INTERNATIONAL JOURNAL OF CURRENT MEDICAL AND PHARMACEUTICAL RESEARCH

IJCMPR

Available Online at http://www.journalcmpr.com

RESEARCH ARTICLE

PLAGIARISM DETECTION IN ARABIC USING TRANSLATION AND MEDICAL ONTOLOGY

[1]Khaled Omar., [2]Bassel Alkhatib and [3]Mayssoon Dashash

[1,2]Department of Artificial Intelligence, Faculty of Informatics Engineering, Damascus University
[3]Department of Pediatric Dentistry, Faculty of Dentistry, Damascus University

ARTICLE INFO

Article History:
Received 19th June, 2016
Received in revised form 16th July, 2016 Accepted 18th August, 2016 Published online 19th September, 2016

Key words:

Plagiarism Detection, Semantic Web, Arabic Word Net, Medical Ontologies, Automatic Translation

ABSTRACT

The huge increase in documents on the world-wide web and the availability to reach and download them has led to a dangerous problem which is using others' works without giving them credits.
Although a number of methods have been developed to discover popular cases of plagiarism in Arabic Language, as changing sentence structure or replacing words with their synonyms, it is still difficult to diagnose plagiarism when modifying deliberately quoted sentences.
In this paper a Semantic Similarity Algorithm System is proposed for detecting plagiarism in medical Arabic papers using semantic networks in Arabic language (Arabic Word Net), automatic translation to English language, and international medical Ontologies(in English Language). The developed algorithm depends on determining the degree of semantic similarity between original documents and suspected documents by calculating the intersection of the semantic information between files, the proposed algorithm uses Arabic Word Net to detect sentences concepts with their synonyms, and the medical Ontologies are used to expanding the sentences of origin and suspected texts, and calculates the semantic similarity between them, the automatic translation was used to translate text from Arabic Language to English Language to benefit from International Medical Ontologies, because of the lack of Medical Ontologies in Arabic Language.
The proposed algorithm has showed a good results by determining the similarity between the origin documents and the suspected documents using semantic detection score, that it has discovered the plagiarism cases even if the user replace some words with their synonyms, and if the user restructured the sentences of the plagiarized texts.

INTRODUCTION

Plagiarism is the process of reusing ideas or writings of a person or several persons by others, without referring to the source of information [1], [2]. Research related to plagiarism detection has been started in the seventies of the last century. Several algorithms and methods have been designed to determine the unusual similarities between student’s assignments. Recently efforts of researchers in natural languages processing tended to determine the similarities between natural languages texts. However, it is not an easy task because of the ambiguity in natural languages and the unconstrained vocabulary size in natural languages. This study aimed to investigate a new semantic method for plagiarism detection in Arabic medical papers.

Taking in consideration the absence of diacritics in Arabic text which causes a lot of problems when analysing the texts morphological analyses, in addition the absence Arabic Medical Ontologies which is very essential to exist when developing semantic plagiarism detection methods for Arabic medical papers. The structure of this paper is related work, proposed method, system design and implementation, tests and results, discussion, conclusion and discusses future work.

Related Work

A lot of research has been done about plagiarism detection and many algorithms have been developed to detect plagiarism, but a few of it take into consideration semantic meaning, that many of it depends on string matching, In general plagiarism detection methods classified into two main types: finger printing plagiarism detection algorithms [3] and content based plagiarism detection algorithms[4], fingerprinting algorithms generate a code for each file called text fingerprint and then comparing between generated fingerprint codes, the most famous fingerprint algorithm called Winnowing fingerprint algorithm[5], there are types of fingerprint algorithms (character based, phrase based, statement based)[6], and the second category of plagiarism detection methods is content based algorithms which contains string matching algorithms (mainly depends on string matching and NLP techniques), these types of algorithms have a weakness point called “split match problem”[7] which determines the optimal string length to be matched or compared between files , and tree matching algorithms which depends on generating text trees based on syntactic texts analyses, and then comparing between generated trees for origin documents and suspected ones.

Recently Citation-based Plagiarism Detection algorithms were developed and these algorithms could be applied to any text containing citations – this includes academic documents, scientific publications [8]. This approach overcomes the shortcoming of existing text-based plagiarism detection methods, that existing methods typically fail to detect translated and strongly disguised plagiarism instances, since they only examine words (i.e. text overlap) in documents to detect suspicious similarity.

In contrast, Citation-based Plagiarism Detection makes use of the semantic information implied by the citations within documents. The approach identifies and analyses similar patterns in the citation sequences of academic documents to compute similarity [9].

