large number of alignment issues were faced. These issues were mostly related to wrongly placed punctuation marks and Urdu sentence boundary detection. English sentences were ending on 'full-stop' while Urdu sentence had a 'comma'. Some of the English sentences had exclamation marks whereas Urdu text had no punctuation mark. One of the major problems was the detection of sentence boundaries for the Urdu sentences as a considerable number of them were not ending on standard termination markers. These alignment issues were removed manually and the corpus has been saved in a standard format for public release.

Table 3.14 reports detailed statistics of the final set of parallel sentence data for the EUPC-20 Corpus. It contains 154,258 properly aligned English-Urdu parallel sentences extracted from different online sources (72,566 parallel sentences, Section 3.4.2) combined with already available parallel data (81,692 parallel sentences, Section 3.4.1) [Jawaid and Zeman, 2011, Post et al., 2012, Lison and Tiedemann, 2016]. These statistics are reported after applying standard pre-processing, text normalisation, removal of punctuation marks, deleting extra white spaces, and special characters. Furthermore, long (> 40 words) and duplicate sentences were filtered from each source.

[39] The script is available at https://github.com/muhmmadsharjeel/PhD-Work

Corpus	Sentences	Tokens		Vocabluary	
		English	Urdu	English	Urdu
Newly compiled data					
Bible translation	26,803	597,707	608,146	25,361	13,506
HRCP website	1,983	27,214	34,091	4,496	4,883
MIT TRP website	5,704	79,525	99,909	10,074	12,238
Tanqeed blog	3,002	53,293	59,906	9,484	10,221
TED Talks transcripts	9,886	136,884	159,068	13,117	13,253
Daniel Pipes articles	6,454	104,402	134,181	16,822	19,444
General sentences	1,737	20,327	22,075	4,442	3,671
Novels	16,997	163,493	198,087	13,143	12,623
Sub-total	72,566	163,493	198,087		
Existing data					
OS-18	26,810	164,183	186,815	10,686	12,143
Tatoeba	1,274	8,098	9,294	1,934	1,655
CLE-PC	3,318	60,596	77,272	10,261	11,654
IPC	38,525	478,719	579,863	16,741	34,455
EURPC	11,765	244,796	264,223	14,934	14,798
Sub-total	81,692	163,493	198,087		
Grand total	154,258	2,139,237	2,432,930		

Table 3.14: Statistics of EUPC-20 Corpus

3.5 English-Urdu bi-lingual dictionaries

Bi-lingual dictionaries are a fundamental and useful resource for computerised language processing tasks [Van Der Eijk et al., 1992]. They potentially provide, for each source language word or phrase, a set of translations in the target language. Moreover, they may also include syntactic information, sense division, usage examples, semantic fields, usage guidelines, etc. These dictionaries are crucial for multiple NLP tasks like CLIR [Pirkola et al., 2001], cross-lingual knowledge induction [Lu et al., 2004, Peirsman and Padó, 2010], and cross-lingual text reuse and plagiarism

detection [Schafer and Yarowsky, 2002].

The manual construction of bi-lingual dictionaries is a labour-intensive task, therefore, a considerable body of work has focused on methods for their automatic induction. Automatic bi-lingual dictionary induction is the task of finding words or phrases across natural languages that share a common meaning. This induction can be approached with a parallel corpus or comparable corpora [Caseli et al., 2006, Li and Gaussier, 2010]. For most language pairs, and most domains, parallel data are either scarce or unavailable, and therefore a range of methods has been proposed to find translations directly from the monolingual text [Shezaf and Rappoport, 2010, Irvine and Callison-Burch, 2013].

As of today, there are a few English-Urdu dictionaries available on the Web but are in PDF format. These are soft copies of their paperback versions. There are others, licensed or freely available but linked with translation programs. For cross-lingual text reuse detection experiments (Section 5.1.2) conducted on the TREU Corpus (Section 3.3), an English-Urdu dictionary was required as a supporting resource. With this in mind, the already available English-Urdu dictionaries were utilised as well as new ones which were assembled by using different approaches.

The sections that follow discuss the individual English-Urdu dictionaries one-by-one. All of these dictionaries are saved in XML format and provided as free to download for academic research purposes[40].

Waseem-Shahab: One of the largest English-Urdu dictionary available on the Web in PDF[41] format is by Waseem Siddiqui and Shahab Alam. The dictionary contains the word, its translation, as well as the most probable POS tag for each entry. The PDF file was converted to raw text, however, the conversion process generated a lot of noisy data. A Java program[42] was used to clean

[40] https://github.com/muhmmadsharjeel/PhD-Work
[41] https://www.scribd.com/doc/11342223/Urdu-to-English-Dictionary
[42] The program is available at https://github.com/muhmmadsharjeel/PhD-Work

the entries with the help of regular expressions. During the cleaning process, some of the data were discarded because of the poor format.

Indic Parallel Corpus: The authors of Indic Parallel Corpus (Section 3.4) have also created, apart from the parallel corpus, a free to use English to Urdu dictionary[43] [Post et al., 2012]. The dictionary was built using crowdsourcing from the Amazon Mechanical Turk[44]. The MTurk workers were tasked to translate English words into the Urdu language. For each English word, three reference sentences were given to the workers, which provided the context of the given word. The dictionary contains words and their translations only.

Babylon: Babylon is a well-known translation program, developed by Babylon Software Limited[45]. The software comes with its proprietary dictionaries but there are many third-party free to download dictionaries also available that can be used with the software. Two such dictionaries, (1) 'English-to-Urdu Lughat'[46] and (2) 'One-click English-to-Urdu Dictionary'[47] were downloaded and converted into raw text format. The data were then cleaned for junk entries. English-to-Urdu Lughat contains the word and its translation only whereas One-click English-to-Urdu Dictionary has a POS tag coupled with each entry as well.

Wiki Data: With the expansion of the Web into a multilingual hub, Wiki websites appear to be one of the favourable resources for extracting translation pairs. Three Wiki services, i.e., Omega[48], Wikipedia[49], and Wiktionary[50] were

[43] http://homepages.inf.ed.ac.uk/miles/babel.html
[44] https://www.mturk.com/
[45] https://www.babylon-software.com/translation_software
[46] https://www.babylon-software.com/free-dictionaries/English_To_Urdu_Lughat/66622.html
[47] https://www.babylon-software.com/free-dictionaries/reference/dictionaries-thesauri/One-Click-Engllish-Urdu-Dictionary-v1.3/51182.html
[48] www.omegawiki.org
[49] https://en.wikipedia.org
[50] https://www.wiktionary.org/

utilised to compile an English-Urdu dictionary. Each entry of the dictionary contains an English word and its Urdu translation.

Ur-GIZA: A well-known method to generate bi-lingual dictionaries is by using a statistical word alignment tool [Aker et al., 2014]. GIZA++ [Och and Ney, 2003, Junczys-Dowmunt and Szał, 2011], a statistical word alignment toolkit uses IBM Models [1−5] [Brown et al., 1993] and HMM [Baum and Petrie, 1966] to map words from a sentence-aligned parallel corpus. It returns words with their possible translation(s) alongside their statistical probabilities. The translation pairs with high probability are assumed to be good. However, due to the probabilistic nature of the method, it returns the wrong translations but with high probabilities. There are various methods proposed in the literature to clean such entries [Aker et al., 2014].

For the creation of Ur-GIZA dictionary, GIZA++ was applied to extract English-Urdu translation phrases from the EUPC-20 Corpus (Section 3.4.2) and a simple filter-based approach was used to clean the wrong entries. A random 100 entries were reviewed to set a threshold value of 0.30. All the entries whose probability was below the threshold were filtered out.

Name	Entries
Waseem-Shahab	60,651
Indic Parallel Corpus	113,911
Babylon	278,453
Wiki Data	3,919
Ur-GIZA	21,819
Total	478,753

Table 3.15: Statistics of English-Urdu bi-lingual dictionaries

Table 3.15 shows total number of entries in each of the English-Urdu bi-lingual dictionaries assembled using different methods.

3.6 Chapter summary

This chapter described the research undertaken in the development of standard evaluation resources for the mono- (Urdu) and cross-lingual (English-Urdu) text reuse and mono-lingual (Urdu) extrinsic plagiarism detection. It also detailed the creation of supporting resources for the cross-lingual (English-Urdu) text reuse detection.

Two Urdu and one English-Urdu standard evaluation corpora are created to promote the text reuse and extrinsic plagiarism detection research in a language that is highly under-resourced, i.e., Urdu. The COUNTER Corpus is a mono-lingual text reuse corpus that contains real examples of Urdu news text reuse. It has 1,200 documents categorised into three levels of reuse, i.e., Wholly Derived, Partially Derived and Non Derived. The UPPC Corpus is an Urdu extrinsic plagiarism corpus that contains simulated cases of mono-lingual plagiarism. The corpus has been created to mimic the real world paraphrase plagiarism practised by the students in academia. It contains 160 text documents, 20 source Wikipedia articles, 75 Paraphrased Plagiarised, and 65 Non-Plagiarised text documents. The TREU Corpus is the first cross-script cross-lingual corpus developed for text reuse detection research in the English-Urdu language pair. It consists of 4,514 documents that are manually tagged into three classes at the document level, i.e., Wholly Derived, Partially Derived, and Non Derived. The corpus is sufficiently representative to serve as a benchmark for developing and evaluating methods for cross-lingual text reuse detection for the English-Urdu language pair.

The chapter also described the creation of two supporting resources for the English-Urdu language pair. The first is EUPC-20, a large-scale multi-domain parallel corpus that contains 154,258 parallel sentences collected from the Web. It contains data from several domains such as religion, technology, politics, literature, and Wikipedia. The second supporting resource is the compilation of several bi-lingual dictionaries for the English-Urdu language pair using different methods

from online and offline sources.

To promote research in the highly under-resourced Urdu language, all the standard evaluation corpora and supporting resources developed in this study are freely available to download for academic research.

"People who copy you will always be one step behind."

Wayne Gerard Trotman

4

Mono-lingual (Urdu) Text Reuse and Extrinsic Plagiarism Detection

Chapter 3 presented details of the two benchmark corpora developed for the mono-lingual (Urdu) text reuse and extrinsic plagiarism detection. The COUNTER Corpus (Section 3.1) [Sharjeel et al., 2017] is an Urdu news text reuse corpus that contains real cases of text reuse from the journalism domain. The UPPC Corpus (Section 3.2) [Sharjeel et al., 2016], on the other hand, is an extrinsic plagiarism corpus that contains manually created simulated cases of Urdu text plagiarism.

This chapter describes Urdu text reuse and extrinsic plagiarism detection experiments performed on the two benchmark corpora i.e., COUNTER Corpus and UPPC Corpus. The chapter aims to make a direct comparison of existing state-of-the-art

mono-lingual text reuse and extrinsic plagiarism detection methods (Section 2.3.1) to investigate their behaviour on the Urdu text. The comparison will enable to understand what methods are most suitable for the Urdu text reuse and extrinsic plagiarism detection tasks. Moreover, it will highlight the strengths and weaknesses of the newly proposed standard evaluation resources. Furthermore, the reason for conducting these experiments on two different types of corpora is to examine how these methods perform on real (COUNTER Corpus) as well as simulated (UPPC Corpus) cases of reuse. As far as we are aware, no previous study has applied these methods and made a detailed comparison for the Urdu language.

The chapter starts with the description of the mono-lingual text reuse and extrinsic plagiarism detection methods and how they were used in the experiments performed on the Urdu text (Section 4.1). After discussing the methods, the experimental setup is presented including corpora, text pre-processing, evaluation methodology and evaluation measures (Section 4.2). Finally, the results of the experiments are presented and discussed (Section 4.3).

4.1 Methods for mono-lingual text reuse and extrinsic plagiarism detection

A range of popular and state-of-the-art mono-lingual methods are applied on both COUNTER (Section 3.1) [Sharjeel et al., 2017] and UPPC (Section 3.2) [Sharjeel et al., 2016] corpora in order to show how the newly proposed standard evaluation resources may be used for the development, evaluation and comparison of text reuse and extrinsic plagiarism detection systems for the Urdu language[1]. The chosen methods are based on different characteristics of text i.e., lexical overlap, string matching, structural similarity, and stylistic similarity (Section 2.3.1). For lexical

[1] Note that same settings are applied for all the methods used in these experiments.

overlap, Word n-grams overlap (Section 2.3.1.1.1) and Vector Space Model (VSM) (Section 2.3.1.1.2) are used. For string matching, Longest Common Subsequence (LCS) (Section 2.3.1.2.1) and Greedy String-Tiling (GST) (Section 2.3.1.2.2) are chosen. For structural similarity, Stop-word n-grams overlap (Section 2.3.1.4.1), and for stylistic similarity, Token ratio (Section 2.3.1.5.2) and Sentence ratio (Section 2.3.1.5.3) are applied. These methods are used to perform a pair-wise comparison of two texts (source and reused text documents in this case), to produce similarity scores based on the features obtained from both texts. A higher score indicates both texts are similar while a low score is the indicator of their dissimilarity [Wise, 1992, Brin et al., 1995, Gitchell and Tran, 1999, Lyon et al., 2001].

In the following sections, first, each of the methods is briefly described (for the detailed working of these methods see Section 2.3.1) and then a detail explanation is provided on how it is used in the mono-lingual (Urdu) text reuse and extrinsic plagiarism detection experiments.

4.1.1 Lexical overlap

4.1.1.1 Word n-grams overlap

Word n-grams overlap is a popular method used to compute the similarity between two texts (Section 2.3.1.1.1). It works by first breaking the texts into fixed-length n-grams and then comparing sets of generated n-grams. The similarity is calculated by counting the common n-grams and dividing the value by the length of one or both texts. It has proven to be successful in detecting text reuse [Clough et al., 2002, Chiu et al., 2010] and plagiarism [Lyon et al., 2001] in the past.

For the experiments performed on Urdu text reuse and plagiarism detection using Word n-grams overlap, the length of n is varied from [1−5] and the Containment similarity measure[2] (Equation 2.4) is used to calculate the similarity between text

[2] Jaccard, Dice and Overlap similarity measures are also tested but the scores were low when

pairs.

4.1.1.2 Vector Space Model

The Vector Space Model is a well-known IR method mostly used by search engines to rank web pages [Grossman et al., 1997, Gravano et al., 1999] (Section 2.3.1.1.2). The method works by creating a high dimensional vector space to represents the text documents. The size of the vector space is equal to the vocabulary of the corpus. The similarity between two vectors (or two documents) is measured using the cosine of the angle between them. The method has provided good results on the detection of text reuse [Clough, 2003, Bendersky and Croft, 2009, Ekbal et al., 2012] and document duplicates [Hoad and Zobel, 2003, Runeson et al., 2007].

For these experiments, vectors are created using source and reused text documents and the similarity is measured using cosine similarity (Equation 2.5). However, before computing the similarity between the two vectors, the popular *tf-idf* [Jurafsky and Martin, 2009, Baeza-Yates and Ribeiro-Neto, 2011] (Equation 2.6) weighting scheme is applied to weight individual terms in both the text documents.

4.1.2 String matching

4.1.2.1 Longest Common Subsequence

Longest Common Subsequence is a string matching method that counts the longest stream of consecutive terms common between a pair of texts (Section 2.3.1.2.1). It is important to mention here that the method is order-preserving, therefore, it is useful for capturing text modifications and word re-ordering. The method has found success previously in detecting document duplicates [Elhadi and Al-Tobi, 2009] and plagiarism detection [Gipp and Meuschke, 2011].

For the experiments on COUNTER and UPPC corpora, a normalised (0−1)

compared to the Containment similarity measure.

LCS similarity score, i.e., LCS_{norm}, between the source and reused text documents, is computed by dividing the length of LCS score with the length of the shorter text document (Equation 4.1).

$$LCS_{norm}(d1, d2) = \frac{|LCS|}{min(|d1|, |d2|)} \tag{4.1}$$

In Equation 4.1, |d1| and |d2| represent the length of the source and reused texts, respectively.

4.1.2.2 Greedy String Tiling

Greedy String Tiling is another well-known string matching method that has been used in this study to compute the similarity between text pairs (Section 2.3.1.2.2). It works by identifying sub-strings of maximal length (or tiles) that are common between a source and reused text document. To prevent matches of small lengths, the method only keeps those tiles whose length is greater than or equal to minimum Match Length (mML) (Section 2.3.1.2.2).

As the text documents in COUNTER Corpus are longer than UPPC Corpus, the minimum Match Length (mML) is set [1−10] for COUNTER Corpus and [1−5] for UPPC Corpus. The normalised GST similarity score (0−1), i.e., GST_{norm}, is computed using the Equation 4.2.

$$GST_{norm}(d1, d2) = \frac{|GST|}{min(|d1|, |d2|)} \tag{4.2}$$

where, |d1| represents the length of the original text and |d2| represents the length of reused text.

4.1.3 Structural similarity

4.1.3.1 Stop-word n-grams overlap

A stop-word is a natural language word that is extremely common in use. It has been argued that during paraphrasing these frequent words are often kept unchanged while replacing content words with synonyms [Bär et al., 2012]. In the past, such structural similarity-based methods have been used to detect text plagiarism [Stamatatos, 2011].

In Stop-words n-grams overlap (Section 2.3.1.4.1) experiments, the n-grams are generated by ignoring the content words from the text document pair. As with Word n-grams overlap, the similarity score is computed by varying the length of n [1−5] and using the Containment similarity measure (Equation 2.4).

4.1.4 Stylistic similarity

4.1.4.1 Sentence ratio

The sentence length and token length is considered as a characteristic of text writing style [Yule, 1939] and such stylistic measures have been a popular choice for authorship attribution in the literature [Craig, 2004, Stamatatos, 2009].

As the text documents in both the COUNTER and UPPC corpora are structured as single paragraph essays, therefore, for Sentence ratio (Section 2.3.1.5.3) experiments the number of sentences [3] per text document is computed and then the ratio between them.

[3] For sentence boundary detection, potential sentence termination markers for the Urdu language such as '۔', '؟', '!' were used.

4.1.4.2 Token ratio

For Token ratio experiments (Section 2.3.1.5.2), the numbers of tokens per document are calculated and then the ratio between both the text documents.

4.2 Experimental setup

This section describes the corpora, text pre-process settings, evaluation methodology, and evaluation measure used to evaluate the methods.

4.2.1 Corpora

For the set of experiments carried out, to assess the performance of methods (Section 4.1), the entire COUNTER Corpus (Section 3.1) [Sharjeel et al., 2017] and the entire UPPC Corpus (Section 3.2) [Sharjeel et al., 2016] are used. There are in total 1,200 text documents (600 source, 600 derived) in the COUNTER Corpus with three levels of text reuse (Wholly Derived = 135, Partially Derived = 288 and Non Derived = 177). The UPPC Corpus contains simulated cases of Urdu plagiarism with a total of 160 documents, 20 source and 140 suspicious documents (Paraphrased Plagiarised = 75, Non Plagiarised = 65).

4.2.2 Text pre-processing

Before applying the methods (Section 4.1), the Urdu text is pre-processed to remove all punctuation marks, newlines, extra white spaces, and illegal characters[4]. The text is tokenised on white space and a dictionary look-up based method is used to treat compound words as single units. Furthermore, to see the effect of stop-words on the performance of the detection task, the experiments are conducted with and

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

without stop-words. A standard list of stop-words[5] is used to filter them out before feature extraction.

4.2.3 Evaluation methodology

The main aim of these experiments is to distinguish between different levels of Urdu text reuse and extrinsic plagiarism at the document level. The problem is tackled as a supervised text classification task. The prime objective of the task is to see whether it is possible to automatically differentiate between the source and reused text and further understand which method(s) perform best on the Urdu text.

