Fuzzy-Semantic Similarity for Automatic Multilingual Plagiarism Detection

Hanane EZZIKOURI
LMACS laboratory, Mathematics Department,
Faculty of sciences and techniques,
Sultan Moulay Slimane University
Beni-Mellal, BP: 523, Morocco

Mohamed ERRITALI
TIAD laboratory, Computer Sciences Department,
Faculty of sciences and techniques,
Sultan Moulay Slimane University
Beni-Mellal, BP: 523, Morocco

Mohamed OUKESSOU
LMACS laboratory, Mathematics Department,
Faculty of sciences and techniques,
Sultan Moulay Slimane University
Beni-Mellal, BP: 523, Morocco

***Abstract*—A word may have multiple meanings or senses, it could be modeled by considering that words in a sentence have a fuzzy set that contains words with similar meaning, which make detecting plagiarism a hard task especially when dealing with semantic meaning, and even harder for cross language plagiarism detection. Arabic is known by its richness, word's constructions and meanings diversity, hence changing texts from/to Arabic is a complex task, and therefore adopting a fuzzy semantic-based approach seems to be the best solution. In this paper, we propose a detailed fuzzy semantic-based similarity model for analyzing and comparing texts in CLP cases, in accordance with the WordNet lexical database, to detect plagiarism in documents translated from/to Arabic, a preprocessing phase is essential to form operable data for the fuzzy process. The proposed method was applied to two texts (Arabic/English), taking into consideration the specificities of the Arabic language. The result shows that the proposed method can detect 85% of the plagiarism cases.**

***Keywords*—CLPD; fuzzy similarity; natural language processing; plagiarism detection; semantic similarity**

I. Introduction

A word may have several possible meanings and senses due to the richness of natural languages, which make detecting plagiarism a hard task especially when dealing with semantic meaning, not just searching for patterns of text that are illegally copied from others (copy and paste texts from digital resources without acknowledging the original resource), this is the most and common plagiarism and it's called literal Plagiarism. The dangerous kind of plagiarism is semantic plagiarism also named obfuscated plagiarism, the plagiarized passages are unseen for existing PD tools like Paraphrasing the text by modifying the structure of the original sentences and changing their syntactical structure and lexical variations such as replacing some of the original words with its synonyms, etc., without proper citation or quotation marks, the other type of plagiarism is cross-language plagiarism and the aim of our work, its importance has grown up recently as semantic content of a document could be discreetly plagiarized through translation (human or machine-based). CLP consists in discriminating semantically similar texts independent of the languages they are written in, i.e. an unacknowledged reuse of a text involving its translation from one language to another and no reference to the original source is given [9]. We can say that semantic plagiarism is an idea plagiarism, because the texts are changed but ideas in the original texts remain the same.

Similarity is a fundamental and extensively used concept. Several similarity measures based on the semantic relatedness of words have been proposed these last years, to recover the luck of traditional PD methods and technics that give good result with literal plagiarism, and do not work with plagiarized texts that are semantically similar.

In this paper, we propose a very detailed fuzzy semantic-based similarity model for analyzing and comparing texts in CLP cases, in accordance with the WordNet lexical database, to detect plagiarism in documents translated from/to Arabic. Arabic is a language known by its complex linguistic structure and translation is often a fuzzy process that is hard to search for, which make CLPD a challenging task. We focus on highly obfuscated plagiarism cases which are translated and rephrased into another text and no reference to the original source is given.

An important task in any text analysis application is the creation of a suitable target data set to which models and algorithms can be applied is preprocessing, such as tokenization, part-of-speech (POS) tagging, lemmatization and stop words removal for deleting meaningless words and text segmentation is done using word 3gram.

Fuzzy semantic-based approach is obtained based on the fact that words from two translated compared texts have, in general, a Strong fuzzy similarity words of the meaning from the second language.

