Design and Development of a Plagiarism Corpus in Thai for Plagiarism Detection

Santipong Thaiprayoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
santipong.tha@nectec.or.th

Pornpimon Palingoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
pornpimon.pal@nectec.or.th

Kanokorn Trakultaweekoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Pathumthani, Thailand
kanokorn.tra@nectec.or.th

***Abstract*—One of the main problems of creating a plagiarism corpus in Thai is that it is quite a difficult task to acquire the plagiarized documents with real cases due to the copyright issue. To solve the problem, we present a design and development of a Thai plagiarism corpus to evaluate and compare plagiarism detection algorithms for Thai. The corpus is developed by using the simulated plagiarism method based on Thai Wikipedia articles and web page articles. For this method, we provide a Thai plagiarism annotation tool and a Thai plagiarism guideline for assisting human annotators to plagiarize text passages. Our corpus contains simulated cases of plagiarized documents based on four classes of Thai plagiarism and linguistic mechanisms including copy-based change, lexicon-based change, structure-based change, and semantic-based change. We show that the suspicious documents in the corpus are manually created by using different obfuscation strategies, which make the suspicious documents more realistic and challenging. We then believe that the corpus developed in this paper will be a valuable contribution in the development, comparison, and evaluation of plagiarism detection algorithms. Moreover, our corpus is free and publicly available for research purposes.**

Keywords—Thai plagiarism corpus, corpus construction, plagiarism detection, obfuscation strategies, natural language processing

I. INTRODUCTION

In recent years, plagiarism is a crucial problem that attracts a lot of attention in academic and educational communities [1], [2]. With the ability to easily access academic information, this make it easy to plagiarize some documents by copying and modifying texts obtained from the online academic sources without proper acknowledgements. As a result, the number of plagiarisms has increased dramatically in higher education institutions. To overcome plagiarism problems, a myriad of plagiarism detection algorithms and tools have been developed to deal with this issue. Previous research studies on plagiarism detection algorithms [3] presented that two approaches are generally used to detect plagiarism: intrinsic and external approaches. In the intrinsic approach, plagiarized texts in a suspicious document are detected by analyzing writing style without comparing to source documents. The external approach, on the other hand, adopts different techniques to find plagiarized texts in a suspicious document by comparing them against a set of source documents. To develop and evaluate the algorithms of external plagiarism detection in Thai, we need a benchmark corpus containing example cases that resemble the real cases of plagiarism.

In many languages, benchmark corpora [4], [5], [6], [7] have been developed previously to meet the proposes such as Arabic, Persian, and English. Most corpora are available in English language. The benchmark corpora in Thai language are the lack of a standardized evaluation resources. Moreover, the simulated plagiarism construction usually lacks annotation guidelines and tools to help human annotators to imitate plagiarizing the text passages. In general, plagiarism corpus construction can be divided into three major methods: (1) real cases of plagiarism (real plagiarized documents from someone else's work), (2) simulated plagiarism (manually created plagiarized documents by humans to simulate plagiarism cases), and (3) artificial plagiarism (automatically generated plagiarized documents by machine) [6]. However, the construction of the corpus containing cases of real plagiarism is quite expensive and difficult due to copyright issue. Therefore, we choose the simulated plagiarism method because it is more realistic in terms of simulating the real behavior of plagiarists and can create plagiarized text passages that are not much different from the real cases of plagiarism.

To construct a Thai plagiarism corpus, we present a design and development of a Thai plagiarism corpus containing simulated cases of plagiarized documents. However, as previously mentioned, the lack of Thai annotation guidelines makes it hard to implement the simulated plagiarism. To overcome the problem, we provide a Thai plagiarism annotation tool called *PlaTool* and a Thai plagiarism guideline for assisting human annotators to plagiarize the text passages. We aim to design and develop the tool that imitates the plagiarism scenario from plagiarists as closely as possible. This tool is designed to support plagiarizing text passages in four classes of Thai plagiarism. We use Thai Wikipedia articles as the main sources for plagiarizing text passages. In the simulated plagiarism method, we manually create suspicious documents using four classes of Thai plagiarism and linguistic mechanisms including copy-based change, lexicon-based change, structure-based change, and semantic-based change. The results show that the suspicious documents in our corpus are created using different obfuscation strategies, which makes the suspicious documents more realistic and challenging. We believe that our corpus will be a valuable contribution to the evaluation of Thai plagiarism detection algorithm.

