ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

AUTHOR IDENTIFICATION FOR ARABIC TEXT

Feryal Haj Hassan, IT Dep. College of Computer and Information Sciences, King Saud University, Saudi Arabia, feryal@ksu.edu.sa

Abstract

In the recent years, authorship attribution has gained a great attention from researchers, especially with the proliferation of Internet and its technology in our life. However, researches in Arabic authorship attribution for Arabic documents are still limited and few works have relatively been published. One of the main applications of authorship attribution is author identification where an anonymous text is attributed to an author between a predefined set of authors. In this paper, we explore author identification of Arabic texts by adopting a set of lexical and syntactic features. The authors group consists of four different authors. In the training stage, a writing style profile consisting of a set of features is generated for each author. For author identification, the features of the anonymous text are extracted, and then compared with authors' profiles. For comparison, two types of classification methods are used: Neural Network and Naïve Bayes where accurate results were achieved by the first method.

Keywords: Authorship attribution; Author identification; Lexical and syntactic features; Neural Network; Naïve Bayes classifier.

Introduction

Authorship analysis is the task of studying text features in order to derive information about its author. It is categorized into three major fields [8]: authorship attribution, which identifies the similarity of a given text with a set of writings produced by a particular author; authorship characterization, i.e. extracting information about the author (gender, age, education,..); and plagiarism detection i.e. detecting similarity between two texts to determine if they are written by a single person without identifying the author.

Authorship attribution can be considered as a problem of text categorization where a document is assigned to a class from a predefined set of classes. However, in text categorization the classification is based on the contents of the document [1], while in authorship attribution the classification is based on author style. In fact, linguistics confirms that everyone has its own unique writing style, which can be considered as "writing print". This is because each text has a set of latent properties, unique for its author.

The authorship attribution task is about either identifying the author of a text among a list of candidate authors, or verifying if a specific author did or did not write the text [6]. In both cases, one of the main concerns is the search for quantifiable features that remain relatively constant among a number of writings by a particular author.

For author identification system, there are two steps: extraction of writing style features, and classification.

Text features can be divided into 4 groups [4, 7]:

- Lexical features (character-based and word-based). For example: number of words, number of sentences, and average word length.
- Syntactic features, including punctuation and function words.
- Structural features, such as paragraph length and use of indentation.
- Content-specific features, which refer to keywords in a specific topic.

Once a feature set has been chosen, the second step of author identification is to apply classification methods to identify the author between the candidate authors. Hence, the features set and the classification techniques may significantly affect the performance of authorship identification.

Most authorship attribution researchers address English texts while researches for Arabic documents are still limited and very few works have been published. In our work, we are interested in authorship identification for Arabic texts. Mainly we are focusing on defining a suitable set of Arabic text features that can be used to define an author profile based on his writings.

The rest of the paper is organized as follows. Section 2 is an overview of some related works. Section 3 describes our approach. Finally, section 4 contains the conclusion and future work.

Related works

Jiexun L. and et al. [7] proposed in their study a Generic algorithm (GA)-based feature selection model to identify text features for English and Chinese online messages. They used different types of features: lexical (total number of upper-case letters/total number of characters and 2-letter word frequency), Syntactic (frequency of punctuation "!" and ":" and frequency of function word "if" and "can"), Structural (number of sentences per paragraph) and Content-specific (frequency of word "check" and "sale"). Their study showed that syntactic features might be more reliable than lexical features in authorship identification.

TAŞ T. and Görür A. K [10] presented a fully automated approach to the identification of the authorship of unrestricted text by adopting a set of style markers to the analysis of the text. The selected style markers were the number of words and sentences, some of word types (verb, noun), punctuation marks and some of word based features. They used 15 different Machine Learning Algorithm and obtained maximum identification rate (80%) with Naïve Bayes multinomial classifier.

