HAT - A new corpus for experimental stylometric evaluation in Arabic

Halim Sayoud, Siham Ouamour
EDT, FEI, USTHB University, Algeria

https://doi.org/10.36505/ExLing-2021/12/0052/000525

Abstract

Stylometry is a research field of NLP dealing with the identification of the actual author of a piece of text. Even though it does exist several datasets for some occidental languages, it is quite difficult to find an adequate corpus in Arabic. This problem motivated us to build a natural Arabic corpus dedicated to the task of experimental stylometric evaluation. Our corpus is composed of 100 groups of Arabic texts that are extracted from different Arabic books, where the main topic is "Travel". The books are written by 100 different authors and each group contains 3 different texts that are written by the same author. We called this corpus "HAT" (Hundred of Arabic Travelers). Furthermore, we propose it as a free corpus that may represent a reference dataset for author style analysis in Arabic, which could be used for a purpose of experimental evaluation. To evaluate our dataset, we conducted some baseline experiments using an SVM classifier with a Leave-One-Out cross validation technique.

Keywords: natural language processing, computational linguistics, author style analysis; stylometry, digital libraries

Introduction

As per definition, the task of author recognition can be divided into several fields that are:

- Authorship attribution (AA) or identification: it consists in identifying the author(s) of a set of different texts;
- Authorship verification: in this case, the main task is checking whether a piece of text is written or not by an author who claimed to be the writer;
- Authorship discrimination: it consists in checking if two different texts are written by a same author or not (Sayoud 2012);
- Plagiarism detection: in this research field we look for the sentences or paragraphs that are taken from another author (Küppers 2012);
- Text indexing and segmentation: the main goal is to segment the global text into homogeneous segments (each segment or paragraph contains the contribution of only one author) by giving the name of the appropriate author in each text segment (paragraph) (Forest 2006).

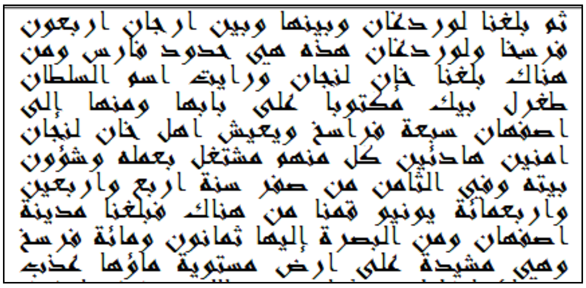
In this paper, a new text dataset is proposed to the scientific community for the experiments of authorship attribution in Arabic.

Although several works and datasets were cited for the English (Küppers 2012) (Juola 2006) and Greek (Tambouratzis 2003) (Tambouratzis 2004) languages, the authors did not found a lot of corpora in Arabic.

The proposed corpus contains several texts written by 100 Arabic authors on the topic of travel. The authors wrote several documents describing their travels. So a specific Arabic corpus/baseline has been built for a purpose of comparative authorship attribution.

HAT corpus

Our textual corpus is composed of 100 groups of Arabic texts that are extracted from 100 different Arabic books. The books are written by 100 different authors and each group contains 3 different texts that are written by the same author, which means that each group belongs to only one author. This set of 300 text documents has been collected in 2019 from "Alwaraq digital library"; we caled it HAT corpus (i.e. Hundred of Arabic Travelers). Furthermore, this corpus could represent a reference dataset for author style analysis in Arabic, which could be used by researchers in this field for a purpose of comparative evaluation. For concreteness, here is a piece of text belonging to Author #92 (figure 1).



ثم بلغنا لوردغان وبينها وبين ارجان اربعون
فرسخا ولوردغان هذه هي حدود فارس ومن
هناك بلغنا خان لنجان ورايت اسم السلطان
طغرل بيك مكتوبا على بابها ومنها إلى
اصفهان سبعة فراسخ ويعيش اهل خان لنجان
امنين هادئين كل منهم مشتغل بعمله وشؤون
بيته وفي الثامن من صفر سنة اربع واربعين
واربعمائة يونيو قمنا من هناك فبلغنا مدينة
اصفهان ومن البصرة إليها ثمانون ومائة فرسخ
وهي مشيدة على ارض مستوية ماؤها عذب

Figure 2. Example of Arabic text belonging to Author #92 (*N. Khasru*).

