This is a repository copy of An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE.

White Rose Research Online URL for this paper: http://eprints.whiterose.ac.uk/94770/

Version: Accepted Version

Article:

https://doi.org/10.1109/TCBB.2016.2542803

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE

Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough

Abstract—The identification of duplicated and plagiarised passages of text has become an increasingly active area of research. In this paper we investigate methods for plagiarism detection that aim to identify potential sources of plagiarism from MEDLINE, particularly when the original text has been modified through the replacement of words or phrases. A scalable approach based on Information Retrieval is used to perform candidate document selection - the identification of a subset of potential source documents given a suspicious text - from MEDLINE. Query expansion is performed using the UMLS Metathesaurus to deal with situations in which original documents are obfuscated. Various approaches to Word Sense Disambiguation are investigated to deal with cases where there are multiple Concept Unique Identifiers (CUIs) for a given term. Results using the proposed IR-based approach outperform a state-of-the-art baseline based on Kullback-Leibler Distance.

Index Terms—Natural Language Processing, Information Retrieval, Extrinsic Plagiarism Detection, MEDLINE, UMLS Metathesaurus, Query Expansion

I. INTRODUCTION

PLAGIARISM generally refers to the unacknowledged copying of existing information, such as documents and programs [1], [2]. This can include the reuse of one’s own material (known as self-plagiarism [3]), as well as that produced by others. In higher education, plagiarism is acknowledged as a significant problem and has been reported to be on the increase [4], [5], [6]. Sheard et. al. [7] reported a summary of three different surveys in which 88%, 90% and 91.7% of the students admitted that they were involved in cheating or academic dishonesty at least once during their study. Plagiarism is not restricted to students, but has also surfaced amongst academics [8]. For example, Citron & Ginsberg [9] analyze text reuse within the ArXiv.org scientific corpus and Errami et al. [10] identify duplication in PubMed abstracts. Consequently, plagiarism and its detection has recently received significant attention [11], [12] and automated systems are now routinely used by higher education institutions and publishers to identify potential cases of plagiarism.

Determining whether plagiarism has occurred is ultimately a human action; however, automated tools can assist with the process [13]. Various factors can signal plagiarism, such as inconsistencies in writing style, unexpected use of advanced vocabulary, incorrect references and shared similarities with existing materials. Broadly speaking, approaches to detecting plagiarism (whether manual or automatic) can be categorised into two main problems. Intrinsic plagiarism detection relates to identifying stylistic inconsistencies within a text that give rise to questions regarding its authorship; extrinsic plagiarism detection relates to identifying the possible sources of a suspicious document [14].

MEDLINE (Medical Literature Analysis and Retrieval System Online) contains a large number of publications in the area of medicine and related fields. New publications are being added at such a rate that it becomes difficult for individuals or groups to keep abreast of the information contained within it. As a result, it is possible that people may reproduce the same research carried out by others without the connection being noticed and resulting in duplication (and potential plagiarism). Errami et. al. [10] examined a set of over 62,000 citations in MEDLINE to identify highly similar citation pairs. They found that 1.39% of the citations were highly similar. A number of these (1.35%) had shared authors and were similar enough to be considered as duplicate publications. The remaining (0.04%) had no shared author and could be considered as potential cases of plagiarism. Although the highly similar documents identified in this study are a small portion of the documents examined, given the size of MEDLINE it would suggest that as many as 117,500 citations are duplicate publications and 3,500 citations are potentially plagiarised. (Note that these figures were reported in 2007 and are likely to be higher now.)

The process of plagiarism detection from large document collections, such as MEDLINE, is commonly treated as a two-stage process [14]. The first stage, called candidate document selection, involves identifying a set of candidate sources from a document collection for a given suspicious document. This is followed by the second stage, referred to as detailed analysis, which makes an exhaustive comparison of the suspicious document with all candidates to identify (and align) similar sections. The focus of this paper is the first stage of the extrinsic plagiarism detection process - candidate document selection - that can improve the overall speed and accuracy of extrinsic plagiarism detection systems [15]. The set of “candidate documents” should be carefully chosen from the
document collection because any source document missed at this stage will not be identified in the detailed comparison stage.