As the COUNTER Corpus has three levels of text reuse, the task is further divided into binary and ternary classifications. In the former case, the target is to differentiate between 2 classes (i.e., Derived (D) and Non Derived (ND)) while in the latter case, the target is to differentiate between 3 classes (i.e., Wholly Derived (WD), Partially Derived (PD), and Non Derived (ND)). For the binary classification task, the text documents categorised as Wholly Derived and Partially Derived are coupled to make the "Derived" class while the text documents categorised as Non Derived are part of the "Non Derived" class. For the UPPC Corpus, the text documents are labelled as "Paraphrased Plagiarised" (PP) and "Non Paraphrased" (NP) and the target is to train the classifier to distinguish between the two classes.

Similarity scores generated by applying various methods (Section 4.1) are used as input features for the classifier. The WEKA's[6] [Holmes et al., 1994, Hall et al., 2009, Witten et al., 2016] implementation of the Bayes theorem based Naïve Bayes classifier (with default parameter settings) using 10-fold cross validation, is used for the classification task. Naïve Bayes is appropriate for these kinds of experiments as it can handle the numeric features generated by applying the methods. Weighted average F_1 (Equation 2.21) results are computed and reported for both binary and

[5] The stop-words list is available with the corpus download (Section 3.1).
[6] http://www.cs.waikato.ac.nz/ml/weka/

ternary classifications for the COUNTER Corpus (Section 4.3.1) and binary classification for the UPPC Corpus (Section 4.3.2).

4.3 Results and analysis

This section presents results of the experiments performed on the COUNTER (Section 3.1) and UPPC (Section 3.2) corpora.

4.3.1 Results using the COUNTER Corpus

Table 4.1 presents Naïve Bayes classifier reported weighted average F_1 results on the COUNTER Corpus for both binary and ternary classification tasks using Word n-grams overlap, VSM, LCS, GST, Stop-word n-grams overlap, Sentence ratio and Token ratio methods (Section 4.1). *Word Uni-grams* means that the results are obtained using word 1-gram as a single feature for the classifications task. Similarly, *Word Bi-grams*, *Word Tri-grams*, *Word Four-grams*, and *Word Five-grams* means that the results are obtained using word 2-grams, 3-grams, 4-grams and 5-grams respectively as a single feature. *Word n-gram Combined* means that results are obtained by similarity scores of word unigrams, 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-word Uni-grams* means that the results are reported using stop-word 1-gram, *Stop-word Bi-grams* means stop-word 2-grams, *Stop-word Tri-grams* means stop-word 3-grams, *Stop-word Four-grams* means stop-word based 4-grams, *Stop-word Five-grams* means stop-word based 5-grams, and *Stop-word n-grams Combined* means that similarity scores of Stop-word 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 the Vector Space Model method, *LCS* means results obtained using the Longest Common Subsequence method, and *GST* means results

obtained using Greedy String Tiling method. For GST, *mML1* to *mML10* means results with minimum match lengths of tiles from 1 to 10, respectively. Again, *SWR* after each method means results computed after stop-words removal. *Sentence Ratio* and *Token Ratio* means that the results are obtained after applying the Sentence ratio and Token ratio method, respectively. 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 (41 features). The best results obtained overall are presented as bold letters whereas the best results obtained category-wise are underlined in the table.

Table 4.1: Weighted average F_1 results obtained for binary and ternary classification of COUNTER Corpus using different text reuse detection methods

	Binary (F_1)	Ternary (F_1)
Lexical overlap		
Words Uni-grams	<u>0.80</u>	**0.73**
Word Uni-grams + SWR	<u>0.80</u>	0.72
Word Bi-grams	0.66	0.64
Word Bi-grams + SWR	0.70	0.68
Word Tri-grams	0.57	0.56
Word Tri-grams + SWR	0.60	0.64
Word Four-grams	0.52	0.52
Word Four-grams + SWR	0.55	0.57
Word Five-grams	0.49	0.52
Word Five-grams + SWR	0.50	0.53
Word n-grams Combined	0.56	0.54
Word n-grams Combined + SWR	0.57	0.57
VSM	0.66	0.54
VSM + SWR	0.64	0.53
String matching		
LCS	0.77	0.70
LCS + SWR	0.77	0.71
GST mML1	**0.81**	0.72
GST mML1 + SWR	**0.81**	**0.73**
GST mML2	0.77	0.71

GST mML2 + SWR	0.74	0.67
GST mML3	0.70	0.65
GST mML3 + SWR	0.63	0.60
GST mML4	0.63	0.60
GST mML4 + SWR	0.60	0.57
GST mML5	0.58	0.59
GST mML5 + SWR	0.55	0.53
GST mML6	0.56	0.53
GST mML6 + SWR	0.53	0.51
GST mML7	0.54	0.52
GST mML7 + SWR	0.48	0.50
GST mML8	0.51	0.50
GST mML8 + SWR	0.46	0.50
GST mML9	0.47	0.49
GST mML9 + SWR	0.44	0.47
GST mML10	0.46	0.49
GST mML10 + SWR	0.43	0.45
Structural similarity		
Stop-word Uni-grams	0.58	0.40
Stop-word Bi-grams	<u>0.63</u>	0.42
Stop-word Tri-grams	0.47	0.44
Stop-word Four-grams	0.41	<u>0.46</u>
Stop-word Five-grams	0.35	0.34
Stop-word n-grams Combined	0.40	0.37
Stylistic similarity		
Sentence Ratio	0.58	0.32
Token Ratio	<u>0.68</u>	<u>0.45</u>
All features combined	0.70	0.68

From Table 4.1, as expected, overall, best results are lower for the ternary classification task ($F_1 = 0.73$) compared to the binary classification task ($F_1 = 0.81$). For both classification tasks, the same pattern of differences in the results can be seen across all the methods. This demonstrates that, in Urdu text reuse detection problem, it is easier to distinguish between two levels of reuse than three.

For binary classification, best score is obtained using GST mML1 ($F_1 = 0.81$), nearly matching the result with Words Uni-gram ($F_1 = 0.80$). It can also be noticed that both of these results did not improve after removal of stop-words. For ternary classification, the highest score is obtained for both GST mML1 + SWR and Words Uni-gram ($F_1 = 0.73$) and we can see a small effect of stop-words removal on both methods (improvement of 0.01 in GST mML1 while a decline of 0.01 in Words Uni-gram). 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 Urdu 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 the majority of examples in the corpus are rewritten).

GST (with $mML = 1$) outperformed all other methods for binary classification task and its performance for ternary classification task is the same as the Word n-grams overlap (with $n = 1$) 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., 2002], 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.

The results using the VSM method, for both binary ($F_1 = 0.66$) and ternary classifications ($F_1 = 0.54$) are lower compared to the Word n-grams overlap. This is likely to happen because VSM aims to identify topical similarity among text document pairs and used mostly for IR, whereas in text reuse detection the aim is to identify the overlap between document pairs. Moreover, the removal of stop-words

did not improve VSM results for both classification tasks.

As expected, performance using the LCS ($F_1 = 0.77$ for binary and $F_1 = 0.71$ for ternary classification) is lower compared to the GST method because it is not able to deal with *block move* problem. Furthermore, the removal of stop-words did not show any improvement in the LCS results for the binary classification task, however, there is a slight improvement of 0.01 for the ternary classification task.

The performance of Stop-word n-grams overlap ($F_1 = 0.63$ (Stop-words Bi-gram) for binary classification; $F_1 = 0.46$ (Stop-words Four-gram) for ternary classification) and stylistic similarity methods ($F_1 = 0.68$ for binary and $F_1 = 0.45$ for ternary classification both with token ratio), is low overall and they demonstrated poor results in both classification tasks. This shows that both structural and stylistic methods are comparatively not suitable for the Urdu text reuse detection task.

The results for the combination of features, "Word n-grams Combined" and "Stop-word n-grams Combined", does not improve performance.

For both classification tasks, from all the methods used on the COUNTER corpus, 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 the Wilcoxon signed-rank test ($p < 0.05$) [Wilcoxon et al., 1970]. LCS also demonstrated slightly better results, for the ternary classification task, on pre-processed text with stop-words removed. However, results using the VSM and GST methods do not show improvement after the removal of stop-words. This highlights the fact that such pre-processing is useful in some cases for text reuse detection on the Urdu text.

The experiments by combining all the features from all the methods (*All features combined* method) used in this study. The 12 features of *Word n-grams overlap*, 20 features of *GST*, 6 features of *Stop-word n-gram overlap*, and 2 features of each *VSM*, *LCS* and *Sentence/Token Ratio* methods are combined and best feature selection method is applied on the combination of all features. However, the *All features combined* method does not improve performance ($F_1 = 0.70$).

Table 4.2 shows the confusion matrix for the "GST mML1" method (it produced the best results for both classification problems). The columns and rows of the matrix represent the instances in the predicted and actual classes, respectively.

Table 4.2: Confusion matrix for ternary classification using GST mML1 on the COUNTER Corpus

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

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, it is 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.

4.3.2 Results using UPPC Corpus

Table 4.3 shows the Naïve Bayes classifier reported weighted average F_1 (Section 2.21) results on the UPPC Corpus. It should be noted that the results are obtained using the same set of methods used on the COUNTER Corpus i.e., Word n-grams overlap, VSM, LCS, GST, Stop-word n-grams overlap, Token ratio and Sentence ratio (Section 4.1). Nevertheless, the text documents in the UPPC are categorised as "PP" and "NP" (Section 3.2), therefore, the results are only obtained and reported for the binary classification task.

Table 4.3: Weighted average F_1 results obtained for binary classification of UPPC Corpus using different extrinsic plagiarism detection methods

	Binary (F_1)
Lexical overlap	

Word Uni-grams	0.88
Word Uni-grams + SWR	0.91
Word Bi-grams	0.86
Word Bi-grams + SWR	0.84
Word Tri-grams	0.82
Word Tri-grams + SWR	0.79
Word Four-grams	0.78
Word Four-grams + SWR	0.74
Word Five-grams	0.70
Word Five-grams + SWR	0.51
Word n-grams Combined	0.85
Word n-grams Combined + SWR	0.87
VSM	0.81
VSM + SWR	0.80
String matching	
LCS	0.88
LCS + SWR	0.90
GST mML1	0.90
GST mML1 + SWR	**0.92**
GST mML2	0.90
GST mML2 + SWR	0.87
GST mML3	0.85
GST mML3 + SWR	0.82
GST mML4	0.81
GST mML4 + SWR	0.76
GST mML5	0.74
GST mML5 + SWR	0.65
Structural similarity	
Stop-word Uni-grams	0.66
Stop-word Bi-grams	0.75
Stop-word Tri-grams	0.74
Stop-word Four-grams	0.72
Stop-word Five-grams	0.66
Stop-word n-grams Combined	0.74
Stylistic similarity	
Sentence Ratio	0.76

Token Ratio	0.37
All features combined	0.89

Overall, the results portray a similar pattern to that of the COUNTER Corpus (Table 4.1). The best result is obtained using GST mML1 ($F_1 = 0.92$) and second-best with Word Uni-grams ($F_1 = 0.91$). Furthermore, a decrease in performance can easily be spotted as the length of n or mMl increases. However, in contrast to the COUNTER Corpus, the best results for both methods are obtained with SWR, i.e., after removal of stop-words.

It is noteworthy that the results reported by all the methods are higher comparatively than the COUNTER Corpus. This indicates that the performance of these simple surface-level similarity estimation methods falls short when evaluated on the real cases of text reuse. It is worth recalling that the news stories in the COUNTER Corpus are written/edited by journalists, who are experts in their field, using different paraphrasing mechanisms while the text documents in the UPPC Corpus are intentionally plagiarised by under-grad students in a controlled environment. As a result, the reused examples in the COUNTER Corpus contain heavily paraphrased text, and these basic methods, which are designed for word-to-word matching, reported lower scores when evaluated on these examples.

The best score, among all the methods used, is obtained by GST and the second-best by Word n-grams overlap. Notably, the best results are obtained after removal of stop-words from the text. This shows that GST is the best method to distinguish between Urdu paraphrased and non-paraphrased text documents. In addition, it is able to capture reshuffling of words better when the stop-words are removed. However, as expected, both GST and Word n-grams overlap show a decline in performance when the length of tokens ($mMl > 1$ or $n > 1$) is increased. This is due to the reason that it becomes hard to find a matching of longer chunks in the plagiarised text. It is worth mentioning here that these findings compare well with the results of the METER Corpus [Clough et al., 2002], COUNTER Corpus [Shar-

jeel et al., 2017] and the experiments on Urdu short text reuse detection (USTRC Corpus) [Sameen et al., 2017].

Similar to the COUNTER Corpus, VSM did not perform well on the UPPC Corpus ($F_1 = 0.81$). Again, the probable reason is that the method is best suited for the text documents classification task. Moreover, its performance further decreases after the removal of stop-words.

The performance of the LCS method on the UPPC Corpus is comparatively better ($F_1 = 0.88$) than the COUNTER Corpus and it further increases after stop-words removal from the text ($F_1 = 0.90$). This highlights that the volunteers who created the plagiarised text documents for the UPPC Corpus have reused longer chunks of text verbatim or with light paraphrasing, without much word reordering, from the source text documents. These longer chunks are captured by the LCS method and hence it performed relatively better. Moreover, both GST and LCS results demonstrate that the string matching methods were able to detect longer portions of Urdu paraphrased text and, consequently, performed better.

As expected, the structural ($F_1 = 0.75$ (Stop-word Bi-grams)) and stylistic ($F_1 = 0.76$ (Sentence ratio)) similarity methods reported lower results and are not suitable for the extrinsic plagiarism detection on the Urdu text. Besides that, the Sentence ratio method performed somewhat better. The possible reason might be that when creating plagiarised documents, the volunteers have used a sentence by sentence approach to generate paraphrased text from the source text. This has resulted in the equal number of sentences in both source and plagiarised text documents, and hence, the performance of Sentence ratio method is better in comparison to the COUNTER Corpus.

The results obtained using the combined features for Word n-grams overlap ($F_1 = 0.87$) and Stop-word n-grams overlap ($F_1 = 0.74$) and similarly, a combination of all features ($F_1 = 0.89$), did provide competitive results. This indicates that combining different features for Urdu extrinsic plagiarism detection might provide fruitful results in some cases.

In contrast to the COUNTER Corpus, the removal of stop-words has a negative effect on the performance of Word n-grams overlap for $n > 1$, GST for $mMl > 1$ and VSM. However, in the case of LCS, it is the opposite as it improves the score. Also, as discussed before, the best results overall are obtained without the stop-words. This implies that the removal of stop-words has a mixed effect on the performance of methods in the Urdu extrinsic plagiarism detection task using the UPPC Corpus.

Table 4.4: Confusion matrix for binary classification using GST mML1 + SWR on the UPPC corpus

	PP	NP
PP	71	4
NP	7	58

Table 4.4 shows the confusion matrix for the method that produced the best result i.e., "GST mML1 + SWR". The columns and rows of the matrix represent the instances in the predicted and actual classes, respectively.

It can be observed that slightly more instances of NP are misclassified as PP (7) than PP to NP (4). Overall, the results using binary classification on the UPPC Corpus are higher than COUNTER Corpus and it reflects in the confusion matrix too.

4.4 Chapter summary

This chapter presented the experiments conducted on two benchmark Urdu corpora for the text reuse and extrinsic plagiarism detection. The main findings of the reported results are summarised in the following points.

- The state-of-the-art methods reported higher results on simulated cases of Urdu text plagiarism whereas their performance decreases on the real cases of Urdu text reuse.

- The best results are obtained with GST and Word n-grams overlap methods indicating that these two are the best-suited methods for the Urdu text reuse and extrinsic plagiarism detection task.
- GST mMl1 and Word Uni-grams with Containment similarity measure are the most distinguished features among all used in our experiments.
- LCS, VSM, Stop-word n-grams overlap, Sentence and Token ratio methods could not perform well on the Urdu text reuse and extrinsic plagiarism detection task.
- The best results are obtained when a shorter length of tokens (mML = 1 or $n = 1$) are used.
- The combination of features does not improve performance in Urdu text reuse detection, however, they were competitive enough for Urdu plagiarism detection.
- For the COUNTER Corpus the methods performed consistently better after stop-words are removed from the text with the increasing length of tokens ($n > 1$ or mMl > 1), however, for UPPC, it is the opposite.

"You can steal a man's bolts, but you can't steal his thunder."

Ed Zerne

5

Cross-lingual (English-Urdu) Text Reuse Detection

The previous chapter described the mono-lingual (Urdu) text reuse and extrinsic plagiarism detection experiments conducted on the two benchmark corpora, i.e., COUNTER Corpus and UPPC Corpus. The aim was to evaluate and compare the performance of state-of-the-art mono-lingual text reuse and extrinsic plagiarism detection methods on the Urdu text.

This chapter presents the cross-lingual (English-Urdu) text reuse detection experiments performed on the TREU Corpus (Section 3.3). A range of methods are applied on the corpus to show its usefulness and how it could be utilised in the development and evaluation of cross-lingual (English-Urdu) text reuse detection systems.

To the best of our knowledge, this is the first study that has applied these diverse methods on an English-Urdu cross-lingual text reuse corpus at the document level. Moreover, the applied methods and the use of supporting resources in these experiments provide in-depth analysis and set a strong baseline for the text reuse detection task in a low resource language pair, i.e., English-Urdu. Furthermore, these methods could easily be extended to other similar language pairs (e.g., English-Arabic, English-Persian, etc.) for the cross-lingual text reuse detection.

The rest of this chapter is organised as follows: Section 5.1 describes the methods used to extract features for the cross-lingual (English-Urdu) text reuse detection. Section 5.2 describes the experimental setup including the corpus, text preprocessing, evaluation methodology, and evaluation measure. Finally, results and their analysis are presented in Section 5.3.

5.1 Methods for cross-lingual text reuse detection

This section describes the cross-lingual (English-Urdu) text reuse detection methods applied on the TREU Corpus. The methods used are broadly categorised into three types, (1) Translation + Mono-lingual Analysis (Section 5.1.1), (2) cross-lingual Vector Space Model (Section 5.1.2), and (3) cross-lingual embeddings (Section 5.1.3). The chosen methods are most appropriate for the cross-lingual cross-script text reuse detection, especially for the English-Urdu language pair. There are other methods proposed in the literature (Section 2.3.2), however, they only work with language pairs belonging to similar language families (e.g., English-Spanish, English-German, etc.) or require knowledge bases or supporting resources that are not available for the Urdu language.

5.1.1 Translation + Mono-lingual Analysis

The Translation + Monolingual Analysis (T+MA) method is based on MT for the task of cross-lingual text reuse detection and has been very popular and widely used because of its simplicity [Barrón-Cedeño et al., 2013a]. The method first translates the source or derived text documents into one language and then addresses the task as mono-lingual text reuse detection. The translation is usually performed using an automatic MT system.

For the cross-lingual (English-Urdu) text reuse detection experiments performed on the TREU Corpus using T+MA method, the derived text documents are translated from Urdu to English using Google Translate[1]. Afterwards, the similarity score between a text pair is obtained by applying a diverse range of mono-lingual text reuse detection methods. The applied methods are classified under five categories, (1) lexical overlap, (2) string matching, (3) structural similarity, (4) mono-lingual word embeddings, and (5) mono-lingual sentence embeddings. For lexical overlap, Word n-grams overlap (Section 5.1.1.1.1) and Vector Space Model (Section 5.1.1.1.2) are applied. For string matching, Longest Common Subsequence (Section 5.1.1.2.1) and Greedy String Tiling (Section 5.1.1.2.2) are used. For structural similarity, Stop-word n-grams overlap (Section 5.1.1.3.1) is chosen. For mono-lingual word embeddings, averaged embeddings (Section 5.1.1.4.1), weighted averaged embeddings (Section 5.1.1.4.2), and weighted maximum embeddings (Section 5.1.1.4.3) variants are applied. For the more recent mono-lingual sentence embeddings, Sent2Vec (Section 5.1.1.5.1), InferSent (Section 5.1.1.5.2), Universal Sentence Encoder (Section 5.1.1.5.3), and LASER (Section 5.1.1.5.4) are used.