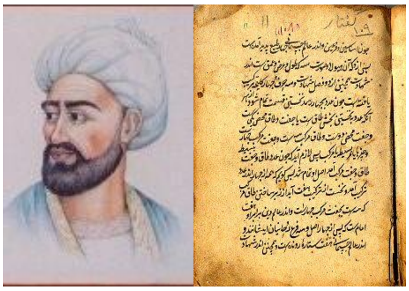


Figure 2. Ancient portrait and sheet of paper containing a text of Author #92.

The texts have a medium/short size: the average text length is about 1100 words per document and there are 3 documents per author, which corresponds to 300 documents in the total corpus. This situation involves severe experimental conditions, since it has been shown in previous research works (Eder 2010) that the minimum number of words per text should be at least 2500 words to get good attribution performances. In this investigation, the use of relatively short texts is interesting in order to evaluate the different classifiers with small documents in Arabic. In fact, when short texts are used, the AA performances decrease and it becomes difficult to make an efficient identification.

Experiments

In order to set a baseline evaluation for our new corpus, some experiments of author style analysis have been conducted by using an SMO-SVM classifier and by employing character-N-grams as features. Furthermore, the different experiments are made using a Leave-One-Out (LOO) cross-validation for a purpose of significance and consistency. The results of this baseline evaluation are summarized in table 1.

As one can see in table 1, the authorship attribution accuracy tends to increase by increasing the size of the N-gram: hence, the best accuracy was obtained by characters 4-grams (0.95), which is higher than the accuracy of 3-grams (0.81). This last one is almost similar to the accuracy of characters 2-grams (0.83), which is again higher than the accuracy of character 1-gram (0.71).

Table 1. Baseline results of authorship attribution on the HAT corpus.

Feature	LOO Accuracy
Characters	0.71
Char bi-grams	0.83
Char tri-grams	0.81
Char tetra-grams	0.95

In particular, with character tetra-grams, the performances revealed by the LOO validation show that 95% of the texts are well attributed, which represents a good score since we deal with a large number of authors: 100 authors.

Discussion

In this research work, we proposed a new corpus for authorship attribution, where we conducted a baseline experimental evaluation.

During this evaluation, we noticed that the best accuracy was obtained by character-tetra-grams. Right now, the first results appear interesting even though the investigation is not completely finished yet.

On the other hand, the following problems were noticed:

- The text documents are not sufficiently long to get a fair classification, as reported by previous works of Eder (Eder 2010);
- The noisy nature of those travel notes makes the task more complicated, since we did not apply any cleaning preprocessing;
- The Arabic language is very specific since it tends to employ very common introductory sentences that are commonly used by authors (eg. *Albasmala).*

However, the HAT corpus has the advantage to be quite large (100 authors) and the advantage to respect all the conditions required in stylometry. Several experiments of authorship identification were tested and the reported results have shown that one can reach an accuracy of 0.95. Since we did not reach the score of 100%, the HAT-corpus remains still interesting in order to try getting better results by the competing researchers. Finally, one of our objectives is to make this corpus a reference in experimental stylometric evaluation in Arabic.

References

Eder M. 2010. Does size matter? : autorship attribution, short samples, big problem. In Digital humanities 2010 conference, pp. 132-135, London.

Forest D. 2006. Application de Techniques de Forage de Textes de Nature Prédictive et Exploratoire à des Fins de Gestion et d'Analyse Thématique de Documents Textuelles Non Structurés. PhD, Université du Québec à Montréal.

Juola P. 2006. Authorship attribution. Foundations and Trends in Information Retrieval, pp. 233–334, Now publishing.

Küppers R., Conrad S. 2012. A Set-Based Approach to Plagiarism Detection. PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2012, 17-20 September, Rome, Italy.

Sayoud H. 2012. Author Discrimination between the Holy Quran and Prophet's Statements. Literary and Linguistic Computing, Volume 27 Issue 4, pp.427-444.

Tambouratzis G., Markantonatou S., Hairetakis N., Vassiliou M., Carayannis G., Tambouratzis D. 2004. Discriminating the Registers and Styles in the Modern Greek Language-Part 1. Literary & Linguistic Computing, 19(2), 197-220.