This paper uses an Information Retrieval (IR)-based approach to retrieve candidate documents that is scalable to large document collections, such as MEDLINE. However, if an exact match approach is used in the retrieval process then such an approach may fail to identify similarity between document pairs when the original text has been rewritten. Therefore, we investigate the use of query expansion techniques to deal with situations in which the source text has been rewritten as often plagiarists attempt to disguise their behavior by altering the text in some way (obfuscation), for example by paraphrasing or summarising text [16], [17], [18]. Barrón-Cedeño et. al. [19] investigated different strategies for mono-lingual paraphrasing to identify the paraphrases most difficult to detect. They used simulated (manually paraphrased) cases of plagiarism in the PAN-PC-10 Corpus [20]. Their analysis showed that lexical substitution is the most common editing operation used in paraphrasing for plagiarism and plagiarised text is often a summarised version of the original text. Therefore, to capture the most common paraphrasing phenomenon for plagiarism, the content words of the document which is suspected of containing plagiarised text are expanded with synonymous words from the UMLS Metathesaurus. We find that the proposed IR-based approach is suitable for the candidate document selection problem and outperforms a previously reported approach. The use of query expansion further improves retrieval performance, particularly when the source text has been obfuscated.

The rest of this paper is organised as follows: Section II discusses different existing techniques for plagiarism detection and query expansion; Section III presents our proposed approach; Section IV describes the experimental setup (implementation details, dataset and evaluation measures); Section V present the results and analysis of the experiments; Finally Section VI concludes the paper and discusses avenues for future work.

II. BACKGROUND

A. Plagiarism Detection

Alzahrani et. al. [21] summarise the range of approaches commonly used to detect plagiarism, ranging from simpler lexical methods to more complex semantic-based methods. Potthast et. al. [22] also describe different types of approaches for producing exact and modified copies. The detection of plagiarism in cases involving little or no modification of the original sources has been shown to be straightforward [23], [24], [25]. However, situations in which the source text has been rewritten, for example paraphrased, is far more challenging to detect [26], [27], [28].

For example, Maurer et. al. [26] paraphrased a passage with an Anti-Anti Plagiarism System² - a simple automatic tool for word replacement. The paraphrased passage was analysed by two well-known commercial plagiarism detection services and both failed to detect plagiarism. In addition, the best system [29] in the 2nd International Competition on Plagiarism Detection [27] achieved a recall of more than 0.99 and precision of 0.95 when detecting verbatim (exact copy) plagiarism. However, none of the systems which took part in the competition achieved a recall of more than 0.28 for manually paraphrased (simulated) cases of plagiarism (the precision score varied). Uzuner et. al. [30] extracted syntactic features using a context free grammar to identify modified text. Results showed an improvement in performance using these features. Recently, Chong et. al. [31] applied various pre-processing and NLP techniques (e.g., tokenization, sentence segmentation, Part Of Speech (POS) tagging, chunking and dependency parsing) to normalise documents and found that it improves the performance of existing plagiarism detection approaches. Mozgovoy et. al. [32] also showed that applying parsing to normalize the effect of word reordering improves performance for plagiarism detection. However, these approaches fail to identify semantic similarity between a pair of documents.

B. Candidate Document Selection

A number of approaches have been proposed for the candidate document selection problem. One approach retrieves candidate documents using the Kullback-Leibler Symmetric Distance method. KLd (see Equation 1) [15]. Documents are modelled as probability distributions and compared using KLd. Documents are converted into probability distributions by removing stop words, stemming [33] and then computing tf.idf weights for the remaining word unigrams. Assume Pd is the probability distribution generated from d, a document in the reference collection, and that Qs is the equivalent distribution for s, a suspicious document. The Kullback-Leibler Symmetric Distance between them (over a feature vector X) is computed as follows:

\[
KL_d(P_d||Q_s) = \sum_{x \in X} (P_d(x) - Q_s(x)) \log \frac{P_d(x)}{Q_s(x)}
\]  

(1)

Results showed that the overall accuracy and speed of the plagiarism detection system improved by applying the Kullback-Leibler Symmetric Distance to reduce the plagiarism detection search space. The system’s performance without search space reduction was precision 0.73 and recall 0.63. When the search space reduction step was applied performance improved to a precision 0.75 and recall 0.74. The execution time also reduced substantially from 2.32 seconds to 0.19 seconds.

