IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, p-ISSN: 2278-8719, www.iosrjen.org
Volume 2, Issue 9 (September 2012), PP 01-06

Assessment on Stylometry for Multilingual Manuscript

Sushil Kumar[1], Mousmi A. Chaurasia[2]
[1]*Department of Electrical & Electronics, BIT Durg (C.G.) INDIA.*
[2]*Research Scholar, INDIA, Member ACM*

ABSTRACT— *Linguistics and stylistics have been studied for author identification / verification for few years but recently, we have testified a remarkable development in the quantity with which lawyers, courts, applicable in cyber crime, detective agencies etc. etc. have called upon the expertise of linguists in cases of disputed authorship. This inspires researchers to look to the problem of author identification / verification from a different outlook. This paper pact shows text author verification problem using character n-gram information(final n-gram & initial n-gram) for both English & Arabic Text. Experiments demonstrate that author profiles generated with initial bi-gram and initial tri-gram are effective in verifying texts authors. A doorsill value has been set using dissimilarity measure that separate dissimilarity of same author texts from texts written by different authors.*

KEYWORDS— *Author Verification, Character N-gram, Dis-similarity measure*

I. INTRODUCTION

In the typical authorship attribution problem, a text of unknown authorship is assigned to one author, given a set of authors for whom text samples of undisputed authorship are available. The authorship attribution is a process of an anonymous text authorship recognition based on text samples written by a group of already known authors. It relies on a set of features deduced from text documents, and attempts to establish whether texts of an unknown authorship are significantly similar to any of the known text samples. This task is called authorship identification or author verification techniques. There are various types of stylometric features using computational methods such as word n-grams, vocabulary richness, character n-gram (fixed and variable length), content or language specific features. The N-gram method has been found useful in a wide variety of natural language-processing applications, including spelling error detection and correction [1],[2],[3], text compression [4], language identification [5],[6],[7], text categorization [8],[9],[10], text searching and retrieval [8], text retrieval from document images [11], and other information retrieval related applications [12],[13]. Character n-gram is noise tolerant feature. Texts seems to be noisy, containing grammatical faults or short use of punctuation, usually occurs in e-mail or in online forum messages. Vividly, character *n*-gram representation does not affect [14]. An N-gram is a token consisting of a series of characters or words. N-gram approach is a encouraging alternative manuscript representation system for stylistic purposes. It has been shown that sub-word units as character n-gram can be very effective for capturing the distinctions of an author's style.

Stamatatos [14] explores computational necessities and setting that can be applied authorship applications. Keselj et al [15] presented a programmed authorship attribution on character n-gram profiles on byte level n-grams. This approach is tested on three different languages: English, Greek and Chinese. Stamatatos [16] proposed method for intrinsic plagiarism detection using character n-gram profiles. In our effort, we examine character n-gram based author's profile. We focused on identification process based on initial, medial, final and total bi-grams and tri-grams which - for our knowledge - have not been used so far [17]. Amasyali and Diri [18] used character n-gram in text categorization to identify Author, Genre and Gender. The success in determining the author of the text, genre of the text and gender of the author was obtained as 83%, 93% and 96%, respectively.

Sanderson and Guenter [19] described the use of several sequence kernels based on character n-grams of variable length, and the best results for short English texts were achieved when examining sequences of up to 4-grams. An evaluation experiment has been prosed by Ogawa[20] which speed up the process of document retrieval using n-gram indexing. Authors proposed minimum-cost selection method and the redundant n-gram method which reduces retrieval time by using n-grams .Furthermore which also minimizes the processing cost estimated by the document histogram. Kanaris[21] presented a content-based approach to spam detection based on a predefined *n*-gram length (n=3, 4, or 5). In past study [22], three types of variant has been specified in which each variant is adapted by machine learning algorithm to resolve author various issues like author's demographics, verification of mysterious document and in needle-in-a-haystack problem in which many thousands of candidates for each of whom we might have a very limited writing sample. SVM and Bayesian

regression machine learning method in conjunction with character n-gram technique offers efficient and real resolutions to the ordinary authorship attribution problem. N-gram text representation technique is used in Chinese text categorization. It has been shown that combination of 1-, 2-grams is little better than that of 1-, 2-, 3-grams for Chinese text classification [23]. In recent study [24], it has been shown that hybrid algorithm is good way to attribute the author with reliability and accuracy based on short text messages. The correctness of author identity detection, depending of writing style, is surrounded in the range from 71% to 100%.

