Article:
Nawab, R.M.A., Stevenson, M. and Clough, P. (2017) An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14 (4). pp. 796-804. ISSN 1545-5963

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

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 ULMS 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.

Rao Muhammad Adeel Nawab is with the Department of Computer Science, COMSATS Institute of Information Technology, Defence Road, Off Raiwind Road, Lahore, Pakistan e-mail: adeelnawab@ciitlahore.edu.pk (see http://www.ciitlahore.edu.pk/PL/profile.aspx?employeeid=74).

Mark Stevenson is with the Natural Language Processing Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP United Kingdom e-mail: mark.stevenson@sheffield.ac.uk (see http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/).

Paul Clough is with the Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP United Kingdom e-mail: p.d.clough@sheffield.ac.uk (see http://ir.shef.ac.uk/cloughie/).

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.[1] 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

[1] http://www.ncbi.nlm.nih.gov/pubmed/ Last visited: 21-04-2015

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 System[2] - 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.

[2]http://sourceforge.net/projects/aaps/ Last visited: 21-04-2015

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, KL_δ (see Equation 1) [15]. Documents are modelled as probability distributions and compared using KL_δ. Documents are converted into probability distributions by removing stop words, stemming [33] and then computing *tf.idf* weights for the remaining word unigrams. Assume P_d is the probability distribution generated from d, a document in the reference collection, and that Q_s 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:

$$KL_\delta(P_d \| Q_s) = \sum_{x \in X} (P_d(x) - Q_s(x)) log \frac{P_d(x)}{Q_s(x)} \quad (1)$$

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* [**?**], [**?**]. 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 when 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.

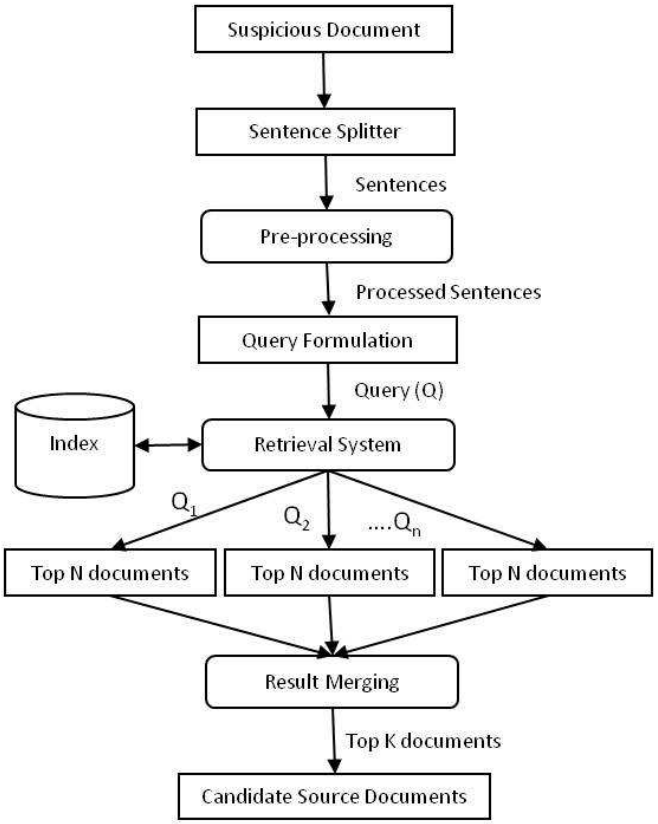

Fig. 1. Process of candidate document retrieval

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, $S_{finalscore}$, is obtained by adding the similarity scores of source documents obtained against each query q:

$$S_{finalscore} = \sum_{q=1}^{N_q} S_q\,(d) \tag{2}$$

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 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.

Input Text
Gamma-glutamyl transpeptidase.
MRCONSO Table Entries in English for the CUI C0202035
C0202035 ENG Gamma glutamyl transferase measurement
C0202035 ENG Gamma glutamyl transpeptidase measurement
C0202035 ENG GTP measurement

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

TABLE II
SIMPLIFIED EXAMPLE OUTPUT FROM METAMAP WITH AND WITHOUT WORD SENSE DISAMBIGUATION (WSD) BEING APPLIED. EACH ENTRY (E.G. "C0202035:GAMMA-GLUTAMYL TRANSPEPTIDASE") IS COMPOSED OF A CUI (E.G. "C0202035") AND INPUT TERM (E.G. "GAMMA-GLUTAMYL TRANSPEPTIDASE").