Pasqualoni A. study [9] presented a neural network designed to determine the authors of English sonnets written by William Shakespeare and other poets of the same period. The network is trained to attribute a given sonnet to either Shakespeare or a poet from a set of three authors. The input data for the neural network are three lexical features from the text: individual word counts and word pair counts both horizontally and vertically. The used neural network is a multilayer perceptron with two hidden layers of twenty nodes each and one output node indicating whether the input vector belonging to the sonnets written by Shakespeare or by another author. An average over 90 out of 100 sonnets were correctly attributed. This experiment has shown that it is possible to obtain high accuracy rates in authorship attribution using simple lexical measurements, if enough data are provided and a limited author set is used.

In Amasyali M. F. and et al. study [2], a text classification using n-gram model has been realized for Turkish text. They showed whether the modeling of Turkish texts with n-grams is successful approach or not for determining the author of the text, genre of the text and gender of the author. Four different classifiers (Naive Bayes, Support Vector Machine, C 4.5 and Random Forest) were trained on bi-gram and tri-gram models. They found that Naïve Bayes classifier gave the best result in identifying the author of text and bi-gram model is more successful than tri-gram in determining the author of the text.

In our previous works [5, 6], we examined character n-gram based English document author's profile. We investigated total bi-gram and trigram, and n-gram subsets: initial bi-gram and tri-gram, medial and final bi-gram. Results obtained for total bigram and tri-gram were not encouraging, in opposite to result for initial bi-gram and tri-gram where an accurate author identification and verification rates were achieved. For initial bi-gram and tri-gram a threshold is found that separates dissimilarity of same author texts from texts written by different authors.

Previous research's has shown that neural networks and Naïve Bayes are effective in stylometry. They show also that it is possible to obtain high accuracy rates in authorship attribution using simple lexical and syntactic features for style description. In our work we use a set of lexical and syntactic features with neural networks and Naïve bayes classifiers to identify an Arabic novel's author among a set of four candidates.

The approach considered

This approach aims at identifying the author of an Arabic anonymous text between the candidate set of authors. We consider four authors: Abdulkaream Naseaf, Abdullah Tayeh, Ahlam Mustagmani and Nadia Khoust. We use six novels per author as training data and four novels per author for testing. In training phase we generate a "profile" for each author consisting of vector of "n" features.

In identification and testing, we calculate the features of a text and compare the resulted vector with the "nth" predefined author's profiles to identify the author of the text.

The steps of the training stage are given in figure 1 and author identification's steps are shown in figure 2.

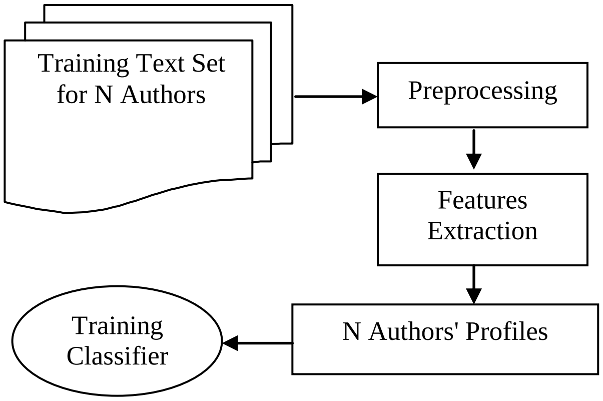

Figure 1. Training stage

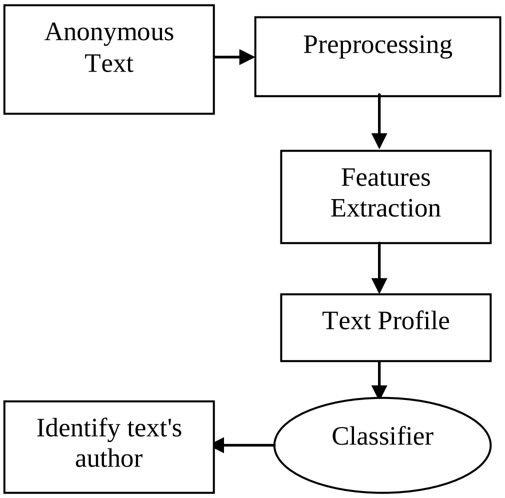

Figure 2. Identification stage

The text to be analyzed must be prepared before performing feature extraction process. The objective of text preprocessing is to convert each line sequence of space into a single space and delete the elongation in every word.