[1] https://translate.google.com
The text documents are translated using the current version of Google Translate that supports Neural Machine Translation (NMT) technology which is better than the previous Statistical Machine Translation (SMT). However, the translation accuracy of Asian and African languages is still imperfect [Freitas and Liu, 2017]. The results might change/improve if Google provides better support for the Urdu language in the future.

The working details of some of these methods are already discussed in Section 2.3.1 and Section 4.1. In what follows, a brief description of each of these methods, their working, and how they are used in the experiments performed, is presented.

5.1.1.1 Lexical overlap

5.1.1.1.1 Word n-grams overlap The Word n-grams overlap method tries to estimate the number of common n-grams between source and derived text documents (Section 2.3.1.1.1). It is one of the simplest methods used in text reuse detection that could easily be applied to a large collection of texts because of its low complexity.

For the experiments performed on the TREU Corpus, word n-grams are generated from the source and derived text documents by varying the lengths of n from [1–5]. Moreover, the similarity between the sets of unique n-grams is computed using four different similarity measures, i.e., Containment (Equation 2.4), Jaccard (Equation 2.1), Overlap (Equation 2.3), and Dice (Equation 2.2).

5.1.1.1.2 Vector Space Model The Vector Space Model is another method used for calculating the degree of similarity between a given text pair (Section 2.3.1.1.2). Using this method, the source and derived text documents are represented in a high dimensional vector space and similarity between them is calculated using the cosine similarity.

For these experiments, Vector Space Model is applied in two ways i.e., (1) Bag-of-Words (VSM-BoW) and (2) Character n-Grams (VSM-CnG).

For VSM-BoW, each source and derived text document is first converted into its BoW representation. The individual terms (words) are then weighted using the *tf-idf* weighting scheme (Equation 2.6). After that, the text documents are converted into vectors and similarity between the vectors is calculated using cosine similarity (Equation 2.5).

For VSM-CnG, in the first step, all white space characters in the source and

derived text documents are replaced with *hyphen* "-" and then the text is codified into character n-grams of size [3−5]. These n-grams are then weighted using *tf-idf* (Equation 2.6) and converted into vectors. Subsequently, the similarity score between source and derived text document vectors is estimated using the cosine similarity (Equation 2.5).

5.1.1.2 String matching

5.1.1.2.1 Longest Common Subsequence The Longest Common Subsequence is a string matching method that computes the longest group of elements (words) that are common between the two texts and are in the same order in each text (Section 2.3.1.2.1).

For the experiments conducted on the TREU Corpus, the normalised LCS score (LCS_{norm}), between each source and derived text document, is calculated by dividing the length of LCS on the length of the shorter text document (Equation 4.1).

5.1.1.2.2 Greedy String Tiling The Greedy String Tiling identifies the longest rewritten sequence of substrings from the source text and returns the sequence (as tiles) paired with the derived text (Section 2.3.1.2.2). To avoid very short matching lengths, a minimum match length (mML) value is used. It is a powerful algorithm that may detect matches even if some of the text is deleted or if additional text has been inserted.

For these experiments, the well-known Running Karp-Rabin Matching and Greedy String Tiling implementation is used [Wise, 1993] and the length of mML is varied [1−5]. The normalised GST similarity score (GST_{norm}) is calculated by taking the ratio of the length of GST and length of the shorter text document (Equation 4.2).

5.1.1.3 Structural similarity

5.1.1.3.1 Stop-word n-grams overlap Similar to the Word n-grams overlap method, Stop-word n-grams overlap is used to measure the degree of stop-words overlap be-

tween a text document pair (Section 2.3.1.4.1).

For these experiments, the source and derived text documents are first filtered to remove content words[2]. Subsequently, n-grams are generated for the remaining stop-words in the text by varying the length of n [1−5]. Eventually, the similarity between the sets of unique stop-word n-grams is computed using four different similarity measures i.e., Containment (Equation 2.4), Jaccard (Equation 2.1), Overlap (Equation 2.3), and Dice (Equation 2.2).

5.1.1.4 Mono-lingual word embeddings

The main idea of mono-lingual word embeddings is to represent words as continuous vectors in a multidimensional vector space [Mikolov et al., 2013]. This representation enables the capture of the semantic and syntactic properties of the text. The underlying assumption, from the domain of distributional semantics, is that the words which occur close to each other are semantically similar or have similar meanings.

Several mono-lingual word embeddings models are available, e.g., Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], fastText [Bojanowski et al., 2017], etc. that are trained on large corpora using unsupervised methods, i.e., Continuous Bag-of-Words (CBOW) and Skip-gram. The CBOW predicts the word based on the context of its surrounding words whereas Skip-gram predicts the context word(s), surrounding the word itself. These models are capable of capturing some elements of the context of a word, its semantics, and relation with other words, although their precise properties are still being evaluated [Yaghoobzadeh et al., 2019]. Consequently, they have been shown to benefit performance for a number of NLP tasks including IR [Vulić and Moens, 2015], text similarity [Kenter and De Rijke, 2015], topic modelling [Li et al., 2016], sentiment analysis [Yu et al., 2017], and authorship analysis [Sari et al., 2017].

The commonly used text reuse detection methods (n-gram overlap, LCS, GST,

[2] Standard English stop-words list from NLTK is used [Bird et al., 2009].

etc.) rely on the surface form of the text only, whereas word embeddings could be used to estimate the semantic similarity between pair of words (or vectors) [Kenter and De Rijke, 2015, Meng et al., 2017]. Therefore, in this work, mono-lingual word embeddings based methods are used to capture the semantic level similarities between source and derived text documents.

For the experiments performed on the TREU Corpus using mono-lingual word embeddings, both pre-trained and custom trained models are used. The pre-trained models are Google Word2Vec [Mikolov et al., 2013], Stanford NLP GloVe [Pennington et al., 2014], and Facebook fastText [Bojanowski et al., 2017]. Table 5.1 provides details of these pre-trained models.

Model	Domain	Words	Vocab	Dim
Google Word2Vec [3]	News	3B	3M	300
Stanford NLP GolVe [4]	Common Crawl [5]	840B	2.2M	300
Facebook fastText [6]	Common Crawl	600B	2M	300

Table 5.1: Details of the word embeddings pre-trained models

Moreover, all three models are also custom trained on English news data. For training, 105k text documents collected during the development of TREU Corpus (Section 3.3.1) are used. These are the English news reports, in plain text format, released by the news agencies (henceforth called Pakistan English News (PEN) Corpus). The PEN Corpus contains 16,120,843 words and 139,634 types. The corpus text is pre-processed (Section 5.2.2) and all the three models (i.e., Word2Vec, GloVe, and fastText) are trained using Gensim ("Generate Similar") toolkit [Řehůřek and Sojka, 2010] with same parameter settings, i.e., dimension 300[7], min-count 5, and

[3] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing
[4] http://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip
[5] https://commoncrawl.org/
[6] https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M-subword.zip
[7] Different dimensions (50, 100, 300) were tested and we determined that 300 works the best.

windows-size 10.

To estimate similarity between source and derived text documents using monolingual word embeddings, three different methods are used, (1) averaged embeddings (Section 5.1.1.4.1), (2) weighted averaged embeddings (Section 5.1.1.4.2), and (3) weighted maximum embeddings (Section 5.1.1.4.3). Each of the methods is explained in the following sections.

5.1.1.4.1 Averaged embeddings For the averaged embeddings method, a simple average of all the word embedding vectors in a document is calculated to generate the document vector. For instance, a text document d, composed of words $\{w_1, w_2, w_3, ..., w_n\}$, the word embedding vectors for each word are $\{v_{w_1}, v_{w_2}, v_{w_3}, ..., v_{w_n}\}$. The averaged embedding vector V_d for document d is calculated using Equation 5.1.

$$V_d = \frac{1}{n} \sum_{i=1}^{n} v_{w_i} \tag{5.1}$$

In Equation 5.2, w_i is the ith word of the document d and n is the number of words in the document.

For the experiments performed on the TREU Corpus, the source and derived text documents are first converted to their BoW representations and word embedding vectors are obtained for the set of unique words in both text documents. For each text document, all the word embedding vectors are averaged to obtain the resultant document vectors. Finally, the source and derived averaged embedding document vectors are normalised and the degree of similarity between them is computed using cosine similarity (Equation 2.5).

5.1.1.4.2 Weighted averaged embeddings Taking the simple average of the word embedding vectors of constituent words in a text document tends to give too much weight to words that are semantically irrelevant. This can possibly be addressed, to some extent, by taking a weighted average of the word embedding vectors. The

weights to individual words may be assigned using *pos* weights, *idf* weights, etc.

The weighted averaged embedding vector WV_d of a document d is calculated using Equation 5.2.

$$WV_d = \frac{1}{n}\sum_{i=1}^{n}(idf(w_i).v_{w_i}) \tag{5.2}$$

In Equation 5.2, *idf* is the function that returns the *idf* value of the ith word w_i, v_{w_i} is the word embedding vector of the ith word w_i and $\cdot$ is the scalar product.

Once the weighted averaged embedding document vectors for both source and derived text documents are generated, the process of computing similarity is similar to the averaged embedding method (Section 5.1.1.4.1). Moreover, for these experiments, *idf* weights for each word are computed using the PEN Corpus (Section 5.1.1.4).

5.1.1.4.3 Weighted maximum embeddings Averaged embeddings and weighted averaged embeddings are computationally cheap and based on BoW representations. However, one major drawback of the BoW representation is the loss of word order which results in corrupting the semantics of the text. Though weighting schemes give importance to individual words, they also suffer from the same word order issue. Moreover, for large text documents, using an averaging or linear summation of word vectors, the resultant document vectors ultimately start to approximate each other.

To overcome these issues and to efficiently use word embedding vectors for text reuse detection, a new method is proposed. The proposed weighted maximum embeddings method works as follows.

Consider a source text document s containing words $\{w_1, w_2, w_3, ..., w_n\}$, and a derived text document d containing words $\{w'_1, w'_2, w'_3, ..., w'_m\}$. In the first step, sets of unique words from both text documents are converted to their respective word vectors, i.e., $\{v_{w_1}, v_{w_2}, v_{w_3}, ..., v_{w_n}\}$ and $\{v_{w'_1}, v_{w'_2}, v_{w'_3}, ..., v_{w'_m}\}$, respectively. After that, cosine similarity (Equation 2.5) is computed for each normalised word vector

from the derived text document paired with every normalised word vector in the source text document {cos-sim($v_{w'_1} \leftrightarrow v_{w_1}$), cos-sim($v_{w'_1} \leftrightarrow v_{w_2}$), ..., cos-sim($v_{w'_1} \leftrightarrow v_{w_n}$), and so on}. However, only the maximum similarity is recorded for each source-derived word pair (vector). The resultant maximum cosine similarity scores are multiplied with the *idf* weights of the words from the derived text document. The final similarity between a source and derived text document pair is computed using Equation 5.3 by taking the ratio of sum of all weighted maximum cosine similarity scores and sum of all derived text document word *idf* weights.

$$sim(s,d) = \frac{\sum_{w'_i \in d} idf(w'_i) \times max_{w_j \in s, w'_i \in d} cosine(v_{w'_i}.v_{w_j})}{\sum_{w'_i \in d} idf(w'_i)} \tag{5.3}$$

In Equation 5.3, w, w′, v_w, and $v_{w'}$ are the sets of words and their respective vectors from the source and derived text documents, respectively. *idf* is the function that returns the *idf* weight, cosine is cosine similarity (Equation 2.5), and · is the scalar product.

The proposed method is further elaborated with the help of an example.

Source text: *A man is playing the guitar*

Derived text: *A young man is playing the guitar and singing*

The set of words, after case-folding, removing stop-words and applying lemmatisation, for source and derived texts are {man, play, guitar} and {young, man, play, guitar, sing}, respectively. Suppose that the *idf* weights for each of the words in the derived text are {young (1.5), man (2.7), play (1.1), guitar (5.3), sing (4.9)}. Using the weighted maximum embeddings method, as a first step, pair-wise cosine similarity is computed for each word embedding vector from the derived text paired with the word embedding vectors of the source text. Table 5.2 shows (hypothetical) cosine similarities of each word (vector) from the derived text paired with the word (vector) from the source text.

After that, the highest cosine similarity for each word vector from the derived

young	man	0.8
young	play	0.5
young	guitar	0.4
man	man	1.0
man	play	0.6
man	guitar	0.6
play	man	0.6
play	play	1.0
play	guitar	0.7
guitar	man	0.6
guitar	play	0.7
guitar	guitar	1.0
sing	man	0.7
sing	play	0.6
sing	guitar	0.9

Table 5.2: Hypothetical cosine similarities of word pairs

text paired with the word vector from the source text is saved i.e., young − man (0.8), man − man (1.0), sing − guitar (0.9) and so on. These similarity scores are weighted with the *idf* weights and the final score is calculated using the Equation 5.4.

$$sim(s,d) = \frac{(0.8 \times 1.5) + (1.0 \times 2.7) + (1.0 \times 1.1)(1.0 \times 5.3) + (0.9 \times 4.9)}{1.5 + 2.7 + 1.1 + 5.3 + 4.9} \tag{5.4}$$

The weighted cosine similarity between word pairs allows for approximate matching. This way the words that are replaced with their synonyms in the derived text document may also be captured. As the proposed method uses the similarity at word-level and does not use averaging of all the word vectors in a text document, it is expected to improve the performance.

For these experiments, sets of unique words from the source and derived text documents are converted to their word embeddings vectors and *idf* weights are calculated using the PEN Corpus (Section 5.1.1.4). The word-level similarity is measured using cosine similarity (Equation 2.5) and the final similarity score is

computed using Equation 5.3.

5.1.1.5 Mono-lingual sentence embeddings

The unsupervised word embeddings are best suited for word-level similarity. However, to better estimate semantic relatedness (meaning of words) between pairs of sentences or documents, contextual information and word order are important. Besides, supervised learning, presumably, can be more effective in learning the actual meaning of a word in a given sentence (or document).

For this purpose, pre-trained supervised and unsupervised sentence embedding models are available which are similar to word embeddings but for sentences. These models are pre-trained (some of them have an option to fine-tune or custom train) on large corpora to capture as much semantic and syntactic information of lexical units (words) as possible. They produce a fixed-length vector for a given input text (normally a sentence) of any length and have been used in a number of downstream applications such as opinion-polarity, Semantic Textual Similarity (STS), paraphrase identification, question-type classification, and sentiment analysis [Conneau and Kiela, 2018, Wang et al., 2018].

To capture similarity between source and derived text documents from the TREU Corpus, this study uses four sentence embedding models, (1) Sent2Vec (Section 5.1.1.5.1), (2) InferSent (Section 5.1.1.5.2), (3) Universal Sentence Encoder (Section 5.1.1.5.3), and (4) LASER (Section 5.1.1.5.4). Each of these models outputs a fixed-length sentence embeddings vector on a given input sentence of any length. Using these models, the degree of similarity between a source and derived text document is computed as follows: All the sentences [8] from each source and derived text document are converted to their respective sentence vectors using one of the sentence embedding models. These sentence embedding vectors are then summed to produce the document level vector representation. Each vector is normalised and

[8] Stanford sentence tokeniser [Manning et al., 2014] is used for sentence boundary detection.

the closeness between the source and derived document vectors is estimated using cosine similarly (Equation 2.5).

In the following sections, each of the sentence embedding models and its working is described.

5.1.1.5.1 Sent2Vec Sent2Vec is an unsupervised sentence embedding model that learns the distributed representations of sentences (or short texts) using the CBOW approach [Pagliardini et al., 2018]. The model simply combines (by averaging) word embeddings with n-grams embeddings of each word in a sentence. The method has proven to be beneficial in many NLP tasks such as sentiment analysis [Lee et al., 2017], IR [Allot et al., 2019], word similarity [Gupta et al., 2019], and text classification [Agibetov et al., 2018].

For these experiments, the pre-trained as well as, custom trained mono-lingual Sent2Vec models are used. The pre-trained model[9] used is based on the Toronto Books Corpus [Zhu et al., 2015] with bi-grams and 700-dimensions. For custom training, all the sentences from the PEN Corpus (Section 5.1.1.4) are used to train the model with exactly the same parameters i.e., bi-grams and 700-dimensions.

5.1.1.5.2 InferSent InferSent is a pre-trained supervised model developed by Facebook [Conneau et al., 2017]. It is neural network based, trained on 570k human-written English sentence pairs from the Stanford Natural Language Inference (SNLI) corpus [Bowman et al., 2015]. It uses biLSTM (bidirectional Long Short-Term Memory) with max pooling architecture (Figure 5.1). The model has recently found success in sentiment analysis [Bai et al., 2018], text summarisation [Daiya and Singh, 2018], and question answering systems [Choi et al., 2018] tasks.

Figure 5.1 shows the architecture of the InferSent training model. It takes two sentences as input. Each word of a sentence is converted to its word embedding

[9] https://drive.google.com/open?id=0B6VhzidiLvjSdENLSEhrdWprQ0k

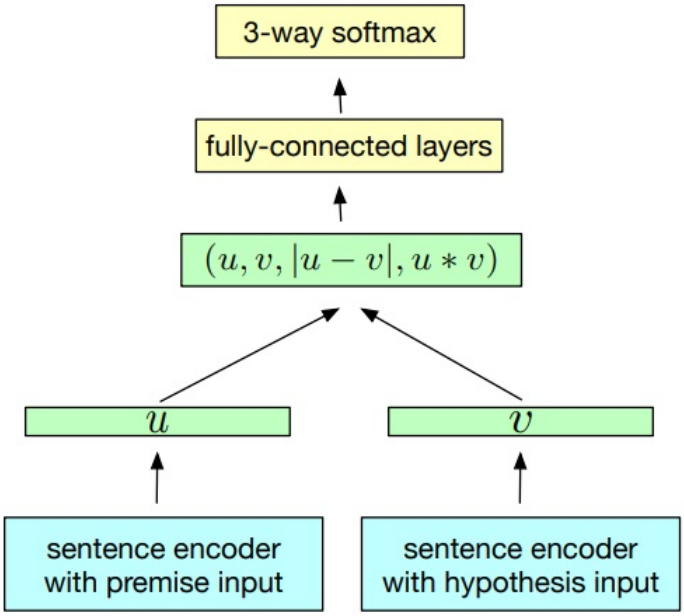

Figure 5.1: InferSent architecture [Conneau et al., 2017]

vector. These word vectors are passed to the biLSTM with a max pooling encoder which transforms them into a fixed-length sentence vector. In the next phase, three matching methods i.e., concatenation, product, and difference are applied to extract relations between the two sentence vectors. The output of this phase is fed into a 3-class Multi-Layer Perceptron (MLP) classifier (as the SNLI corpus has three classes i.e., entailment, contradiction, and neutral) and finally into a softmax layer.