The rest of the paper is organized as follows. In the next section, we review some related works on construction of plagiarism corpus in any languages, existing corpora, and plagiarism cases and linguistic mechanisms in Thai. In section 3, the process of corpus construction is sequentially described. In section 4, we show the details of Thai plagiarism corpus. Finally, the conclusion and suggestions for future works are presented.

II. RELATD WORK

Recently, the research in plagiarism pays a lot of attention to the construction of plagiarism corpora in different languages. In general, research shows that plagiarism corpora are constructed with the purpose of comparison and evaluation

978-1-7281-3003-3/19/$31.00 ©2019 IEEE

of plagiarism detection algorithms. Most of these algorithms rely on plagiarism corpora.

A. Existing Plagiarism Corpora

The short answer corpus [1] contains plagiarized and non-plagiarized texts in English language. The corpus is manually created by human consisting of 100 documents of length between 200-300 words. The documents are created with four levels of reuse, namely, near copy, light revision, heavy revision, and non-plagiarism. The corpus has 5 source documents which are used to create 57 plagiarized and 38 non-plagiarized documents. The PAN-PC corpora [5] are based on Project Gutenberg1 books and largely contain automatically generated artificial plagiarism cases. However, the later versions contain sufficient number of manually paraphrased cases, though in English language only PAN-PC-10 [5] and PAN-PC-11 [8] corpus contains 3,671 and 4,609 cases respectively). The P4P corpus [10] was built using examples of simulated plagiarism passages found in the PAN-PC-10 Corpus. It contains 847 paraphrase sentence pairs in English language of length 50 words or less. The METER corpus [10] was manually annotated with three different levels of text reuse: verbatim, rewrite, and new. The corpus consists of news stories collected during the period of 12 months between 1999 and 2000 in law and business domains.

B. Thai Plagiarism Corpus

In previous research studies of plagiarism corpus in Thai, Supawat [2] described a design and process in creating academic Thai plagiarism corpus, a corpus that collects simulated academic plagiarism texts in Thai. The corpus uses two main methods: manually created by participants and automatically generated by a program. This corpus consists of two main types of texts: plagiarized and non-plagiarized texts. Plagiarized texts are categorized into four types based on the degree of linguistic mechanisms used in plagiarism.

C. Plagiarism Cases and Linguistic Mechanisms in Thai

This section provides the classification of Thai plagiarism cases, which are used to obfuscate text passages from the source document to become plagiarized texts in the suspicious document. Barron [9] described that plagiarism cases are classified as morpholexicon-based, structure-based, semantic-based, and miscellaneous changes. Their theoretical framework is appropriate for our study because it attempted to systematically capture general related-paraphrase patterns identified as plagiarism cases. Moreover, it is useful for the study of plagiarism cases in Thai which is characterized as the context-oriented language containing multiple types of paraphrasing and semantic-based modifications. We adopted three main classes of the framework, i.e., morpholexicon-based, structure-based, and semantic-based, with a slight modification to the first class. This is because the morpholexicon-based implies morphological patterns (e.g., inflectional, modal verb, and derivation changes) which is present in English but absent from Thai. Therefore, we prefer the term lexicon-based class as it shows characteristics of the Thai language more accurately. Moreover, we deleted Barron's last class, the miscellaneous changes (e.g., change of order) as it can be considered part of the structure-based class which contains syntactic and discourse patterns (e.g., reordering words in sentences). We categorize the Thai plagiarism into four classes: *copy-based change*, *lexicon-based change*, *structure-based change*, and *semantic-based change*. For the copy-based change, this class is word-by-word copying that a text passage is copied from a set of source documents without any modifications and pasted into a document. For the lexicon-based change, a document is created from document sources by rewriting. This class contains three plagiarism cases, namely, lexical substitution (or replacing), addition (or inserting), and deletion (or removing). For structure-based change, this class contains two main changes, namely, syntactic and discourse changes, which establish five plagiarism cases, namely, part of speech changes, coordination changes, subordination and nesting changes, change of order, and topic changing. For semantic-based, this class is about copying the idea or concept such as the result of experiment, summary or discussion from other studies in the same discipline.