II. Related Work

From the review of literature, several works have been made to detect the likeness between texts documents, limited researches have concentrated on obfuscated plagiarism detection that integrate the semantic relationships between two candidate texts, thus a few researches in CLPD especially in Arabic. Therefore, this section presents several recently proposed plagiarism detection techniques founded on semantic similarity measures and fuzzy semantic-based models based on lexical taxonomies such as WordNet.

Alzahrani et al. [1] presented a semantic based plagiarism detection technique, which used fuzzy membership function to calculate the degree of similarity. The method developed in

four main stages. First is preprocessing, contains tokenization, stop words removal and stemming. Then the use of Jaccard coefficient and shingling algorithm to retrieve candidate documents list for each suspicious document. Detailed comparison is carried out next between the suspicious document and the corresponding candidate documents. Fuzzy similarity is calculated, it varies between 0 to 1; 0 for completely different sentences and 1 for duplicate sentences. The decision is based on the calculated fuzzy similarity compared to a threshold. In the end a post-processing is carried out where consecutive sentences are combined to form paragraphs.

Osman et al. [2] proposed an approach based on a Fuzzy Inference System and Semantic Role Labeling (FIS-SRL), the technique analyses and compares text based on a semantic allocation for each term inside the sentence. The proposed method generate arguments for each sentence semantically, and then chooses for each argument generated by the FIS in order to select important arguments. The FIS select the most important arguments, and uses the results in the similarity calculation process. Authors evaluate the method using PAN-09 corpus and found that gave good results; but it is required a lot of calculation.

Gupta, et al. [3] uses different preprocessing methods based on NLP techniques, authors shows that similarity calculation could be improved using fuzzy semantic similarity measures and introduce a measure that provide an important amelioration in the efficiency and accuracy of the system compared to the original method offered by Alzahrani et al. The system evaluated using PAN 2012 data set.

Ahangarbahan et al. [4] proposed a method based on lexical and semantic features of Persian texts. A necessary first step gather preprocessing, stop word removing and dividing the text into two parts: general and domain-specific knowledge words, after that the system was designed to measure text similarity.

Alzahrani et al. [5] presented a fuzzy semantic-based model for plagiarism detection based on fuzzy rules and semantic information from words in compared texts. Firstly, extracting features from texts to implement n-gram/sentence segments and POS-related semantic spaces. Secondly, evaluating fuzzy rules to judge the similarity in compared texts wherein word-to-word semantic similarity was studied based on Wu and Palmer similarity measure, a learning method that combines a permission and a variation threshold is used to decide true plagiarism cases.

III. Fuzzy Sets Theory for CLPD

Fuzzy logic is an extension of Boolean logic by Lofti Zadeh in 1965 based on his mathematical theory of fuzzy sets, which is a generalization of the theory of classical sets. The membership of an elements in a classical set is evaluated in binary terms—an element either belongs to the set (membership is1) or does not belong (membership is 0). Fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval [0, 1] [6]. Fuzzy set theory can be used in a wide range of domains especially for handling uncertain and imprecise data.

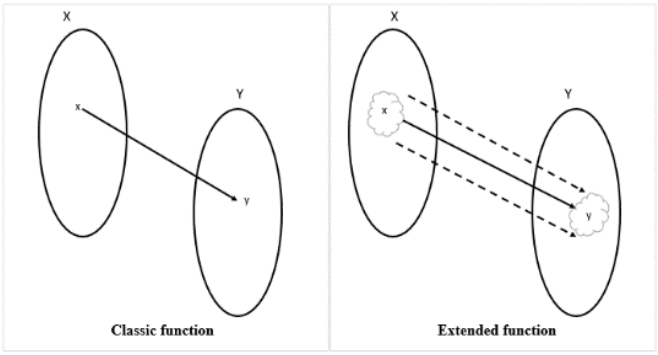

Fig. 1. The extension principle of Zadeh.

Cross Language Plagiarism might be further than we could expect, a fuzzy complex operation. So the use of fuzzy sets theory in CLPD can be modeled by considering that each word in a document is associated with a fuzzy set that contains words with same meaning, and there is a degree of similarity between words in a document and the fuzzy set [7].

