Bensalem, I.; Rosso, P.; Chikhi, S. (2013). A new corpus for the evaluation of arabic intrinsic plagiarism detection. En Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer Verlag (Germany). 53-58. doi:10.1007/978-3-642-40802-1_6.

A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

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

[1]MISC Lab., Constantine 2 University, 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. The present paper introduces the first corpus for the evaluation of Arabic intrinsic plagiarism detection. The corpus consists of 1024 artificial suspicious documents in which 2833 plagiarism cases have been inserted automatically from source documents.

Keywords: Arabic intrinsic plagiarism detection, evaluation corpus, automatic plagiarism generation

1 Introduction

"Plagiarism occurs when someone presents the work of others (data, text, or theories) as if they were his/her own and without proper acknowledgment" [1]. One may uncover plagiarism in a text document by observing similarities between it and other documents (external plagiarism detection), or by noticing a sort of heterogeneity in the writing style (intrinsic plagiarism detection) [2]. Automatic methods of plagiarism detection are inspired by these two traditional approaches. In the external approach, it is necessary to hold a collection of documents representing the source of plagiarism; whereas, in the intrinsic approach, there is no need for source documents. Indeed, the importance of this approach emerges when the plagiarism source is unknown or does not have a digital version. In this paper, we are interested in the intrinsic plagiarism detection in Arabic texts. Concretely, we will describe the first corpus for the evaluation of Arabic intrinsic plagiarism detection. The remainder of the paper is structured as follows: Sections 2 and 3 provide a brief overview of the intrinsic plagiarism detection in English and Arabic languages respectively. In this overview we focus on the evaluation aspect. Section 4 presents the methodology adopted in the construction of our corpus and provides statistics on it. Finally, Section 5 concludes the paper.

2 Intrinsic Plagiarism Detection in English Text

In the last years, a great effort has been made to standardize the evaluation of the automatic plagiarism detection with its external and intrinsic approaches. As a result,

an evaluation framework has been developed. It consists in a set of quality measures and a series of evaluation corpora involving automatically created suspicious documents [3]. This evaluation framework was used in the plagiarism detection task of PAN competition[1] from 2009 to 2011 [2] [4] [5]. The part of PAN 2011 corpus, used to evaluate the intrinsic approach, contains 4753 suspicious documents with 11443 plagiarism cases.

In PAN 2012, another evaluation framework has been introduced [6]. Unlike the previous corpora, all the suspicious documents of PAN 2012 corpus were created manually through crowdsourcing. This new corpus was used to evaluate only the external approach, while the intrinsic one has been considered as an authorship clustering problem and therefore, has been evaluated within PAN authorship attribution task using another evaluation corpus, which is very small in comparison with the former (less than 10 documents) [7].

3 Intrinsic Plagiarism Detection in Arabic Text

Although the broad spread of plagiarism in the Arab world [8], plagiarism detection in the Arabic text is still in its infancy, especially when it concerns the intrinsic approach. We think that the main reason behind this fact is the lack of an evaluation corpus. Moreover, there are very few works on Arabic authorship analysis [9–11] which is one of the most related disciplines to intrinsic plagiarism detection. To the best of our knowledge, the only work in this area is ours [12] where we used a toy corpus composed of 10 documents with 63 plagiarism cases.

With regard to the external approach, some detection methods were proposed in the last few years. Nonetheless, it is difficult to draw a clear conclusion on the performance of these methods since they were evaluated, using different strategies and corpora. Jadalla and Elnagar [13] compared their web-based system with a baseline method using a number of documents that have been presumed to be suspicious. Alzahrany and Salim [14] as well as Menai [15] evaluated their methods using respectively 15 and 300 suspicious documents constructed by rewording and restructuring sentences. Jaoua et al. [16] created 76 suspicious documents by the manual insertion of text fragments obtained by queries to search engine, using keywords in relation with the subject of the document that will host the plagiarism.

The next section describes the building of the first Arabic corpus for intrinsic plagiarism detection evaluation. We think that the creation of such a corpus will encourage researchers to investigate this unexplored area.