A further common approach to the problem of candidate document selection involves the use of techniques from IR. For example, in many of the International Competitions on Plagiarism Detection [34], [27], [28], [35], [36], [37] IR-based approaches were used by the majority of the participating groups for the candidate document retrieval task. Using this method, documents in the reference collection are converted to fixed length word n-grams and indexed. N-gram representations of the suspicious document are also created in the same way and used to query the index. If the number of matching fingerprints between suspicious-source document pair is

above some pre-defined threshold then the source document is marked as potential candidate document. However, these approaches only aim to detect candidate documents that have been copied verbatim with minor changes.

C. Identifying Duplicates in MEDLINE

Lewis et. al. [38] proposed a vector-based text similarity search algorithm (called eTBLAST) to identify highly similar citation pairs (potential cases of plagiarism) in MEDLINE. A query is formed from the title and abstract of a MEDLINE citation (stop words are removed and remaining keywords are weighted using a term weighting scheme). eTBLAST computes the similarity score between title and abstract query and MEDLINE citations and returns a list of highly similar citations ranked by their similarity scores. The top 400 citations returned by eTBLAST are re-ranked using a sentence-alignment algorithm to generate a final ranked list of highly similar citations. Errami et. al. [39] reported an improvement in performance over eTBLAST on the same MEDLINE dataset. Their proposed approach computes the number of common “Statistically Improbable Phrases” (SIP), essentially word 6-grams, between a pair of MEDLINE documents. SIPs were weighted using language modeling probability scores, which were computed using the entire MEDLINE database.

A limitation of both eTBLAST and using SIPs is that they are unable to identify similar MEDLINE citations when the original text has been substantially altered, such as by paraphrasing or replacing words with synonyms [10], [39]. The authors suggest the use of such approaches which can identify ‘smart duplication’ [10] as well as to “analyse grammar and extract meaning from sentences rather than rely on word comparisons only” [39].

D. Query Expansion

Query expansion, the process of adding search terms to a query, has been previously used in IR to deal with problems of *vocabulary mismatch* [2], [2]. Applying query expansion will typically improve retrieval performance, particularly recall [40], [41]. For instance, the query ‘car’ could be expanded to ‘car cars automobile vehicle’. The process of query expansion can be applied to an initial query, reformulated query or both. Moreover, the addition of expansion terms to original query terms can be combined with term re-weighting. For example, expansion terms can be assigned less weight than original ones.

For plagiarism detection, methods based on query expansion have also been proposed to identify plagiarism when the original text has been heavily paraphrased. For example, Nawab et.al. [42] applied various query expansion approaches (pseudo relevance feedback, query expansion using WordNet and a paraphrase lexicon) to retrieve candidate documents when the source text has been heavily paraphrased. Results showed that query expansion based on WordNet and the paraphrase lexicon improves candidate document retrieval performance. Nawab et. al. [43] demonstrated an improvement in performance when word n-grams were expanded with synonymous words from knowledge-bases. Chen et. al. [44] used WordNet synsets and relationships (hypernyms/hyponyms) between synsets to identify semantic similarity between a plagiarized and source document. Ceska [45] used WordNet’s first sense, all senses and sense selection after word sense disambiguation to detect synonym replacement in a suspicious document. However, results with WordNet did not show any significant improvement as compared to a baseline approach.

Our proposed framework for candidate document selection (see Section III-A) uses an IR-based approach and incorporates query expansion to identify obfuscated documents. Previous studies have attempted to take into account the modifications in the documents for identifying text reuse and plagiarism [44], [46], [43]. However, these have not been applied to MEDLINE citations. The approach most similar to the one presented here [42] was used to retrieve candidate plagiarised documents in free text. As far as we are aware, the proposed IR-based approach using query expansion based on UMLS Metathesaurus has not been previously used for retrieving candidate documents from MEDLINE.

III. PROPOSED APPROACH

This section presents the IR-based approach to the identification of candidate source documents (Section III-A) followed by a description of how it can be extended by query expansion using resources from the medical domain (Section III-B).