N-gram technique examines the performance of the bi-gram and trigram term conflation techniques in the context of Arabic free text retrieval. The N-gram approach does not appear to provide an efficient conflation approach due to the idiosyncrasy imposed by the Arabic infix structure that reduces the rate of correct N-gram harmonizing [25]. Experimentation with Arabic language retrieval is still in its immature stage. Researchers in Arabic Information Retrieval can benefit greatly from the work that has been done in English, which has led to the development of many new techniques [26].

In this paper we mainly focused on author verification (i.e. to define if a specific author did or did not write the text) in English & Arabic Language. The rest of the paper is organized as follows. Section 2 describes our approach. The experimental results and discussion are included in section 3. Finally section 4 contains conclusion.

II. METHODOLOGY AND THE ALGORITHM

Our approach is based on character bi-gram and tri-gram. We look into 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 author (i) profile 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 (i). The comparison is performed based on the dissimilarity measure algorithm [15]. In this algorithm, we generate the bi- and tri-grams from author's training sample text (author (i) 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:

$$\sum_{neprofiles} (f_1(n) - f_2(n))^2 \quad (1)$$

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

$$sum = \sum_{neprofiles} \left(\frac{f_1(n) - f_2(n)}{\frac{f_1(n) + f_2(n)}{2}} \right)^2 \quad (2)$$

$$sum = \sum_{neprofiles} \left(\frac{2(f_1(n) - f_2(n))}{f_1(n) + f_2(n)} \right)^2 \quad (3)$$

***Verification Process*:** For a new text (new document) the profile is generated and dissimilarity with the author (i) profile is calculated. If it is less than the author (i) dissimilarity threshold "thi" then the new text belongs to the author (i). If not, the new text belongs to another author (Fig-2).

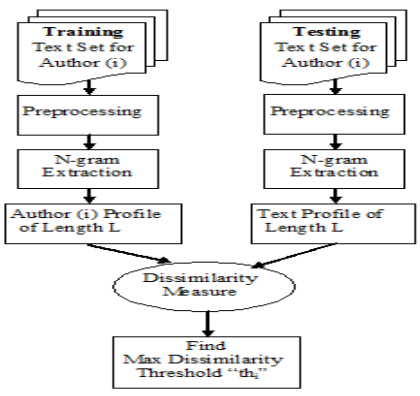

Figure-1: Dataflow for finding Dissimilarity Threshold

Fig. 1 describes the method to find threshold value of a particular author after processing training and testing text. In this scheme, we do compare the author training profile and test profile of same author to find dis-similarity measure. Among all those values of dis-similarity, maximum value could be taken as the door-step (***author dissimilarity threshold "thi")*** of that particular author.

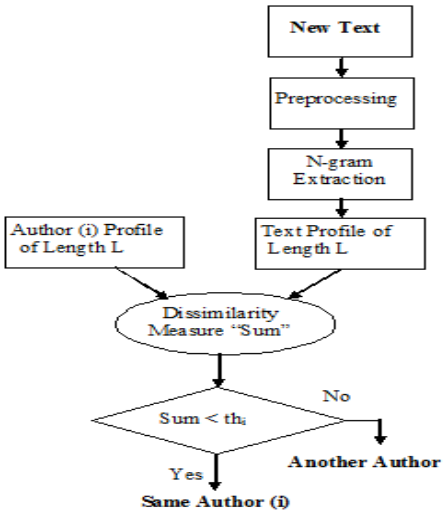

Figure-2: Dataflow for N-gram Author Verification

Again, Fig. 2 after all pre-processing steps, a test profile of different author is being compared with the training profile to find dis-similarity measure which is further compared with particular author's threshold value to verify the authentication of test profile.

