Vol. 1, No. 1 (2013) 35-43

JOURNAL of COMPUTER SCIENCE
journals.sjp.ac.lk/jcs

Finding Efficient Linguistic Feature Set for Authorship Verification

R.V.S.P.K. Ranatunga

Center for Computer Studies,Sabaragamuwa University of Sri lanka, Belihuloya, 70140, Sri Lanka

Email: spkr@sab.ac.lk

Abstract

Authorship verification rely on identification of a given document to verify whether it is written by a particular author or not. Internally, analyzing the document itself with respect to variations in writing style of the author and identification of the author's own idiolect is the main context of the authorship verification. Mainly, the detection performance depends on the used feature set for clustering the document. Linguistic features and stylistic features have been utilized for author identification according to the writing style of a particular author. Disclosing the shallow changes of the author's writing style is the major problem which should be addressed in the domain of authorship verification. It motivates the computer science researchers to do research on authorship verification in the field of computer forensics and this research also focuses on this problem. The contributions from the proposed research are two folded: Former is introducing a new feature extracting method with Natural Language Processing (NLP) and latter is proposing a novel and more efficient linguistic feature set for verification of the author of the given document. Experiments were carried out on a corpus composed of freely downloadable genuine 19th century English text. Each word segment obtained from the corpus is subjected to feature extraction and 49 stylistic features are used for clustering the text. Other than the standard stylistic features, 19 linguistic features are used as new feature set for the experiments. Generated parse trees by the Stanford Parser are utilized for extracting these linguistic features. Self organizing maps have been used as the classifier to cluster the documents. Proper word segmentation is also introduced in this work which helps us to demonstrate that the proposed strategy can produce promising results. Finally, it is realized that more accurate classification is generated by the proposed strategy with the extracted linguistic feature set.

Keywords: Authorship Verification, Style Markers, Natural Language Processing, Self Organizing Maps.

1. INTRODUCTION

Every author personifies his own diverse and individual version of his own idiolect. His active individual vocabulary, a part of his idiolect, is built up over many years since his birth, which differs from the vocabularies of others. Any writer can, as a principle, use any word at any time in his documents from co-selection of the individual vocabulary. This disclosure is possible to be utilized as a signature of an author for detecting authorship of a document or at least can obtain the basic semblance of the authorship [1, 2]. However, authorship verification is concerned on determining whether a text is written by the same author. This can be a critical problem that very small variations must be taken into account when verifying the author, since the authors'

shallow changes should be caught. Thus the verification of author is one class classification problem [3].

Section 2 will demonstrate the existing technologies already studied and tested in authorship verification. Section 3 introduces the approach with the unsupervised learning with SOM and the utilization of SOM for authorship verification. In Section 4 results obtained will be discussed and the section 5 presents the conclusion and feature work.

2. PREVIOUS WORK

The basic assumption of the Stylometry is that every author has a set of quantifiable characteristics (i.e. style markers). These characteristics are inevitable in all the work of a particular author. However, they can be changed in different authors [4]. Analyzing Stylometry features of the sub divisions such as paragraphs, sentences, words, etc. within the document is the main characteristic of authorship verification. Basically, extraction of style markers of such divisions is counterpart of these analyzing and a large amount of style markers are included with it. These style markers are used for the identification of the writing style of the particular author(s) of a document. The purpose of extracting writing styles to identify the stylistic anomalies and then is used to make the decision on authorship verification.

Mayer and Stain have categorized the style markers into five categories such as (i) character level statistics, (ii) sentence-level text statistics, (iii) part-of-speech (POS) features, (iv) count of special words, (v) structural features and finally, introduced a new feature called Average Word Frequency Class which is successfully used as a vocabulary richness measure. Another five variations of style markers which have been implemented in lexical level of the documents are simple ratio measures, readability scores, vocabulary richness measures, frequency lists, and relative entropy [5].

Simple ratio measures are simply text variables such as Average Word Length, Sentence Length, and Syllables per Sentence etc. [6,7]. There are several kinds of readability measuring scores. Flecsh Index which is calculated by using average sentence length and number of syllables per word is useful for measuring easiness of reading text [8]. If this index is in a higher value it denotes the text is easy to read. Another index is Kincaid Index which also uses same data to calculate the index. The FOG Index is another variation of readability score which further uses the complex words which are use more than three syllables [9]. Other resolutions of this kind of measures are SMOG' Formula, FORCAST formula and Fry Readability Graph. These formulae use complex calculations and need long text to give proper scores.

