2012 International Conference on Innovation and Information Management (ICIIM 2012)
IPCSIT vol. 36 (2012) © (2012) IACSIT Press, Singapore

N-Gram Based Text Author Verification

Feryal I. Haj Hassan [1, a+], Mousmi A. Chaurasia[2, b]

[1, 2]Department of Information Technology, King Saud University,
Saudi Arabia

Abstract. The main goal of authorship attribution is to identify a set of features that remain relatively constant among a number of writings by a particular author. Recently, the common n-gram analysis method for text classification has been successfully used in automatic authorship attribution. This paper deals with text author verification problem using character n-gram information. In addition to usually used total n-gram we studied initial, medial and final bi-grams and tri-grams. Experiments show that author profiles generated with initial bi-gram and initial tri-gram are effective in verifying texts authors. A threshold in the used dissimilarity measure is found for these n-grams that separate dissimilarity of same author texts from texts written by different authors.

Keywords: Author Verification, N-Gram, Dissimilarity Measure

1. Introduction

Authorship analysis tries to find out an author writing style which can be used as his "writeprint". In recent years, practical applications for authorship attribution have grown in areas such as intelligence, criminal law, civil law, and computer security [1].

Stylometry is a linguistic discipline that applies statistical analysis to literary style. It is the basis for authorship analysis, which evaluates writing characteristics to make inferences about who wrote it. Authorship attribution task is about either identifying the author of a text between a list of candidate authors, or verifying if a specific author did or did not write the text. In both cases, one of the main concerns is the search for quantifiable features that are able to differentiate between authors. Many types of features have been investigated including various measures of vocabulary richness and function word frequencies [2].

A promising alternative text representation technique for stylistic purposes makes use of word and/or character n-grams. It has been shown that sub-word units as character n-gram can be very effective for capturing the nuances of an author's style. The most frequent n-grams of a text provide important information about the author's stylistic choices on the lexical, syntactical, and structural level [2]. Generating n-gram author's profile has been adopted by several researchers. Keseljet all [3] presented an automated Type Style and Fonts authorship attribution based on byte n-gram profiles. Razaet all [4] applied word n-gram based authorship attribution method to Urdu poetry. Frantzeskouet all [5] presented an approach to source code author identification based on byte level n-gram. Tbou-Assalehet all [6] explored the idea of automatic n-gram based detection of new malicious Code. Stamatatoset all [7] and Amasyali and Diri [8] used character n-gram in text categorization. Stamatatos [9] proposed method for intrinsic plagiarism detection using character n-gram profiles.

In our work we investigate character n-gram based author's profile. We have studied bi-gram and tri-gram; then we expanded our studies to initial, medial and final bi-grams and tri-grams which - for our knowledge - have not been used so far. In a previous work we focused on the identification process [10]. In this paper we are interested in author verification i.e. to define if a specific author did or did not write the text.

+ Corresponding author
E-mail address: [a]feryal@ksu.edu.sa, [b] mousmi.ksu@gmail.com

The rest of the paper is organized as follows. Section II describes our approach. The experimental results are included in section III. Finally section IV contains conclusion and future work.

2. Methodology and the Algorithm

Our approach is based on character bi-gram and tri-gram. We investigated character n-gram and subsets of n-grams; the initial, medial and final n-grams. Digits and punctuation marks are removed. The profile is defined as a set of length L of the most frequent n-grams with their normalized frequencies.

An authorprofile is generated from a *training author text*. The n-gram profile of the text document to be classified (document profile) is compared with the profile of the corresponding author. The comparison is performed based on the dissimilarity measure algorithm [3]. In this algorithm, we generate the bi- and tri-grams from author's training sample text (author profile). A similar profile is generated for the test data. Let f_1 (n) be the frequency of the nth bi- or tri-grams in the Author's Profile. Let f_2 (n) be the frequency of the nth bi- or tri-grams in the test data. The dissimilarity between the two profiles is calculated using the following formula:

$$\sum n \in profiles \left(f_1(n) - f2(n) \right)^2 \tag{1}$$

In order to normalize these differences, we divide them by the average frequency for a given n-gram

$$sum = \sum n \in profiles \left(\frac{f_1(n) - f_2(n)}{\frac{f_1(n) + f_2(n)}{2}} \right)^2 \tag{2}$$

$$sum = \sum n \in profiles \left(\frac{2 f_1(n) - f_2(n)}{f_1(n) + f_2(n)} \right)^2 \tag{3}$$

In the *testing phase*, first we find out, for each author, the maximum dissimilarity measure between testing texts written by him and his profile i.e. *author dissimilarity threshold*.

Verification Process: For a new text (new document) the profile is generated and dissimilarity with the specific author profile is calculated. If it is less than the author dissimilarity thresholdthen the new text belongs to this author. If not, the new text belongs to another author.

3. Experimental Results

First Corpus. We considered English data set ″A″ including four authors: Eva Gale, Ruth Ann Nordin, Ross Beckmann, and Payton Lee (Table 1). Authors' profiles are obtained using training texts. In bi-gram, we used profile size L= 50, 100 & 200 and for tri-gram we take profile size of 100, 200, 500 & 700.

Table 1. DATA SET- A

Author name	Training text size (words)	Testing text size (words)
Eva -Gale	22068	22068
Payton Lee	278303	275585
Ruth Ann Nordin	235807	109254
Ross Beckmann	124047	124047

Total N-Gram. Results obtained for total bi-gram and tri-gram show that there is no appropriate threshold, which can be used to verify the author. Tables 2 and Table 3 show the results for author verification using the dissimilarity threshold obtained from testing phase.