III. EXPERIMENTAL RESULTS & DISCUSSION

3.1 First Corpus:

We considered English data set 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. English Data-set

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

I. ENGLISH NOVELS:-

Table 2. Final bi-gram

PROFILE SIZE	50	100	200
PERCENTAGE	66.66	75	72.91

Table 3.Final tri-gram

PROFILE SIZE	100	200	500	700
PERCENTAGE	68.7	64.6	62.5	64.6

II. ENGLISH NOVELS:-

Table 4. Initial Bi-gram

PROFILE SIZE	50	100	200
PERCENTAGE	83.33	93.75	100

Table 5.Initial tri-gram

PROFILE SIZE	100	200	500	700
PERCENTAGE	95.8	100	100	97.91

Table 2 & Table 3 give you an idea about verification of authors using final n-gram. Results of final n-grams are satisfactory. Table 4 & Table 5 demostrate that initial n-grams has more added accuracy in identifying the authors. The outcome of initial n-gram reaches to 100% accuracy with the profile size L = 200 in both bi- & tri-grams whereas it also present the same accuracy level with L = 500 in initial tri-gram. Along with this, higher profile size of n-grams also provides comparatively more accuracy than lower profile size.

3.2 Second Corpus:

We reflected Arabic data set including four authors: Abdul Naseef, Adbullah Tayeh, Ahlam Mestiganmy, Nadia Quest (Table 6). Authors' profiles are achieved using Arabic training texts. For both bi- & tri-gram profile size (L), we used L = 200, 500 & 700.

Table 6. Arabic Data-set

Author name	Training text size (words)	Testing text size (words)
Abdul Naseef	81480	239456
Abdullah Tayeh	38230	55329
Ahlam Mestiganmy	70064	109071
Nadia Quest	80766	254837

The main focus in Arabic novels is selection of profile size (L). We empirically revealed that higher profile size L = 200, 500 & 700 present good accuracy level than lower profile size L = 50 & 100. Furthermore, we decided to analysis Arabic novels with higher profile size.

III. ARABIC NOVELS:-

Table 7.Final bi-gram

PROFILE SIZE	200	500	700
PERCENTAGE	89.74	100	100

Table 8.Final tri-gram

PROFILE SIZE	200	500	700
PERCENTAGE	100	100	100

IV. ARABIC NOVELS:-

Table 9.Initial bi-gram

PROFILE SIZE	200	500	700
PERCENTAGE	84.61	84.61	94.87

Table 10.Initial tri-gram

PROFILE SIZE	200	500	700
PERCENTAGE	100	100	100

Table 7& Table 9 present a good quality inspirational results about verification of authors using bi-gram technique. Results of bi-grams in arabic novels are acceptable. Table 8 & Table 10 demostrate that tri-grams is precise or exact in identifying the authors. The outcome of tri-gram reaches to 100% accuracy with all the profile size L = 200 , 500 & 700 in both final & initial tri-grams.

IV. CONCLUSIONS

In this paper we presented character n-gram based author verification approach. Final n-gram & initial n-gram has been investigated in two languages: English & Arabic. In both the languages, results give confidence for further analysis. Our comparison shows the initial n-gram has more power in verifying the author rather than Final n-gram.

REFERENCES

[1] Harding, S.M., Croft, W.B., & Weir, C. (1997). Probabilistic retrieval of OCR degraded text using N-grams. Research and Advanced Technology for Digital Libraries. *Proceedings of the First European Conference (ECDL), Pisa, Italy.* Retrieved June 2001, from http://citeseer.nj.com/ harding97probablistic.html

[2] Peterson, J.L. (1980). Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, 23, 676–687.

[3] Zamora, E.M., Pollock, J.J., & Zamora, A. (1981). The use of trigram analysis for spelling error detection. *Information Processing and Management*, 17(6), 305–316.

[4] Wisniewski, J.N. (1987). Effective text compression with simultaneous digram and trigram encoding. *Journal of Information Science: Principles and Practice, 13(3),* 159–164.

[5] Damashek, M. (1995). Gauging similarity with N-grams: Language-independent categorization of text. *Science, 267,* 843–848.