Most of the previous studies have focused on constructing of plagiarism corpora in any languages. To enhance research of creating Thai plagiarism corpus, we present a design and construction on development of a benchmark plagiarism corpus in Thai.

III. Corpus Construction Process

In this section, we describe a process of designing and developing the plagiarism corpus in Thai. The details of corpus construction process are explained in the following subsections.

A. The process overview of constructing corpus

For the process overview of Thai plagiarism corpus construction, we use the simulated plagiarism method for creating Thai plagiarism corpus. We aim to design and develop the tool that imitates the plagiarism scenario from plagiarists as closely as possible. Our corpus consists of three steps of constructing the corpus. The details of each step are illustrated in Figure 1.

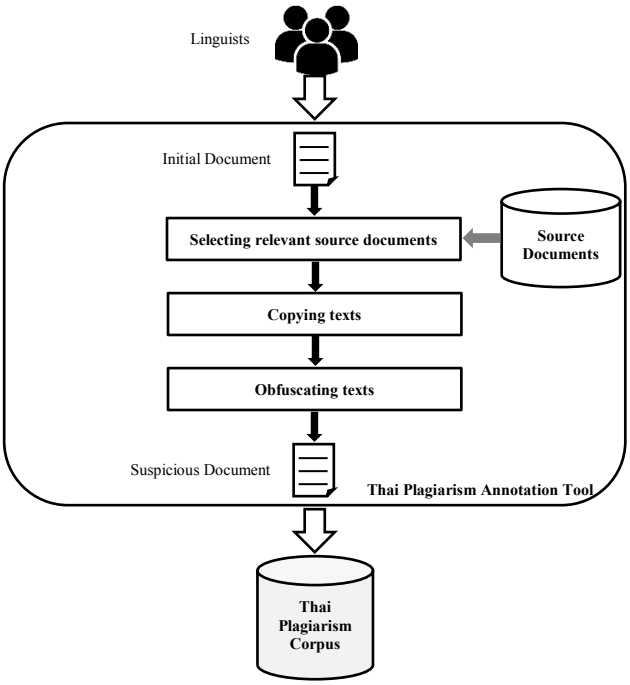

Fig. 1. Overview of Thai plagiarism corpus construction

For the simulated plagiarism method, a suspicious document containing simulated plagiarism cases is manually created by linguists based on four classes of Thai plagiarism and linguistic mechanisms described in section 2. The Thai plagiarism classes are adopted to create plagiarized text

passages and inserted them into an initial document. To help linguists easily make the plagiarized text passages including Thai plagiarism cases, we provide a Thai plagiarism annotation tool called *PlaTool* and a Thai plagiarism guideline for assisting human annotators to plagiarize the text passages from Thai Wikipedia articles and web page sources. *PlaTool* is a web-based application which consists of DBMS and GUI. In short, constructing the Thai plagiarism corpus consists of three steps: (1) selecting relevant source documents (2) copying texts, and (3) obfuscating texts. First, a linguist selects a relevant source document from Thai Wikipedia articles and web page sources. Second, a selected text passage is copied and obfuscated with only one in the Thai plagiarism cases or mix many cases together. After the step of obfuscating a text passage is finished, the obfuscated text passage is pasted into the initial document and saved as the suspicious document. The suspicious document is then imported into the Thai plagiarism corpus.

B. *The construction of simulated Thai plagiarism document*

The preliminary creation of simulated Thai plagiarism documents consists of the initial and source documents. The initial document is prepared for being plagiarized documents. The source documents are a collection of source documents that is prepared for searching articles corresponding to given keyword title in the initial document. To create the initial and source documents, we collected 111,000 articles from Thai Wikipedia and web pages. These articles are divided into two groups, namely, 110,000 articles for source documents and 1,051 articles for the initial documents selected by linguists from several domains such as science, sports and travel. These articles are then imported to *Platool* for the linguists to manually create the plagiarized text passages easily as shown in Figure 2.

