Using Deep Learning for Intelligent and Cross-Lingual Plagiarism Detection

Oumaima Hourrane and El Habib Benlahmar
Faculty of Sciences Ben Msik, Hassan II University of Casablanca

Abstract

The huge availability of textual data and documents via the Internet makes it easier and easier to regain ideas. Therefore, authors, rights holders, publishers, or even institutions issuing diplomas or certificates after validation of a textual rendering,must guarantee the originality of these contents. It is to meet this need that anti-plagiarism software has emerged. There are more and more of them being used in higher education institutions and other academic institutions. Thus, most of these tools are still insufficient for the detection of intelligent plagiarism, as the internet also facilitates access to plagiarism tools and back-translation tools.

Objectives:

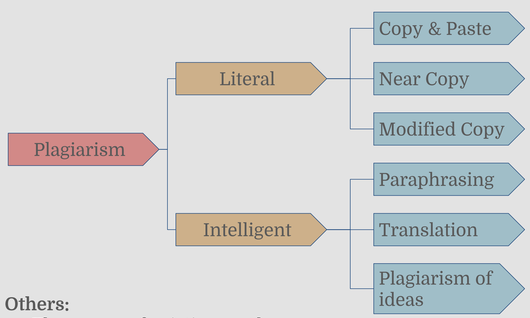
- ➔ Apply Deep Learning techniques which make it possible to achieve high performances on several Computational Linguistic tasks, to detect intelligent forms of plagiarism.
- ➔ Reduce academic dishonesty among students and the research community.
- ➔ Fortify the quality of pedagogical and scientific productions in Moroccan universities, then in the world in general
- ➔ Guarantee the value of the degrees they deliver..

Introduction

What is Plagiarism?

- Plagiarism is the act of taking or using the works of another author such as his own, without references or citations, either wholly or in part. It can include "copy & paste" directly, modify or change certain words of the original text. In another perspective.
- "... A document is said to be plagiarized when it is obtained by applying a series of transformations on an original document. The plagiarized document must retain the same function as the original but may have a different shape. One can suspect a duty to be plagiarized when a reasonably small number of transformations has been applied from another document in the corpus..." [1]

Types of Plagiarism



Others:

- The re-use of existing works.
- Purchase of schoolwork.
- Buying assignments.
- Accidental Plagiarism...

Literature reviews

1. Character-based approaches

Those methods are the most used for plagiarism detection, that rely specifically on character-based lexical features, n-grams based lexical features, syntax features, and fingerprinting features, to compare the query document with each candidate document.

N-grams	- The use of Scatter Plot - The use of Stop words
Fingerprinting	- Winnowing algorithm - Hash-breaking algorithm - DTC algorithm - Simhash technique
Clustering-based	- Winnowing fingerprint extraction algorithm and the clustering technique.

Table.1. Character-based methods

2. Style-based approaches

Based on stylometric features, this method can be constructed to quantify the characteristics of the writing style, of a document, aiming at recognizing the style of a particular author.

- Stein & al. [2] are interested in the analysis undeclared changes in the writing style of a document. Starting with constructing segments to operationalize the verification of the author. Then, analyzing the Style variance of these sections, at the end, Style measurements and quantization functions are used to analyze the variance of the different style characteristics...
- Oberreuter & al. [3] have explored a model for the quantification of writing style, by exploring words as linguistic traits using Text-Mining.
- Rexha & al. [4] suggests the use of existing PDF extraction techniques, by adapting a text segmentation algorithm TextTilling with Stylometrics features.

3. Semantic-based approaches

Those methods take in consideration that two sentences can be semantically the same but differ in their structure.

Latent Semantic analysis LSA	Examining the similarity between the contexts in which a word appears
Semantic Role Labeling SRL	Transforming the sentences into groups of arguments based on the location of each term in the sentences.
Fuzzy semantic-based	Calculating the degree of fuzzy similarity between two documents, that lies between two edges 0 and 1.
Vector-based	Applying the vector regression SVR to distinguish the semantic similarity score from the given sentence pairs.

Table.2. Semantic-based methods

4. Citation-based approaches

This approach was introduced by Gipp et al. [5]. Based on a sequential pattern analysis the authors compare different algorithms, such as the longest common citation sequence of two documents or citation chunking.
CbPD algorithms : Greedy Citation Tiling, Longitudinal Citation Sequence, and Citation Chunking.