[6] Schmitt, J.C. (1990). Trigram-based method of language identification. U.S. Patent 5,062,143. Washington, DC: U.S. Trademark and Patent Office.

[7] Sibun, P., & Reynar, J. Language identification: Examining the issues. *Proceedings of the Symposium on Document Analysis and Information Retrieval* (pp. 125–135), Las Vegas, NV.

[8] Cavnar, W.B., & Trenkle, J.M. (1994). N-gram-based text categorization. *Proceedings of the Third Symposium on Document Analysis and Info* (pp. 161–175). Las Vegas, NV: UNLV Publications/Reprographics.

[9] Huffman, S. (1995). Acquaintance: Language-independent document categorization by N-grams. *Proceedings of the Fourth Text Retrieval Conference (TREC-4)* (pp. 359–372). Gaithersburg, MD: National Institute of Standards and Technology.

[10] Huffman, S., & Damashek, M. (1994). Acquaintance: A novel vector-space N-gram technique for document categorization. *Proceedings of the Third Text Retrieval Conference (TREC-3)* (pp. 305–

310). Gaithersburg, MD: National Institute of Standards and Technology.
[11] Tan, C.L., Sung, S.Y., Yu, Z., & Xu, Y. (2000). Text retrieval from document images based on N-gram algorithm. *PRICAI 2000 Workshop on Text and Web Mining* (pp. 1–12), Melbourne, Australia.
[12] Cavnar, W.B., & Vayda, A.J. (1992). Using superimposed coding of N-gram lists for efficient inexact matching. *Proceedings of the Fifth USPS Advanced Technology Conference*. Washington, DC.
[13] Cavnar, W.B., & Vayda, A.J. (1993). N-gram-based matching for multifield database access in postal applications. *Proceedings of the 1993 Symposium on Document Analysis and Information Retrieval*. Las Vegas, NV: University of Nevada.
[14] E. Stamatatos (2009), A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology (ACM), Vol. 60 issue 3 march 2009*.Wiley. pp. 538-556
[15] V. Keselj , F. Peng, N. Cercone, C. Thomas, " N-gram-based author profiles for authorship attribution" , *Pacific Association For Computational Linguistics 2003*
[16] E. Stamatatos, "Intrinsic Plagiarism Detection Using Character n-gram Profiles",*PAN09*, Volume: 2, Pages: 38-46, (2009)
[17] 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.
[18] M. F. Amasyali, 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*
[19] Sanderson, C., &Guenter, S. (2006). Short text authorship attributionvia sequence kernels, Markov chains and author unmasking: An investigation. *In Proceedings of the International Conference on EmpiricalMethods in Natural Language Engineering* (pp. 482–491). Morristown,NJ: Association for Computational Linguistics.
[20] Ogawa Y. and Matsuda T. (2002) *An Efficient Document Retrieval Method Using n-gram Indexing, Systems and Computers in Japan, Vol. 33, No. 2*, 2002
[21] I. Kanaris, K. Kanaris, and E. Stamatatos, Spam Detection Using Character N-Grams, *SETN 2006, LNAI 3955*, pp. 95 – 104, 2006, Springer-Verlag Berlin Heidelberg 2006
[22] M. Koppel and J.Schler, S.Argamon, Computational Methods in Authorship Attribution, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 60(1)*:9–26, 2009, Wiley Inter-Science.
[23] Z. Wei, D. M., J. Chauchat, and C. Zhong, Feature Selection on Chinese Text Classification Using Character N-Grams, *RSKT 2008, LNAI 5009*, pp. 500–507, 2008, Springer-Verlag Berlin Heidelberg 2008
[24] Monika Nawrot, Automatic Author Attribution for Short Text Documents, *LNAI 6562*, pp. 468–477, 2011, Springer-Verlag Berlin Heidelberg 2011
[25] S. H. Mustafa and Q. A. Al-Radaideh, Using N-Grams for Arabic Text Searching, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 55(11)*:1002–1007, 2004, Wiley InterScience
[26] lbrahim Abu El-Khair, *Arabic Information Retrieval* (Book), Annual Review of Information Science and Technology (2006), ISBN 978-3-8443-1300-0