The result of several research focuses on the importance of preprocessing on Part-Of-Speech (POS) level and its integration with fuzzy based methods for an efficient identification of similar documents [8].

As mentioned before, fuzzy semantic-based approach can be modeled by considering that words in a sentence (from two compared texts) have a fuzzy set that contains words with similar meaning with a degree of similarity (usually less than 1) (Yerra & Ng 2005) which could be considered as the application of the extension principle of Zadeh for fuzzy set (Fig. 1).

Fuzzification is one of the main components in Fuzzy inference systems. In the fuzzifier process, relationships between the inputs and linguistic variables are defined by a fuzzy membership functions, for this work to fuzzify the relationship of word pairs (from text pairs), we proposed Wu and Palmer (1994) semantic similarity metric as a fuzzy membership function.

IV. Wu and Palmer

Lexical and semantic-based features and similarity metrics have been widely used in plagiarism detection to assess the extent of similarity between two the texts. An important number of similarity measures have been proposed in the last few years, lch (Leacock and Chodorow, 1998), wup (Wu and Palmer, 1994), res (Resnik, 1995), lin (Lin, 1998), lesk (Banerjee and Pedersen, 2003), and hso (Hirst and St Onge, 1998) [12]-[15] metrics which we discussed and used for CLPD in [9].

In this paper, we used Wu & Palmer (1994) [11] which has been widely used (Lin et al., 1998; Lee, 2011; Alzahrani et al. 2015). WUP metric relates the depth of the words' synsets in the DAG taxonomy and the depth of their LCS (or the most specific ancestor).

Semantic Similarity refers to similarity between two concepts in a taxonomy such as the WordNet (Miller, 1995) [16], where lexes are arranged into groups called synsets

(synonyms sets), synsets that share a common property are linked with more general words called hypernyms, and most specific words called hyponyms. The proposed algorithm (Wu&Palmer) uses WordNet to automatically evaluate semantic relations between words, in WordNet, a word may have one to many synset, each corresponding to a different meaning.

The WUP measure calculates similarity by considering the depths of the two concepts in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer (LCS) (Fig. 2), the formula is [11] :

$$Score = 2 \times \frac{depth(LCS)}{depth(S1)+depth(S2)} \quad (1)$$

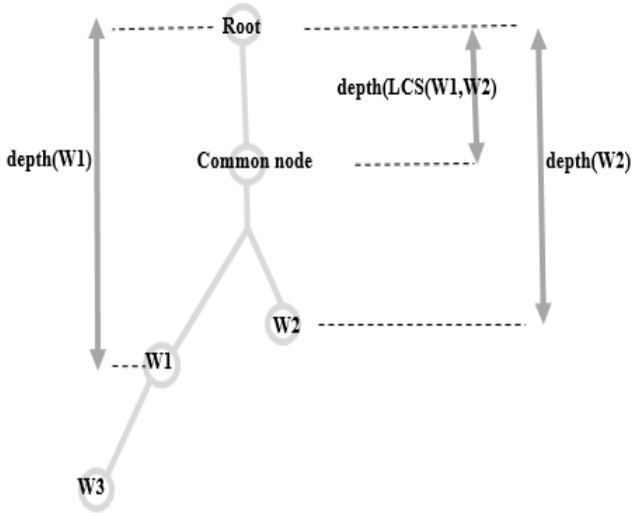

Fig. 2. Directed-Acyclic-Graph (DAG) for WordNet.

V. PREPROCESSING

The work presented in this paper treat intelligent multilingual plagiarism detection using Fuzzy-Semantic Similarity based methods. Input texts are from two different languages, the creation of a suitable target data of each document is elementary, and various preprocessing methods based on NLP techniques are implemented (Fig. 3).