4 Methodology

A corpus of plagiarism detection evaluation should be composed of two collections of documents: suspicious documents and source documents. A suspicious document

[1] http://pan.webis.de/

contains fragments of texts plagiarized from one or more source documents. These latter are omitted from the corpus if the evaluation concerns the intrinsic approach.

Due to the difficulty (for ethical and feasibility reasons) of owning a document collection containing actual plagiarism cases, suspicious documents have to be built. Two approaches have been used in the state-of-the-art researches: manual and automatic. The manual approach [17] is the more realistic in terms of simulating the real plagiarist behaviour. It consists in charging people to write essays on designated topics with allowing the text reuse from different references. However, the automatic approach [3] follows two steps: (1) Compilation of target and source documents. Documents of both collections must be tagged with their author names and topics to prepare them for the second step; (2) Insertion of plagiarism: this task tries to simulate the act of plagiarism by borrowing automatically text segments from source documents and inserting them randomly in a target document. The target document and their sources of plagiarism must have the same topic but different authors.

Although the automatic approach is less realistic and suffers from many shortcomings [6], we adopted it to build our corpus for two main reasons. First, the automatic approach is acceptable since it has been used to build PAN 2009-2011 corpora. Second, the manual approach is costly in terms of human and material resources [17]. The following subsections provide details on the steps of our corpus construction which are text compilation and plagiarism insertion.

4.1 Text Compilation

Criteria of Texts: We set a number of criteria that should be verified in the target documents (documents where plagiarised fragments will be inserted).

C1. Each target document must be written by one author only. Otherwise, the document will contain many writing styles which may complicate the intrinsic plagiarism detection even further.

C2. Target documents should not include much of text reuse or many quotations. In fact, this is a feature of Arabic religion books which include many quotations from Holy texts. The purpose of this criterion is to avoid altering the evaluation by texts that are likely to be detected as plagiarism cases, although they are actually legitimate cases of text reuse.

C3. Target documents should not be too short. Indeed, we presume that the stylistic analysis becomes unreliable with short Arabic texts as it is with short English text (less than half a page approx.) [18].

C4. Texts should be punctuated because they will undergo a style analysis where the punctuation is an important feature. This criterion seemed obvious, but we decided to mention it because in a late stage of the text compilation, we noticed a lack of quality of some Arabic online texts. Effectively, we discarded many of the collected docu-

ments because they were poorly edited in terms of punctuation as well as section separations (no new line character between sections)[2].

Source of Text. Since we plan to make the corpus publicly available, it was primordial to gather texts from a copyright-free source. For this reason along with the specific desired criteria, sources of text have become very limited. We finally decided to build our corpus from Arabic Wikisource which is a library of heritage books and public domain texts. Furthermore, most of its documents are tagged with topics and author names (see our paper [19] for further details on the text compilation from Wikisource). We also added some texts from other sources, after making sure that they are without copyright. Table 1 presents the sources of our document collection.

Table 1. Our corpus sources of text.

Source of text	Percentage of documents in the corpus
Arabic Wikisource[3]	98%
Create your own country blog [4] KSUCCA corpus[5] Islamic book web site[6]	2%

4.2 Insertion of Plagiarism

Inspired by the PAN 2009-2011 corpora methodology, the suspicious documents were created automatically according to two parameters: the percentage of plagiarism per document and the length of plagiarism fragments. The main steps of the plagiarism insertion are:

1. Indexing source documents as fragments of different lengths to be used as plagiarism cases.
2. Selection of plagiarism sources for each target document according to its topic and its author name.
3. Random selection of segments from the source documents indices and their insertion in a random position in the target document.
4. Annotation of the plagiarism cases in an XML document following PAN corpora scheme.

To generate the suspicious documents with a variety of the plagiarism percentage and the case lengths, we split the target documents into 6 sets according to the document

[2] It is particularly the case of old books which represent an important part of the copyright-free text available online.
[3] http://ar.wikisource.org
[4] http://diycountry.blogspot.com
[5] Al-Rabiah, M.: King Saud University Corpus of Classical Arabic (KSUCCA), http://ksucorpus.ksu.edu.sa (2012).
[6] http://www.islamicbook.ws

lengths. Each set was divided arbitrary into 4 equal subsets. Finally, plagiarism was inserted in each subset with a fixed percentage limit and a list of plagiarism case lengths. Statistics on the obtained corpus are provided in Table 2.