A. IR-Based Approach

Figure 1 shows the process of retrieving candidate source documents using the proposed IR-based approach. The source collection is indexed with an IR system (an off-line process). In the IR-based framework, the candidate retrieval process can be divided into four main steps: (1) pre-processing, (2) query formulation, (3) retrieval and (4) results merging. These steps are described as follows:

1) **Pre-processing**: Each suspicious document is split into sentences using NLTK [47]. The terms in each sentence are converted to lower case. Stopwords\(^3\) and punctuation marks are removed. Stemming (using the Porter Stemmer [33]) is applied to the remaining terms prior to indexing.

2) **Query Formulation**: Sentences from the suspicious document are used to form multiple queries. The length of a query can vary from a single sentence to all sentences appearing in a document as reused text can be sourced from one or more documents and vary from a single sentence to an entire document. A long query is likely to perform well in situations where large portions of text are reused for plagiarism; on the other hand small portions of plagiarised text are likely to be effectively detected by a short query. Therefore, the choice of query length is important in obtaining effective results.

3) **Retrieval**: Terms are weighted using the *tf.idf* weighting scheme and then text forming the query is used to retrieve similar documents (and potentially the source documents of the suspicious text) from the index.

\(^3\)A list of 127 English stop words from NLTK [47] was used.
4) Results Merging: The top \(N\) documents returned against multiple queries are merged to generate a final ranked list of source documents. A standard data fusion approach, CombSUM [48], is used to generate the final ranked list of documents by combining the similarity scores of source documents retrieved against multiple queries. In CombSUM the final similarity score, \(S_{final}\), is obtained by adding the similarity scores of source documents obtained against each query \(q\):

\[
S_{final} = \sum_{q=1}^{N_q} S_q(d) \tag{2}
\]

where \(N_q\) is the total number of queries to be combined and \(S_q(d)\) is the similarity score of a source document \(d\) for a query \(q\). The top \(K\) documents in the ranked list generated by the CombSUM method are marked as potential candidate source documents.

B. Query Expansion

The Unified Medical Language System\(^4\) (UMLS), a set of tools and resources to assist with the development of biomedical text processing systems, is used to carry out query expansion. Our approach uses two main UMLS resources (the Metathesaurus and MetaMap) which are now described, followed by an explanation of how they are used for query expansion.

1) UMLS Metathesaurus: The UMLS Metathesaurus is a large database of more than 100 multi-lingual controlled source vocabularies and classifications, which contains information about concepts (related to biomedical and health), concept names and relationships between concepts. The basic units of the Metathesaurus are concepts, whereby the same concept can be referred to using different terms. One of the main goals of Metathesaurus is to group all the equivalent terms (synonyms) from different source vocabularies into a single concept. Thus, a concept is a collection of synonymous terms. Each concept in Metathesaurus is assigned a unique identifier called a CUI (Concept Unique Identifier).

<table>
<thead>
<tr>
<th>Table I: Example showing some of the MRCONSO Table entries in English for the phrase “Gamma-glutamyl transpeptidase”, whose CUI is C0202035. “ENG” means that entry is in English language. Additional information from the table is omitted for brevity.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input Text</strong></td>
</tr>
<tr>
<td>Gamma-glutamyl transpeptidase</td>
</tr>
<tr>
<td><strong>MRCONSO Table Entries in English for the CUI C0202035</strong></td>
</tr>
<tr>
<td>C0202035 ENG Gamma glutamyl transferase measurement</td>
</tr>
<tr>
<td>C0202035 ENG Gamma glutamyl transpeptidase measurement</td>
</tr>
<tr>
<td>C0202035 ENG GTP measurement</td>
</tr>
</tbody>
</table>

The information about concept names, key features associated to each concept name (e.g., language, name type and source vocabulary) and concept identifiers is stored in the MRCONSO table. The entire concept structure of Metathesaurus, therefore, is stored in this file that contains information in multiple languages and each entry is either marked as suppressed or preferred. Table I shows three entries in English in the MRCONSO table for the term “Gamma-glutamyl transpeptidase” with CUI C0202035. In this example, “Gamma glutamyl transferase measurement”, “Gamma glutamyl transpeptidase measurement” and “GTP measurement” terms can be used as synonyms for the original term “Gamma-glutamyl transpeptidase”.