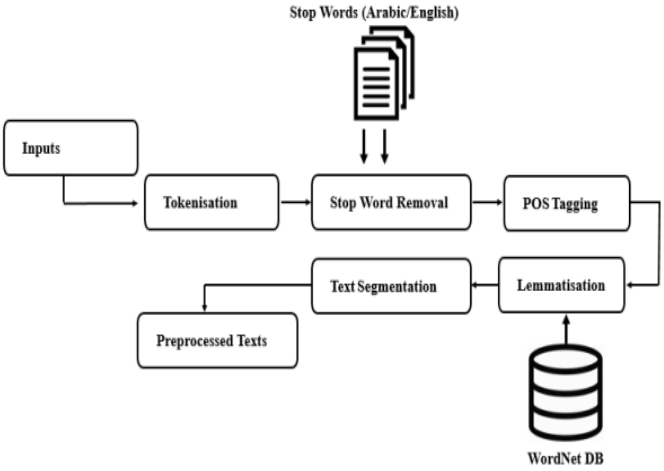

Fig. 3. Texts preprocessing for CLPD.

The important ones could be described as follows:

Tokenization – This is the process of segmenting running text into linguistic units, called tokens, such as words and sentences. A token is more than simply identifying strings delimited on both sides by spaces or punctuation. It is a linguistically significant word and methodologically useful Word tokenization is generally considered as easy relative to other tasks in natural language, especially in language that separates words by a special 'space' character. However, for our case subject and verb may be as one single word in Arabic language.

Stop words refer to the most common words in a language, usually does not contribute to the semantic meaning of the sentence, such as “a”, “an”, “the”, “is”, “are”, etc. for English and “في”, “و”, “إلى”, “من”, “ثم”, etc. for Arabic. Since all the work is semantic based, so to reduce computation time and avoid any meaningless comparison, stop words will be removed from the two documents. Stop words list contain the 173 most frequent words of the English and 104 word for Arabic language.

In the proposed system Part-of-speech disambiguation (or POS tagging) with Stanford CoreNLP is used [10]. POS tagging is the process of assigning a part-of speech marker to each word in an input text.

Lemmatization – This is to remove inflectional and derivationally related forms of a word to a common base form or dictionary form using vocabulary and morphological analysis (called lemma). The use of lemmatization and not stemming is based on two facts, first that a lemma is the base form of all its inflectional forms. However, the stem can be the same for the inflectional forms of different lemmas, providing then noise to our search results, results found also in (Alzahrani and Salim, 2010) [1] work, the second is WordNet is based on “lemmas” rather than “stems” which should facilitate finding the appropriate synset. The produced dictionary base forms (lemmas) are more appropriate for semantic comparisons of two sentences based on their (lemmatized) words derived from the WordNet.

Text segmentation – Text is segmented into word 3-grams (W3G) and sentences based on [5] authors compared several segmentation (word 3-grams (W3G), word 5-grams (W5G), word 8-grams, and sentences (S2S)) to see which approach can better handle intelligent plagiarism cases with the proposed fuzzy semantic-based similarity method, and concluded that W3G gives better results.

VI. PROPOSED METHOD

A fuzzy semantic-based approach can be modeled by considering that words in a sentence (from two compared texts) have a fuzzy set that contains words with similar meaning (approximate or vague) with a degree of similarity (usually less than 1) between words (in a sentence) and the fuzzy set (Yerra & Ng 2005). Word-to-word relationships can be based on different assumptions (Fig. 4).

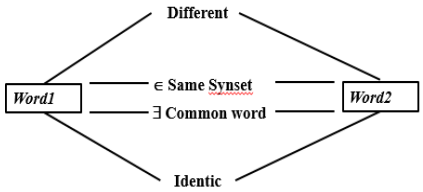

Fig. 4. Relationships between two words.

Various semantic similarity metrics of words have been proposed regarding their relationship in the WordNet [16] lexical database [9], based on (Alzahrani et al. 2015) [1] work and (Ezzikouri et al. 2016) [9]; Wu & Palmer gives interesting results. Therefore, to fuzzify the relationship of word pairs (from input texts), we used Wup measure as a fuzzy membership function, (1) will be expressed as follow:

$$\mu_{aibj} = Wup(a_i, b_j) \quad (2)$$

The fuzzy relationship between two words ranges between 1, for words that are identical or have the same meaning (i.e. synonyms), and 0 for words that are totally different (i.e., do not have any semantic relationship). A fuzzy inference system was constructed to evaluate the similarity of two texts and infer about plagiarism.