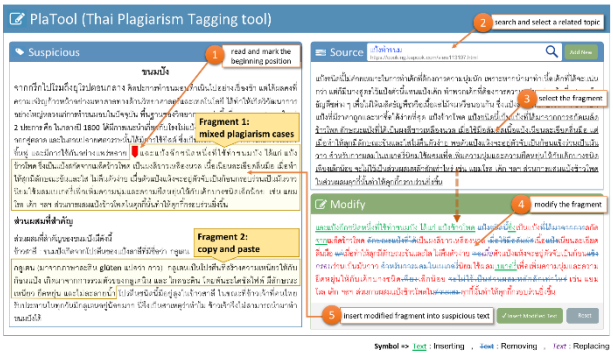

Fig. 2. PlaTool (Thai plagiarism annotation tool)

From Figure 2, an initial document is first displayed in the left column. The process of annotation consists of many steps. A human annotator starts reading the article and marking the position of text where he or she needs to put the plagiarized text passage in the initial document (see number 1 in the Figure 2). Then, the annotator searches and selects a related article and then the annotator copies some parts of text in the related article approximately 150 - 250 characters, called text passage, from the source documents (see number 2 and 3 in Figure 2). Next, the annotator intentionally and freely obfuscates the text passage by using only one in Thai plagiarism cases or mix them together (see number 4 in Figure 2). After completing the obfuscations of the text passages, the annotator inserts the obfuscated text passage into the initial document (see number 5 in the Figure 2). After plagiarizing has finished, the initial document becomes suspicious document and then the suspicious document is imported to the database. The suspicious documents generated by the simulated method are written in the standard XML format following PAN-PC-10 corpus schema [4]. The collection of all suspicious documents in the database is the Thai plagiarism corpus.

IV. The Details of Thai Plagiarism Corpus

In this section, we show the details of Thai plagiarism corpus. This section is divided into three subsections including suspicious document format, general statistics of corpus, and examples of Thai plagiarism cases. Each of the subsections is explained below.

A. *Suspicious Document Format*

A suspicious document is written in an XML format which contains general information about the cases of plagiarism. The suspicious document is illustrated in Figure 2.

```
<document language="th" reference="suspicious-document-129.txt">
<plagiarized source_reference="10257.txt" source_offset="502"
source_length="247" this_offset="2351" this_length="217" />
</document>
```

Fig. 3. Example of anotation schema for a suspicious document

- *this_length*: The number of characters in a plagiarized text of the suspicious document
- *this_offset*: Starting offset, the position of the first character corresponding to a plagiarized text in the suspicious document
- *source_reference*: Name of source document in the corpus
- *source_length*: The number of characters in a source fragment of the source document
- *source_offset*: Starting offset, the position of the first character corresponding to a source fragment in the source document

B. *General Statistics of Corpus*

Our corpus consists of a set of source documents and a set of suspicious documents. The number of source documents is 110,000 articles, 109,200 Thai Wikipedia articles and 800 web pages articles. The number of suspicious documents contains 1,051 articles, in which 8,979 plagiarized cases are inserted. The results of constructing the Thai plagiarism corpus are shown in Table 1 and 2.

TABLE I. Corpus Statistics

Corpus Statistics	
Number of source documents	110,000
Number of suspicious documents	1,051

TABLE II. Plagiarism Cases Statistics

Plagiarism Cases Statistics	
Number of Plagiarized Fragments	
Small (1-3 fragments)	485
Medium (4-6 fragments)	323
High (7-10 fragments)	243
Plagiarized Fragment Length	

Short (15-50 words)	350
Medium (50-100 words)	351
Long (100-150 words)	350
Thai Plagiarism Classes	
Copy-Based Change	
Copying	2,137
Lexicon-Based Change	
Inserting	2,042
Removing	2,123
Replacing	1,961
Structure-Based Change	
Part of speech	117
Coordination	91
Subordination and nesting	104
Change of order	93
Topic changing	102
Semantic-Based Change	
Paraphrasing	209

C. *Examples of Thai Plagiarism Cases*

In this subsection, we show the examples of creating the Thai plagiarism cases from our annotation tool in Table 3. In addition, we skip the example of copy-based change because this class is straightforwardly copying an original text without any modifications.