For the experiments performed on the TREU corpus, InferSent mono-lingual model pre-trained[10] on the SNLI corpus [Bowman et al., 2015] is used. The model is trained with the biLSTM encoder with max pooling, batch-size 64, and word embeddings dimension 300. It outputs a 2,048-dimension sentence embedding vector for input sentence of any length. Moreover, two variations of the input word embed-

[10] Only pre-trained model is used as custom training is not possible because of the nature of the training corpus, i.e., SNLI [Bowman et al., 2015]

dings are used, (1) Glove [11] [Pennington et al., 2014] and (2) fastText [12] [Bojanowski et al., 2017]. Both word embedding models are custom trained on the PEN Corpus (Section 5.1.1.4) with dimension 300, min-count 5, and windows-size 10.

5.1.1.5.3 Universal Sentence Encoder The Universal Sentence Encoder, developed by Google, is a supervised sentence embedding model that takes a sentence of any length as input and converts it into a 512-dimension fixed-length vector [Cer et al., 2018]. Two versions of the model are available, both mono-lingual, trained on a variety of data sources, i.e., news websites, discussion groups, Wikipedia, and the SNLI corpus [Bowman et al., 2015]. First is the advanced transformer based architecture that uses attention to calculate context-aware embeddings of words in a sentence. These embeddings are then averaged to obtain sentence embeddings. The attention architecture takes care of the ordering and identity of words in the text. The second variant, called Deep Averaging Network (DAN), averages the uni-gram and bi-gram embeddings of all words together. The embeddings are then passed through a deep neural network to generate sentence embeddings. The transformer model has outperformed the DAN model on a number of tasks on the SentEval [Conneau and Kiela, 2018] and GLUE [Wang et al., 2018] benchmarks.

For these experiments, both transformer[13] and DAN[14] pre-trained models [15] are used.

5.1.1.5.4 LASER LASER (Language-Agnostic SEntence Representations) is an encoder-decoder architecture (Figure 5.2), released by Facebook, that converts multi-lingual sentences to fixed-length vector representations [Artetxe and Schwenk, 2018]. It is pre-trained on 223M parallel texts of 90+ languages. The multi-lingual model

[11] https://dl.fbaipublicfiles.com/infersent/infersent1.pkl
[12] https://dl.fbaipublicfiles.com/infersent/infersent2.pkl
[13] https://tfhub.dev/google/universal-sentence-encoder-large/3
[14] https://tfhub.dev/google/universal-sentence-encoder/2
[15] There is no option to custom train the models.

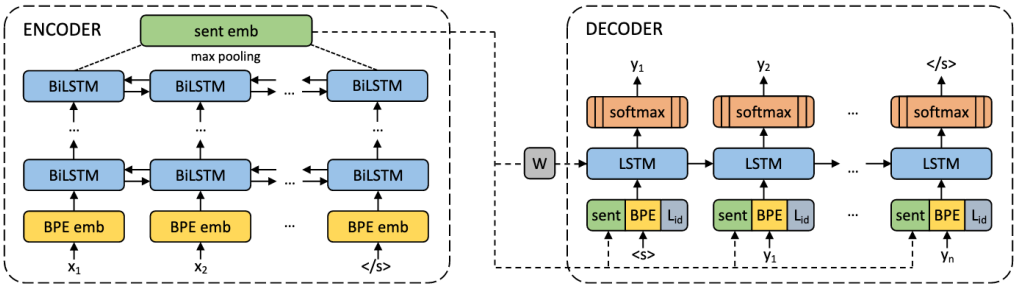

Figure 5.2: LASER architecture [Artetxe and Schwenk, 2018]

follows a sequence-to-sequence design where the output of the encoder is used as input to the decoder. The encoder is an enhanced version of InferSent (Section 5.1.1.5.2), language independent, pre-trained [16] on multi-lingual text, and the one responsible for constructing sentence embeddings. It uses a 5-layer biLSTM (each 512-dimension) with max pooling over the final states of the last layer. It takes a sentence as input and outputs a 1,024-dimension fixed-length vector.

For these experiments, the pre-trained [17] encoder module is used.

5.1.2 Cross-lingual Vector Space Model

The cross-lingual Vector Space Model for cross-lingual text reuse detection overcomes the language boundary by using bi- or multi-lingual dictionaries. The method works by first translating words from source or derived text documents into a common language using a dictionary or thesaurus. Once the text documents are in the same language they are converted to their BoW representations and projected on a high dimensional vector space. The size of the vector space is equal to the total vocabulary of the text documents. The similarity between these vectors is measured by the angle between them, typically using cosine similarity (Equation 2.5).

For these experiments, the six bi-lingual dictionaries compiled using different approaches (Section 3.5) are used to translate the Urdu words from the derived

[16] There is no option to custom train the encoder.
[17] https://dl.fbaipublicfiles.com/laser/models

text documents to the English language. Each of the six dictionaries, i.e., Waseem-Shahab, Urdu Lughat, One Click, Indic Parallel Corpus, Wiki Data, and Ur-GIZA, comes from different sources and contains Urdu words and their translation(s).

In the experiments performed, to observe the effect of lexical coverage, the dictionaries are used separately as well as combined as a single resource. Moreover, as a word may have multiple senses or a single word may translate into multiple words during translation, two types of experiments are performed, 1) using 'first word' as translation and 2) using 'all words' as translation. The former means that if an Urdu word has multiple English word translations in the dictionary, only the first translation word is used whereas the latter means that all the available translations of a word are used.

Once the words from the derived text documents are translated from Urdu to English, the source and (translated) derived text documents are converted to their BoW representations and each word (term) is assigned weights using *tf-idf* (Equation 2.6). The text documents are then transformed into vectors and similarity between them is computed using cosine similarity (Equation 2.5).

5.1.3 Cross-lingual Embeddings

Similar to the mono-lingual word and sentence embeddings (Section 5.1.1.4), distributed representations of words or sentences can be extended to the cross-lingual context. The cross-lingual embeddings are language independent representations that try to map words (or sentences) from multiple languages into one semantic (embedding) space [Upadhyay et al., 2016]. It uses the analogy that most of the words in different languages refer to common concepts or have same meanings. For example, 'horse' (English), 'caballo' (Spanish), and 'گھوڑا' (Urdu) are all different language words having the same meaning.

The following sections explain cross-lingual word and sentence embeddings methods that are used in the experiments performed on the TREU Corpus.

5.1.3.1 Cross-lingual word embeddings

Mono-lingual word embeddings trained separately for two different languages might be difficult to compare in one embedding space. However, there are different approaches that map two (or more) mono-lingual word embeddings into one shared embedding space using some form of alignment signal. These alignments can be at document-level [Vulić and Korhonen, 2016, Mogadala and Rettinger, 2016], sentence-level (parallel or comparable) [Levy et al., 2017, Gella et al., 2017] or word-level [Artetxe et al., 2016, Hauer et al., 2017]. Such embeddings learned using a shared space can be used in a number of downstream applications, e.g., cross-lingual word similarity, cross-lingual dictionary induction, cross-lingual document classification and cross-lingual dependency parsing [Upadhyay et al., 2016].

Similar to the mono-lingual word embeddings, for the experiments performed on the TREU Corpus using cross-lingual word embeddings, the same three types of word embeddings models are used i.e., Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], and fastText [Bojanowski et al., 2017]. However, only the fastText pre-trained model is available for the Urdu language. Besides, results from the mono-lingual word embeddings experiments (Section 5.3) showed that custom trained models performed better than pre-trained models. Therefore, for all the experiments performed using cross-lingual word embeddings, only custom trained models are used.

To train the Word2Vec, GloVe, and fastText word embedding models, English and Urdu news data, as well as parallel sentences from the EUPC-19 Corpus (Section 3.4.2) are used. In the first phase, mono-lingual English and Urdu models are trained separately. For English models, the training is performed by combining the PEN Corpus (Section 5.1.1.4) with the English sentences from the EUPC-19 Corpus. For Urdu models, the Urdu news archives collected during the creation of TREU Corpus (Section 3.3.1) are used. These are new stories published in leading Urdu newspapers of Pakistan (henceforth called Pakistan Urdu News (PUN) Corpus).

Moreover, they are merged with the Urdu sentences from the EUPC-19 Corpus. In the second phase, these mono-lingual English and Urdu word embeddings models are mapped to a shared embedding space using a powerful self-learning semi-supervised tool called VecMap[18] [Artetxe et al., 2018]. The tool utilises a small seed dictionary to generate a mapping between two mono-lingual word embeddings. The output is two mapped word embeddings for the two input languages where similar words vectors from both languages are analogous to each other.

Likewise mono-lingual word embeddings, three methods have been applied to estimate similarity between source and derived text documents using cross-lingual (mapped) word embeddings, 1) averaged embeddings (Section 5.1.3.1.1), 2) weighted averaged embeddings (Section 5.1.3.1.2), and 3) weighted maximum embeddings (Section 5.1.3.1.3). For each method, to compute the similarity between the source and derived text documents, word embedding vectors for English words are extracted from the mapped English word embedding models whereas Urdu word embedding vectors are obtained from the mapped Urdu word embedding models.

5.1.3.1.1 Averaged embeddings For these experiments, source and derived text documents are converted to their BoW representations and sets of unique words are extracted. For each word, its word embeddings vector is obtained from the respective embedding model. These word vectors are then averaged (Equation 5.1) to obtain the document vectors. Eventually, the similarity between two document vectors is computed using cosine similarity (Equation 2.5).

5.1.3.1.2 Weighted averaged embeddings For these experiments, *idf* weighting for each English word is calculated using the PEN Corpus while the PUN Corpus is used to weight the Urdu words. The weighted averaged embeddings for each source and derived document are then calculated using Equation 5.2. Finally, the similarity

[18] https://github.com/artetxem/vecmap

between two document vectors is estimated using cosine similarity (Equation 2.5).

5.1.3.1.3 Weighted maximum embeddings Similar to its mono-lingual counterpart (Section 5.1.1.4.3), cross-lingual weighted maximum embeddings tries to estimates the similarity between the source (English) and derived (Urdu) text documents using word-level alignment and *idf* weighting. The *idf* weights for English words are calculated using the PEN Corpus and for Urdu words using the PUN Corpus. The final similarity score is computed using Equation 5.3.

5.1.3.2 Cross-lingual sentence embeddings

The cross-lingual sentence embeddings try to map sentences from multiple languages into the same shared embedding space. This way the sentence embeddings from various languages become comparable. Similar to cross-lingual word embeddings, it is argued that sentences that are close to each other in a shared embedding space are semantically related.

For the experiments conducted on the TREU Corpus, two cross-lingual sentence embedding models[19] are used, 1) Sent2Vec (Section 5.1.3.2.1) and 2) LASER (Section 5.1.3.2.2).

5.1.3.2.1 Sent2Vec For these experiments, two mono-lingual Sent2Vec sentence embedding models (Section 5.1.1.5.1) are used, i.e., one for the English language and second for the Urdu language. The English Sent2Vec model is trained using the sentences from the PEN Corpus (Section 5.1.1.4) and English sentences extracted from the EUPC-19 Corpus (Section 3.4.2). The Urdu Sent2Vec model is trained using the PUN Corpus (Section 5.1.3.1) and the sentences from the Urdu part of the EUPC-19 Corpus. Moreover, the same parameter settings are used during the

[19] Pre-trained InferSent and Universal Sentence Encoder do not support the Urdu language. Moreover, they do not provide an option for custom training. Hence they are ruled out in these experiments.

training of both models, i.e., bi-grams and 700-dimensions.

To estimate the degree of overlap between the source and derived text documents, the sentences from the source text document are converted to sentence vectors using the English Sent2Vec model whereas sentences from the derived text document are converted to sentence vectors[20] using the Urdu Sent2Vec model. The sentence embedding vectors from the respective documents are then summed to produce the document level representation. Each document vector is normalised and the closeness between the source and derived document vectors is estimated using cosine similarity (Equation 2.5).

5.1.3.2.2 LASER For the experiments performed on the TREU Corpus, the pre-trained multi-lingual LASER model (Section 5.1.1.5.4) is used to encode the sentences from source (English) and derived (Urdu) text documents to their respective sentence vectors. It should be noted that the LASER encoder is trained on multi-lingual parallel data simultaneously for 90+ languages, therefore, the same encoder is used to generate both English and Urdu sentence embedding vectors. For each source and derived text document, all the sentence embedding vectors are added together to generate the respective document vectors. The degree of similarity between a source and derived normalised document vector is then computed using cosine similarity (Equation 2.5).

5.2 Experimental setup

This section describes the corpus, evaluation methodology, and evaluation measure used to evaluate the various cross-lingual text reuse detection methods.

[20] For Urdu sentence boundary detection, potential sentence termination markers for the Urdu language such as '۔', '؟', '!' are used.

5.2.1 Corpus

The entire TREU Corpus (Section 3.3) is used for the set of experiments carried out in this study. There is a total of 4,514 text documents (2,257 source, 2,257 derived) in the corpus with three levels of text reuse. The text documents tagged as "Wholly Derived" are 672, "Partially Derived" are 888, and "Non Derived" are 697.

5.2.2 Text pre-processing

For all the methods used in the Translation + Mono-lingual Analysis (Section 5.1.1) experiments, the English text is first pre-processed to remove the punctuation marks, extra white spaces, new line characters, foreign characters, numbers, and single alphabet tokens. It is then lemmatised using the Stanford Lemmatiser [Manning et al., 2014]. NLTK is used for word tokenisation and stop-words removal from the text [Bird et al., 2009]. Lastly, case-folding is applied to convert the text to lower case. Moreover, the same pre-processing settings are used on the PEN Corpus (Section 5.1.1.4) text for the training of mono-lingual word and sentence embeddings models.

For the experiments conducted using cross-lingual Vector Space Model (Section 5.1.2) and cross-lingual embeddings (Section 5.1.3), for both the English and Urdu text, punctuation marks, extra white spaces, new line characters, foreign characters, numbers, and single alphabets are removed from the text. Moreover, the English text is case folded and NLTK is used for word tokenisation and stop-words removal [Bird et al., 2009]. For Urdu text, stop-words are removed using a standard stop-words list, the text is tokenised on white space, and a dictionary look-up based method is used to treat compound words as single units.

It is worth noting here that the same pre-processing settings are used on the PEN Corpus, PUN Corpus and EUPC-19 Corpus used for the training of cross-lingual word and sentence embedding models.

5.2.3 Evaluation methodology

For the set of experiments conducted, the main objective is to distinguish between different levels of cross-lingual (English-Urdu) text reuse at the document level. The tag assigned to a document pair reflects the level of text reuse it contains. Wholly Derived (WD) means that the derived text document is the translation of the source text document with minor changes, (2) Partially Derived (PD) means that the derived text is paraphrased after translation, and (3) Non Derived (ND) means that the derived text document is written without considering the source text document.

To differentiate between multiple levels of text reuse, the problem is approached as a supervised text document classification task. The prime objective of the task is to see whether it is possible to automatically differentiate between the source and derived text at the document level and further understand which method(s) performs best.

Two variations of the task are used: (1) binary classification and (2) ternary classification. In the first case, the "Wholly Derived" (672 instances) and "Partially Derived" (888 instances) text documents are combined to make the "Derived" class (1,560 instances) and the "Non Derived" text documents remains part of the "Non Derived" class (697 instances). In the second case, the target is to distinguish between three levels of text reuse i.e., "Wholly Derived", "Partially Derived", and "Non Derived" classes.

For the set of experiments, the performance of a number of ML classifiers are investigated i.e., (1) Naïve Bayes, (2) Random Forest, (3) J48, (4) Support Vector Machine, (5) Multilayer Perceptron, and (6) Logistic Regression. All of these classifiers take numeric features as inputs and therefore are suitable for the experiments performed on the TREU Corpus. Similarity scores generated by applying various methods (Section 5.1) are used as input feature(s) for the classifiers. The Python's

Scikit-learn 0.23[21] [Pedregosa et al., 2011] based implementation of all the classifiers, with their default parameter settings, is used.

10-fold cross-validation is applied to better estimate the performance of the methods used in the study. The evaluation results are computed for both binary and ternary classes and reported using the weighted average F_1 (Equation 2.21) score.

5.3 Results and analysis

5.3.1 Results using Translation + Mono-lingual Analysis

Table 5.3 shows the results for both binary and ternary classification tasks obtained after applying Translation + Mono-lingual Analysis (Section 5.1.1) on the TREU Corpus. Note that only the best results are reported for each method applied[22].

The "Method" columns list the name of the methods which produced the highest result. "lo-wno-d-cmb" refers to the Word n-grams overlap method with Dice similarity measure and by combining n-grams of length [1–5] (5 features). Similarly, "lo-wno-j-cmb" refers to the Word n-grams overlap method with Jaccard similarity measure and by combining n-grams of length [1–5] (5 features) for the classification tasks. "lo-vsm-bow", "lo-vsm-c4g", and "lo-vsm-c5g" refers to the Vector Space Model method applied with Bag of Words, Character 4-grams and Character 5-grams, respectively. "sm-lcs" refers to the Longest Common Subsequence while "sm-gst-cmb" refers to the Greedy String Tiling method applied by combining the mML length [1−5] (5 features). "ss-sno-j-cmb" and "ss-sno-d-cmb" refers to the Stop-word n-grams overlap method with n-grams of length [1−5] (5 features) and similarity measures Jaccard and Dice, respectively. "we-w2v-ct-ae" refers to the custom trained mono-lingual Word2Vec model with averaged embeddings method. Likewise, "we-w2v-ct-wae" and "we-w2v-ct-wme" refers to the custom trained mono-

[21] https://scikit-learn.org
[22] The complete results are available in Appendix A

lingual Word2Vec model with weighted average embeddings and weighted maximum embeddings methods. "se-laser-pt" refers to the results reported by pre-trained mono-lingual LASER sentence embeddings method. "lo-sm-ss-cmb" refers to the combinations of lexical overlap, string matching, and Structural similarity methods. Similarly, "we-se-cmb" refers to the combination of mono-lingual word and sentence embeddings methods, and lastly, "all-methods-comb" refers to the experiments performed by combining all variants of all methods used in the study. The "Classifier" columns list the Machine Learning (ML) classifiers which produced the highest score among all the classifiers used in this study. "nb" is used as short for Naïve Bayes, "rf" as Random Forest, "mlp" as Multilayer Perceptron , "lr" as Logistic Regression.

Overall, best results for both classification tasks are obtained using "all-methods-cmb" method ($F_1 = 0.66$ ternary, $F_1 = 0.78$ binary) which shows that combining a range of features from different methods helps discriminate between various levels of cross-lingual text reuse in the TREU corpus. It can be noted that these results are not very high and need further improvement. This highlights the fact that detection of cross-lingual (English-Urdu) text reuse at the document level is a challenging task.

For the ternary classification task, an F_1 score of 0.66 seems lows, however, it is in line with the METER Corpus (best $F_1 = 0.66$, ternary) which is a gold standard mono-lingual (English) text reuse detection corpus [Clough et al., 2002]. Nevertheless, the TREU corpus is larger in size (2,257 text document pairs) than the METER Corpus (945 text documents pairs), yet the result remains the same. This shows that T+MA method used in this study is effective in detecting cross-lingual (English-Urdu) text reuse at the document level to a larger extent. Moreover, these results further support the stance that the T+MA method performs better at longer texts (document level) but its performance declines on short cases of text reuse (at the sentence level) [Barrón-Cedeño et al., 2013a, Muneer et al., 2019].