The selection of text features is based on previous works in author attribution for English text and on our observation of the novels of our authors set. We use three lexical features: number of words with length 3, 4 and, 5 characters, 18 syntactic features including number of punctuation marks and some function words. The feature set is presented in table 1.

Table 1: Features set

Features	Description
f_1	Number of Words of length = 3 characters
f_2	Number of Words of length = 4 characters
f_3	Number of Words of length = 5 characters
f_4	Number of punctuation mark “.”
f_5	Number of punctuation mark " , "
f_6	Number of punctuation mark ":"
f_7	Number of punctuation mark "?"
f_8	Number of punctuation mark "!"
f_9	Number of punctuation mark "؟"
f_{10}	Number of punctuation mark "("
f_{11}	Number of punctuation mark "-"
f_{12}	Number of punctuation mark "‘"
f_{13}	Number of function word "من"
f_{14}	Number of function word "قال"
f_{15}	Number of function word "قالت"
f_{16}	Number of function word "لكن"
f_{17}	Number of function word "أن"
f_{18}	Number of function word "عن"
f_{19}	Number of function word "على"
f_{20}	Number of function word "في"
f_{21}	Number of function word "كان"

To avoid the dependence of feature values on the length of the analyzed text, we take their average by 2000 words text length.

To identify the author we investigate two classification methods: Neural Network and Naïve Bayes.

A. Neural Network Classifier

Feed-forward back propagation neural network consisting of three layers is used. The first layer has twenty-one neurons of the logsigmoid type, the second layer neurons are of the tansigmoid type, and the third one has four (number of authors) purelinear neurons. The input vector is the feature vector for each author. There are f input features (f1, f2,…, f21) equal to the number of neurons. The output is four classes corresponding to our selected authors. The network was trained on the 21 features of each author. A correct identification rate of 95% was achieved.

B. Naïve Bayes Classifier

For Naïve Bayes classification we use the algorithm proposed by Conigliaro J. [3]. We consider each data instance (author style) to be as a n dimensional vector of feature values:

$$F = (f1, f2, f3,\ldots, fn) \qquad (1)$$

Where n is the number of features.

A data instance F is assigned to the class for which it has the highest posterior probability conditioned on F, i.e. F is assigned to class ai if:

$$P(ai/F) > P(aj/F) \text{ for all j such that: } 1 \leq j \leq N$$

Where: N is the number of classes (Authors).

According to Bayes theorem:

$$P(ai/F) = P(F/ai)\ P(ai)/P(F) \qquad (2)$$

Since P(F) is the same for all classes, we need to maximize the numerator P(F/ai) P(ai) for classification.

Assuming the features for each class (author ai) are conditionally independent, we have [3]:

$$P\left(\frac{F}{ai}\right) = \sum_{k=0}^{n} -\log10[\mathrm{P}\left(\frac{\mathrm{fk}}{\mathrm{ai}}\right)] \qquad (3)$$

Where:

fk is the kth feature.

ai is the ith category.

n is the number of features.

The Naïve Bayes classifier may be summarized with the following equation:

$$\hat{a} = \text{argmax } [-\text{Log10 } [P(ai)+P(F/ai)] \qquad (4)$$

Where:

â is the estimated classification (identified author).

A is the set of all possible categories (Authors set).

In other words, the Naïve Bayes classifier estimates the probability that an anonymous text (data) belongs to a category (author) ai in A, estimating the probability of the given features being present. The value with the highest probability indicates the resulting author.

Conditional probabilities can be estimated directly as relative frequencies:

$$P(fk/ai) = fk/ni \qquad (5)$$

Where:

ni is the total number of training instances with class, ai and fk is the number of instances with feature fk and class, ai (i.e. number of occurrences of feature fk in class ai).

