Automatic Authorship Attribution of Noisy Documents

H. Sayoud, S. Khennouf, H. Benzerroug, Z. Hamadache, H. Hadjadj, S. Ouamour

halim.sayoud@uni.de, salah.khennouf@uni.de, hadjadj.has@gmail.com, siham.ouamour@uni.de

FEI, USTHB University

Abstract

In this survey, we conduct an investigation on the robustness of several features and classifiers in automatic authorship attribution. Our corpus consists in 25 different documents written by 5 different American philosophers in English.

The different documents pass throw a digital conversion into grey-scaled images and several levels of noise are added to corrupt those image documents. The noise consists in a "Salt & Pepper" type, which is randomly added on the surface of the images with the following noise levels: 0%, 1%, 2%, 3%, 4%, 5%, 6% and 7%. Thus, each image goes throw an OCR program (Optical Character Recognition) to extract the text from the image. Then, the obtained text document is kept to be used during the experiments of authorship attribution.

Several features and classifiers are employed and evaluated with regards to the classification performances. Results are quite interesting and show that the most robust feature in authorship attribution is the character-tetragram, which provides a score of 100% even at a noise level of 7%.

Introduction

Stylometry is a research field concerned with the recognition of the actual author of a text document (Sayoud, 2012). It has been employed several centuries before in order to try discovering some political or criminal writers.

Nowadays, it has gained a lot of interest due to the number of related applications, especially in security and history purposes.

For instance, the discovery of the author of very ancient documents is always challenging to highlight some obscures historical points or check some religious paragraphs. However, most of the old documents are noisy, due to the degradation of the paper and/or the ink (Bronzato, 2015). That is why, it could be interesting to see the performances of Authorship Attribution (AA) in such conditions.

Another interesting point is that many printed documents from the 16th to the early 20th century, are quite noisy (see figure 1) and present bad material for OCR (Optical Character Recognition) systems (Patel, 2012).

Very few researchers were interested in the effect of noise and we did not find very serious related works except the work of Juola in 2012 (Juola, 2012) and the work of Eder in 2012 (Eder, 2012). Both works concerned a simulated text noise and not a real noise in the document. So, for instance, Eder simulated a random word change in the text, while Juola simulated a random character change in the text. The two previous works are interesting but unfortunately do not represent real noisy conditions.

For instance, in the work of Eder, one can find the word "ok" replaced by "experimentally", which is impossible in practice due to the different lengths of those words. So the proposed error remains a simulation only but is not very realistic in practice.

Unfortunately, in real conditions, one can find two different words concatenated into one unique word, such as the text "far, away", which can easily be recognised as the word "faraway" by an OCR system.

Again, in the work of Juola, one can find the character "i" replaced by the character "O", or the character "M" replaced by the character "t", which is unrealistic, as commented by Juola himself, due to the different shapes of those letters.

Unfortunately, in real conditions, OCR systems generate several types of errors (Afli, 2016). For instance, some OCR systems can concatenate two close letters into a unique one, such as the very common mistake we can get in practice: the bigram "fi" recognised as character "h" or "-i" by OCR systems.

Another real problem, which one can meet in practice, is the apparition of hidden text (characters, words and sentences) from the back of the paper. This hidden text can be mistakenly added to the scanned text document and consequently may cause some text errors in the document.

Again, any line caused by graphical noise or originally present in the document may be recognised as a character, number, coma, slash or anything else. Hence, a small vertical line can be identified as "i" or "l" for example.

Moreover, when we observe the old documents, we usually notice several black dots within the text, leading to an incorrect punctuation and/or a complete transformation of the character (eg. "O " becoming "Q").

On the other hand white dots or lines may have another type of errors such as character transformation (eg. "l" becoming "!") or the erase of an entire word in the text (eg. fold of paper).

Again, concerning real noisy documents, a strange fact due to the effect of random graphical noise present in the image document can draw any form that may be similar to a specific character or word, as one can notice in some old documents that are strongly corrupted.

So, all those real errors cannot be only represented by simple simulations in the text, as proposed by Eder and Juola, but must be realised at the image pixel level (low level) and evaluated after the process of a real OCR system.

That is, we have conducted several experiments of AA on noisy documents issued from real image documents and a real OCR recognition process.

Real OCR based noisy Corpus

Since we did not find a real noisy OCR based corpus, we decided to conceive it (we called it OCR5P) (Benzerroug, 2016). Our corpus contains documents written by the following 5 American philosophers: Chauncey Wright, Corliss Lamon, Henri Bergson, Michael James and Solomon Ibn Gabriol. All the documents have the same genre and theme (philosophy). For every author, we chose 5 different texts of about 850 words each.

The different texts are converted into images in "jpg" format with high coding quality. Thereafter, each image documents is randomly corrupted (in Matlab) at the pixel level by the "Salt & Pepper" noise type and with different noise levels (1%, 2%, 3%, 4%, 5%, 6% and 7%).