Input Text
Gamma-glutamyl transpeptidase.
MetaMap Output without WSD
Phrase: "Gamma-glutamyl transpeptidase."
Meta Mapping:
C0202035:Gamma-glutamyl transpeptidase
Meta Mapping:
C0017040:gamma glutamyl transpeptidase
MetaMap Output with WSD
Phrase: "Gamma-glutamyl transpeptidase."
Meta Mapping:
C0202035:Gamma-glutamyl transpeptidase

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 these 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*).

TABLE III
EXAMPLES OF EXPANDED QUERY USING UMLS METATHESAURUS

Query Sentence	hbf was correlated with total hemoglobin concentration and with serum afp concentration in hepatoma and bladder carcinoma
WSD	`hbf fetal^w hemoglobin^w was correlated correlation^w with total hemoglobin concentration finding^w of^w hemoglobin^w concentration^w and with serum afp alpha^w 1^w fetoprotein^w measurement^w concentration concentration^w measurement^w in hepatoma liver^w carcinoma^w and bladder carcinoma carcinoma^w of^w bladder^w`
Without WSD	`hbf foetal^w hemoglobin^w was correlated correlation^w with total of^w total^w hemoglobin concentration finding^w of^w hemoglobin^w concentration^w and with serum afp alpha^w 1^w fetoprotein^w measurement^w concentration concentration^w measurement^w in hepatoma carcinoma^w of^w liver^w and bladder carcinoma carcinoma^w bladder^w`
WSD Phrase	`hbf ''fetal hemoglobin''^w was correlated ''correlation''^w with total ''hemoglobin concentration'' ''finding of hemoglobin concentration''^w and with ''serum afp'' ''alpha 1 fetoprotein measurement''^w concentration ''concentration measurement''^w in hepatoma ''liver carcinoma''^w and ''bladder carcinoma'' ''carcinoma of bladder''^w`
Without-WSD Phrase	`hbf ''foetal hemoglobin''^w was correlated ''correlation''^w with total ''of total'' ''hemoglobin concentration'' ''finding of hemoglobin concentration''^w and with ''serum afp'' ''alpha 1 fetoprotein measurement''^w concentration ''concentration measurement''^w in hepatoma ''carcinoma of liver''^w and ''bladder carcinoma'' ''carcinoma bladder''^w`

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.[6] Each duplicate citation pair was classified into four categories:[7] (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, 265 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.

[6]http://dejavu.vbi.vt.edu/dejavu/duplicate/ Last visited: 21-04-2015

[7]There are also other categories but these four are more relevant to plagiarism.

TABLE IV
EXAMPLE DUPLICATE CITATION PAIR FROM 265 MANUALLY EXAMINED AND VERIFIED DUPLICATE CITATION PAIRS IN THE *Deja vu* DATABASE.

MEDLINE Corpus
Source: Gammaglutamyl transpeptidase **is an enzyme primarily located in the brush border of the proximal convoluted** tubules of the kidney. Its unique localisation in the renal **cells most easily damaged by ischaemia and its ease of assay** provides **the rationale for its use in the measurement** of renal ischaemic injury. Using a standard experimental animal model, canine urinary gamma-GT activity was shown to be increased up to 70-fold following 90 min of unilateral renal ischaemia and was significantly raised following only 5 min ischaemia. The urinary gamma-GT was used as a measure of ischaemic injury associated with renal transplantation in man and 20 consecutive patients undergoing kidney transplant were studied by daily 24-hour urinary gamma-GT estimations and **excellent correlation was obtained between raised** enzyme activity **and the clinical diagnosis of transplant rejection**.
Rewrite: The sites of ischaemic injury within the kidney are reviewed and the diagnostic value of measurements of plasma and urinary enzymes in renal ischaemic injury and in renal homotransplant rejection in experimental animals and man is examined. Gamma-glutamyl transpeptidase (gamma-GT) **is an enzyme primarily located in the brush border of the proximal convoluted** tubule of the kidney. Its unique localization in the **cells most easily damaged by ischaemia and its ease of assay** provide **the rationale for its use in the measurement** and diagnosis of renal ischaemic injury. gamma-GT activity was measured in dogs undergoing varying periods of renal ischaemia and under conditions of local renal hypothermia and was shown to be a sensitive indicator of ischaemic injury. Twenty consecutive patients undergoing renal homotransplantation were studied by daily estimation of their 24-h urinary gamma-GT activity; **excellent correlation was obtained between raised** levels of this enzyme **and the clinical diagnosis of transplant rejection**.

