Intrinsic Plagiarism Detection in Arabic Text: Preliminary Experiments

Imene Bensalem[1], Paolo Rosso[2], Salim Chikhi[1]

[1] MISC Lab., Mentouri University, Constantine, Algeria
bens.imene@gmail.com, chikhi@misc-umc.org
[2] Natural Language Engineering Lab. – EliRF, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. This paper presents a preliminary study on intrinsic plagiarism detection in Arabic textual documents. A set of experiments were conducted to gain an insight into the effect of some well-known language-independent stylistic features on Arabic text discrimination. We used Stylysis tool to measure these features on our small-sized corpus.

Keywords: Intrinsic plagiarism detection, Arabic text discrimination, Stylistic features

1 Introduction and Related Work

Automatic plagiarism detection (APD) is a classical problem which has been tackled with a range of researches in the last decades. Nowadays, this task becomes more challenging because the widespread use of electronic documents and the advent of Internet facilitate access to knowledge and make easier its re-use.

Plagiarism detection approaches fall into two main categories: Intrinsic and External. Intrinsic APD methods uncover changes in the writing style, and use them as an evidence of a plagiarism act. In contrast, external APD methods do not rely on internal evidence, but rather they are based on detecting similarities between the suspicious document and other documents from a reference corpus [1].

Although many researches were conducted on plagiarism detection in the last years, those concerning the Arabic language text remain quite limited. To the best of our knowledge, the only works in this area are those of Alzahrani et al. [2, 3], Menai et al. [4] and Jaoua et al. [5]. All of them addressed the external approach. However, intrinsic approach has not been yet applied to Arabic text.

This paper addresses Arabic intrinsic plagiarism detection. Our work consists in testing whether some language independent stylistic features are effective or not to discriminate between plagiarized and not plagiarized sentences.

The rest of the paper is organized as follows. Section 2 describes the process of building and pre-processing the evaluation corpus. Sections 3 and 4 describe experiments and discuss their results. Section 5 contains concluding remarks.

2 Corpus Building and Pre-processing

Since we are the first to deal with Arabic intrinsic plagiarism detection, there are no evaluation corpora for this respect. Thus, we built manually a small corpus of 10 documents with different sizes and topics. Some statistics about this corpus are provided in Table 1.

Table 1. Corpus statistics.

Number of documents	10
Number of all sentences	336
Number of plagiarized sentences	63
Number of words	13309
Number of tokens	6765

We tried to simulate real plagiarism in terms of inserting in each document sentences in relation with its topic. However, we did not obfuscate them (with paraphrasing for example) since the purpose of our study is not to detect similarity between plagiarized sentences and their sources, but rather, detect abnormal sentences (outliers) with regard to the document general writing style.

After the building step, the corpus was pre-processed with the intention to prepare it for the stylistic analysis using Stylysis [6]. Two tasks were performed in this regard.

First, the corpus was transliterated with the Buckwalter scheme[1] [7], because Stylysis do not support the Arabic language. Second, each document was split into sentences considering full stop as the separation character. Stylysis considers lines as the text segments on which features are computed regardless their lengths and the punctuation they contain. For example, if a line (all words before newline character) contains a paragraph, features such as sentence length and average word length represent the paragraph length and the average length of its words. Hence, each line in our documents contains exactly one sentence.

3 Experiments 1: Insight into some stylistic features

3.1 Description

As we have mentioned above, we used Stylysis to analyse the writing style of each document of the corpus. Stylysis computes six lexical features for each sentence, namely Gunning Fog Index, sentence length, average word length, Honore's R function, Yule's K function and Flesch-Kincaid readability test. Its results are

[1] The Buckwalter transliteration replaces each Arabic letter by a Latin letter or a symbol. For example the transliteration of the Arabic word: انتحال (Plagiarism) is: AntHAl.

displayed as graphs with upper and lower boundaries. Sentences over and under these boundaries –respectively– are considered outliers of the general writing style in terms of the considered feature (discriminator).

We recorded the state (outlier or not) of each sentence from the graphs of four discriminators: word average length, sentence average length, and R and K functions. We did not take the results of the two remaining features because their calculation is based on the number of syllables which could not be computed faithfully from a transliterated Arabic text without diacritics since those latter represent vowels[2].

Therefore, each sentence is recorded as a vector of four Boolean values which represent the sentence states with regard to the four considered discriminators. A sentence is considered plagiarized by a discriminator if it is an outlier.

3.2 Results and Discussion

Two measures were used to evaluate the performance of discriminators: Precision (equation 1) and Recall (equation 2). Results are shown in Table 2.

$$Precision = \frac{|\text{ sentences detected as plagiarized } \cap \text{ actual plagiarized sentences }|}{|\text{ sentences detected as plagiarised }|} \quad (1)$$

$$Recall = \frac{|\text{sentences detected as plagiarized } \cap \text{ actual plagiarized sentences }|}{|\text{ actual plagiarized sentences}|} \quad (2)$$

Table 2. Performance evaluation.

Discriminator	*Precision*	*Recall*
Average word length	16.0%	19.0%
Average sentence length	19.4%	19.0%
R function	20.5%	**41.3%**
K function	**24.4%**	34.9%

As can be seen from these results, average word length (W) has the lower performance; therefore it is an unreliable stylistic discriminator of Arabic text. This is consistent with Abbasi et al.'s research [8] on authorship analysis[3]. These authors attributed the unreliability of W to the small range over which the lengths of Arabic words are distributed.