Unlike character-based approaches, this approach does not rely on text comparisons alone, but analyses citation patterns within documents to form a language-independent "semantic fingerprint" for similarity assessment. The practicability of Citation-based Plagiarism Detection was proven by its capability to identify so-far non-machine detectable plagiarism in scientific publications [10], and this type of detection algorithms is classified as a semantic plagiarism detection algorithms.

There is a new approach for plagiarism detection using fuzzy information retrieval but this approach required a lot of text pre-processing before applying it (corpus collecting and stemming) [11].

The Proposed Method

Before go in details about the proposed algorithm we will explain about the resources used by the proposed algorithm, the proposed algorithm users the following resources:

- Arabic Word Net for Arabic language that At the time Arabic Word Net consists of 9228 synsets (6252 nominal, 2260 verbal, 606 adjectival, and 106 adverbial), containing 18,957 Arabic expressions. This number includes 1155 synsets that correspond to Named Entities which have been extracted automatically and are being checked by the lexicographers [12].
- English Word Net: The main relation among words in Word Net is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets). Each of Word Net's 117 000 synsets is linked to other synsets by means of a small number of "conceptual relations." Additionally, a synset contains a brief definition ("gloss") and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique [13].
- Medical Ontologies: the algorithm uses a huge set of medical Ontologies which are grouped into one Ontology lookup service OLS, this OLS provides a web service interface to query multiple ontologies from a single location with a unified output format[14]

The proposed algorithm mainly contains five stages:

- analyzing stage,
- translation stage,
- expansion stage,
- comparison stage,
- printing stage

As shown in the following figure below:

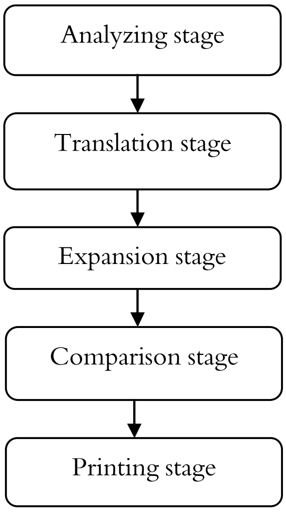

Fig. 1 Proposed algorithm main stages

The algorithm takes tow files as input (origin file, suspected file) and it analyses the tow files (which are in Arabic language), then it translates the two files to English language, then it expands the translated files concepts, then it compares between the files data and finally it print the plagiarism detection results, here we will go in details for each stage.

Analyzing stage: in this stage the algorithm takes the tow input files and analyses them according to the following steps (the Analyzing stage is done on Arabic texts before translation):

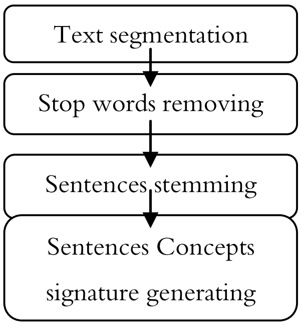

Fig. 2 Analyzing stage steps

the segmentation is done to detect text sentences, and it done basing on the punctuation marks in the text, stop words are removed basing on a list for stop words in Arabic Language[15], sentences stemming is done using Stanford for Natural Language Processing Tools[16], in Sentences concepts signature generating step of the analyzing stage the algorithm iters throw text sentences and for each sentence it gets the stemmed sentences words, , and then these words are expanded using Arabic WordNet, The expansion here is only by adding each word synonyms, by the ending of this step every sentence has a set of stemmed words and their synonyms and this set is called sentence concept signature.

Translation stage: the aim of this stage is to benefit from International Medical Ontologies which describe medical concepts, and to benefit from English WordNet which describes general knowledge concepts, the translation is done using Bing Translation API [17], that the algorithm iters throw texts sentences and translates every sentence to English language, the translated sentences of the origin and suspected texts are mapped to Arabic origin sentences.

Expansion stage: the aim of this stage is to enrichment of the translated sentences of the origin text and the suspected one after analyzing these sentences and extracting concepts from it, the aim of extracting and enriching of sentences concepts is to overcome of the actions which taken by plagiarizer such as word synonyms replacements and sentences re-structuring this stage contains the following steps as shown in the figure below:

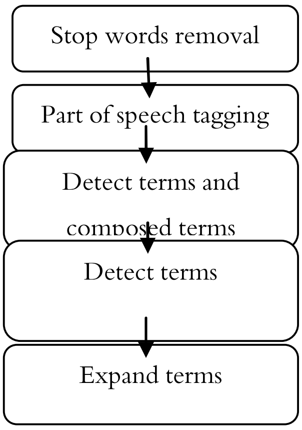

Fig. 3 Expansion stage steps

Stop words removal is done basing of stop words list in the English language [18], the algorithm uses Stanford Natural language processing tools to execute part of speech tagging for translated sentences, the algorithm deals only with detected nouns , and composed nouns(the composed nous are detected using Stanford relations dependency)[19],after the nouns and composed nouns are detected (called in this step terms) the algorithm begins to enrichment these terms by firstly detect home ontology for each term, as mentioned that two types of Ontologies are used.