TABLE III. THE EXAMPLES OF THAI PLAGIARISM CASES

Thai Plagiarism Classes	Examples
Lexicon-Based Change	
Inserting	**[In Thai]** ***Original document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรมประเภทหนึ่ง ***Modified document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรม (หน่วยความจำ) ประเภทหนึ่ง **[Translation]** ***Original document*** A computer consists of processing elements, typically a central processing unit, and some form of memory. ***Modified document*** A computer consists of processing elements, typically a central processing unit, and some form of memory (RAM).
Removing	**[In Thai]** ***Original document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรมประเภทหนึ่ง ***Modified document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรม~~ประเภทหนึ่ง~~ **[Translation]** ***Original document*** A computer consists of processing elements, typically a central processing unit, and ~~some form of~~ memory. ***Modified document***
Replacing	**[In Thai]** ***Original document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรมประเภทหนึ่ง ***Modified document*** เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยปกติเป็นหน่วยประมวลผลกลาง และแรมชนิดหนึ่ง **[Translation]** ***Original document*** A computer consists of processing elements, typically a central processing unit (CPU), and some form of memory. ***Modified document*** A computer consists of processing elements, normally a central processing unit (CPU), and some type of memory.
Structure-Based Change	
Part of speech	**[In Thai]** ***Original document*** ข้อนี้ถือว่าช่วยให้พนักงานเก็บขยะลำเลียงสิ่งปฏิกูลเหล่านี้ไปยังสถานที่กำจัดได้อย่างสะดวกและรวดเร็วขึ้น ***Modified document*** ข้อนี้ถือว่าช่วยให้พนักงานเก็บขยะลำเลียงสิ่งปฏิกูลเหล่านี้ไปยังสถานที่กำจัดได้สะดวกและรวดเร็วขึ้น **[Translation]** ***Original document*** This subject helps a garbage man to quickly transfer these sewages to refuse disposal place. ***Modified document*** This subject helps a garbage man to quick transfer these sewages to refuse disposal place.
Coordination	**[In Thai]** ***Original document 1*** ขยะมูลฝอยทำให้ถนนหนทางไม่สะอาด ***Original document 2*** ขยะส่งผลเสียต่อสุขภาพของผู้คน ***Modified document*** ขยะมูลฝอยทำให้ถนนหนทางไม่สะอาดและขยะส่งผลเสียต่อสุขภาพของผู้คน **[Translation]** ***Original document 1*** The garbage made the road dirty. ***Original document 2*** The garbage affected people's health. ***Modified document*** The garbage made the road dirty and the garbage affected people's health.
Subordination and nesting	**[In Thai]** ***Original document*** ขยะมูลฝอยทำให้ถนนหนทางไม่สะอาด ***Modified document*** ขยะมูลฝอยที่เพิ่มปริมาณมากขึ้นทุกวันทำให้ถนนหนทางที่เราใช้สัญจรไปมาไม่สะอาด **[Translation]** ***Original document*** The garbage made the road dirty. ***Modified document*** The garbage that was increased everyday made the road that we are roaming dirty.
Change of order	**[In Thai]** ***Original document*** ผลไม้มีความสำคัญต่อมนุษย์ในแง่ของการบริโภคเป็นอาหาร

Thai Plagiarism Classes	Examples
	Modified document นนุษย์ให้ความสำคัญกับผลไม้ในแง่ของการบริโภคเป็นอาหาร **[Translation]** ***Original document*** A fruit is important to a man in terms of food assumption. ***Modified document*** A man is important to a man in terms of food assumption.
Topic changing	**[In Thai]** ***Original document*** การศึกษาการเปลี่ยนแปลงสภาวะอากาศบนดาวเคราะห์ดวงอื่นนั้น$_1$ มีประโยชน์ในการพยากรณ์สภาวะเปลี่ยนแปลงสภาพลมฟ้าอากาศบนโลก$_2$ ***Modified document*** การพยากรณ์สภาวะเปลี่ยนแปลงลมฟ้าอากาศบนโลก$_2$ เรารู้ได้โดยการศึกษาการเปลี่ยนแปลงสภาวะอากาศบนดาวเคราะห์ดวงอื่น$_1$ **[Translation]** ***Original document*** The study on changing the weather on other planets.$_1$ is useful for forecasting the weather on the earth.$_2$ ***Modified document*** To forecast the weather on the earth.$_2$, we can learn it by studying on changing the weather on other planets.$_1$
Semantic-Based Change	
Paraphrasing	**[In Thai]** ***Original document*** นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ***Modified document*** นอกจากผลการวิเคราะห์ปัจจัยพื้นฐาน$_{synonym}$ของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ซึ่ง$_{synonym}$ ได้มา$_{insert}$จากการศึกษาภาคสนามดังกล่าว$_{delete}$ข้างต้น **[Translation]** ***Original document*** In addition to the componential analysis results of social capital in the case of a sustainable lifelong learning society, which are obtained from field studies mentioned above. ***Modified document*** In addition to the results from an analysis of the basic factors of social capital in the case of a sustainable lifelong learning society, which are derived from field studies.