Finally each image document goes throw an OCR system, which transforms it into a textual document that is kept for the task of AA and evaluation.

Authorship Attribution Methods

The general classification process is divided into two methods: Training Model based Classification and Nearest Neighbor based Classification. In the first type, a training step is required to build the model or the centroid (*in case of similarity measures*); afterward, the testing step could be performed by using the resulting model. In the second type, the training is not required, since a simple similarity distance is computed between the unknown document and each referential text: the smallest distance gives an indication on the most probable class. Furthermore two types of measures are employed: a simple distance and a centroid based distance. The first type is known to be inaccurate, while the second one (*i.e. centroid*) is more accurate and robust against noises.

The first classification type includes the following classifiers: Centroid based Similarity measures, Multi-Layer Perceptron, Support Vector Machines and Linear Regression; whiles the second classification type includes only the nearest neighbor similarity measures.

After every identification test, a score of good authorship attribution is computed in order to get estimation on the overall classification performances. Concerning the features, the following features have been employed: Characters, Character-bigrams, Character-trigrams, Character-Tetragrams and Words. In our experiments, we kept only the 500 most frequent features to speed up the computation process.

Experiments of AA on Noisy Documents

The results of AA on the different noisy documents have led to the following results (see figures 1 and 2, and tables 1 to 6).

Table 1: AA accuracy obtained with Characters

Accuracy								
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Manhattan Centroid	80	90	80	80	80	70	60	50
Linear Regression	90	90	70	60	60	50	40	20
SVM	80	80	80	70	50	50	30	20
MLP	90	90	80	60	60	60	60	30
Average per column	85	87.5	77.5	67.5	62.5	57.5	47.5	30

Table 2: AA accuracy obtained with Character-Bigrams

Accuracy								
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Manhattan Centroid	100	100	100	90	90	90	80	60
Linear Regression	100	100	90	90	80	90	30	50
SVM	90	90	90	90	60	90	30	50
MLP	100	100	90	90	60	80	30	40
Average per column	97.5	97.5	92.5	90	72.5	87.5	42.5	50

Table 3: AA accuracy obtained with Character-Trigrams

Accuracy								
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Manhattan Centroid	100	100	100	100	100	100	90	90
Linear Regression	100	100	100	100	100	100	70	40
SVM	100	100	100	100	100	100	90	90
MLP	100	100	100	100	100	100	90	80
Average per column	100	100	100	100	100	100	85	75

Table 4: AA accuracy obtained with Character-Tetragrams

Accuracy								
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Manhattan Centroid	100	100	100	100	100	100	100	100
Linear Regression	100	100	100	100	100	100	90	100
SVM	100	100	100	100	100	100	90	100
MLP	100	100	100	100	100	100	90	100
Average per column	100	100	100	100	100	100	92.5	100

Table 5: AA accuracy obtained with Words

Accuracy								
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Manhattan Centroid	100	100	100	100	100	100	90	80
Linear Regression	100	100	100	100	100	100	100	90
SVM	100	100	100	100	100	100	100	80
MLP	100	100	100	100	100	100	100	90
Average per column	100	100	100	100	100	100	97.5	85

Table 6: Average AA accuracies for all the features

	Average accuracy							
Noise Level in %	0%	1%	2%	3%	4%	5%	6%	7%
Character	85	87.5	77.5	67.5	62.5	57.5	47.5	30
Character-Bigram	97.5	97.5	92.5	90	72.5	87.5	42.5	50
Character-Trigram	100	100	100	100	100	100	85	75
Character-Tetragram	100	100	100	100	100	100	92.5	100
Word	100	100	100	100	100	100	97.5	85