- WordNet: to enrichment general knowledge concepts.
- Medical Ontologies : to enrichment medical concepts, this enrichment is done after detecting home ontology for each term, after that the algorithm executes the enrichment as the following:

For general knowledge concepts: the enrichment is done by adding term synonyms, parents, Childs.

For medical concepts: the enrichment is done by adding concept parents, Childs, related concepts.

at the end of this step the translated sentences are mapped to the origin Arabic sentences with enrichment vectors represent the translated texts sentences, this vectors are weighted that every origin concept has the weight value /1/ and every concept which generated from the enrichment have the weight /0.5/ .

Example includes (analyzing, translation, expansion) stages: let us take a small paragraph from Arabic medical paper abstract as the following:
خلفية البحث وهدفه: إن ذات الرئة المرافقة للتنبيب الرغامي من أخطر الاختلاطات التي قد تودي بحياة المريض في وحدة العناية الجراحية بأنواعها، ولا سيما الإسعافية منها.

Applying the analyzing stage

The segmentation of the Arabic text results three following sentences:

Sentence 1:

خلفية البحث و هدفه.

Sentence 2:

إن ذات الرئة المرافقة للتنبيب الرغامي من أخطر الاختلاطات التي قد تودي بحياة المريض في وحدة العناية الجراحية بأنواعها.

Sentence 3:

ولا سيما الإسعافية منها.

Removing the stop words in the three sentences as the following:

Sentence 1:

خلفية البحث هدفه.

Sentence 2:

الرئة المرافقة للتنبيب الرغامي أخطر الاختلاطات تودي بحياة المريض وحدة العناية الجراحية بأنواعها.

Sentence 3:

ولا سيما الإسعافية منها.

Stemming the words in the first sentence as the following:

خلف بحث هدف

Generating the concept signature for the first sentence as the following:

To generate the sentence (1) concept signature the algorithm expands every stemmed term as the following

The term "بحث" is expanded by adding its synonyms from Arabic WordNet, its synonyms set is:

طلب,استفسر,درس,نظر,راعى,دراسة,تفتيش,فتش,تحقيق,استقصاء,نشد,تحر, استعلم .

And for the term "هدف" its synonyms set is:

تسجيل الهدف,غاية,رمى إلى,منطقة الهدف,قصد,غرض,دريئة,نطاق الهدف,تسجيل النقطة.

And for the term "خلف" its synonyms set is:

قفا,ذرية, خلفي,ظهر,ترك ,نسل

The algorithm stems the synonyms sets for all sentence words and aggregates them into sentence concept signature.

Applying the translation stage:

The algorithm translates the text as the following:
Sentence (1) translation:
Research background and purpose.
Sentence (2) translation:
The accompanying tracheal intubation pneumonia is one of the most serious complications that may kill the patient in the surgical care unit types.

Sentence (3) translation:
Especially ambulatory including.
-Applying the expansion stage:
The algorithm begins expansion stage by iteration for all translated sentences and for each sentence the algorithm detects terms and composed terms, and detect every term home ontology and then expands it.

For the first translated sentence “Research background and purpose” which is mapped to the origin sentence

“.خلفية البحث وهدفه” the algorithm detect terms in this sentence, which are: Research, background, purpose

Then the algorithm detects the home ontology for each term as the following:

purpose: WN ,background: WN, Research: WN the detection of term home Ontologies showed that all terms in this sentence are belong to Word Net Ontology.

Then the algorithm expand the terms by adding childs, Parents, synonyms, after the expansion is done every translated sentence is represented with a weighed vector contains its origin concepts and the concepts resulting from the expansion step,(every origin concept has the weight value/1/, and every new concept from expansion has the weight value/0.5/.