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 fromed 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 [33].Terms are weighted using the *tf.idf* weighting scheme.

Lucene computes the similarity score between query and document vectors using the cosine similarity measure:

$$sim(d,q) = \frac{\overrightarrow{q} \bullet \overrightarrow{d}}{|\overrightarrow{q}| \times |\overrightarrow{d}|} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2 \times \sum_{i=1}^{n}(d_i)^2}} \quad (3)$$

[8]http://lucene.apache.org/ Last visited: 21-04-2015
[9]NLTK [47] stop word list of 127 words in English was used.

where $|\overrightarrow{q}|$ and $|\overrightarrow{d}|$ 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:

$$R@K_{avg} = \frac{1}{|N|} \sum_{i=1}^{N} R@K_i \quad (4)$$

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.

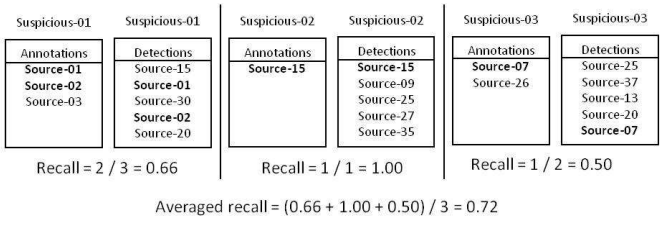

Fig. 2. Example showing calculation of averaged recall score

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 V
PERFORMANCE FOR THE MEDLINE CORPUS

	Avg. Recall for top K documents				
Approach	1	5	10	15	20
Kullback-Leibler	0.7596	0.8154	0.8442	0.8558	0.8596
No Query Expansion	0.8769	0.9173	0.9250	0.9288	0.9288
WSD	0.9077	0.9519	0.9558	0.9558	0.9596
Without-WSD	0.9035	0.9519	0.9519	0.9558	0.9558
WSD Phrase	**0.9219**	**0.9595**	0.9595	**0.9652**	0.9652
Without-WSD Phrase	0.9115	0.9558	**0.9596**	0.9634	**0.9673**

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 (W) 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
QUERY BY QUERY PERFORMANCE. NUMBER OF QUERIES FOR WHICH THE RANKING IS HIGHER, LOWER OR REMAINED SAME USING A QUERY EXPANSION

		No. of Queries (%) effecting Rank		
Corpus	Approach	Higher	Lower	Same
MEDLINE	WSD	14(5.38)	2(0.77)	234(90.00)
	Without-WSD	17(6.54)	5(1.92)	230(88.46)
	WSD Phrase	13(5.00)	4(1.54)	234(90.00)
	Without-WSD Phrase	15(5.77)	4(1.54)	233(89.62)

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

[1] B. Martin, "Plagiarism: A misplaced emphasis," *Journal of Information Ethics*, vol. 2, pp. 36–47, 1994.
[2] M. Joy and M. Luck, "Plagiarism in programming assignments," *IEEE Transactions of Education*, vol. Vol. 42(2), pp. 129–133, 1999.
[3] P. Samuelson, "Self-plagiarism or fair use?" *Communications of the ACM*, vol. Vol. 37(8), pp. 21–25, 1994.
[4] C. Park, "In other (people's) words: plagiarism by university students," *literature and lessons, Assessment and Evaluation in Higher Education*, vol. 5, no. 8, 2003.
[5] D. McCabe, "Research report of the center for academic integrity," Tech. Rep., 2005.
[6] G. Judge, "Plagiarism: Bringing economics and education together (with a little help from it)," *Computers in Higher Education Economics Review*, vol. 20, pp. 21–26, 2008.
[7] J. Sheard, M. Dick, I. M. S. Markham, and M.Walsh, "Cheating and plagiarism: Perceptions and practices of first year it students," in *in: ACM SIGCSE Bulletin*, vol. 34, 2002, pp. 183–187.