To evaluate the relationship of a word in one text with regard to words in the other text, we can use the fuzzy PROD operator as in the following formulas:

$$\mu_{a_1,B}=1-\prod_{b_j\in B, j\in[1,m]}(1-Wup(a_1,b_j))$$
$$\mu_{a_n,B}=1-\prod_{b_j\in B, j\in[1,m]}(1-Wup(a_n,b_j)) \quad (3)$$

Then we calculate the average sum:

$$\mu_{A,B} = (\sum_{i=1}^{n} \mu_{a_,B)/n} \quad (4)$$

The inputs are two texts from two different languages passed by various step before obtaining the result, as mentioned before in section (5) preprocessing is an important step and contain several NLP processes and W3G segmentation, so it is obvious that the first step of our system is the preprocessing step. The inputs texts passes to preprocessing with some differences view that Arabic is a rather difficult language to treat. The resulting texts are used as inputs to the fuzzy inference system, then Wu and Palmer semantic similarity measurement is modeled as a membership function. The output is a similarity score between input texts. This can be modeled and resumed in the algorithm below:

```
Algorithm: FCLPD
Inputs: Text A , Text B
Output: CPD(A,B)
BEGIN
Preprocessing for Text A
Preprocessing for Text B
For each segment Ai ∈A do
For each Segment Bj ∈B do
Input Ai and Bj to fuzzy inference system
Compute Wup(Ai,Bj)
If CPD(Ai,Bj) is true
Add (Ai,Bj) to Output
                End If
            End of loop For
            End of loop For
END
```

EXAMPLE

In this example, the second text is translated and reworded from the first one, but the meaning has remained almost the same. Texts Ar and En pass first by preprocessing and then to the fuzzy system.

" ويكيبيديا هو مشروع موسوعة عالمية متعددة اللغات على الانترنت، تهدف لتوفير محتوى يمكن إعادة استخدامه بحرية وموضوعية وقابل للتحقق، جميع محرري مقالات ويكيبيديا هم من المتطوعين، ويمكن للجميع تعديل وتحسين المحتوى."

"Wikipedia is an online, universal, multilingual and wiki-based encyclopedia project. Wikipedia aims to provide freely reusable, objective and verifiable content that everyone can modify and improve. All the editors of Wikipedia articles are volunteers."

An important task in any plagiarism detection/natural language processing application is the adaptation and the formatting of a suitable data to which plagiarism detection processes and algorithms could be applied. This is particularly important in this paper due to the characteristics of Arabic language and translation operation.

The purpose of preprocessing is to keep only the useful information for the PD analysis.

The two texts after the preprocessing process (tokenization, stop words removal, post-tagging…) are shown in Fig. 5 and 6.

```
Sentence #1 (11 tokens):
ويكيبيديا
مشروع
موسوعة
رقمية
متعددة
اللغات
حرة
المحتوى
[Text=ويكيبيديا PartOfSpeech=NN Lemma=ويكيبيديا]
[Text=مشروع  PartOfSpeech=NN Lemma=مشروع]
[Text=موسوعة  PartOfSpeech=NN Lemma=موسوعة]
[Text=رقمية  PartOfSpeech=NN Lemma=رقمية]
[Text=متعددة PartOfSpeech=JJ Lemma=متعددة]
[Text=اللغات PartOfSpeech=NN Lemma=لغة]
[Text=حرة  PartOfSpeech=NN Lemma=حرة]
[Text=المحتوى PartOfSpeech=NN Lemma=محتوى]
Sentence #2 (19 tokens):
يستطيع
شخص
```