Comparison stage: in this stage the algorithm compares between origin document text and suspected document text, the comparison is done by comparing between texts sentences. That the algorithm iters throw origin text sentences and for each sentence it compares it with suspected text sentences and search for the best matching sentences of the suspected sentences, the comparison is done at tow levels, firstly the algorithm calculates the intersection between sentences concept signature for the sentences under comparison, and if the intersection score is smaller than a threshold, the algorithm stop the comparison action with the current sentence, The sentences concept signature comparison is done according to the function:

$$\mathbf{f} = \begin{cases} 1: \text{ if the intersection is greater than threshold.} \\ 0: \text{ if intersection is smaller than threshold.} \end{cases}$$

If the previous f function returns the value /0/, the algorithm stops the comparison between the two sentences and it iters to the next sentence to compare with it.

But if the f function returns the value /1/, then the algorithm calculates the semantic similarity between two sentence enrichment vectors using the following formula [20]:

$$similarity = \cos(\bar{\ }) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where:
A: is the first terms vector,
B: is the second terms vector.
N: is the count of the shared terms.
the algorithm complete the iteration for each sentence in the origin document and keep only the best matching sentence from the suspected document, at the end of this stage every sentence from the origin document has its best matching sentence from the suspected document.

Printing stage: at this stage the algorithm shows the plagiarism detection results, by colouring the plagiarized text of the origin document with red colour, that the algorithm does not change the origin structure of the origin document or the of the suspected one, so that makes printing the plagiarism detection results an easy and accurate task.

System Design and Implementation

In the implementation of the proposed System we use Arabic Word Net Ontology files, English Word Net Ontology files, Ontology lookup service(OLS) which provides a web service to get access to the Medical Ontologies, this web service provides methods like get Parents, get Childs, get Relations of the current Concept.

The Developing language was java and the IDE was NetBeans7.1, and we use SqlServer 2008 as a DataBase environment to store concepts and their related concepts to make system faster by time, so the algorithm searches for concepts firstly into System local database if it does not find it then it uses OLS web service.

The Developed system consist of the following Modules: Searcher Module which uses Bing search API[21] to search for the possible plagiarism sources on the internet, file downloader module to download the search results, translator which uses Bing translation API to translate Arabic texts to English language, semantic detector module which contains expanding the translated texts calculates the scores of semantic similarity between origin files sentences and suspected files sentences, result viewer module which show the plagiarism detection results.

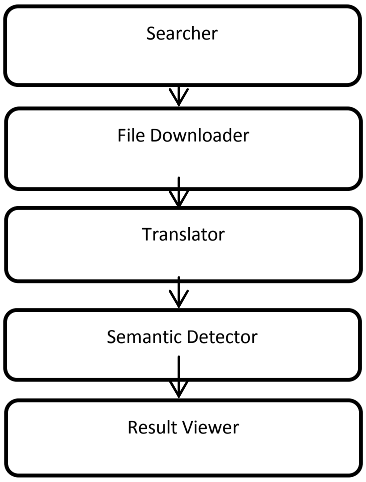

Fig 4 System structure

Test and Results

The proposed algorithm was tested on a sample of files. About of 50 medical papers in Arabic from the health magazine of Damascus university archive www.**damascus university**. edu.sy/mag/**health**/.

The detection algorithm which is based on Arabic WordNet, English Word Net, Medical Ontologies for enrichment text sentences(enrichment include adding concept synonyms, parents, childs, related concepts) was very effective algorithm for plagiarism detection, This is because it segments the text according its origin sentences and them comparing between sentences at two levels (concept signature level between sentences in Arabic Language, expanded vector comparisons in English translated language, this comparison technique overcomes the problem when the user change some text words with its synonyms,or when the suspected sentences structure are changed.

The evaluation of our system done by the Precision factor [22], which is described below:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

for the Arabic Medical publications the precision was 73%, and this value affected by the resolution of the translation from Arabic to English, and affected by the concepts which don't belong to any medical ontology to enrich them, that the experience from medical papers text analyses showed that there is about 30% from texts concepts are not found in English Word Net or in any Medical Ontology used by our Proposed system.

DISCUSSION

In this study we developed a system to detect plagiarism in medical papers in Arabic languages using "Bing" search engine.

The testing on developed algorithm approved that is very effective in plagiarism detection and more effective than string based plagiarism detection algorithms and traditional plagiarism detection methods that the developed algorithm overcomes on weakness points such as word re-ordering , sentence re-structuring, words replacements with their synonyms, the comparison complexity of proposed algorithm is efficient that the algorithm stop comparison stage if the intersection between sentences concept signature is smaller than the algorithm intersection threshold (this threshold is not constant ,it adaptive threshold basing on the count of concepts between sentences under comparison)

And there was a weakness in our developed algorithm causing from that there some words are not include in Arabic WordNet so the algorithm cannot find their synonyms.