[8] M. Lesk, "How many scientific papers are not original?" *Proceedings of the National Academy of Sciences*, vol. 112, no. 1, pp. 6–7, 2015. [Online]. Available: http://www.pnas.org/content/112/1/6.short

[9] D. T. Citron and P. Ginsparg, "Patterns of text reuse in a scientific corpus," *Proceedings of the National Academy of Sciences*, vol. 112, no. 1, pp. 25–30, 2015. [Online]. Available: http://www.pnas.org/content/112/1/25.abstract

[10] M. Errami, J. Hicks, W. Fisher, J. W. D. Trusty, T. Long, and H.Garner, "vu - a study of duplicate citations in medline," *Bioinformatics*, vol. 24, pp. 243–249, 2008.

[11] R. Boisvert and M. Irwin, "Plagiarism on the rise," *Communications of the ACM*, vol. 49, pp. 23–24, 2006.

[12] D. McCabe, K. Butterfield, and L. Trevino, "Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action," *Academy of Management Learning and Education*, vol. 5, no. 3, pp. 1–294, 2006.

[13] F. Culwin and T. Lancaster, "Plagiarism issues for higher education," *Vine*, vol. Vol. 31(2), pp. 36–41, 2001.

[14] B. Stein, S. Eissen, and M.Potthast, "Strategies for retrieving plagiarized documents," in *in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 825–826.

[15] A. Barrón-Cedeño, P. Rosso, and J.Benedi, "Reducing the plagiarism detection search space on the basis of the kullback-leibler distance," in *in: Proceedings of 10th International Conference on Computational Linguistics and Intelligent Text Processing*, 2009, pp. 523–534.

[16] C. Campbell, "Writing with other's words: Using background reading text in academic compositions," in *in: In B. Kroll (Ed.) Second language writing: research insights for the classroom*. Cambridge: Cambridge University Press, 1990, pp. 211–230.

[17] A. Johns and P. Myers, "An analysis of summary protocols of university esl students," *Applied Linguistics*, no. 11, pp. 253–271, 1990.

[18] C. Keck, "The use of paraphrase in summary writing: A comparison of l1 and l2 writers," *Journal of Second Language Writing*, no. 15, pp. 261–278, 2006.

[19] A.Barrón-Cedeño, M. Vila, M. Marti, and P. Rosso, "Plagiarism meets paraphrasing: Heading to the next generation in automatic plagiarism detection," *(Submitted)*, 2012.

[20] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in *in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010, pp. 997–1005.

[21] S. Alzahrani, V. Palade, N. Salim, and A. Abraham, "Using structural information and citation evidence to detect significant plagiarism cases in scientific publications," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 286–312, 2012. [Online]. Available: http://www.cs.ox.ac.uk/people/vasile.palade/papers/JASIST_final.pdf

[22] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 45–62, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1007/s10579-009-9114-z

[23] N. Shivakumar and H. Garcia-Molina, "Scam: A copy detection mechanism for digital documents."

[24] P. Lane, C. Lyon, and J. Malcolm, "Demonstration of the ferret plagiarism detector," in *in: Proceedings of the 2nd International Plagiarism Conference*, 2006.

[25] P. Clough and M.Stevenson, "Developing a corpus of plagiarised short answers," *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis*, vol. 1, no. 45, pp. 5–24, 2011.

[26] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - a survey," *Journal of Universal Computer Science*, vol. 8, no. 12, pp. 1050–1084, 2006.

[27] M. Potthast, B. Stein, A. Eiselt, A. Barr, and P.Rosso, "Overview of the 2nd international competition on plagiarism detection," in *in: Proceedings of the CLEF10 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, 2010.

[28] M. Potthast, A. Eiselt, A. Barr, B.Stein, and P.Rosso, "Overview of the 3rd international competition on plagiarism detection," in *Notebook Papers of CLEF 11 Labs and Workshops*, 2011.

[29] J. Kasprzak and M. Brandejs, "Improving the reliability of the plagiarism detection system," in *in: Proceedings of the 4th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, Lab Report for PAN*, at CLEF, 2010.

[30] O. Uzuner, B. Katz, and T. Nahnsen, "Using syntactic information to identify plagiarism," in *in: Proceedings of the 2nd Workshop on Building Educational Applications using Natural Language Processing, ACL,*, 2005, pp. 37–44.

[31] M. Chong, L. Specia, and R. Mitkov, "Using natural language processing for automatic detection of plagiarism," in *in: Proceedings of the 4th International Plagiarism Conference (IPC-2010)*, 2010.