2) MetaMap: MetaMap\(^5\) is a key supporting tool for the UMLS [49]. The objective of this program is to efficiently link terms mentioned in input text to concepts in UMLS Metathesaurus. MetaMap performs syntactic/lexical analysis of the input text to map Metathesaurus concepts to input terms. During the mapping process, it also includes the option of carrying out Word Sense Disambiguation (WSD) to attempt to select between candidates when there are multiple possible CUIs for a term [50]. Table II shows example output generated by MetaMap with and without WSD. It can be noted that there are two Meta Mappings when the WSD option is not used, while there is only one Meta Mapping with WSD. Also, during parsing, MetaMap treats the phrase “Gamma-glutamyl transpeptidase” as a single term instead of treating it as

\(^4\)http://www.ncbi.nlm.nih.gov/books/NBK9676/

\(^5\)MetaMap2010 was used for experiments
two separate terms: “Gamma-glutamyl” and “transpeptidase”. MetaMap treats many multi-word phrases as single terms.

3) Query Expansion using the UMLS Metathesaurus: Input terms are mapped to UMLS CUIs using MetaMap. The UMLS Metathesaurus’s MRCONSO table is then consulted to identify synonymous terms for each CUI and are used for query expansion. Two approaches are used for mapping input terms to UMLS CUIs: (1) CUI mapping with WSD and (2) CUI(s) mapping without WSD. In the former case, synonymous terms for query expansion are selected from only one mapped CUI; whereas in the latter case, additional search terms can be selected from any of the mapped CUIs. Once input terms are mapped to CUIs, synonymous terms in English that are marked as preferred are selected as additional search terms from the MRCONSO table. We were unable to find a suitable resource find a suitable resource to decide with synonymous term(s) should be used to create expanded queries. Therefore, each input term is expanded with a single additional search term which is selected at random.

Table III shows examples of expanded queries created using the UMLS Metathesaurus (where w is the weight assigned to an additional search term/phrase). An additional search term is added to a query term in two ways: (1) treating multi-word input and additional search terms as phrases (see examples of WSD Phrase and Without-WSD Phrase) and (2) treating multi-word input and expansion terms as a sequence of single words (see examples of WSD and Without-WSD).

IV. Experimental Setup

This section describes the dataset used for evaluation (Section IV-A), how the approach was implemented (Section IV-B) and the evaluation measure (Section IV-C) used to evaluate the various query expansion methods.

A. Evaluation Dataset

Evaluation is carried out using an existing source of potentially plagiarised publications from Medline. Errami et. al. [10], [39] used an automatic text similarity tool called eTBlast [38], [51] to identify highly similar citation pairs in MEDLINE. The aim of this study was to identify potential cases of plagiarism in the biomedical domain. A total 79,383 highly similar Medline citation pairs were identified and complied in the Deja vu database. Each duplicate citation pair was classified into four categories: (1) duplicate citation pairs having Shared Author (SA), (2) duplicate citation pairs written by Different Authors (DA) i.e. no-shared authors, (3) duplicate citation pairs published in the Same Journal (SJ) and (4) duplicate citation pairs published in Different Journals (DJ) [10]. Out of 79,383 highly similar citation pairs identified using eTBlast [38], [51], only a subset of 2,106 citation pairs have been manually examined and verified as true duplicate citation pairs. Among manually examined duplicate citation pairs, 256 pairs are written by Different Authors (DA) and 1,841 pairs have Shared Authors (SA). Although highly similar citation pairs are identified at title and abstract level, Errami et. al. [10] suggested that highly similar duplicate citation pairs with no shared author are potential cases of plagiarism.


There are also other categories but these four are more relevant to plagiarism.
Table IV shows an example of a potential plagiarism case in the MEDLINE corpus. It can be noted that there are five exact matches in both texts whose length is greater than five tokens (shown in bold). These long exact matches are unlikely to occur by chance. In addition, there are also other, shorter exact matches.

For these experiments, the source collection is from 19,569,568 citations from the 2011 MEDLINE/PubMed Baseline Repository. The collection of suspicious documents contains 260 citations from the Deja vu database that have been manually examined and verified as duplicates. These citation pairs are selected because they do not have a common author, making them potential cases of plagiarism [10].