Vocabulary richness measures are also successfully used as features for authorship attribution. Diversity of the author's vocabulary is determined by these measures. The Type/Token Ratio, Once Occurring Words V1 (hapax legomena) or Twice Occurring Words V2 (hapax dislegomena) are usually used matrices [10]. The problem of these vocabulary richness measures is that those are depended on the document length and present unreliable result for short documents. The Average Word Frequency Class measure has been used by Mayer and Stain which calculate the vocabulary of the author and it does not depend on the length of the document. A document's average word frequency class tells the style complexity and the size of an author's vocabulary and it has very less variance between document lengths [11].

Under syntactic level measures which depend on POS tags also has been used as style markers in authorship attribution [12]. Mostly used measures are count number of passives, count of the frequency of various categories of POS tags [13]. N-grams of syntactic labels from partial parsing also have been used as features of authorship attribution [14]. Functional lexical features are also reliable markers of style. The basic disadvantage that has been identified in these measures is the computational complexity to generate such measures. However, these measures have been presented good result in both long and short texts.

The used Stylometry features can be categorized into several major areas and the linguistic features are in one category. Since the computational complexity for extracting these features the

implementation of feature extraction of linguistic features with Natural Language Processing is more problematic in the context of the authorship verification domain and thus the contribution of this kind of features for the area is less in majority of research. The efficient method of extraction the linguistic features and the set of more significant Stylometry feature set for authorship verification are introduced by this research.

3. MATERIALS AND METHODS

Four books which have been selected on two authors from American essayists in 19th century called Thoreau and Emerson. 5000 word blocks from the beginning of each document have been used for creating the test data. Firstly, the selected word blocks from two different authors were mixed. Two documents called "Walden with Conduct" and "Concord with English Traits" were created by such admixture. Secondly, 5000 word blocks were selected from same author. These word segments of same author were mixed together and created two documents named "Walden with Concord" and "Conduct with English Traits". Experimental corpus was created by these four documents. The classes are made on each created text as 100, 200, 300, 400, and 500 word segmentations.

Forty-nine features were used for the experiments covering almost all the aspects of the previously defined features in the literature. Typically, these features might include simple ratios such as total number of characters, average length per word, number of sentences, words per sentences etc. Six word based features have been used such as words longer than six characters, total number of short words, number of syllables, syllables per word, number of complex words (more than 3 syllables), number of specific words. Nine features have been used to measure the vocabulary richness as standard authorship attribution like Hapax legomena, Hapax dislegomena, Yule's K measure, Simpson's D measure, Sichel's S measure, Harden's V measure, Brunets W measure, Honore's R measure, and Average Word Frequency Class.

Syntactic and POS features are also used and there are 19 such features including number of nouns, number of passive verbs, number of base verbs, number of adjectives, and also number of clauses and number of phrases. Since many authors irregularly attempt to use adverbs as their own pattern, other than these features pertaining to the POS features, adverbs also extracted as domain, duration, frequency, focus, locating, manner, promina, and sequence. Number of articles, number of prepositions, number of coordinate conjunctions, and number of auxiliary verbs are also used as the syntactic features. With respect to readability measures, Flesh Index, Kincaid Index, and Fog Index have been used. In this work we have introduced the punctuation measures which is not in the literature as important and those includes number of commas, number of single quotes (’), number of double quotes (“), number of colons(:), number of semi-colons(;), number of question marks(?), number of exclamation marks(!), and the number of "etc.". Syntactic and POS features have been extracted by using the parse trees of the sentences. These parse trees are obtained by using the Stanford Parser.

In the preprocessing level, the document is converted to a text file. Then, the segmentation of the document is done since authorship verification is based on the analysis of one document and all the evidence should be extracted from the entre document. The segmentation of the document should be done very carefully because the authors writing styles may vary very closely. the segment size θ is declared as a threshold of the experiments. The number of words has been used as the measure of lengths of both document and segment. The number of segments of the document d should be proportionate with the document length. Then $n = \frac{d}{\theta}$ where n is the number of segments in the document d. Each segment of the document d is subjected to extracting its features for further analysis.