For both classification tasks, the combination of different methods "lo-sm-ss-cmd", i.e., lexical overlap, string matching, and structural similarity combined ($F_1 = 0.65$ ternary, $F_1 = 0.76$ binary) and "we-se-cmd" mono-lingual word and sentence

	Ternary			Binary	
Method	**F_1**	**Classifier**	**Method**	**F_1**	**Classifier**
Lexical overlap					
lo-wno-d-cmb	0.60	mlp	lo-wno-j-cmb	0.74	lr
lo-vsm-bow	0.50	lr	lo-vsm-bow	0.68	mlp
lo-vsm-c4g	0.51	rf	lo-vsm-c5g	0.69	mlp
String matching					
sm-lcs	0.56	rf	sm-lcs	0.73	nb
sm-gst-cmb	0.57	lr	sm-gst-cmb	0.74	mlp
Structural similarity					
ss-sno-j-cmb	0.43	mlp	ss-sno-d-cmb	0.61	rf
Mono-lingual word embeddings					
we-w2v-ct-ae	0.47	rf	we-w2v-ct-ae	0.64	rf
we-w2v-ct-wae	0.47	mlp	we-w2v-ct-wae	0.64	nb
we-w2v-ct-wme	0.58	lr	we-w2v-ct-wme	0.75	rf
Mono-lingual sentence embeddings					
se-laser-pt	0.47	mlp	se-laser-pt	0.67	j48
Combination of methods					
lo-sm-ss-cmb	0.65	rf	lo-sm-ss-cmb	0.76	rf
we-se-cmb	0.61	lr	we-se-cmb	0.75	lr
all-methods-cmb	**0.66**	**rf**	**all-methods-cmb**	**0.78**	**rf**

Table 5.3: Weighted average F_1 scores obtained by applying different variants of T+MA method on the TREU Corpus

embeddings combined ($F_1 = 0.61$ ternary, $F_1 = 0.75$ binary) improves performance. This indicates that using a set of features together has proven to be useful in the cross-lingual (English-Urdu) text reuse detection in the TREU Corpus.

In terms of individual methods performance, for ternary classification task, from lexical overlap, Word n-grams overlap performed better ($F_1 = 0.60$) than both variants of VSM method ($F_1 = 0.50$ VSM-BoW, $F_1 = 0.51$ VSM-CnG). Moreover, the best result is obtained using a combination of features [$n = 1{-}5$] (5 features) and the Dice similarity measure. This demonstrates that simple overlap of word n-grams between source and derived text documents is a good indicator of cross-lingual text reuse and a combination of features has further increased its performance. This also shows that combining various lengths of n-grams together contributes better in identifying the cross-lingual (English-Urdu) reuse of text. Besides, the low result of VSM shows that it is better suited for IR or to find topical relevance between text documents instead of overlap between them.

For string matching, GST ($F_1 = 0.57$) reported comparatively better results than LCS ($F_1 = 0.56$). This indicates that both these methods are able to capture the word reordering in reused texts, however, could not beat the simple Word n-grams overlap method. It further shows that during the formulation of newspaper stories (derived text documents), the journalist(s) have not reused longer chunks from the news agency's report (source text document) in the TREU Corpus. A possible reason for GST performing better than LCS is because it does not suffer from the block-move problem. Additionally, it produced the best result when the lengths of mML are combined (1−5) (5 features) for the classification task. This again highlights the advantage of using a combination of features over single feature used in the T+MA experiments performed on the TREU Corpus.

Stop-word n-grams overlap, the only structural similarity method, performed poorly and reported the lowest score ($F_1 = 0.43$) in all the methods used. The rationale being that it is more suitable for the authorship attribution and intrinsic plagiarism detection tasks rather than the text reuse detection.

Among mono-lingual word embeddings, custom trained Word2Vec model with weighted maximum embeddings performed significantly better ($F_1 = 0.58$) than the other two variants ($F_1 = 0.47$ averaged embeddings, $F_1 = 0.47$ weighted averaged embeddings). This shows the usefulness of the proposed approach which takes into account word-level similarities with *idf* weighting instead of averaging individual word vectors. Moreover, among the three word embeddings models used, Word2Vec has outperformed GloVe and fastText. Furthermore, it is worth noting that methods based on custom trained word embeddings have consistently performed better than the pre-trained ones (Appendix A). The most probable reason is that the pre-trained word embeddings use Google News, Common Crawl, Wikipedia etc. for training whereas custom word embeddings are trained on domain-specific text, i.e. PEN Corpus (Section 5.1.1.4). Consequently, custom trained word embeddings are less likely to suffer from Out-Of-Vocabulary (OOV) words. Moreover, using domain specific data, the models could learn representations of words better and ultimately perform better in the downstream task.

For mono-lingual sentence embeddings, LASER ($F_1 = 0.67$) has reported better result than others (Sent2Vec, InferSent, and Universal Sentence Encoder (Appendix A)). There seem to be two possible reasons, 1) the model is trained on a large corpus (221M sentences), and 2) it supports biLSTM based recurrent and deeper architecture. Thereby, it captures the syntactic and semantic properties of a sentence (text) better, which has helped in detecting the similarity between two texts. On the other hand, both Universal Sentence Encoder and Sent2Vec use uni- or bi-grams with averaging to produce sentence embeddings, hence, could not produce good results. It is worth mentioning here that the pre-trained LASER model (encoder) is trained on different domain data (Europarl [Koehn, 2005], United Nations Corpus [Eisele and Chen, 2010], etc.) than the TREU Corpus (journalism), hence its performance could not surpass Word embedding based methods.

Table 5.4 (columns represent the instances in the predicted and rows represent the instances in the actual classes) shows the confusion matrix for "all-methods-cmb"

	WD	PD	ND
WD	468	178	26
PD	100	593	195
ND	10	272	415

Table 5.4: Confusion matrix for ternary classification using all methods combined

method that produced the best result for ternary classification task (Table 5.3). As expected, it is easier to discriminate between WD and ND text documents, whereas it is more difficult to discriminate between WD/PD and PD/ND text document pairs. It is noteworthy that a large number of ND instances (272 out of 697 (39%)) are misclassified as PD, similarly, PD instances are misclassified as ND (195 out of 888 (22%)). This shows that the classifier suffers mostly between discriminating PD and ND classes which resulted in the low performance in the ternary classification task.

Table 5.3 also shows results for binary classification. As expected, all the results are higher than the ternary classification task. This indicates that cross-lingual (English-Urdu) text reuse detection at the document level is easier between two classes than three. Overall, the results corroborate with ternary classification task results. However, these results are still low considering the binary classification task is much simpler as it involves distinguishing between two classes which are relatively distinct. There could be several possible reasons for this. In binary classification, the WD (672 instances) and PD (888 instances) classes are combined to make the "Derived" class (total 1,560 instances). This has resulted in class imbalance (1,560 Derived, 697 Non-Derived) which is one of the reasons for its low result. Moreover, the confusion matrix (Table 5.4) for the best result of ternary classification shows the classifier mainly finds it difficult to distinguish between PD and ND classes. When one of these problematic classes, i.e., PD is combined with WD to make the derived class, it has contributed to the low performance.

Regarding individual method results, for lexical, similar to ternary classification

task, Word n-grams overlap ($F_1 = 0.74$) outperformed VSM-BoW ($F_1 = 0.68$) and VSM-CnG ($F_1 = 0.69$). Once again the best result is obtained using a combination of n-gram features [1−5] (5 features) and Jaccard similarity measure. This shows that the combination of features is helpful in improving performance even in the binary classification task.

The results of the string matching methods show a similar pattern to that of the ternary classification task. GST ($F_1 = 0.74$) performed slightly better than the LCS ($F_1 = 0.73$) method. Again it emphasises the strength of GST which can detect the transposition of tokens (words) better than LCS. Furthermore, the result is obtained by combining mML length [1−5] which highlights that the classifier is better suited for the combination of features.

As expected, and similar to ternary classification, the structural similarity method Stop-word n-grams overlap reported the lowest result ($F_1 = 0.61$). This indicates that it is not an appropriate method to use for the cross-lingual text reuse detection task on the TREU Corpus.

The performance of various word embedding methods also shows a similar trend to that of the ternary classification task. The best result is obtained using a custom trained Word2vec model and using the proposed weighted maximum embeddings method ($F_1 = 0.75$). It is noteworthy that, for the binary classification task, the method has performed better than all the other distinct methods used in the study. This shows that the proposed method is able to capture the semantic word-level overlap better between source and derived text documents than averaging of word vectors.

For sentence embeddings, once again, due to its recurrent neural network architecture (biLSTM) and having been trained on a large data set (221M sentences), LASER reported highest result ($F_1 = 0.67$) among others.

Regarding classifiers, in the majority of the cases, RF performed better than the others. Moreover, the highest results for both classification tasks ($F_1 = 0.78$ for binary and $F_1 = 0.66$ for ternary) are also reported using RF. This shows that the

RF classifier is more appropriate to use for the cross-lingual experiments performed using various T+MA methods on the TREU Corpus.

5.3.2 Results using cross-lingual Vector Space Model

Table 5.5 shows the results for both ternary and binary classification tasks obtained after applying the cross-lingual Vector Space Model (Section 5.1.2) using different dictionaries on the TREU Corpus. The prefix in each name "cl-vsm" refers to cross-lingual Vector Space Method while "fw" and "aw" refer to the 'first word' or 'all words' variants of the method. Furthermore, the post-fix after each name points to the name of the dictionary used in the experiment. "ws" refers to Waseem-Shahad, likewise, "ipc", "giza", "lu", "oc", and "wiki" refers to Indic Parallel Corpora, Ur-GIZA, Lughat, One Click, and Wiki dictionaries, respectively. Lastly, the postfix "cmd" means the combination of different methods together.

Overall, the best results for both classification tasks are obtained using the Ur-GIZA dictionary and using all words "clvsm-aw-giza" method ($F_1 = 0.52$ ternary, $F_1 = 0.70$ binary). This highlights that although using the translation of all words may generate noise but it helps in improving the performance in the experiments performed on the TREU Corpus. Moreover, the Ur-GIZA dictionary has the highest lexical coverage among all the dictionaries individually used in the experiment (Table 5.6).

It is important to note here that the best results obtained using the cross-lingual Vector Space Model are low. A possible reason may be because of the low lexical coverage of various dictionaries used in the experiments (Table 5.6, highest 63.03%). This shows that despite the fact that these dictionaries are reasonably large in size (Table 3.15) they have limited words from the journalism domain (the TREU Corpus is compiled from newspaper data). Moreover, they lack POS information, which is helpful in determining the context of a word. In addition, they do not contain lemmatised forms of words. All these factors collectively contributed to the overall

	Ternary		Binary	
Method	**F_1**	**Classifier**	**F_1**	**Classifier**
First word				
cl-vsm-fw-ws	0.40	mlp	0.56	lr
cl-vsm-fw-ipc	0.49	lr	0.68	j48
cl-vsm-fw-giza	0.45	lr	0.61	lr
cl-vsm-fw-lu	0.41	mlp	0.56	lr
cl-vsm-fw-oc	0.41	nb	0.56	lr
cl-vsm-fw-wiki	0.40	mlp	0.56	lr
cl-vsm-fw-cmb	0.44	j48	0.58	lr
All words				
cl-vsm-aw-ws	0.41	mlp	0.56	lr
cl-vsm-aw-ipc	0.46	lr	0.63	lr
clvsm-aw-giza	**0.52**	**j48**	**0.70**	**mlp**
cl-vsm-aw-lu	0.43	mlp	0.58	lr
cl-vsm-aw-oc	0.40	lr	0.56	lr
cl-vsm-aw-wiki	0.40	mlp	0.56	lr
cl-vsm-aw-cmb	0.50	nb	0.67	mlp
cl-vsm-fw-aw-cmb	0.51	lr	0.69	lr

Table 5.5: Weighted average F_1 scores obtained by applying different variants of the cross-lingual Vector Space Model on the TREU Corpus

Dictionary	**Coverage**
ws	25.18%
ipc	41.10%
giza	47.83%
lu	37.41%
oc	30.52%
wiki	07.27%
cmd	63.03%

Table 5.6: Coverage of different dictionaries used in the cross-lingual Vector Space Model experiment

low performance of cross-lingual vector space model on the TREU Corpus.

It is worth mentioning here that apart from Indic Parallel Corpus, for all the dictionaries used in the experiments, the results are consistently better using all words than the first word. Besides, the result of each method is directly related to the lexical coverage of the dictionary used in the experiment. The dictionaries having better lexical coverage have reported higher results. Surprisingly, when all dictionaries are combined into a single resource, the lexical coverage is considerably increased (63.03%) but it only produced similar results. The probable reason for this is that combining words from all dictionaries together has led to generating more noise during translation, and, as a result, the performance decreases.

Among other results, in both classification tasks, the result obtained using the Indic Parallel Corpus is highest ($F_1 = 0.49$ ternary, $F_1 = 0.68$ binary) when only the first word is used in the translation. Once again, as the dictionary is created using crowd-sourcing and has words from different domains, it is likely to produce a majority of the word translations. On the other hand, the results obtained using Wiki are lowest as it only has 7.27% lexical coverage.

Using a combination of all dictionaries in both first word "clvsm-fw-cmb" ($F_1 = 0.44$ ternary, $F_1 = 0.58$ binary) and all words "clvsm-aw-cmb" ($F_1 = 0.50$ ternary, $F_1 = 0.67$ binary) experiments, does not improve performance. Moreover, combining first word and all words together "clvsm-fw-aw-cmb" ($F_1 = 0.51$ ternary, $F_1 = 0.69$ binary) also does not improve performance.

Among classifiers, in most of the cases "lr" performed better than the others. However, the best results for ternary classification are obtained using "j48" and binary classification using "mlp" classifier.

5.3.3 Results using cross-lingual embeddings

Table 5.7 shows the results for both ternary and binary classification tasks obtained after applying different variants of cross-lingual embeddings (Section 5.1.1) on the

TREU Corpus. Note that only the best results are reported for each method applied[23].

In Table 5.7, "cl-we-w2v-ct-ae" refers to the custom trained cross-lingual Word2Vec model with average embeddings method. Similarly, "cl-we-w2v-ct-wae" and "cl-we-w2v-ct-wme" refer to the custom trained cross-lingual Word2Vec model weighted average embeddings and weighted maximum embeddings methods. "cl-se-laser-pt" refers to the pre-trained cross-lingual LASER method and "cl-we-se-cmb" refers to the experiment performed by combining all the cross-lingual word and sentence embeddings methods.

	Ternary			**Binary**	
Method	**F_1**	**Classifier**	**Method**	**F_1**	**Classifier**
Cross-lingual word embeddings					
cl-we-w2v-ct-ae	0.34	rf	cl-we-w2v-ct-ae	0.58	rf
cl-we-w2v-ct-wae	0.34	rf	cl-we-w2v-ct-wae	0.58	rf
cl-we-w2v-ct-wme	0.38	rf	cl-we-w2v-ct-wme	0.60	rf
Cross-lingual sentence embeddings					
cl-se-laser-pt	0.41	j48	cl-se-laser-pt	0.65	nb
Combination of methods					
cl-we-se-cmb	**0.47**	**mlp**	**cl-we-se-cmb**	**0.66**	**mlp**

Table 5.7: Weighted average F_1 scores obtained by applying different variants of cross-lingual embeddings on the TREU Corpus

Overall, the results show a similar pattern to that of mono-lingual word and sentence embedding methods (Table 5.3). The best results for both classification tasks are obtained using the "cl-we-se-cmb" method ($F_1 = 0.47$ ternary, $F_1 = 0.66$ binary). This again points out that the combination of different methods is beneficial in improving the results on the TREU Corpus. However, comparatively, these results

[23] The complete results are available in Appendix B

are very low, which indicates that cross-lingual cross-script English-Urdu text reuse detection at the document level is a challenging task. There are several possible reasons for this low result. First, although the data used to train the mono-lingual English and Urdu word embeddings models are domain-specific but are not large enough. Word embedding models trained on a small data set would find it difficult to get accurate word representations or lack vocabulary coverage. As a consequence, when these models are used in a downstream task, it has a negative effect on the performance of the result. Second, there are major linguistic differences between English and Urdu languages. English is a Subject-Verb-Object (SVO) order language whereas Urdu follows Subject-Object-Verb (SOV) order. Moreover, Urdu has a very rich morphological system where a word could have up to sixty different forms [Rizvi and Hussain, 2005]. These differences have probably made it difficult to map the English and Urdu word embeddings to a shared embeddings space or bringing similar word vectors together, which has eventually affected the performance.

Interestingly, in both classification tasks, the cross-lingual sentence embeddings method LASER has performed ($F_1 = 0.41$ ternary, $F_1 = 0.65$ binary) relatively better than word embeddings methods. A possible reason is that because it is trained on a large multi-lingual parallel corpus (223M sentences) hence it has learned the syntax and semantics of multi-lingual text better. Moreover, the encoder part of the model used in the experiment is based on a recurrent architecture (biLSTM) whereas Sent2Vec uses simple averaging of uni- and bi-grams to generate sentence embeddings. LSTM memory units help in learning the contextual information of words in a sentence, which results in producing better sentence embedding vectors and eventually helps in capturing similarity between two texts.

Among cross-lingual word embeddings, the proposed method, i.e., weighted maximum embeddings, again performed fairly better ($F_1 = 0.38$ ternary, $F_1 = 0.60$ binary) than the averaged and weighted averaged embeddings methods. This again highlights the strength of the proposed method and making use of word-level similarities combined with term weighting using *idf*. It is worth noting that, similar

to the mono-lingual word embeddings, both GolVe and fastText models reported unsatisfactory results (Appendix B).

Regarding classifiers, although the highest results for both classification tasks are reported by "mlp" classifier, in majority of the cases "rf" performed better than the others. This demonstrates that using cross-lingual embeddings, both "mlp" and "rf", are suitable for cross-lingual (English-Urdu) text reuse detection on the TREU Corpus.

5.4 Chapter summary

	Ternary			**Binary**	
Method	**F_1**	**Classifier**	**Method**	**F_1**	**Classifier**
Translation + Monolingual Analysis					
all-methods-cmb	**0.66**	**rf**	**all-methods-cmb**	**0.78**	**rf**
Cross-lingual Vector Space Model					
cl-vsm-aw-giza	0.52	j48	cl-vsm-aw-giza	0.70	mlp
Cross-lingual embeddings					
cl-we-se-cmb	0.47	mlp	cl-we-se-cmb	0.66	mlp

Table 5.8: Summary of the results

This chapter presented the cross-lingual (English-Urdu) experiments conducted on the TREU Corpus. Table 5.8 shows a summary of the results obtained after applying a diversified range of methods on the corpus. The results portray a clear picture that, in both ternary and binary classification tasks, Translation + Monolingual Analysis has outperformed both cross-lingual Vector Space Model and cross-lingual embeddings. This shows that it is the best suited method to discriminate between different levels of cross-lingual (English-Urdu) text reuse at document level. It should be noted that the best result is obtained using a combination of all T+MA

methods used. This indicates that combining different methods for cross-lingual text reuse detection on the TREU Corpus is helpful. Even though evaluation results are reported using only F_1 measure, the detailed evaluation shows that most of the methods used in this study are recall oriented methods. The high recall and overall low results in both classification tasks highlights that cross-lingual (English-Urdu) text reuse detection is a challenging task and needs further research.

"Do something others will have the desire to plagiarise but will find difficult to do."

Robert Genn

6

Conclusions and Future Directions

Text reuse is defined as a borrowing procedure to create a new text(s) using the one(s) already available. Unlike plagiarism, defined as the unacknowledged reuse of text, it is a common practice in journalism or collaborative authoring. Text reuse or plagiarism can be mono-lingual (in the same languages) or cross-lingual (across languages) and may take the form of copy-paste, paraphrasing, or reuse of ideas. Due to the rapid increase in readily available online text reservoirs (especially multilingual), there is a sharp rise in both mono- and cross-lingual text reuse and plagiarism cases. As a consequence, developing reliable and efficient methods for their detection has become an interesting research area. However, for the development, evaluation, and comparison of state-of-the-art mono- and cross-lingual text reuse and plagiarism detection methods, a major bottleneck is the unavailability of standard evaluation

resources containing real reuse cases, especially for under-resourced languages (e.g., Urdu).