Methodology

General framework

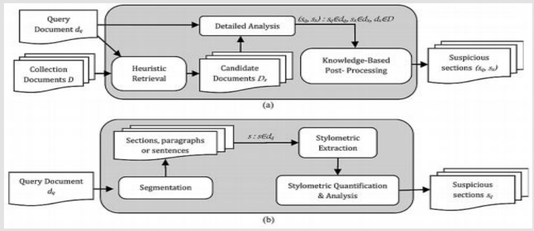


Fig. 1. Framework for different plagiarism detection systems. (a) White-box design for extrinsic plagiarism detection system. (b) White-box design for intrinsic plagiarism detection system.

Our Previous work

we have previously [6] proposed a method that captures similarity between citations on a corpus of scientific papers, by using the TF-IDF weighted average word embeddings. the proposed technique considers the effect of word embeddings on sentence meaning. To assess our similarity calculation, we take a huge dataset of NeurIPS papers, which contains an a huge number of citations sets and an a large number of words from an variety of articles in Neural Network subject.

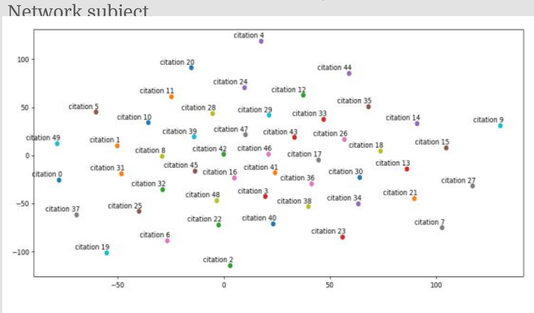


Fig. 2. Citation embeddings visualization with t-SNE.

Our method, show the significant result. However, this kind of sentence embedding is hard to capture full semantics since the context of a sentence is limited. Furthermore, this method can only account for a very small part of the sentence, since most of the sentences are compositional. In contrast, our method attempts to learn the semantic vector representation for any sentence.

Our current work

- ➔ Step 1: sentence embeddings: that captures semantic similarities even for multilingual sentences and apply this in plagiarism detection.
- ➔ Step 2: Making our approach more specific to detect intelligent forms of plagiarism, thereby leveraging syntactic-based methods to extract task-specific textual information.
- ➔ Step 3: Finding relations and similarities between documents and passages; Detecting either intrinsic plagiarism, which means observing changes in writing style in a given text, or detecting extrinsic plagiarism by comparing the suspect documents and their representations with other sources in a given corpus.

Conclusion

The need for anti-plagiarism systems becomes very important in proportion to the increase in the phenomenon of plagiarism, of which several research studies have been approached to solve this problem by adopting several techniques of plagiarism detection, However, there is no approach capable of detecting 100% plagiarism, each with advantages and limitations, depending on their characteristics and performance.

Since recent advances of Deep Learning techniques which make it possible to achieve high performances on several Computational Linguistic tasks. The aim of our work is to apply these techniques to detect intelligent forms of plagiarism. We will focus on the plagiarism that takes ideas from the source text, precisely, by paraphrasing of back-translating the original text.

Therefore, our final solution will reduce academic dishonesty among students and the research community, which can lead in the first place to fortify the quality of pedagogical and scientific productions in Moroccan universities, then in the world in general, and to guarantee the value of the degrees they deliver.

References

1- Brixtel, R., Lesner, B., Bagan, G., & Bazin, C. (2009, November). De la mesure de similarité de codes sources vers la détection de plagiat : le" Pomp-O-Mètre".

2- Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources and Evaluation.

3- Oberreuter, G., & VeláSquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. Expert Systems with Applications, 40(9), 3756-3763.

4- Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2015). Towards Authorship Attribution for Bibliometrics using Style Features. In Proc. of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI).

5- Gipp, B., & Meuschke, N. (2011). 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.

6- Hourrane, O., Mifrah, S., Bouhriz, N., Rachdi, M. (2018, April). Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers. In International Conference on Big Data, Cloud and Applications (pp. 185-196). Springer, Cham.

Contact Info

Email : oumaima.hourrane@gmail.com
Website: https://oumaimahourrane.com/
Linkedin: https://www.linkedin.com/in/oumaima-hourrane/