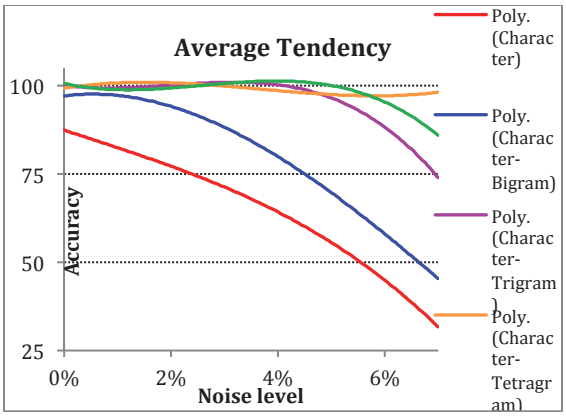

Figure 1: Mean Accuracy of AA vs Noise level

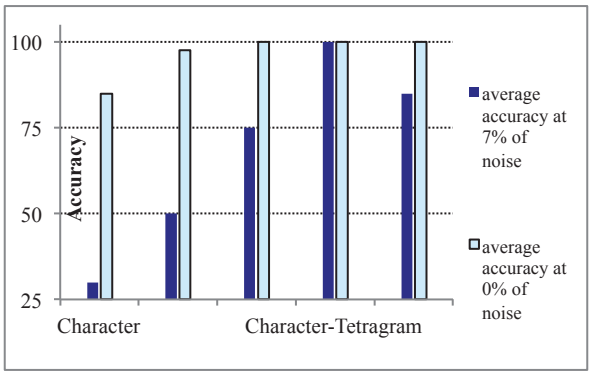

Figure 2: Accuracy at 0% and 7% of noise level

The different experiments of authorship attribution AA have been conducted on a noisy dataset containing 5 different philosophers (OCR5P), where each author is represented by 5 different documents. We used 5 different features namely: characters, character bigrams, character trigrams, character tetragrams and words. Furthermore, three machine learning classifiers (SVM, MLP and Linear regression) and a centroid based distance (Manhattan distance) are employed to attribute the documents to their corresponding authors. The results of this investigation are summarized by the previous tables and figures.

Tables 1 to 5 present the scores of AA at different noise levels. We notice that in overall, when the noise level increases the accuracy decreases for all classifiers and features, except for the character tetragrams, which seem to be robust against the added noise, the score of AA reaches 100% for the three classifiers and the Manhattan distance for almost all the levels of noise (table 4 and figure 1). On the other hand, we notice that the best scores are given by the character tetragrams followed by the words, character trigrams, character bigrams and finally characters, which give the worst score (30% in average at 7% of noise).

Hence, we can conclude that the larger the character Ngram is, the more robust the feature is (table 6 and figure 1). In fact, when we introduce the noise in the document, we add a number of black points to the text, thus these black points are pasted to some text characters, and they may transform them to other new characters (eg. character "r" becoming "n"). That is why, the characters are very sensitive to the noise, however, when we increase the number of characters in the feature representation, we decrease the noise influence and enhance the AA score.

Concerning the classifiers and the distance tested, we notice that the more accurate classifier is the Manhattan distance giving a score of 100% when using character tetragrams, even at a high level of noise. However, with clean text (without noise) all the classifiers presented high performances except when the character feature is used, which confirms again that the use of characters as features is not suitable in our application.

Discussion

In this investigation, we conducted several experiments of automatic authorship attribution on noisy text documents. The main purpose was to assess the robustness of the different features and classifiers in authorship attribution.

We recall that our corpus consists in 25 different documents written by 5 different American philosophers in English and that the noise was randomly added in the image format of each document, before the OCR recognition process (Salt & Pepper noise), with the following noise levels: 0%, 1%, 2%, 3%, 4%, 5%, 6% and 7%. The obtained OCR text was finally used to recognise the documents authors.

During our experiments, we have noticed that, in overall, when the noise level increases the accuracy decreases for all classifiers and features, except for the character tetragrams, which seem to be very robust against the added noise: the score of AA reaches 100% of good attribution even at 7% of noise level.

This is probably due to the low likelihood to have simultaneously 4 characters (of the same word) altered during the noise process. In other words, it is more likely to have one character changed (by noise) than two; and it is more likely to have two characters changed (by noise) than three in the same word; etc. That is, by using character-4grams, one can ensure a certain guaranty to have a very low quantity of noise embedded within the features.

In practice one can verify that fact by observing the performances of the different features: so, we notice that the best scores are given by character tetragrams followed by words, then character trigrams, then character bigrams and finally character monograms, which provide the worst results (30% in average at 7% of noise).

In general, for most of the features, we noticed that from 5% of noise, the performances begin to decrease considerably. So, it appears that a noise level of 5% (in the image document) can cause quite serious problems in AA systems. Finally, according to the results of this investigation, we highly recommend the use of character-tetragrams in case of noisy documents.

References

Afli H., Qiu Z., 2016, A Way, Sheridan P. Using SMT for OCR Error Correction of Historical Texts. Proceedings of LREC-2016, Portorož, Slovenia.

Benzerroug H., 2016. Master Thesis, June 2016. Etude et Analyse de l'Effet d'Acquisition Optique des Documents Textuels sur l'Attribution d'Auteurs. University of Msila.

Bronzato M., Calvini P., Federici C., Dupont A. L., Meneghetti M., D. M. Valerio, Biondie B. and Zoleo A., 2015. Degradation by-products of ancient paper leaves from wash waters. Anal. Methods, 2015,7, 8197-8205. DOI: 10.1039/C5AY01114K

Eder M., 2012. Mind your corpus: systematic errors in authorship attribution. Digital Humanities 2012 was hosted by the University of Hamburg. From 16 to 20 July 2012.

Juola P., Noecker J. I. Jr, Ryan M. V., 2012. Authorship Attribution and Optical Character Recognition Errors. TAL. Volume 53 - no 3/2012, pages 101 à 127.

Patel C., Patel A., Patel D. 2012. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications55.10 (2012).

Sayoud H., 2012, Author Discrimination between the Holy Quran and Prophet's Statements. LLC journal, Literary and Linguistic Compting, pp 427-444, Vol. 27, No. 4, 2012.