The author's writing style attributes can be quantified by the style markers. Let $s_1, s_2, .. s_n$ be the segmentation of d into n contiguous, non-overlapping segments. Let m denotes the number of styles makers and $\tau_1, \tau_2, .., \tau_m$ be the quantified styles of the segment s and $s = (\tau_1 \cdots \tau_m)$

denotes the segment of the document d. The input feature space of the model can be represented by a n × m vector space per document d as in the equation 1.

$$d = \begin{pmatrix} s_1 : \tau_1, \tau_2, .., \tau_m \\ \vdots \\ s_n : \tau_1, \tau_2, .., \tau_m \end{pmatrix} \quad (1)$$

In the classification procedure; Self Organizing Maps (SOM) is applied to cluster the text [15]. While preprocessing, tables and figures are removed from all documents. AutoSOME tool has been used for conducting the experiments[1]. The tool consists of several parameters which can be used to enhance the clustering performance such as Ensemble Runs, SOM Iterations, SOM Grid Size, SOM Topology, SOM Error Exponent, SOM Distance Metric, Cartogram XY Size, Clustering Method, MST P-value. Other than those parameters there are several normalization techniques also that can be applied for normalize the dataset. Log2 Scaling, Unit Variance, Median Center, Sum of Squares=1 for both column and rows are some normalizing facilities available in the tool.

Four documents were segmented into 100, 150, 200, 250, 300, 350, 400, 450, 500 word segments. Each document under each segment was fed to extract the features and these extracted feature values were analyzed by using a genetic algorithm based feature selection to filter the best feature set from all features of each document. The mutation rate of the genetic algorithm may vary in each segmentation between 0.1 to 1.0 and finally identifies the good clustering performance which can be obtained at 1.0 mutation rate. Finally, each selected feature file under each segmentation has been input to the AutoSOME tool and the result was obtained for further analysis.

4. RESULT AND DISCUSSION

The main consideration of the experiments is to identify a set of significant linguistic style markers which is more suitable for verify the author's writing style. Two main experiments have been employed for testing. First experiment is mainly divided into two as same author and different author. The two documents which are already mixed together called "Walden and Conduct" and "Walden and Concord" are used in this experiment. The former is used for testing 'different author' and later is used for testing 'same author'. Only one cluster should appear in the former and in the two clusters for the latter. Nine successful SOM runs are included in each file and finally there are eighteen files. Parameters set up in both testing are equal.

Exact result is given in all segmentations and provides the numbers of clusters according to the number of authors. The divided clusters of each sub experiments are shown in Tables 1 and 2.

TABLE 1: Clusters obtained by the models in each segmentation for two different authors

Segment	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
No of Clusters	2	2	2	2	2	2	2	2	2

[1] (http://jimcooperlab.mcdb.ucsb.edu/autosome/)

TABLE 2: Clusters obtained by the model in each segmentation for same author

Segment	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
No of Clusters	1	2	1	1	2	1	1	2	2

TABLE 3: Clustering performance of the model in each segmentation on two different authors

	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
Precision	0.90	0.88	1.00	1.00	0.94	1.00	1.00	1.00	1.00
Recall	0.53	0.52	0.74	0.56	0.54	0.56	0.52	0.5	0.59
F measure	0.67	0.65	0.85	0.72	0.69	0.72	0.68	0.67	0.74

Tables 4 and 5 shows the result of experiment 01 on different authors and the same author respectively.

TABLE 4: Clustering performance of the model in each segmentation on two same authors

	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
Precision	0.91	0.97	1.00	1.00	0.94	1.00	1.00	1.00	1.00
Recall	0.78	0.94	1.00	1.00	0.54	1.00	0.5	0.5	0.53
F Measure	0.84	0.95	1.00	1.00	0.69	1.00	0.67	0.67	0.69

Inaccurate results are given by 150 and 300 segments and also 450 and 500. The reason for having two clusters in the 450 and 500 segmentations is, when the number of words are increased in the segment the model attempts to discriminate the styles of the same author according to the topic or the theme. It is clearly shown by the experiment. Optimum clustering performance is given in the segmentations of 200 in both sub experiments. The F measure is 0.85 and 1.0 respectively.