[32] M. Mozgovoy, T. Kakkonen, and E. Sutinen, "Using natural language parsers in plagiarism detection," in *in: Proceedings of SLaTE'07 Workshop*, 2007.

[33] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 3, no. 14, pp. 130–137, 1980.

[34] B. Stein, P. Rosso, E.Stamatatos, M.Koppel, and E. Agirre, "3rd pan workshop on uncovering plagiarism, authorship and social software misuse," in *in: 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2009, pp. 1–177.

[35] M. Potthast, T. Gollub, M. Hagen, J. Graßegger, J. Kiesel, M. Michel, A. Oberländer, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso, and B. Stein, "Overview of the 4th International Competition on Plagiarism Detection," in *Working Notes Papers of the CLEF 2012 Evaluation Labs*, P. Forner, J. Karlgren, and C. Womser-Hacker, Eds., Sep. 2012. [Online]. Available: http://www.clef-initiative.eu/publication/working-notes

[36] M. Potthast, T. Gollub, M. Hagen, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein, "Overview of the 5th International Competition on Plagiarism Detection," in *Working Notes Papers of the CLEF 2013 Evaluation Labs*, P. Forner, R. Navigli, and D. Tufis, Eds., Sep. 2013. [Online]. Available: http://www.clef-initiative.eu/publication/working-notes

[37] M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, and B. Stein, "Overview of the 6th International Competition on Plagiarism Detection," in *Working Notes Papers of the CLEF 2014 Evaluation Labs*, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, Eds. CLEF and CEUR-WS.org, Sep. 2014. [Online]. Available: http://www.clef-initiative.eu/publication/working-notes

[38] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. Garner, "Text similarity: An alternative way to search medline," *Bioinformatics*, vol. 18, no. 22, pp. 2298–2304, 2006.

[39] M. Errami, Z. Sun, T. L. A. George, M. Skinner, J. Wren, and H. Garner, "Identifying duplicate content using statistically improbable phrases," *Bioinformatics*, vol. 11, no. 26, pp. 1453–1457, 2010.

[40] H. Fang, "A re-examination of query expansion using lexical resources," in *in: Proceedings of Association of Computational Linguistics*, 2008, pp. 139–147.

[41] K. Lu and X. Mu, "Query expansion using umls tools for health information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 1, no. 46, pp. 1–16, 2009.

[42] R. Nawab, M. Stevenson, and P. Clough, "Retrieving candidate plagiarised documents using query expansion," in *in: Proceedings of the 34th European Conference on Information Retrieval (ECIR)*, Springer, 2012, pp. 207–218.

[43] R. Nawab, P. Clough, and M. Stevenson, "Detecting text reuse with modified and weighted n-grams," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics*, 2012, pp. 54–58.

[44] C. Chen, J. Yeh, and H. Ke, "Plagiarism detection using rouge and wordnet," *in: Journal of Computing*, vol. 2, pp. 34–44, 2010.

[45] Z. Ceska, *Plagiarism Detection Based on Singular Value Decomposition*. Springer: in: Lecture Notes in Computer Science, 2008, vol. 5221.

[46] M. Chong and L.Specia, "Lexical generalisation for word-level matching in plagiarism detection," in: Proceedings of the Recent Advances in Natural Language Processing (RANLP), pp. 704–709, 2011.

[47] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, in: Proceedings of the ACL-02 Workshop, 2002, pp. 63–70.

[48] E. Fox and J. Shaw, "Combination of multiple searches," In Proceedings of TREC-2, 1994, pp. 243–249.

[49] A. Aronson and F. Lang, "An overview of metamap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 3, no. 17, pp. 229–236, 2010.

[50] S. Humphrey, W. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. Rindflesch, "Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment," *Journal of the American Society for Information Science and Technology*, vol. 1, no. 57, pp. 96–113, 2006.

[51] M. Errami, J. Wren, J. Hicks, and H. Garner, "etblast: A web server to identify expert reviewers, appropriate journals and similar publications," *Nucleic Acids Research*, no. 35, pp. 12–15, 2007.

[52] F. Wilcoxon, S. Katti, and R. Wilcox, *Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Tank Test*, ser. Selected Tables in Mathematical Statistics, 1973, no. 1.

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.