Fig. 5. Preprocessed Arabic text.

```
Sentence #1 (11 tokens):
wikipedia
online
universal
multilingual
online
encyclopedia
project
[Text=wikipedia  PartOfSpeech=NN Lemma=wikipedia ]
[Text=online CharacterOffsetBegin=11 CharacterOffsetEnd=17 PartOfSpeech=NN
Lemma=online  ]
[Text=universal PartOfSpeech=JJ Lemma=universal  ]
[Text=multilingual  PartOfSpeech=JJ Lemma=multilingual  ]
[Text=encyclopedia  PartOfSpeech=NN Lemma=encyclopedia  ]
[Text=project  PartOfSpeech=NN Lemma=project  ]
Sentence #2 (13 tokens):
wikipedia
aims
provide
freely
reusable
objective
verifiable
content
everyone
modify
improve
[Text=aims CharacterOffsetBegin=101 CharacterOffsetEnd=105 PartOfSpeech=NNS
```

Fig. 6. Preprocessed English text.

The analysis of both texts means that every segment in text Ar will be compared with every segment in text En. It is clear that both texts are identic and segments of the first sentences are almost the same by a percentage of 88.13%, and the segment Ar2 and En2 are similar to a high degree of 64.86 %, also Ar4 is the same as En3 to a degree of 99.96%, same thing could be noticed for the last segments. If we compare the two whole texts, it will give a percentage of 76.75% semantic similarity, which is a high rate of plagiarism.

VII. CONCLUSION

Fuzzy-Semantic based automatic multilingual plagiarism detection is presented in this paper. Different pre-processing methods based on NLP techniques were used, principally lemmatization, stop word removal and POS tagging for both Arabic and English languages. Texts were segmented to Ngram segmentation (3G is the best for this case). Wu and Palmer similarity measure is used to evaluate the similarity in compared texts. It also shows how similarity calculation can be enhanced using fuzzy-semantic similarity measures. Future works will be extended using fuzzy-semantic based with other measures from our previous work [9].

REFERENCES

[1] Alzahrani, S., & Salim, N. (2010). Fuzzy semantic-based string similarity for extrinsic plagiarism detection. Braschler and Harman, 1-8.

[2] Osman, A. H., Salim, N., Kumar, Y. J., & Abuobieda, A. (2012, January). Fuzzy Semantic Plagiarism Detection. In AMLTA (pp. 543-553).

[3] Gupta, D., Vani, K., & Singh, C. K. (2014, September). Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 2694-2699). IEEE.

[4] Ahangarbahan, H., & Montazer, G. A. (2015, June). A Mixed Fuzzy Similarity Approach to Detect Plagiarism in Persian Texts. In International Work-Conference on Artificial Neural Networks (pp. 525-534). Springer, Cham.

[5] Alzahrani, S. M., Salim, N., & Palade, V. (2015). Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University-Computer and Information Sciences, 27(3), 248-268.

[6] Dubois, D., & Prade, H. (2000). General Introduction. In Fundamentals of Fuzzy Sets (pp. 1-18). Springer US.

[7] Yerra, R., & Ng, Y. K. (2005). A sentence-based copy detection approach for web documents. Fuzzy systems and knowledge discovery, 481-482.

[8] Gupta, D., Vani, K., & Singh, C. K. (2014, September). Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 2694-2699). IEEE.

[9] Ezzikouri, H., Erritali, M., & Oukessou, M. (2016). Semantic Similarity/Relatedness for Cross Language Plagiarism Detection. Indonesian Journal of Electrical Engineering and Computer Science, 1(2), 371-374.

[10] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The stanford corenlp natural language processing toolkit. In ACL (System Demonstrations) (pp. 55-60).

[11] Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

[12] Satanjeev Banerjee, Ted Pederson. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. 2002.

[13] Claudia Leacock, Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense

[14] Identification. WordNet: An Electronic Lexical Database, Publisher: MIT Press. 265-283.

[15] Dekang Lin. An Information-Theoretic Definition of Similarity. ICML. 1998: 296-304.

[16] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.