Table 2. Statistics on the Arabic intrinsic plagiarism detection corpus

Document statistics			
Total number of documents	1024		
Plagiarism percentage per document		**Document length**	
Null (0%)	20%	Very Short (1-3 pages)	46%
Hardly]0% 10%]	24%	Short (3-15 pages)	37%
Few]10% 30%]	32%	Medium (15-100 pages)	12%
Medium]30% 60%]	24%	Long (>100 pages)	05%
Plagiarism cases statistics			
Total number of plagiarism cases	2833		
Plagiarism cases length		**Number of plagiarism cases per document**	
Very short (some sentences)	09%	Null (0)	20%
Short (some paragraphs)	40%	Few]0 5]	69%
Medium (around 1 page)	21%	Medium]5 15]	08%
Long (many pages)	30%	Much]15 45]	03%

5 Conclusion

In this paper we described the first evaluation corpus for Arabic intrinsic plagiarism detection. The corpus was built automatically and it follows standards in the annotation of plagiarism cases. The main difficulty we encountered during the construction of this corpus is the lack of good quality copyright-free Arabic text. This fact has limited the text sources of our corpus. We believe that the release of such a free corpus will foster research in intrinsic plagiarism detection in Arabic.

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 Constantine 2 University in Algeria.

References

1. Springer Policy on Publishing Integrity. Guidelines for Journal Editors.
2. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E. (eds.) SEPLN 2009Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9 (2009).

3. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Huang, C.-R. and Jurafsky, D. (eds.) Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10). pp. 997–1005. ACL, (2010).
4. Potthast, M., Barrón-cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler, M. and Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops (2010).
5. 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 (2011).
6. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Forner, P., Karlgren, J., and Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop –Working Notes Papers (2012).
7. Juola, P.: An Overview of the Traditional Authorship Attribution Subtask Notebook for PAN at CLEF 2012. In: Forner, P., Karlgren, J., and Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs andWorkshop –Working Notes Papers (2012).
8. Yakout, M.M.: Examples of Plagiarism in Scientific and Cultural Communities (in Arabic), http://www.yaqout.net/ba7s_4.html.
9. Abbasi, A., Chen, H.: Applying Authorship Analysis to Arabic Web Content. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., and Merkle, R.C. (eds.) ISI 2005. LNCS, vol. 3495. pp. 183–197. Springer, Heidelberg (2005).
10. Shaker, K., Corne, D.: Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In: 2010 UK Workshop on Computational Intelligence (UKCI). pp. 1–6. IEEE (2010).
11. Ouamour, S., Sayoud, H.: Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier. In: 2012 International Conference on Communications and Information Technology (ICCIT). pp. 44–47. IEEE (2012).
12. Bensalem, I., Rosso, P., Chikhi, S.: Intrinsic Plagiarism Detection in Arabic Text : Preliminary Experiments. In: Berlanga, R. and Rosso, P. (eds.) 2nd Spanish Conference on Information Retrieval (CERI 2012). , Valencia (2012).
13. Jadalla, A., Elnagar, A.: A Plagiarism Detection System for Arabic Text-Based Documents. In: Chau, M., Wang, G.A., Yue, W.T., and Chen, H. (eds.) PAISI 2012. LNCS vol. 7299. pp. 145–153. Springer, Heidelberg (2012).
14. 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), Johor Bahru, Malaysia. pp. 267–268 (2009).
15. Menai, M.E.B.: Detection of Plagiarism in Arabic Documents. International Journal of Information Technology and Computer Science. 10, 80–89 (2012).
16. 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).
17. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: 51st Annual Meeting of the Association of Computational Linguistics (ACL 13) (to appear). ACM (2013).
18. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. Language Resources and Evaluation. 45, 63–82 (2010).
19. Bensalem, I., Rosso, P., Chikhi, S.: Building Arabic Corpora from Wikisource. In: 10th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'13). IEEE (2013).