6.1 Thesis Summary

The primary objective of this research was to explore and provide solutions to the open problem of mono- (Urdu) and cross-lingual (English-Urdu) text reuse and extrinsic plagiarism detection. Urdu, the official language of Pakistan and spoken by around 175 million people, is a resource-poor language in terms of NLP with very limited annotated corpora and basic language processing tools available. To contribute to this under-resourced language, this thesis has produced evaluation corpora, supporting resources, and methods to detect mono- (Urdu) and cross-lingual (English-Urdu) text reuse and extrinsic plagiarism with an aim to encourage and support research in Urdu and English-Urdu language pair.

(Chapter 3) Two mono-lingual (Urdu) and one cross-lingual (English-Urdu) gold standard benchmark corpora have been developed for the text reuse and extrinsic plagiarism detection tasks. (1) The COUNTER Corpus is an Urdu text reuse corpus containing real reuse cases at the document level (in total 1,200 text documents). The corpus text is compiled from journalism and manually annotated at three levels of text reuse, i.e., Wholly Derived, Partially Derived, and Non-Derived. (2) The UPPC Corpus is an Urdu extrinsic plagiarism corpus that contains manually created simulated cases of Urdu paraphrased plagiarism (in total 160 text documents). (3) The TREU Corpus is an English-Urdu cross-lingual document-level text reuse corpus. It contains text from the journalism domain and manually annotated at three levels of text reuse, i.e., Wholly Derived, Partially Derived, and Non-Derived. It includes real cases of text reuse (in total 4,514 text documents) from English to Urdu language. A large-scale publicly available English-Urdu Parallel Corpus has also been created as a supporting resource for cross-lingual (English-Urdu) text reuse detection experiments. It is mined from the Web and contains 154,258 par-

allel sentences. Moreover, a number of bi-lingual dictionaries have been assembled for the English-Urdu language pair using different methods from online and offline sources.

Chapter 4 - Mono-lingual (Urdu) text reuse and extrinsic plagiarism detection experiments have been performed on the COUNTER Corpus and UPPC Corpus. The objective was to make a direct comparison of the existing state-of-the-art mono-lingual text reuse and extrinsic plagiarism detection methods to investigate their behaviour on the Urdu text and to highlight the strengths and weaknesses of the proposed standard evaluation resources. The same set of methods were applied on both corpora with the purpose of differentiating between different levels of text reuse and extrinsic plagiarism at the document level. Supervised classification has been used with two variations, (1) ternary classification with an aim to distinguish between three levels of reuse and (2) binary classification with an aim to distinguish between two levels of reuse. Results showed that the methods performed relatively well on the simulated examples of extrinsic plagiarism whereas their performance declined on the real examples of text reuse. The best results were obtained with the GST (mML = 1) and Word n-grams overlap ($n = 1$) methods indicating that these two are the best-suited methods for the Urdu text reuse and extrinsic plagiarism detection. Moreover, it was observed that removing stop-words from the Urdu text has a positive effect on the performance in the case of COUNTER Corpus, whereas, in the case of the UPPC Corpus, it is the opposite. Surprisingly, the combination of different methods did not improve the performance of the methods applied on both corpora.

Chapter 5 - Cross-lingual (English-Urdu) text reuse detection experiments have been performed on the TREU Corpus. The main aim of these experiments was to provide a direct comparison and detailed analysis of the cross-lingual text reuse detection methods for the English-Urdu language pair at document-level. A diversified range of methods were applied on the corpus to show its usefulness and how it can be utilised in the evaluation of cross-lingual text reuse detection systems in general

and specifically for the English-Urdu language pair. A new method is also proposed to detect cross-lingual (English-Urdu) cases of text reuse. The benchmark experiments conducted on the corpus provided a strong baseline for the cross-lingual text reuse detection task in a low-resourced language pair, i.e., English-Urdu. For the cross-lingual (English-Urdu) text reuse detection, the problem is tackled as a supervised classification task and both ternary, as well as binary classification variants, were used. Results revealed that, in both classification tasks, Translation + Monolingual Analysis using a combination of all methods evidently performed better than the rest. However, overall, the results obtained were very low. This implies that cross-lingual text reuse detection is a challenging task especially when the languages involved have non-identical syntax.

6.2 Contributions revisited

The summary of the multiple contributions made by this PhD thesis is revisited below.

- Development of benchmark mono-lingual (Urdu) standard evaluation corpora for the Urdu paraphrased text reuse and extrinsic plagiarism detection.
- Development of a benchmark cross-lingual gold standard text reuse corpus for the English-Urdu language pair.
- Development of supporting lexical resources for the English-Urdu language pair.
- Evaluation and comparison of state-of-the-art mono-lingual methods for text reuse and extrinsic plagiarism detection for the Urdu language.
- Evaluation of state-of-the-art and newly proposed methods for cross-lingual (English-Urdu) text reuse detection.

- Newly proposed method for cross-lingual (English-Urdu) text reuse detection.
- Use of supporting lexical resources for cross-lingual (English-Urdu) text reuse detection.
- Custom training of multiple word and sentence embeddings models on an Urdu news corpus.

6.3 Research goals revisited

This thesis outlined six research goals initially when the research work was started (Section 1.2). In this section, these goals have been reviewed to assess how well they have been satisfied.

- **Explore the problem of text reuse and extrinsic plagiarism detection for an under-resourced Urdu language and English-Urdu language pair.**

  This research goal has been addressed in Chapter 1 and Chapter 2. Chapter 1 defined the basic concepts of text reuse and extrinsic plagiarism detection, its types, classifications, importance, and applications. Chapter 2 presented an extensive review of the existing corpora and methods for the mono- and cross-lingual text reuse and extrinsic plagiarism detection. It categorised and then described in detail each of the corpora and methods. Moreover, it also discussed the evaluation measures commonly used to estimate the performance of the text reuse and extrinsic plagiarism detection systems, i.e., precision, recall and F measure. Furthermore, it highlighted the scarcity of standard evaluation corpora and supporting resources (required by the majority of the methods) for the Urdu (or similar) language.

- **Develop benchmark gold standard mono-lingual text reuse and extrinsic plagiarism corpora for the Urdu language.**

- **Develop benchmark gold standard cross-lingual text reuse corpora for the English-Urdu language pair.**

  The above two research goals have been met in Chapter 3 which defined the efforts in developing benchmark standard evaluation corpora for the mono- (Urdu) and cross-lingual (English-Urdu) text reuse and extrinsic plagiarism. Two mono-lingual (Urdu) and one cross-lingual (English-Urdu) benchmark standard evaluation corpora have been created to promote the text reuse and extrinsic plagiarism research in an under-resourced language, i.e., Urdu. These resources were developed following standard practices, manually annotated, encoded in XML format, and are made publicly available to download for future research and replication.

- **Create supporting lexical resources that assist in the detection of cross-lingual (English-Urdu) text reuse cases.**

  This research goal has been achieved in Chapter 3 which described the creation of supporting resources for cross-lingual (English-Urdu) reuse text detection. A large-scale multi-domain English-Urdu parallel corpus has been developed by scrapping parallel text documents from the Web. Moreover, six bi-lingual (English-Urdu) dictionaries are also complied using different methods from online and offline sources. Both these supporting resources are saved in a standard format and made freely available to download for academic research.

- **Evaluate and compare the performance of state-of-the-art mono-lingual text reuse and extrinsic plagiarism detection methods on the Urdu corpora.**

  This research goal has been addressed in Chapter 4 which presented the mono-lingual (Urdu) text reuse and extrinsic plagiarism detection experiments performed on the COUNTER Corpus and UPPC Corpus. A variant set of state-of-the-art mono-lingual methods (i.e., Word n-grams overlap, Vector Space Model, Longest Common Subsequence, Greedy String Tiling, Local alignment,

Global alignment, Stop-word n-grams overlap, Sentence ratio, and Token ratio) are evaluated on both corpora with an aim to make a direct comparison and to find out which method works best for the Urdu language. Moreover, the rationale for using the variety of corpora (real and simulated) in the experiments was to better examine the performance of these methods. Furthermore, the effect of text pre-processing on the Urdu text reuse and extrinsic plagiarism detection is also investigated.

- **Develop or fine-tune methods for the mono- and cross-lingual text reuse and extrinsic plagiarism detection.**

  This research goal has been achieved in Chapter 5 which described the cross-lingual (English-Urdu) text reuse detection experiments performed on the TREU Corpus. A diverse range of cross-lingual text reuse detection methods classified under 3 categories, i.e., (1) Translation + Mono-lingual Analysis, (2) Cross-lingual Vector Space Model, and (3) Cross-lingual Embeddings, are used in the experiments. The main objective is to show how the corpus can be used in the development and evaluation of cross-lingual (English-Urdu) text reuse detection systems. A new method is also proposed to deal with cross-lingual (English-Urdu) text reuse cases. The method makes use of bi-lingual word embeddings and estimates the degree of overlap between the source and derived text documents using cosine similarity between word vectors. However, rather than averaging of all word vectors it only counts the maximum weighted cosine similarity between word pairs.

6.4 Future directions

The main focus of this thesis was on the development of corpora and methods for mono- (Urdu) and cross-lingual (English-Urdu) text reuse and extrinsic plagiarism detection. Although this research work made substantial contributions, text reuse

and extrinsic plagiarism detection is a vast field and there are improvements and refinements that could still be made. The following are some potential points for the future avenues of this research work.

- The methods used in this thesis attempt to estimate the degree of overlap between the source and derived texts at document-level. A possible future direction could be to explore and develop methods to identify the portion(s) of a source text document that is reused to create the derived text document.
- Urdu being an under-resourced language lacks basic NLP tools such as a word tokeniser, stemmer, and lemmatiser. If these tools are available [Shafi, 2019], the Urdu text reuse and extrinsic plagiarism detection results could be refined.
- The dictionaries used as supporting resources in the cross-lingual experiments do not consider POS information. One possible future direction could be to associate POS tag information when translating a word from the dictionary. The resulting filtering is likely to reduce the noise generated (where one word translates to many) during translation and would help in improving performance.
- Another potential future work is to explore the UCREL Semantic Analysis System (USAS) [Rayson et al., 2004] to measure the degree of overlap between the source and derived text documents. Moreover, as the USAS is now being extended to multi-lingual framework (including Urdu) [Piao et al., 2016], it could also be employed in the cross-lingual (English-Urdu) text reuse detection experiments.
- Word embeddings learn (and then predict) similar words that are close to each other (in similar contexts), however, they may not be synonyms (for instance, "spend" and "save" might appear close to each other, but have completely different meanings). On the contrary, the most commonly used strategy to

rewrite text is by replacing words with their synonyms. Therefore, another area of future work could be to explore the word embeddings as a way to get synonyms or to filter out antonyms from the list of similar words predicted by the word embedding models.

A

Complete Results using Translation+Mono-lingual Analysis

	Binary (F_1)						Ternary (F_1)					
method/classifier	**nb**	**j48**	**rf**	**svm**	**lr**	**mlp**	**nb**	**j48**	**rf**	**svm**	**lr**	**mlp**
Lexical overlap												
wno-uni-j	0.73	0.71	0.71	0.72	0.73	0.73	0.60	0.59	0.58	0.60	0.60	0.59
wno-bi-j	0.71	0.71	0.71	0.56	0.71	0.72	0.54	0.55	0.55	0.55	0.55	0.54
wno-tri-j	0.60	0.56	0.56	0.56	0.56	0.56	0.44	0.49	0.49	0.37	0.49	0.45
wno-four-j	0.53	0.56	0.56	0.56	0.56	0.56	0.38	0.37	0.37	0.35	0.35	0.42
wno-five-j	0.51	0.56	0.56	0.56	0.56	0.56	0.32	0.33	0.34	0.31	0.34	0.40
wno-c-j	0.62	0.72	0.70	0.73	0.74	0.72	0.51	0.57	0.54	0.60	0.59	0.59

wno-uni-d	0.73	0.70	0.71	0.73	0.73	0.73	0.59	0.58	0.58	0.59	0.59	0.59
wno-bi-d	0.70	0.70	0.70	0.56	0.70	0.72	0.55	0.56	0.56	0.55	0.56	0.54
wno-tri-d	0.62	0.56	0.56	0.56	0.56	0.57	0.46	0.49	0.48	0.38	0.48	0.46
wno-four-d	0.54	0.56	0.56	0.56	0.56	0.56	0.37	0.36	0.39	0.34	0.36	0.43
wno-five-d	0.42	0.56	0.56	0.56	0.56	0.56	0.32	0.34	0.33	0.32	0.34	0.41
wno-c-d	0.61	0.73	0.71	0.73	0.73	0.73	0.50	0.57	0.56	0.60	0.59	0.61
wno-uni-o	0.70	0.69	0.70	0.73	0.70	0.71	0.54	0.53	0.52	0.55	0.55	0.53
wno-bi-o	0.68	0.69	0.68	0.56	0.67	0.70	0.50	0.51	0.51	0.50	0.52	0.50
wno-tri-o	0.58	0.56	0.56	0.56	0.56	0.56	0.45	0.45	0.45	0.37	0.45	0.45
wno-four-o	0.56	0.56	0.56	0.56	0.56	0.56	0.38	0.36	0.41	0.33	0.35	0.42
wno-five-o	0.56	0.56	0.56	0.56	0.56	0.56	0.33	0.33	0.33	0.32	0.33	0.40
wno-c-o	0.56	0.71	0.67	0.56	0.71	0.71	0.44	0.52	0.49	0.54	0.55	0.53
wno-uni-c	0.61	0.64	0.63	0.56	0.64	0.67	0.47	0.45	0.48	0.39	0.48	0.48
wno-bi-c	0.57	0.58	0.63	0.56	0.64	0.67	0.51	0.50	0.50	0.39	0.50	0.50
wno-tri-c	0.58	0.56	0.56	0.56	0.56	0.56	0.45	0.46	0.47	0.37	0.44	0.44
wno-four-c	0.56	0.56	0.56	0.56	0.56	0.56	0.40	0.37	0.37	0.34	0.36	0.43
wno-five-c	0.60	0.56	0.56	0.56	0.56	0.56	0.39	0.33	0.34	0.31	0.33	0.40
wno-c-c	0.52	0.65	0.66	0.56	0.65	0.68	0.43	0.47	0.46	0.39	0.50	0.49
vsm-bow	0.67	0.68	0.66	0.56	0.68	0.69	0.49	0.49	0.47	0.49	0.50	0.49
vsm-c3g	0.62	0.63	0.64	0.56	0.63	0.66	0.47	0.46	0.46	0.37	0.47	0.46
vsm-c4g	0.62	0.61	0.65	0.56	0.64	0.68	0.48	0.49	0.51	0.40	0.48	0.49
vsm-c5g	0.58	0.61	0.65	0.56	0.64	0.69	0.49	0.49	0.49	0.38	0.49	0.47
String matching												
lcs	0.73	0.72	0.72	0.56	0.72	0.72	0.55	0.55	0.56	0.54	0.55	0.55
gst-mml-1	0.74	0.73	0.72	0.73	0.74	0.74	0.54	0.54	0.54	0.54	0.54	0.53
gst-mml-2	0.73	0.74	0.73	0.56	0.72	0.73	0.54	0.54	0.54	0.54	0.54	0.53
gst-mml-3	0.64	0.62	0.65	0.56	0.61	0.63	0.48	0.48	0.48	0.37	0.49	0.48
gst-mml-4	0.58	0.61	0.61	0.56	0.56	0.56	0.40	0.44	0.42	0.36	0.37	0.44
gst-mml-5	0.42	0.57	0.58	0.56	0.56	0.56	0.32	0.36	0.36	0.32	0.34	0.40
gst-mml-c	0.62	0.74	0.73	0.74	0.74	0.75	0.49	0.54	0.50	0.56	0.57	0.55
Structural similarity												
sno-uni-j	0.56	0.56	0.59	0.56	0.57	0.56	0.34	0.35	0.38	0.32	0.35	0.40
sno-bi-j	0.56	0.56	0.57	0.56	0.56	0.56	0.34	0.34	0.33	0.32	0.35	0.41
sno-tri-j	0.56	0.56	0.56	0.56	0.56	0.56	0.31	0.33	0.37	0.31	0.33	0.39
sno-four-j	0.56	0.56	0.56	0.56	0.56	0.56	0.31	0.28	0.29	0.31	0.31	0.36

sno-five-j	0.56	0.56	0.56	0.56	0.56	0.56	0.29	0.27	0.28	0.30	0.29	0.35
sno-c-j	0.49	0.57	0.60	0.56	0.57	0.56	0.40	0.41	0.40	0.33	0.36	0.43
sno-uni-d	0.57	0.56	0.60	0.56	0.57	0.56	0.36	0.36	0.39	0.31	0.35	0.39
sno-bi-d	0.56	0.56	0.57	0.56	0.56	0.56	0.34	0.35	0.36	0.31	0.35	0.41
sno-tri-d	0.56	0.56	0.57	0.56	0.56	0.56	0.33	0.35	0.41	0.32	0.33	0.40
sno-four-d	0.56	0.56	0.56	0.56	0.56	0.56	0.30	0.29	0.33	0.30	0.32	0.36
sno-five-d	0.56	0.56	0.56	0.56	0.56	0.56	0.28	0.27	0.27	0.30	0.28	0.35
sno-c-d	0.47	0.57	0.61	0.56	0.57	0.57	0.41	0.39	0.39	0.32	0.36	0.42
sno-uni-o	0.57	0.56	0.58	0.56	0.56	0.57	0.25	0.22	0.36	0.22	0.23	0.32
sno-bi-o	0.56	0.56	0.56	0.56	0.56	0.56	0.27	0.22	0.33	0.22	0.24	0.34
sno-tri-o	0.56	0.56	0.56	0.56	0.56	0.56	0.29	0.31	0.34	0.22	0.29	0.37
sno-four-o	0.56	0.56	0.56	0.56	0.56	0.56	0.28	0.27	0.31	0.26	0.29	0.35
sno-five-o	0.56	0.56	0.56	0.56	0.56	0.56	0.28	0.28	0.28	0.29	0.28	0.35
sno-c-o	0.43	0.56	0.60	0.56	0.56	0.56	0.36	0.33	0.35	0.26	0.29	0.34
sno-uni-c	0.57	0.58	0.59	0.56	0.56	0.56	0.34	0.30	0.36	0.22	0.27	0.37
sno-bi-c	0.57	0.56	0.57	0.56	0.56	0.57	0.31	0.35	0.35	0.22	0.30	0.37
sno-tri-c	0.56	0.56	0.57	0.56	0.56	0.56	0.32	0.32	0.36	0.28	0.32	0.36
sno-four-c	0.56	0.56	0.56	0.56	0.56	0.56	0.30	0.28	0.30	0.30	0.31	0.36
sno-five-c	0.56	0.56	0.56	0.56	0.56	0.56	0.28	0.26	0.27	0.29	0.28	0.35
sno-c-c	0.52	0.57	0.59	0.56	0.56	0.57	0.37	0.41	0.39	0.28	0.36	0.37
Mono-lingual word embeddings												
w2v-ae-pre	0.58	0.57	0.57	0.56	0.56	0.57	0.37	0.39	0.38	0.22	0.25	0.35
w2v-wae-pre	0.57	0.58	0.58	0.56	0.56	0.57	0.35	0.37	0.37	0.22	0.25	0.31
w2v-wme-pre	0.70	0.70	0.70	0.61	0.68	0.71	0.52	0.51	0.52	0.52	0.53	0.52
w2v-ae-cs	0.63	0.63	0.64	0.56	0.61	0.62	0.43	0.46	0.47	0.42	0.45	0.46
w2v-wae-cs	0.64	0.63	0.63	0.56	0.61	0.62	0.44	0.46	0.46	0.41	0.46	0.47
w2v-wme-cs	0.73	0.74	0.75	0.69	0.70	0.73	0.56	0.55	0.55	0.57	0.57	0.55
glv-ae-pre	0.58	0.57	0.57	0.56	0.57	0.57	0.33	0.35	0.34	0.22	0.22	0.33
glv-wae-pre	0.58	0.57	0.57	0.56	0.56	0.57	0.32	0.35	0.37	0.22	0.22	0.31
glv-wme-pre	0.70	0.70	0.71	0.59	0.65	0.70	0.50	0.51	0.51	0.51	0.51	0.52
glv-ae-cs	0.59	0.59	0.58	0.56	0.57	0.57	0.38	0.41	0.41	0.22	0.23	0.39
glv-wae-cs	0.60	0.58	0.59	0.56	0.57	0.58	0.41	0.43	0.42	0.22	0.36	0.42
glv-wme-cs	0.71	0.71	0.71	0.70	0.70	0.72	0.53	0.54	0.54	0.54	0.54	0.55
ft-ae-pre	0.57	0.57	0.58	0.56	0.57	0.56	0.32	0.35	0.38	0.22	0.24	0.31
ft-wae-pre	0.57	0.57	0.58	0.56	0.57	0.56	0.31	0.37	0.36	0.22	0.23	0.31