B. Implementation

Lucene\(^8\), a popular and freely available IR system, is used for the experiment. The source collection is indexed. Documents are pre-processed by converting the text into lower case and removing all non-alphanumeric characters. Stopwords\(^9\) are removed and stemming is carried out using the Porter Stemmer and removing all non-alphanumeric characters. Stopwords are pre-processed by converting the text into lower case for the experiment. The source collection is indexed. Document vectors are obtained using the cosine similarity measure:

\[
sim(q, d) = \frac{q \cdot d}{|q| \times |d|} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n} (q_i)^2} \times \sqrt{\sum_{i=1}^{n} (d_i)^2}}
\]

where \(|q|\) and \(|d|\) represent the lengths of the query and document vectors, respectively.

Our approach requires three parameters to be set: the number of sentences used to formulate a query \((Q)\), the number of source documents retrieved against each query \((N)\) (see Section III-A) and the weights assigned to the term added by the query expansion approach \((W)\) (see Section III-B).

Optimal values for these parameters were set automatically using three fold cross validation. The suspicious collection of the MEDLINE Corpus was split into three folds with two being used to identify the optimal values for the parameters and the remaining third for evaluation. The results of the three runs are then averaged.

C. Evaluation Measure

The goal of the candidate document retrieval task is to identify all the source document(s) for each suspicious document while returning as few non-source documents as possible. It is important for all source documents to be included in the top ranked documents returned by the system since otherwise they will not be identified during later stages of processing. Consequently, recall is more important than precision for this problem.

Recall for the top \(K\) document, averaged across queries is used as the evaluation measure for these experiments. For a single query the Recall at \(K\) \((R@K)\) is 1 if the source document appears in the top \(K\) documents retrieved by the query, and 0 otherwise. For a set of \(N\) queries, the averaged recall at \(K\) score is calculated as:

\[
R@K_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} R@K_i
\]

where \(R@K_i\) is the recall at \(K\) score for query \(i\).

Figure 2 shows an example of calculating averaged recall score for the candidate document selection \((K = 5)\). Sets of relevant and retrieved documents are represented by Annotations and Detections respectively (source documents which are identified are in bold). It can be noted from this example that the rank of a source document in the top \(K\) documents is unimportant. As long as a source document appears in the top \(K\) documents the averaged recall score will be 1, regardless of whether it appears in the first or \(K_{th}\) rank.

V. RESULTS AND ANALYSIS

Table V shows the results of the experiments for the top 1, 5, 10, 15 and 20 candidate source documents. As expected,
retrieval performance increases as the number of retrieved documents increases. Overall it can be noted that our proposed IR-based approach for retrieving candidate documents performs well in identifying real cases of plagiarism. Performance further improves when query expansion is applied.

<table>
<thead>
<tr>
<th>Approach</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler</td>
<td>0.7956</td>
<td>0.8154</td>
<td>0.8442</td>
<td>0.8588</td>
<td>0.8396</td>
</tr>
<tr>
<td>No Query Expansion</td>
<td>0.8769</td>
<td>0.9173</td>
<td>0.9250</td>
<td>0.9288</td>
<td>0.9288</td>
</tr>
<tr>
<td>WSD</td>
<td>0.9077</td>
<td>0.9519</td>
<td>0.9558</td>
<td>0.9558</td>
<td>0.9596</td>
</tr>
<tr>
<td>Without-WSD</td>
<td>0.9035</td>
<td>0.9519</td>
<td>0.9519</td>
<td>0.9558</td>
<td>0.9558</td>
</tr>
<tr>
<td>WSD Phrase</td>
<td>0.9219</td>
<td>0.9595</td>
<td>0.9595</td>
<td>0.9652</td>
<td>0.9652</td>
</tr>
<tr>
<td>Without-WSD Phrase</td>
<td>0.9115</td>
<td>0.9558</td>
<td>0.9596</td>
<td>0.9634</td>
<td>0.9673</td>
</tr>
</tbody>
</table>

Performance is compared against the the Kullback-Leibler Distance method (see Section II). This approach is based in pairwise comparison of documents which would be computationally expensive for the source collection of over 19 million citations used by the IR-based approach. Consequently a randomly selected subset of 3 million citations, which include the sources for the 260 plagiarised citations, is used as source collection for experiments with the Kullback-Leibler Distance approach. Note that an implication of this decision is that the Kullback-Leibler Distance approach has the advantage of a significantly smaller search space from which to identify source documents.