Table 2.TotalBi-gram Results

Profile Length	100	200	430
Result(%)	39.6	62.5	68.8

Table 3. TotalTri-gram Results

Profile Length	100	200	400	700
Result(%)	72.9	68.8	70.8	70.8

Initial N-gram: Table 4 gives the results for testing on initial bigrams. As we can see with profile size 200 for initial bi-gram the dissimilarity for the same author texts is always less than dissimilarity for different authors, which gives a 100% verification ratio. Results for testing on initial trigram (Table 5) show that both profiles 200 and 500 can be used to verify the author.

Table 4.Initial Bi-gramResults

Profile Length	50	100	200
Result(%)	83.3	93.8	100

Table 5. Initial Tri-gram Results

Profile Length	100	200	500	700
Result(%)	95.8	100	100	97.9

Medial and Final N-grams: The experiments results for medial and final bi-grams (Table 6 and Table 7) and for medial and final tri-grams (Table 8 and Table 9) are not encouraging. There are overlaps between the dissimilarities for the same author and others. We decided not to perform further experiments on these bi-grams.

Table 6.Medial Bi-Gram Results

Profile Length	50	100	200
Result(%)	68.8	66.7	75

Table 7.Final Bi-Gram Results

Profile Length	50	100	200
Result(%)	66.7	75	72.9

Table 8.Medial Tri-Gram Results

Profile Length	500	700	1000
Result(%)	75	66.7	68.8

Table 9. Final Tri-Gram Results

Profile Length	100	200	500	700
Result(%)	68.7	64.6	62.5	64.6

Enlarged Corpus. For moretesting, we enlarged our corpus with other group of authors (Data Set “B”): Austin Jane, Dickens Charles, Bronte Anne, and Bronte Charlotte (Table 10). First, we tested the new data sets for initial bi-grams and tri-grams. The results are shown in tables 11 and 12.

Table 10. DATA SET - B

Author name	Training text size (words)	Testing text size (words)
Austin Jane	317804	401208
Bronte Anne	236038	236038
Bronte Charlotte	275597	275597
Charles Dickens	711187	993551

Table 11.Initial Bi-gram Results

Profile Length	100	200
Result(%)	94.1	98.03

Table 12. Initial Tri-gram Results

Profile Length	200	500	700
Result(%)	92.2	100	100

Then we tested the data set for the eight authors (both data sets A and B combined) only for initial bi- and tri-gram only (Tables 13 and 14).

Table 13.Initial Bi-gram Result

Profile Length	100	200
Result(%)	95.2	99.1

Table 14. Initial Tri-gram Result

Profile Length	200	500	700
Result(%)	98.3	100	99.6

Experiments results are quite good; the initial bi-gram giveshigh verification accuracy with profile size 200 and so did initial tri-gram with the profile size 200,500 & 700 for all authors.

4. Summary

In this paper we presented character n-gram based author verification approach. Total bi-grams and tri-gram, initial bi-gram and tri-gram, medial and final bi-grams are investigated. Results obtained for total bi-gram and tri-gram were not encouraging, in opposite to result for initial bi-gram and tri-gram where an accurate author verification rates were obtained. For initial bi-gram and tri-gram a threshold is found that separate dissimilarity of same author texts from texts written by different authors.

Initial bi-gram and/or tri-gram can be used for author identification problem i.e. predict the most likely author of a text between a predefined set of candidate authors. This will be our future work. In addition, we are intending to perform experiments on Arabic language.

5. Acknowledgment

This research project was supported by a grant from the Research Center of the Female colleges for Medical and Scientific Studies in King Saud University, Riyadh, Saudi Arabia.

6. References

[1] E. Stamatatos: A Survey of Modern Authorship Attribution Methods, Information Processing and Management: an International Journal, Volume 44, Issue 2 (March 2008)

[2] J. Houvardas and E. Stamatatos: N-Gram Feature Selection for Authorship Identification, AIMSA 2006, LNAI 4183, pp. 77 – 86, 2006

[3] V. Keselj ,F. Peng, N. Cercone and C.Thomas: N-gram-based author profiles for authorship attribution, Pacific Association For Computational Linguistics 2003

[4] A. A. Raza, A. Athar and S. Nadeem: N-Gram Based Authorship Attribution in Urdu Poetry, Proceedings of the Conference on Language & Technology 2009 , p 88-93

[5] G. Frantzeskou, E. Stamatatos, S. Gritzalis and S. Katsikas: Effective identification of source code authors using byte-level information, International Conference on Software Engineering, Shanghai, China, 2006.

[6] A. Tbou-Assaleh, N. Cercone, V. Keselj and R. Sweidan: N-gram based Detection of New Malicious Code, Proceedings of the 28th Annual International Computer Software and Applications Conference - Workshops and Fast Abstracts - Volume 02, IEEE Computer Society ,2004.

[7] E. Stamatatos, N. Fakotakis and G. Kokkinakis: Automatic Text Categorization in Terms of Genre and Author, Published in Computational Linguistics, Volume 26 , Issue 4, December 2000, pp. 471-495

[8] M. F. Amasyali and B. Diri: Automatic Turkish Text Categorization in Terms of Author, Genre and Gender, 11th international conference on application of Natural Language Processing and Information System NLDB 2006, Austria

[9] E. Stamatatos: Intrinsic Plagiarism Detection Using Character *n*-gram Profiles, PAN09, Volume: 2, Pages: 38-46, 2009.

[10] F. Haj Hassan and M. A. Chaurasia: Author Assertion of Furtive Write Print Using Character N-Grams, 2011 International Conference on Future Information Technology, September 2011, Singapore.