If fk = 0 then the whole posterior will be zero, to solve this problem the m-estimate of probabilities is used:

$$P(fk/ai) = (fk+m*p)/(ni+m) \qquad (6)$$

Where:

p is the prior estimate of the probability.

m is the equivalent sample size (total number of features in vocabulary).

ni is the number of features in corpus i.

A uniform distribution of word use is assumed, so p=1/m. The estimate of the probability for a given feature fk of corpus j is defined as:

$$P(fk/ai) = (fk+1)/(ni+m) \qquad (7)$$

$$P(ai) = nd/Nd \qquad (8)$$

Where:

nd is the number of documents in author corpus.
Nd is the total number of documents.
The training documents are used to generate two corpora: one consisting of author's novels, the second consisting of other authors' novels.
To create the vocabulary all unique features from both corpora are extracted. The probabilities for each word in each corpus are then computed.
For test document, the probability P for each author is calculated and the author with the maximum P is chosen.
By applying this classifier in our work the achieved identification correct rate is 50% only

Conclusion

Researches in the field of author identification of Arabic texts are still limited. In this paper, we aim to identify the author of an anonymous Arabic text by attributing it to one of four Arabic authors: Abdulkaream Naseaf, Abdullah Tayeh, Ahlam Mustagmani and Nadia Khoust.
Considered approach includes two major steps: the first is to extract text's features; the second is to identify the author of the text by comparing its features with style features of a set of authors.
We used 21 lexical and syntactic text features and studied two kinds of classifiers to identify the author of anonymous text.
Using Naïve Bayes classifier, we get only 50% correct identification rate. While using feed-forward back propagation networks 95% rate of correct identification was achieved. This proves the accuracy of selected features and encourages continuing the work to include experiments on a larger data collection with increasing number of authors, and experiments on short texts.

Acknowledgments

This research project was supported by the Research Center of the Computer and Information Science College in King Saud University, Riyadh, Saudi Arabia. The author expresses his gratitude for this valuable support.
The author is also thankful to Gusti S. , Al_gahtani R., Omier L., Al_harbi N., Al_shehri M., Al_shahrani H., Al_mahmoud W. for their help in the implementation.

References

[1] Alhutaish R., Omar N., "Arabic Text Classification Using K-Nearest Neighbour Algorithm", The International Arab Journal of Information Technology (IAJIT First Online Publication), vol.12, no. 2, 2014.

[2] Amasyali M. F. and Diri B., "Automatic Turkish Text Categorization in Terms of Author", NLDB'06 Proceedings of the 11th international conference on Applications of Natural Language to Information Systems, 2006, pp. 221-226, Springer-Verlag Berlin, Heidelberg.

[3] Conigliaro J., “Author Identification Using Naïve Bayes Classification" available at: http://www.researchgate.net/publication/228942446_Author_Identification_Using_Nave_Bayes_Classification
retrieved September, 2014.

[4] Elayidom M. S. Jose C., Puthussery A. Sasi N. K. "Text Classification for Authorship Attribution Analysis", Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September, 2013.

[5] Haj Hassan F., Chaurasia M. A., “Author Assertion of Furtive Write Print Using Character N-Grams", International Conference on Future Information Technology, Singapore, September, 2013.

[6] Haj Hassan F., and Chaurasia M., "N-Gram Based Text Author Verification", Proceedings of Computer Science and Information Technology ISSN: 2010-460X, Volume 36, 2012, pp 74-78.

[7] Jiexun L. , Rong Z., Hsinchun C., "From Fingerprint to Writeprint", Communications of the ACM, Vol. 49, No. 4, April, 2006, pp. 76-82.

[8] Rong Z. Jiexun L., Hsinchun C., and Zan H. "Framework for Author Identification of online Message Writing Style Feature and Classification Technique", Journal of the American Society for Information Science and Technology,V57, Issue 3, February, 2006, Pages 378-393.

[9] Pasqualoni A., "Author attribution using neural networks", June 27, 2006, available at: http://home.southernct.edu/~pasqualonia1/sonnet/report.html
retrieved October, 2014.

[10] Taş T., Görür A. K., "Author Identification for Turkish Texts", Journal of Arts and Sciences Sayi: 7, May (2007).