The IR-based approach proposed here achieves higher results than the Kullback-Leibler Distance approach. Highest recall achieved by this method is 0.8596 for top 20 candidate documents, although it is expected that performance will drop when the entire MEDLINE database is used. The proposed approach (without query expansion) achieves a recall of 0.8769 for \(K = 1\), which is still higher than the maximum recall obtained using the Kullback-Leibler Distance method. This high recall score indicates the strength of the proposed method in detecting potential real cases of plagiarism from large reference collections. As expected, retrieval performance improves when query expansion is applied. Improvement in performance is statistically significant for all query expansion approaches (Wilcoxon signed-rank test, \(p < 0.05\)) [52].

The best results are obtained when input and additional search terms are used as phrases in the query expansion process. A possible reason is that there are many multi-word phrases in biomedical text which are treated as a single term. When similarity is computed between a query term and a source document higher similarity scores are obtained for matching phrases and therefore sources of plagiarised documents are detected. Regarding, WSD and Without-WSD, there is little difference in performance. This is likely because additional search terms are randomly selected and an appropriate resource is not used for the selection of additional search terms (see Section III-B).

Regarding optimal parameter values (see Section IV-B), the best results are obtained using a single sentence as a query \(Q\). The optimal value for the number of source documents retrieved against each query \(N\) is 10. The optimal value for the weight assigned to an expansion term \((\lambda)\) is 0.1.

### A. Query-by-Query Analysis

We carried out an analysis to determine the percentage of queries for which the ranking is “higher”, “lower” or remains the “same” when query expansion is applied (see Table VI). The rank of a query (suspicious document) was considered in the top 20 documents.

**TABLE VI**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>No. of Queries (%) effecting Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Higher</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>WSD</td>
</tr>
<tr>
<td></td>
<td>Without-WSD</td>
</tr>
<tr>
<td></td>
<td>WSD Phrase</td>
</tr>
<tr>
<td></td>
<td>Without-WSD Phrase</td>
</tr>
</tbody>
</table>

For query expansion approaches in the MEDLINE Corpus most of the queries are at the “same” rank and there is little difference in number of queries for “lower” and “higher” ranks. A possible reason for this is that there is little performance difference between various query expansion methods (see Table V).

### VI. Conclusion

This paper describes and evaluates a new query expansion approach to the problem of candidate document selection for extrinsic plagiarism detection. In particular we have focused on cases when the plagiarised version has been highly obfuscated as this presents the greatest challenge to automated plagiarism detection systems. Evaluation was carried out using the MEDLINE Corpus, which contains potential real cases of plagiarism. Results show that the IR-based approach using query expansion outperforms a state-of-the-art approach, Kullback-Leibler Symmetric Distance, for candidate document retrieval task. Query expansion using UMLS Metathesaurus was applied to deal with paraphrased cases of plagiarism. In future work, we would like to further explore different methods for rank fusion and dealing with causes of obfuscation beyond term substitution, such as syntactic changes.

**REFERENCES**


Rao Muhammad Adeel Nawab Dr. Rao is serving as Assistant Professor in the Computer Science Department of COMSATS Institute of Information Technology, Lahore, Pakistan. He holds a PhD in Computer Science from University of Sheffield, UK. His research interests include text reuse and plagiarism detection, author profiling, Information Retrieval (IR) and Natural Language Processing (NLP).

Mark Stevenson Dr. Mark is a senior lecturer within the Natural Language Processing group of Sheffield University. His research areas include lexical semantics, word sense disambiguation, semantic similarity, information extraction and text retrieval. His publications include a monograph, two edited volumes and over one hundred papers in journals, collected volumes and international conferences. He has previously worked at Stanford University, Reuters Ltd and British Telecom’s research centre (Adastral Park).

Paul Clough Paul Clough is Professor in Information Retrieval at the Information School, University of Sheffield. His research interests mainly revolve around developing technologies to assist people with accessing and managing information. He has published work in the areas of multilingual information retrieval, information access to digital cultural heritage, evaluation of IR systems, geo-spatial search, text-based image retrieval, plagiarism detection, text re-use, and search analytics. Paul is co-author of a book on multilingual information retrieval and contributor to over 100 peer-reviewed publications.