In contrast to its performance with English text [9], average sentence length (S) seems to be a poor stylistic discriminator of Arabic text as shows our results based on our small corpus. We believe this result is still a hypothesis that should be verified on a large dataset.

---

[2] For example the word كتب (write) is transliterated as ktb; if it is with diacritics it will be كَتَبَ and its transliteration will be kataba.

[3] Stylistic analysis is the core task of the authorship analysis which makes the latter a very close domain to intrinsic plagiarism detection.

R and K functions which are used to measure vocabulary richness are known as not robust stylistic features with short English text [9]. Surprisingly, our results claim that they are relatively the most prominent detectors in terms of recall despite the small length of our corpus sentences.

It should be noted that the average sentence length and R and K functions have already been used among other features within a machine learning method to detect anomalous Arabic text in a mid-sized corpus [10], but unfortunately, their effectiveness was not tested separately.

4 Experiment 2: Combining Discriminators

In this section we propose to combine the considered discriminators using voting schemes.

4.1 Description

In order to have a basis of comparison, we introduced a lenient baseline method which considers a sentence as plagiarized if it is detected as an outlier, at least, by one of the four considered discriminators. The baseline method's recall is the best we can obtain from combining these four lexical discriminators. Moreover, we experimented using other voting schemes in the aim to raise the baseline precision and keep the recall relatively high (over 45%).

Beside the baseline results, Table 3 shows the two best combinations in terms of Precision. Voting schemes used to combine discriminators are presented as logical expressions.

Table 3. Combination's results: baseline vs. the most precise voting schemes

Discriminator	*Precision*	*Recall*
Baseline: Or(S,W,R,K)	21.5 %	**73.0%**
And(Or(S,R,K),Not(W))	24.5%	54.0%
And(Or(R,K),Not(W))	**24.8%**	46.0%

4.2 Discussion

It is clear that both of the experimental schemes share the pattern "Not(W)" which means sentences detected by average word length are discarded even though they are positive with the remaining discriminators. Since W is the most imprecise detector, many false positive sentences was eliminated by using it as a filter which yielded an increase of 3% in precision.

An examination of W outliers, led us to the two following remarks. First, sentences with a great average word length are mainly: sentences with diacritics (generally Quran verses), and sentences that contains foreign words. Second, most outliers under the average word length are long sentences composed of several

phrases linked by the conjunction “And”. This latter is in Arabic a one-letter word “و” which leads to a decrease in the sentence average word length. Therefore, we can conclude that the average word length is not a feature of discrimination between different writing styles of Arabic text, and then it is not effective as a plagiarism detector. Nonetheless, it may be useful to filter false positive sentences with the characteristics mentioned above.

5 Conclusion

We presented in this paper a set of preliminary experiments on intrinsic plagiarism detection in Arabic text using Stylysis tool. The conclusion that could be drawn from these experiments is: average word length and average sentence length are not reliable stylistic discriminator of Arabic text. Whereas, the vocabulary richness measures R and K might be prominent ones. These experiments were carried out using a small corpus; hence the need to build large corpora to be able to confirm the finding above and conduct further studies.

Acknowledgements. This work is the result of the collaboration in the framework of the bilateral research project AECID-PCI AP/043848/11 (Application of Natural Language Processing to the Need of the University) between the Universitat Politècnica de València in Spain and the Université Mentouri Constantine in Algeria. The research work of Paolo Rosso was done also in the framework of the European Commission WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People, the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03(Plan I+D+i), and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

1. Potthast, M., Eiselt, A., Barrón-cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Petras, V., Forner, P., and Clough, P. (eds.) Notebook Papers of CLEF 2011 LABs and Workshops, CLEF-2011. , Amsterdam (2011).
2. Alzahrani, S., Salim, N.: On the Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents. In: 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT’09). pp. 539-544. IEEE, London (2009).
3. Alzahrani, S., Salim, N.: Statement-Based Fuzzy-Set Information Retrieval versus Fingerprints Matching for Plagiarism Detection in Arabic Documents. In: 5th Postgraduate Annual Research Seminar (PARS’09). pp. 267-268. , Johor Bahru (2009).
4. Menai, M.E., Bagais, M.: APlag: a Plagiarism Checker for Arabic Texts. In: Proceedings of the 6th IEEE International Conference on Computer Science and Education (ICCSE’11). pp. 1379-1383. IEEE, Singapore (2011).
5. Jaoua, M., Jaoua, F.K., Hadrich Belguith, L., Ben Hamadou, A.: Automatic Detection of Plagiarism in Arabic Documents Based on Lexical Chains (in Arabic). Arab Computer Society Journal. 4, 1-11 (2011).

6. Barrón-Cedeño, A., Vallés-Balaguer, E., Rosso, P.: Stylysis, http://memex2.dsic.upv.es:8080/StylisticAnalysis/en/index.jsp.
7. Buckwalter, T.: Arabic Buckwalter Transliteration, http://www.qamus.org/transliteration.htm.
8. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist- Group Web Forum Messages. IEEE Intelligent Systems. 20, 67-75 (2005).
9. Meyer zu Eissen, S., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Decker, R. and Lenz, H.-J. (eds.) Advances in data analysis. pp. 359-366. Springer Berlin Heidelberg (2007).
10. Abouzakhar, N., Allison, B., Guthrie, L.: Unsupervised Learning-based Anomalous Arabic Text Detection. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D. (eds.) Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08). pp. 291-296. ELRA, Marrakech (2008).