CONCLUSION AND FUTURE WORK

We concluded that developing a semantic plagiarism detection method is a very hard and important step in plagiarism detection research because using semantic resources is promised matter in the strengthening of the plagiarism detection methods ,our future plan is extending the domain of developed method to be generic for all domains of science by using and including extra semantic Ontologies for all domains, and using the unified medical dictionary[23] (multilingual medical dictionary) which developed by the World Health Organization to get the exact and best translation of the medical Arabic terms to enhance the accuracy of the developed plagiarism detection algorithm in this research .

Acknowledgements

This study is supported by Damascus University, Syria. Damascus University, Syrian Arab Republic.

References

1. Vinod K.R., Sandhya.S, Sathish Kumar D, Harani A, David Banji and, Otilia JF Banji, "Plagiarism history, detection and prevention ", Hygeia: *journal for drugs and medicines*, Vol.3-Issue.1-pp. 1- 4, 2011
2. Carroll, J. (2002) A Handbook for Deterring Plagiarism in Higher Education. Oxford: Oxford Brookes University
3. Jadalla and A. Elnagar, "A fingerprinting-based plagiarism detection system for Arabic text-based documents," Computing Technology and Information Management (ICCM), 2012 8th International *Conference on*, Seoul, 2012, pp. 477-482.
4. Hoad C., Zobel J. Methods for identifying versioned and plagiarized documents [J]. *Journal of the American Society for Information Science and Technology*, 2003, 54(3): 203-215.
5. Adel Aljohani and Masnizah Mohd, 2014. Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm. Information Technology Journal, 13: 2349-2355.
6. Chow Kok Kent, Naomie Salim, Features Based Text Similarity Detection, *Journal of Computing*, Volume 2, Issue 1, January 2010, ISSN 2151-9617 https://sites.google.com/site/journalofcomputing/
7. Maxim Mozgovoy and Tuomo Kakkonen and Erkki Sutinen, Using Natural Language Parsers in Plagiarism Detection, Speech and Language Technology in Education (SLaTE 2007).
8. B. Gipp and N. Meuschke, "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence," in Proceedings of the 11th ACM symposium on Document engineering (DocEng '11), Mountain, View, CA, USA, 2011.
9. SCIPlore Knowledge Discovery, University. [Online] http://www.sciplore.org/projects/citation-based-plagiarism-detection/ [Accessed 15-March 2016].
10. Bela Gipp. 2014. *Citation-Based Plagiarism Detection: Detecting Disguised and Cross-Language Plagiarism Using Citation Pattern Analysis.* Springer Vieweg
11. Salha Mohammed Alzahrani , and Naomie Salimm, Plagiarism Detection In Arabic Scripts Using Fuzzy Information Retrieval, Proceedings of 2008 Student Conference on Research and Development (SCOReD 2008), 26-27 Nov. 2008, Johor, Malaysia.
12. Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, and Christiane Fellbaum. Arabic WordNet: Current State and Future Extensions in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008. http://nlp.lsi.upc.edu/papers/rodriguez08.pdf
13. Word Net. [online] Available at: https://wordnet.princeton.edu/ [Accessed 1-March 2016].
14. Ontology Lookup Service. [online] Available at: http://www.ebi.ac.uk/ols/v2/init.do [Accessed 10-March 2016].
15. ranks, [online] Available at: http://www.ranks.nl/stopwords [Accessed 10-March 2016].
16. Stanford NLP tools. [online] Available at: http://nlp.stanford.edu/ [Accessed 1-Jan 2016].
17. Microsoft Translator. [online] Available at: https://www.microsoft.com/en-us/translator/getstarted.aspx [Accessed 15-Jan 2016].

18. RANKS NL. [online] Available at: http://www.ranks.nl/stopwords [Accessed 15-Jan 2016].
19. Stanford NLP tools. [online] Available at: http://nlp.stanford.edu/IR-book/html/htmledition/ stemming-and-lemmatization-1.html [Accessed 25-march 2016].
20. Wikipedia, Cosine similarity.[online] https://en.wikipedia.org/wiki/Cosine_similari / [Accessed 1-Jan 2016].
21. Bing Search API. [online] http://www.bing.com/ toolbox/ bingsearchapi [Accessed 1-Jan 2016].
22. wikipedia, [online]Available :https://en.wikipedia.org/ wiki/Precision_and_recall [Accessed 1-Jan 2016].
23. World Health Organization, [online] Available: http://www.emro.who.int/Unified-Medical-Dictionary.html [Accessed 1-Jan 2016].