From Table 3, we can conclude that the lexicon-based change contains three plagiarism cases, namely, lexical substitution (or replacing), addition (or inserting), and deletion (or removing). The structure-based change comprises two main changes, namely, syntactic and discourse changes, which establish five plagiarism cases. The semantic-based change is about copying the idea or concept such as the result of experiment, summary or discussion from other studies in the same discipline.

V. Conclustion and Future Work

We have presented a design and development of a Thai plagiarism corpus that assists for the development and evaluation of Thai plagiarism detection algorithms. The main goal of this paper is to simulate the plagiarized text passages representing the obfuscation strategies. We provided a Thai plagiarism annotation tool called *PlaTool* and a Thai plagiarism guideline for assisting human annotators to plagiarize the text passages. We aim to design and develop the tool that imitates the plagiarism scenario from plagiarists as closely as possible. Our tool is designed to support plagiarizing text passages in four classes of Thai plagiarism. Our corpus is based on four classes of Thai plagiarism and linguistic mechanisms including copy-based change, lexicon-based change, structure-based change, and semantic based change. Our corpus is manually created by linguists containing simulated cases of plagiarized documents that resemble human language. Our corpus consists of a set of source documents and a set of suspicious documents based on Thai Wikipedia articles and web page articles. The number of source documents is 110,000 articles and 1,000 suspicious documents in which 8,979 plagiarism cases. In the future, we plan to use the artificial plagiarism method, automatically generates artificial plagiarism cases, to rapidly increase the corpus size. We also plan to extend our corpus in other languages, and evaluate the performance of plagiarism detection algorithms in Thai. Finally, we will encourage the Thai plagiarism corpus to be available as the standard corpus for research and development of Thai plagiarism detection algorithms.

Acknowledgment

This work is a part of Thai plagiarism detection task on the Twentieth National Software Contest (NSC) 2018. It is partially supported by National Electronics and Computer Technology Center (NECTEC) and NSC. In addition, we would like to thank the linguists for annotating our corpus.

References

[1] P. Clough, and M. Stevenson, “Developing a corpus of plagiarised short answers,” Language Resources and Evaluation, pp. 5–24, 2011.

[2] S. Taerungruang, and W. Aroonmanakun, “Constructing an academic Thai plagiarism corpus for benchmarking plagiarism detection systems,” Journal of Language Studies, vol. 18, no 3, 2018.

[3] A. Barrón-Cedeño1, M. Potthast, P. Rosso, B. Stein, and A. Eiselt, “Corpus and evaluation measures for automatic plagiarism detection,” Proceedings of the Seventh conference on International Language Resources and Evaluation, May 2010.

[4] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, “An evaluation framework for plagiarism detection,” In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10), pp. 997–1005, 2010.

[5] M. Potthast, B. Stein, A. Eiselt, A. Barrón-cedeño, and P. Rosso, “Overview of the 1st international competition on plagiarism detection,” In Proceedings of SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), 2009.

[6] S. Mohtaj, H. Asghari, and V. Zarrabi, “Developing monolingual English corpus for plagiarism detection using human annotated paraphrase corpus,” CLEF 2015, 2015.

[7] M. Sharjeel, P. Rayson, R. Muhammad, and A. Nawab, “UPPC-Urdu paraphrase plagiarism corpus,” Language Resources and Evaluation Conference, 2016.

[8] M. Potthast, A. Barron, A. Eiselt, B. Stein, and P. Rosso, “Overview of the 2nd international competition on plagiarism detection,” CLEF 2010, 2010.

[9] A. Barrón-Cedeño, M. Vila, M. A. Marti. and P. Rosso, “Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection,” Association for Computational Linguistics, 2013.

[10] P. Clough, R. Gaizauskas, S. S. Piao, and Y. Wilks, “METER: MEasuring TExt Reuse,” ACL, pp. 152–159, 2002.