ft-wme-pre	0.70	0.70	0.70	0.60	0.69	0.71	0.51	0.51	0.51	0.50	0.52	0.51
ft-ae-cs	0.61	0.61	0.62	0.56	0.59	0.61	0.42	0.47	0.46	0.37	0.42	0.45
ft-wae-cs	0.62	0.62	0.63	0.56	0.59	0.61	0.42	0.44	0.45	0.36	0.43	0.44
ft-wme-cs	0.72	0.73	0.72	0.69	0.70	0.72	0.54	0.54	0.53	0.55	0.54	0.54
Mono-lingual sentence embeddings												
sent2vec-pre	0.60	0.59	0.59	0.56	0.57	0.57	0.38	0.38	0.39	0.22	0.30	0.36
sent2vec-cs	0.62	0.63	0.62	0.56	0.59	0.61	0.45	0.46	0.45	0.29	0.40	0.45
infersent-ft-pre	0.56	0.57	0.57	0.56	0.56	0.56	0.30	0.29	0.30	0.22	0.22	0.22
infersent-glv-pre	0.56	0.56	0.57	0.56	0.56	0.56	0.23	0.34	0.35	0.22	0.22	0.22
use-dan-pre	0.66	0.65	0.65	0.56	0.58	0.61	0.45	0.42	0.42	0.26	0.36	0.44
use-tra-pre	0.63	0.61	0.61	0.56	0.57	0.57	0.42	0.43	0.44	0.22	0.25	0.38
laser-pre	0.64	0.67	0.65	0.57	0.61	0.64	0.38	0.43	0.43	0.28	0.36	0.47

B

Complete Results using Cross-lingual Embeddings

	Binary (F_1)						Ternary (F_1)					
method/classifier	**nb**	**j48**	**rf**	**svm**	**lr**	**mlp**	**nb**	**j48**	**rf**	**svm**	**lr**	**mlp**
Cross-lingual word embeddings												
w2v-ae-ct	0.29	0.33	0.34	0.22	0.22	0.22	0.56	0.57	0.58	0.56	0.56	0.56
w2v-wae-ct	0..28	0.32	0.34	0.22	0.22	0.22	0.56	0.57	0.58	0.56	0.56	0.56
w2v-wme-ct	0.35	0.36	0.38	0.22	0.22	0.32	0.56	0.58	0.60	0.56	0.57	0.56
glv-ae-ct	0.30	0.33	0.33	0.22	0.22	0.22	0.55	0.55	0.56	0.54	0.56	0.56
glv-wae-ct	0.29	0.32	0.33	0.22	0.22	0.22	0.55	0.56	0.57	0.56	0.56	0.56
glv-wme-ct	0.34	0.35	0.34	0.22	0.22	0.28	0.56	0.57	0.56	0.54	0.56	0.56

ft-ae-ct	0.22	0.32	0.33	0.22	0.22	0.22	0.56	0.57	0.56	0.56	0.56	0.56
ft-wae-pre	0.22	0.33	0.32	0.22	0.22	0.22	0.56	0.57	0.56	0.56	0.56	0.56
ft-wme-pre	0.33	0.33	0.34	0.22	0.22	0.28	0.56	0.56	0.57	0.56	0.56	0.56
Cross-lingual sentence embeddings												
sent2vec-pre	0.22	0.33	0.33	0.22	0.22	0.22	0.56	0.56	0.56	0.56	0.56	0.56
laser-pre	0.39	0.41	0.40	0.31	0.33	0.33	0.65	0.63	0.63	0.56	0.62	0.62

Bibliography

[Agibetov et al., 2018] Agibetov, A., Blagec, K., Xu, H., and Samwald, M. (2018). Fast and Scalable Neural Embedding Models for Biomedical Sentence Classification. *BMC Bioinformatics*, 19(1):541–549.

[Aker et al., 2014] Aker, A., Paramita, M. L., Pinnis, M., and Gaizauskas, R. (2014). Bilingual Dictionaries for all EU Languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 2839–2845.

[Alam et al., 2015] Alam, S., Mehmood, F. U. D. B., and 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*, pages 47–56.

[Alfikri and Purwarianti, 2012] Alfikri, Z. F. and Purwarianti, A. (2012). The Construction of Indonesian-English Cross Language Plagiarism Detection System using Fingerprinting Technique. *Journal of Computer Science and Information*, 5(1):16–23.

[Aljohani and Mohd, 2014] Aljohani, A. and Mohd, M. (2014). Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm. *Information Technology Journal*, 13(14):2349.

[Allot et al., 2019] Allot, A., Chen, Q., Kim, S., Alvarez, R. V., Comeau, D. C., Wilbur, W. J., and Lu, Z. (2019). LitSense: Making Sense of Biomedical Literature at Sentence Level. *Nucleic Acids Research*, 47(1):594–599.

[Alzahrani et al., 2012] Alzahrani, S. M., Salim, N., and Abraham, A. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(2):133–149.

[Anwar et al., 2006] Anwar, W., Wang, X., and Wang, X.-l. (2006). A Survey of Automatic Urdu Language Processing. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 4489–4494.

[Apidianaki, 2008] Apidianaki, M. (2008). Translation-oriented Word Sense Induction based on Parallel Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

[Arefin et al., 2013] Arefin, M. S., Morimoto, Y., and Sharif, M. A. (2013). BAENPD: A Bilingual Plagiarism Detector. *Journal of Computers*, 8(5):1145–1156.

[Artetxe et al., 2016] Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

[Artetxe et al., 2018] Artetxe, M., Labaka, G., and Agirre, E. (2018). A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798.

[Artetxe and Schwenk, 2018] Artetxe, M. and Schwenk, H. (2018). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

[Asghari et al., 2015] Asghari, H., Khoshnava, K., Fatemi, O., and Faili, H. (2015). Developing Bilingual Plagiarism Detection Corpus using Sentence Aligned Parallel Corpus. In *Working Notes Papers of the PAN 2015 Evaluation Labs - CEUR Workshop Proceeding*.

[Baeza-Yates and Ribeiro-Neto, 2011] Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval - The Concepts and Technology behind Search*, volume 2. Addison-Wesley.

[Bai et al., 2018] Bai, M., Han, X., Jia, H., Wang, C., and Sun, Y. (2018). Transfer Pretrained Sentence Encoder to Sentiment Classification. In *IEEE 3rd International Conference on Data Science in Cyberspace*, pages 423–427.

[Baker et al., 2002] Baker, P., Hardie, A., McEnery, T., Cunningham, H., and Gaizauskas, R. J. (2002). EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

[Ballesteros and Croft, 1998] Ballesteros, L. and Croft, W. B. (1998). *Statistical Methods for Cross-Language Information Retrieval*, pages 23–40. Springer US.

[Bär et al., 2012] Bär, D., Zesch, T., and Gurevych, I. (2012). Text Reuse Detection using a Composition of Text Similarity Measures. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 167–184.

[Barrón-Cedeño et al., 2013a] Barrón-Cedeño, A., Gupta, P., and Rosso, P. (2013a). Methods for Cross-Language Plagiarism Detection. *Knowledge-Based Systems*, 50:211–217.

[Barrón-Cedeño et al., 2010] Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010). Plagiarism Detection Across Distant Language Pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45.

[Barrón-Cedeño et al., 2009] Barrón-Cedeño, A., Rosso, P., and Benedí, J.-M. (2009). Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 523–534.

[Barrón-Cedeño et al., 2013b] Barrón-Cedeño, A., Rosso, P., Devi, S. L., Clough, P., and Stevenson, M. (2013b). PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In *Proceedings of the Multilingual Information Access in South Asian Languages*, pages 59–70.

[Barrón-Cedeño et al., 2008] Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008). On Cross-lingual Plagiarism Analysis using a Statistical Model. In *Proceedings of the ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, pages 9–13.

[Barrón-Cedeño et al., 2013c] Barrón-Cedeño, A., Vila, M., Martí, M. A., and Rosso, P. (2013c). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4):917–947.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

[Bell, 1991] Bell, A. (1991). *The Language of News Media*, volume 1. Wiley-Blackwell.

[Bendersky and Croft, 2009] Bendersky, M. and Croft, W. B. (2009). Finding Text Reuse on the Web. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pages 262–271.

[Białecki et al., 2012] Białecki, A., Muir, R., and Ingersoll, G. (2012). Apache Lucene 4. In *Proceedings of the Special Interest Group on Information Retrieval Workshop on Open Source Information Retrieval*, pages 17–24.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, volume 1. O'Reilly Media, Inc.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

[Brin et al., 1995] Brin, S., Davis, J., and Garcia-Molina, H. (1995). Copy Detection Mechanisms for Digital Documents. In *Proceedings of the 1995 ACM International Conference on Management of Data*, pages 398–409.

[Broder, 1997] Broder, A. Z. (1997). On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29.

[Brown et al., 1993] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

[Buckley et al., 2000] Buckley, C., Mitra, M., Walz, J., and Cardie, C. (2000). Using clustering and SuperConcepts within SMART: TREC 6. *Information Processing & Management*, 36(1):109–131.

[Butakov and Scherbinin, 2009] Butakov, S. and Scherbinin, V. (2009). The Toolbox for Local and Global Plagiarism Detection. *Computers & Education*, 52(4):781–788.

[Callan et al., 2009] Callan, J., Hoy, M., Yoo, C., and Zhao, L. (2009). The ClueWeb09 Dataset. http://lemurproject.org/clueweb09/.

[Callison-Burch et al., 2004] Callison-Burch, C., Talbot, D., and Osborne, M. (2004). Statistical Machine Translation with Word and Sentence Aligned Parallel Corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 175.

[Caseli et al., 2006] Caseli, H. M., Maria das Graças, V. N., and Forcada, M. L. (2006). Automatic Induction of Bilingual Resources from Aligned Parallel Corpora: Application to Shallow-Transfer Machine Translation. *Machine Translation*, 20(4):227–245.

[Cer et al., 2018] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 169–174.

[Ceska et al., 2008] Ceska, Z., Toman, M., and Jezek, K. (2008). Multilingual Plagiarism Detection. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 83–92.

[Chen and Vines, 2014] Chen, H. Y. and Vines, P. (2014). Multi Queries Methods of the Chinese-English Bilingual Plagiarism Detection. In *Applied Mechanics and Materials*, volume 462, pages 1158–1162.

[Chiu et al., 2010] Chiu, S., Uysal, I., and Croft, W. B. (2010). Evaluating Text Reuse Discovery on the Web. In *Proceedings of the 3rd Symposium on Information Interaction in Context*, pages 299–304.

[Choi et al., 2018] Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

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

[Church, 1993] Church, K. W. (1993). Char_Align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 1–8.

[Cimiano et al., 2009] Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit Versus Latent Concept Models for Cross-language Information Retrieval. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1513–1518.

[Clough, 2003] Clough, P. (2003). *Measuring Text Reuse*. PhD Dissertation, University of Sheffield, UK.

[Clough, 2010] Clough, P. (2010). *Measuring Text Reuse in the News Industry*, pages 247–259. Cambridge University Press.

[Clough and Gaizauskas, 2009] Clough, P. and Gaizauskas, R. (2009). *Corpora and Text Re-use*, pages 1249–1271. Walter de Gruyter.

[Clough et al., 2002] Clough, P., Gaizauskas, R., Piao, S., and Wilks, Y. (2002). METER: MEasuring TExt Reuse. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pages 152–159.

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

[Cohen, 1960] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

[Cohen, 1968] Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

[Conneau and Kiela, 2018] Conneau, A. and Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

[Cormen et al., 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*, volume 3. MIT Press.

[Craig, 2004] Craig, H. (2004). *Stylistic Analysis and Authorship Studies*, pages 273–288. Blackwell Publishing.

[Culwin and Lancaster, 2001] Culwin, F. and Lancaster, T. (2001). Plagiarism Issues for Higher Education. *Vine*, 31(2):36–41.

[Daiya and Singh, 2018] Daiya, D. and Singh, A. (2018). Using Statistical and Semantic Models for Multi-Document Summarization. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing*, pages 169–183.

[Daud et al., 2017] Daud, A., Khan, W., and Che, D. (2017). Urdu Language Processing: a Survey. *Artificial Intelligence Review*, 47(3):279–311.

[Demner-Fushman and Oard, 2003] Demner-Fushman, D. and Oard, D. W. (2003). The Effect of Bilingual Term List Size on Dictionary-based Cross-language Information Retrieval. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, pages 108–118.

[Diederich, 2006] Diederich, J. (2006). Computational Methods to Detect Plagiarism in Assessment. In *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training*, pages 147–154.

[Dittmar et al., 2012] Dittmar, C., Hildebrand, K. F., Gaertner, D., Winges, M., Müller, F., and Aichroth, P. (2012). Audio Forensics Meets Music Information Retrieval - A Toolbox for Inspection of Music Plagiarism. In *Proceedings of the 20th European Signal Processing Conference*, pages 1249–1253.

[Eaton, 2004] Eaton, L. (2004). A Quarter of UK Students are Guilty of Plagiarism - Survey Shows. *British Medical Journal*, 329(7457):70.

[Ebeling, 1998] Ebeling, J. (1998). Contrastive Linguistics, Translation, and Parallel corpora. *Meta: Journal des traducteurs*, 43(4):602–615.

[Eisele and Chen, 2010] Eisele, A. and Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

[Ekbal et al., 2012] Ekbal, A., Saha, S., and Choudhary, G. (2012). Plagiarism Detection in Text using Vector Space Model. In *Proceedings of the 12th International Conference on Hybrid Intelligent Systems*, pages 366–371.

[Elhadi and Al-Tobi, 2009] Elhadi, M. and Al-Tobi, A. (2009). Duplicate Detection in Documents and Webpages using Improved Longest Common Subsequence and

Documents Syntactical Structures. In *Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology*, pages 679–684.

[Fantinuoli and Zanettin, 2015] Fantinuoli, C. and Zanettin, F. (2015). *Creating and using Multilingual Corpora in Translation Studies*, pages 1–10. Language Science Press Berlin.

[Ferrero et al., 2016] Ferrero, J., Agnes, F., Besacier, L., and Schwab, D. (2016). A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

[Ferrero et al., 2017] Ferrero, J., Besacier, L., Schwab, D., and Agnès, F. (2017). Using word embedding for cross-language plagiarism detection. In *Proceedings of the 15th International Conference of the European Chapter of the Association for Computational Linguistics*, pages 415–421.

[Fortuna, 2004] Fortuna, B. (2004). Kernel Canonical Correlation Analysis with Applications. In *Proceedings of the Conference on Data Mining and Warehouses*, pages 12–15.

[Franco-Salvador et al., 2014] Franco-Salvador, M., Gupta, P., and Rosso, P. (2014). *Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing*, pages 227–236. Springer Berlin Heidelberg.

[Franco-Salvador et al., 2016] Franco-Salvador, M., Rosso, P., and Montes-y Gómez, M. (2016). A Systematic Study of Knowledge Graph Analysis for Cross-Language Plagiarism Detection. *Information Processing & Management*, 52(4):550–570.

[Freitas and Liu, 2017] Freitas, C. and Liu, Y. (2017). Exploring the Differences between Human and Machine Translation. Technical report, WWU Honors Program Senior Projects - Western Washington University.

[Fries, 1997] Fries, U. (1997). Summaries in Newspapers: A Textlinguistic Investigation. *The Structure of Texts*, pages 47–63.

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

[Ganguly et al., 2018] Ganguly, D., Jones, G. J., Ramírez-De-La-Cruz, A., Ramírez-De-La-Rosa, G., and Villatoro-Tello, E. (2018). Retrieving and Classifying Instances of Source Code Plagiarism. *Information Retrieval Journal*, 21(1):1–23.

[Gella et al., 2017] Gella, S., Sennrich, R., Keller, F., and Lapata, M. (2017). Image Pivoting for Learning Multilingual Multimodal Representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845.

[Giguere, 2019] Giguere, M. (2019). Dance Trends: Choreographic Plagiarism: When Does Borrowing Become Stealing? *Dance Education in Practice*, 5(1):29–32.

[Gipp and Meuschke, 2011] Gipp, B. and Meuschke, N. (2011). Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pages 249–258.

[Gipp et al., 2011] Gipp, B., Meuschke, N., and Beel, J. (2011). Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using Gut-

tenPlag. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pages 255–258.

[Gipp et al., 2014] Gipp, B., Meuschke, N., Breitinger, C., Pitman, J., and Nürnberger, A. (2014). Web-based Demonstration of Semantic Similarity Detection Using Citation Pattern Visualization for a Cross Language Plagiarism Case. In *Proceedings of the 16th International Conference on Enterprise Information Systems*, pages 677–683.

[Gitchell and Tran, 1999] Gitchell, D. and Tran, N. (1999). Sim: A Utility for Detecting Similarity in Computer Programs. In *Proceedings of the 30th ACM Special Interest Group on Computer Science Education Technical Symposium*, pages 266–270.

[Gravano et al., 1999] Gravano, L., García-Molina, H., and Tomasic, A. (1999). GlOSS: Text-source Discovery Over the Internet. *ACM Transactions on Database Systems*, 24(2):229–264.

[Grossman et al., 1997] Grossman, D. A., Frieder, O., Holmes, D. O., and Roberts, D. C. (1997). Integrating Structured Data and Text: A Relational Approach. *Journal of the American Society for Information Science*, 48(2):122–132.

[Grozea et al., 2009] Grozea, C., Gehl, C., and Popescu, M. (2009). ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In *Proceedings of the 3rd PAN Workshop - Uncovering Plagiarism, Authorship and Social Software Misuse - SEPLN'09*, pages 10–18.

[Gupta et al., 2016] Gupta, D., Vani, K., and Leema, L. (2016). Plagiarism Detection in Text Documents using Sentence Bounded Stop Word N-Grams. *Journal of Engineering Science and Technology*, 11(10):1403–1420.

[Gupta et al., 2012] Gupta, P., Barrón-Cedeño, A., and Rosso, P. (2012). Cross-language High Similarity Search Using a Conceptual Thesaurus. In *Proceedings*

of the 3rd International Conference on Information Access Evaluation: Multilinguality, Multimodality, and Visual Analytics, pages 67–75.

[Gupta et al., 2019] Gupta, P., Pagliardini, M., and Jaggi, M. (2019). Better Word Embeddings by Disentangling Contextual n-Gram Information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 933–939.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: an Update. *ACM Special Interest Group on Knowledge Discovery and Data Mining - Explorations Newsletter*, 11(1):10–18.

[Hanif et al., 2015] Hanif, I., Nawab, R. M. A., Arbab, A., Jamshed, H., Riaz, S., and Munir, E. U. (2015). Cross-Language Urdu-English (CLUE) Text Alignment Corpus. In *Working Notes Papers of the PAN 2015 Evaluation Labs - CEUR Workshop Proceeding*.

[Hauer et al., 2017] Hauer, B., Nicolai, G., and Kondrak, G. (2017). Bootstrapping Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 619–624.

[Havasi et al., 2007] Havasi, C., Speer, R., and Alonso, J. (2007). ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Proceedings of the 2007 International Conference on Recent Advances in Natural Language Processing*, pages 27–29.

[Hoad and Zobel, 2003] Hoad, T. C. and Zobel, J. (2003). Methods for Identifying Versioned and Plagiarized Documents. *Journal of the Association for Information Science and Technology*, 54(3):203–215.

[Holmes et al., 1994] Holmes, G., Donkin, A., and Witten, I. (1994). WEKA: A Machine Learning Workbench. In *Proceedings of the Australian New Zealnd Intelligent Information Systems Conference*, pages 357–361.

[Hutchins, 2005] Hutchins, J. (2005). Current Commercial Machine Translation Systems and Computer-based Translation Tools: System Types and their Uses. *International Journal of Translation*, 17(1):5–38.

[Irvine and Callison-Burch, 2013] Irvine, A. and Callison-Burch, C. (2013). Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 518–523.

[Jabbar et al., 2018] Jabbar, A., Iqbal, S., Khan, M. U. G., and Hussain, S. (2018). A Survey on Urdu and Urdu like Language Stemmers and Stemming Techniques. *Artificial Intelligence Review*, 49(3):339–373.

[Jawaid and Zeman, 2011] Jawaid, B. and Zeman, D. (2011). Word-Order Issues in English-to-Urdu Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 95(1):87–106.

[Jing and McKeown, 1999] Jing, H. and 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*, pages 129–136.

[Junczys-Dowmunt and Szał, 2011] Junczys-Dowmunt, M. and Szał, A. (2011). SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation. In *Proceedings of the International Joint Conferences on Security and Intelligent Information Systems*, pages 379–390.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Prentice Hall.

[Kazakov and Shahid, 2013] Kazakov, D. and Shahid, A. R. (2013). Using Parallel Corpora for Word Sense Disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 336–341.

[Keck, 2006] Keck, C. (2006). The Use of Paraphrase in Summary Writing: A Comparison of L1 and L2 Writers. *Journal of Second Language Writing*, 15(4):261–278.

[Kent and Salim, 2010] Kent, C. K. and Salim, N. (2010). Web based Cross Language Plagiarism Detection. In *Proceedings of the 2nd International Conference on Computational Intelligence, Modelling and Simulation*, pages 199–204.

[Kenter and De Rijke, 2015] Kenter, T. and De Rijke, M. (2015). Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1411–1420.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

[Köhler and Weber-Wul, 2010] Köhler, K. and Weber-Wul, D. (2010). Plagiarism Detection Test 2010. Technical report, HTW Berlin.

[Kothwal and Varma, 2013] Kothwal, R. and Varma, V. (2013). Cross Lingual Text Reuse Detection Based on Keyphrase Extraction and Similarity Measures. In *Proceedings of the Multilingual Information Access in South Asian Languages*, pages 71–78.

[Lane et al., 2006] Lane, P. C., Lyon, C. M., and Malcolm, J. A. (2006). Demonstration of the Ferret Plagiarism Detector. In *Proceedings of the 2nd International Plagiarism Conference*.

[Lee et al., 2008] Lee, C.-H., Wu, C.-H., and Yang, H.-C. (2008). A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection. In *Proceedings of the 3rd International Conference on Innovative Computing Information and Control*, pages 303–303.

[Lee et al., 2017] Lee, J.-U., Eger, S., Daxenberger, J., and Gurevych, I. (2017). Deep Learning for Aspect Based Sentiment Detection. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 22–29.

[Levy et al., 2017] Levy, O., Søgaard, A., and Goldberg, Y. (2017). A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 765–774.

[Li and Gaussier, 2010] Li, B. and Gaussier, E. (2010). Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652.

[Li et al., 2016] Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2016). Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 366–371.

[Littman et al., 1998] Littman, M. L., Dumais, S. T., and Landauer, T. K. (1998). *Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing*, pages 51–62. Springer US.

[Logue, 2004] Logue, R. (2004). Plagiarism: The Internet Makes it Easy. *Nursing Standard*, 18(51).

[Lu et al., 2004] Lu, W.-H., Chien, L.-F., and Lee, H.-J. (2004). Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems*, 22(2):242–269.

[Lukashenko et al., 2007] Lukashenko, R., Graudina, V., and Grundspenkis, J. (2007). Computer-based Plagiarism Detection Methods and Tools: An Overview. In *Proceedings of the 2007 International Conference on Computer Systems and Technologies*, pages 40:1–40:6.

[Lyon et al., 2001] Lyon, C., Malcolm, J., and Dickerson, B. (2001). Detecting Short Passages of Similar Text in Large Document Collections. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 118–125.

[Manning et al., 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, volume 1. MIT Press.

[Marcus et al., 1993] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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

[Martınez, 2009] Martınez, A. (2009). Wikipedia Usage by Mexican Students. The Constant Usage of Copy and Paste. In *Proceedings of the 5th Annual International Wikimania Conference*.

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

[McCarthy and Jarvis, 2010] McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*, 42(2):381–392.

[McNamee and Mayfield, 2004] McNamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97.

[Meng et al., 2017] Meng, F., Lu, W., Zhang, Y., Cheng, J., Du, Y., and Han, S. (2017). Qlut at SemEval-2017 Task 1: Semantic Textual Similarity based on Word Embeddings. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 150–153.

[Mihalcea et al., 2006] Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st Conference of American Association for Artificial Intelligence*, pages 775–780.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

[Mogadala and Rettinger, 2016] Mogadala, A. and Rettinger, A. (2016). Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-language Text Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 692–702.

[Montes-y Gómez et al., 2001] Montes-y Gómez, M., Gelbukh, A., Lopez-Lopez, A., and Baeza-Yates, R. (2001). Flexible Comparison of Conceptual Graphs. In *Proceedings of the International Conference on Database and Expert Systems Applications*, pages 102–111.

[Muhr et al., 2010] Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010). External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System. In *Working Notes Papers of the PAN 2010 Evaluation Labs - CEUR Workshop Proceeding*.

[Mukund et al., 2010] Mukund, S., Srihari, R., and Peterson, E. (2010). An Information Extraction System for Urdu: A Resource Poor Language. *ACM Transactions on Asian Language Information Processing*, 9(4):15.

[Muneer et al., 2019] Muneer, I., Sharjeel, M., Iqbal, M., Nawab, R. M. A., and Rayson, P. (2019). CLEU-A Cross-language English-Urdu Corpus and Benchmark for Text Reuse Experiments. *Journal of the Association for Information Science and Technology*, 70(7):729–741.

[Munteanu et al., 2004] Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 265–272.

[Navigli and Ponzetto, 2010] Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225.

[Nawab, 2012] Nawab, R. M. A. (2012). *Mono-lingual Paraphrased Text Reuse and Plagiarism Detection*. PhD Dissertation, University of Sheffield, UK.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453.

[Nie et al., 1999] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81.

[Oberreuter et al., 2011] Oberreuter, G., L'Huillier, G., Rıos, S. A., and Velásquez, J. D. (2011). Approaches for Intrinsic and External Plagiarism Detection. In *Working Notes Papers of the PAN 2011 Evaluation Labs - CEUR Workshop Proceeding*.

[Och and Ney, 2003] Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

[Osman et al., 2012] Osman, A. H., Salim, N., and Abuobieda, A. (2012). Survey of Text Plagiarism Detection. *Computer Engineering and Applications Journal*, 1(1):37–45.

[Pagliardini et al., 2018] Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 528–540.

[Pataki, 2012] Pataki, M. (2012). A New Approach for Searching Translated Plagiarism. In *Proceedings of the 5th International Plagiarism Conference*.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, È. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(1):2825–2830.

[Peirsman and Padó, 2010] Peirsman, Y. and Padó, S. (2010). Cross-lingual Induction of Selectional Preferences with Bilingual Vector Spaces. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

[Pereira et al., 2010] Pereira, R. C., Moreira, V. P., and Galante, R. (2010). A New Approach for Cross-Language Plagiarism Analysis. In *Proceedings of the CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation*, pages 15–26.

[Piao et al., 2016] Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., and Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2614–2619.

[Pinto et al., 2009] Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., and Rosso, P. (2009). A Statistical Approach to Cross-lingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60.

[Pirkola et al., 2001] Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4(3):209–230.

[Porter, 2009] Porter, M. (2009). *Beyond Text based Plagiarism: A Paradigm for Tackling Academic Misconduct in the Creative Disciplines*, volume 1. Northumbria University.

[Post et al., 2012] Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. In *Proceedings of the NAACL 7th Workshop on Statistical Machine Translation*, pages 401–409.

[Potthast et al., 2010a] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010a). Overview of the 2nd International Competition on Plagiarism Detection. In *Working Notes Papers of the PAN 2010 Evaluation Labs - CEUR Workshop Proceeding*, pages 1–14.

[Potthast et al., 2010b] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010b). PAN Plagiarism Corpus - PAN-PC-10. http://pan.webis.de/clef10/pan10-web/plagiarism-detection.html.

[Potthast et al., 2011a] Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011a). Cross-language Plagiarism Detection. *Language Resources and Evaluation*, 45(1):45–62.

[Potthast et al., 2011b] Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011b). Overview of the 3rd International Competition on Plagiarism Detection. In *Working Notes Papers of the PAN 2011 Evaluation Labs - CEUR Workshop Proceeding*, pages 1–10.

[Potthast et al., 2012a] Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012a). Overview of the 4th International Competition on Plagiarism Detection. In *Working Notes Papers of the PAN 2012 Evaluation Labs - CEUR Workshop Proceeding*, pages 1–28.

[Potthast et al., 2012b] Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012b). PAN Plagiarism Corpus - PAN-PC-12. http://pan.webis.de/clef12/pan12-web/plagiarism-detection.html.

[Potthast et al., 2015] Potthast, M., Göring, S., Rosso, P., and Stein, B. (2015). Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In *Working Notes Papers of the PAN 2015 Evaluation Labs - CEUR Workshop Proceeding*.

[Potthast et al., 2014] Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., and Stein, B. (2014). Overview of the 6th International Competition on Plagiarism Detection. In *Working Notes Papers of the PAN 2014 Evaluation Labs - CEUR Workshop Proceeding*, pages 1–32.

[Potthast et al., 2013a] Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., and Stein, B. (2013a). Overview of the 5th International Competition on Plagiarism Detection. In *Working Notes Papers of the PAN 2013 Evaluation Labs - CEUR Workshop Proceeding*, pages 1–31.

[Potthast et al., 2013b] Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., and Stein, B. (2013b). PAN Plagiarism Corpus - PAN-PC-13. http://pan.webis.de/clef13/pan13-web/plagiarism-detection.html.

[Potthast et al., 2013c] Potthast, M., Hagen, M., Völske, M., and Stein, B. (2013c). Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1212–1221.

[Potthast et al., 2009a] Potthast, M., Stein, B., Andreas, E., Barrón-Cedeño, A., and Rosso, P. (2009a). Overview of the 1st International Competition on Plagiarism Detection. In *Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 1–9.

[Potthast et al., 2009b] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009b). PAN Plagiarism Corpus - PAN-PC-09. http://pan.webis.de/sepln09/pan09-web/plagiarism-detection.html.

[Pouliquen et al., 2003] Pouliquen, B., Steinberger, R., and Ignat, C. (2003). Automatic Linking of Similar Texts Across Languages. *Recent Advances in Natural Language Processing III. Selected Papers from RANLP'03*, 260:307–316.

[Pupovac et al., 2008] Pupovac, V., Bilic-Zulle, L., and Petrovecki, M. (2008). On Academic Plagiarism in Europe. An Analytical Approach based on Four Studies. *Digithum*, 10:13–19.

[Rayson et al., 2004] Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004). The UCREL Semantic Analysis System. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, LREC-04*, pages 7–12.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

[Resnik and Smith, 2003] Resnik, P. and Smith, N. A. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

[Rizvi and Hussain, 2005] Rizvi, S. J. and Hussain, M. (2005). Analysis, Design and Implementation of Urdu Morphological Analyzer. In *Proceedings of the Student Conference on Engineering Sciences and Technology*, pages 1–7.

[Runeson et al., 2007] Runeson, P., Alexandersson, M., and Nyholm, O. (2007). Detection of Duplicate Defect Reports using Natural Language Processing. In *Proceedings of the 29th International Conference on Software Engineering*, pages 499–510.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

[Sameen et al., 2017] Sameen, S., Sharjeel, M., Nawab, R. M. A., Rayson, P., and Muneer, I. (2017). Measuring Short Text Reuse for the Urdu Language. *IEEE Access*, 6(1):7412–7421.

[Sanchez-Perez et al., 2014] Sanchez-Perez, M. A., Sidorov, G., and Gelbukh, A. (2014). A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In *Working Notes Papers of the PAN 2014 Evaluation Labs - CEUR Workshop Proceeding*, pages 1004–1011.

[Sari et al., 2017] Sari, Y., Vlachos, A., and Stevenson, M. (2017). Continuous n-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 267–273.

[Schafer and Yarowsky, 2002] Schafer, C. and Yarowsky, D. (2002). Inducing Translation Lexicons via Diverse Similarity Measures and Bridge Languages. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–7.

[Schrimsher et al., 2011] Schrimsher, R. H., Northrup, L. A., and Alverson, S. P. (2011). A Survey of Samford University Students Regarding Plagiarism and Academic Misconduct. *International Journal for Educational Integrity*, 7(1).

[Sérasset, 2015] Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-based Multilingual Lexical Resource in RDF. *Semantic Web*, 6(4):355–361.

[Shafi, 2019] Shafi, J. (2019). An Urdu Semantic Tagger - Lexicons, Corpora, Methods and Tools. unpublished thesis.

[Sharjeel et al., 2017] Sharjeel, M., Nawab, R. M. A., and Rayson, P. (2017). COUNTER: Corpus of Urdu News Text Reuse. *Language Resources and Evaluation*, 51(3):777–803.

[Sharjeel et al., 2016] Sharjeel, M., Rayson, P., and Nawab, R. M. A. (2016). UPPC - Urdu Paraphrase Plagiarism Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

[Shezaf and Rappoport, 2010] Shezaf, D. and Rappoport, A. (2010). Bilingual Lexicon Generation using Non-aligned Signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 98–107.

[Shivakumar and Garcia-Molina, 1995] Shivakumar, N. and 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*.

[Simard et al., 1993] Simard, M., Foster, G. F., and Isabelle, P. (1993). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research*, pages 1071–1082.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[Sorg and Cimiano, 2012] Sorg, P. and Cimiano, P. (2012). Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data & Knowledge Engineering*, 74:26–45.

[Sousa-Silva, 2015] Sousa-Silva, R. (2015). Reporter Fired for Plagiarism: A Forensic Linguistic Analysis of News Plagiarism. *Oslo Studies in Language*, 7(1):301–322.

[Stamatatos, 2009] Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

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

[Stein et al., 2007] Stein, B., Zu Eissen, S. M., and Potthast, M. (2007). Strategies for Retrieving Plagiarized Documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–826.

[Steinberger et al., 2014] Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An Overview of the European Union's Highly Multilingual Parallel Corpora. *Language Resources and Evaluation*, 48(4):679–707.

[Steinberger et al., 2002] Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, pages 415–424.

[Steinberger et al., 2006] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A Multilingual

Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147.

[Upadhyay et al., 2016] Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1661–1670.

[Van Der Eijk et al., 1992] Van Der Eijk, P., Bloksma, L., and Van Der Kraan, M. (1992). Towards developing reusable nlp dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics*.

[Vernon et al., 2001] Vernon, R. F., Bigna, S., and Smith, M. L. (2001). Plagiarism and the Web. *Journal of Social Work Education*, 37(1):193–196.

[Vila et al., 2014] Vila, M., Martí, M. A., and Rodríguez, H. (2014). Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology. *Open Journal of Modern Linguistics*, 4(01):205–218.

[Vinokourov et al., 2002] Vinokourov, A., Cristianini, N., and Shawe-Taylor, J. (2002). Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In *Proceedings of the 15th Neural Information Processing Systems Conference*, pages 1497–1504.

[Vinokourov and Girolami, 2002] Vinokourov, A. and Girolami, M. (2002). A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections. *Journal of Intelligent Information Systems*, 18(2-3):153–172.

[Vossen, 1998] Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, volume 1. Springer.

[Vossen, 2004] Vossen, P. (2004). EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17(2):161–173.

[Vulić and Korhonen, 2016] Vulić, I. and Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 247–257.

[Vulić and Moens, 2015] Vulić, I. and Moens, M.-F. (2015). Monolingual and Cross-lingual Information Retrieval Models based on (Bilingual) Word Embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 Empirical Methods in Natural Language Processing*, pages 353–355.

[Weber-Wulff, 2008] Weber-Wulff, D. (2008). On the Utility of Plagiarism Detection Software. In *Proceedings of the 3rd International Plagiarism Conference*.

[Weber-Wulff, 2013] Weber-Wulff, D. (2013). Tests of Plagiarism Software. http://plagiat.htw-berlin.de/software-en.

[Weber-Wulff, 2014] Weber-Wulff, D. (2014). *False Feathers: A Perspective on Academic Plagiarism*, volume 1. Springer Berlin Heidelberg.

[Wilcoxon et al., 1970] Wilcoxon, F., Katti, S., and 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.

[Wilks, 2004] Wilks, Y. (2004). On the Ownership of Text. *Computers and the Humanities*, 38(2):115–127.

[Wise, 1992] Wise, M. J. (1992). Detection of Similarities in Student Programs: YAP'Ing May Be Preferable to Plague'Ing. In *Proceedings of the 23rd ACM Special Interest Group on Computer Science Education Technical Symposium*, pages 268–271.

[Wise, 1993] Wise, M. J. (1993). *Running Karp-Rabin Matching and Greedy String Tiling*, volume 1. University of Sydney.

[Wise, 1995] Wise, M. J. (1995). Neweyes: a System for Comparing Biological Sequences using the Running Karp-Rabin Greedy String-Tiling algorithm. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pages 393–401.

[Wise, 1996] Wise, M. J. (1996). YAP3: Improved Detection of Similarities in Computer Program and Other Texts. In *Proceedings of the 27th ACM Special Interest Group on Computer Science Education Technical Symposium*, pages 130–134.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*, volume 4. Morgan Kaufmann.

[Wood, 2004] Wood, G. (2004). Academic Original Sin: Plagiarism, the Internet, and Librarians. *The Journal of Academic Librarianship*, 3(30):237–242.

[Yaghoobzadeh et al., 2019] Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H. (2019). Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753.

[Youmans, 1990] Youmans, G. (1990). Measuring Lexical Style and Competence: The Type-Token Vocabulary Curve. *Style*, 24(4):584–599.

[Yu et al., 2017] Yu, L.-C., Wang, J., Lai, K. R., and Zhang, X. (2017). Refining Word Embeddings for Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

[Yule, 1939] 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.

[Zhu et al., 2015] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.