Conversely, the utilized feature space for the optimum result also analyzed. The remarkable fact is that most of the traditional features are not selected and the newly introduced linguistic features are selected by the genetic algorithm. For example, some traditional features are used in authorship attribution and verification like Hapax legomena, Hapax dislegomena, Simpson's D measure, Brunets W measure, Average Word Frequency Class (AWFC) etc. are not selected. The adverbial features which have been newly introduced get priority and seven features out of nine are selected. Conversely, both number of clauses and phrases extracted by using NLP are also significant in this experiment. Finally, the punctuation measures play a good roll in this experiment and five measures out of eight are selected as significant for the experiment.

The second sub experiment is done in the same way as the first experiment is done except for the documents used. The two documents which are already mixed together called "English and Concord" and "English and Conduct" are used for this experiment and the former is used for testing different authors and the latter is used for testing for the same author. Only two clusters are expected from the former and one cluster from the latter. Nine successful SOM processing are

included in each file and finally, there are eighteen files. Both tests use equal parameter setup of SOM.

The number of clusters of experiments obtained on "English and Concord" (different authors) are shown in Table 5.

TABLE 5: Clusters obtained by the model in each segmentation for two different authors in the second experiment

Segment	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
No of Clusters	3	2	2	3	2	2	2	2	2

Although the expected number of clusters is two an incorrect result is given by two segmentations only. It is also as same as the above and the four outliers out of one hundred are affected in the 100 segmentation and 03 outliers out of thirty nine affected in the 250 segmentation.

The clustering performance obtained by the model for this experiment is shown in Table 6. The segmentation 450 obtains the optimum clustering performance.

TABLE 6: Clustering performance of the model in each segmentation on two different authors

	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
Precision	0.71	0.91	0.92	0.85	1.00	0.92	1.00	1.00	0.8
Recall	0.58	0.53	0.52	0.47	0.5	0.52	0.52	0.55	0.5
F Measure	0.64	0.67	0.66	0.61	0.67	0.66	0.68	0.71	0.62

The mixed document of the same author is experimented in the second part of the experiment. The expected clusters are one in this experiment. The number of clusters given by the model in each segmentation is shown in Table 7.

TABLE 7: Clusters obtained by the model in each segmentation for the same author

Segment	100 Seg	150 Seg	200 Seg	250 Seg	300 Seg	350 Seg	400 Seg	450 Seg	500 Seg
No of Clusters	1	2	1	2	2	1	1	1	1

Inaccurate clustering is given by 150, 250 and 300 segmentations. Even though more than 300 word segments give significant results on this experiment 200 word segmentation also gives the optimum performance. It is clearly shown in Table 8.

TABLE 8: Clustering performance of the model in each segmentation on two different authors

	100 Seg	**150 Seg**	**200 Seg**	**250 Seg**	**300 Seg**	**350 Seg**	**400 Seg**	**450 Seg**	**500 Seg**
Precision	1.00	0.97	1.00	1.00	0.94	1.00	1.00	1.00	1.00
Recall	1.00	0.91	1.00	0.83	0.6	1.00	1.00	1.00	1.00
F Measure	1.00	0.94	1.00	0.91	0.73	1.00	1.00	1.00	1.00

In the second sub experiment, the same word segmentation provides the most significant and optimum clustering performance and the precision, recall and f-measure obtained 1.00. The features selected in the most significant result are analyzed where it exhibits the same nature. For example, some traditional features like Hapax legomena, Brunets W measure, Sichel's S measure, Harden's V measure, Honore's R measure, Flesh Index, Average Word Frequency Class (AWFC) as well as the features based on a number of words including the average length per word, total number of short words per words are not significantly selected by the genetic algorithm. The adverbial features do not involve significantly in the second experiment and only three features out of seven are selected. Both the number of clauses and the number of phrases extracted by using NLP are also significant in this experiment like the first experiment. Finally, the majority of punctuation measures are also used to identify the clusters.

Figure 1 presents the summarization of categorized features on participation for the clustering of the four major documents in both experiments with four sub experiments.

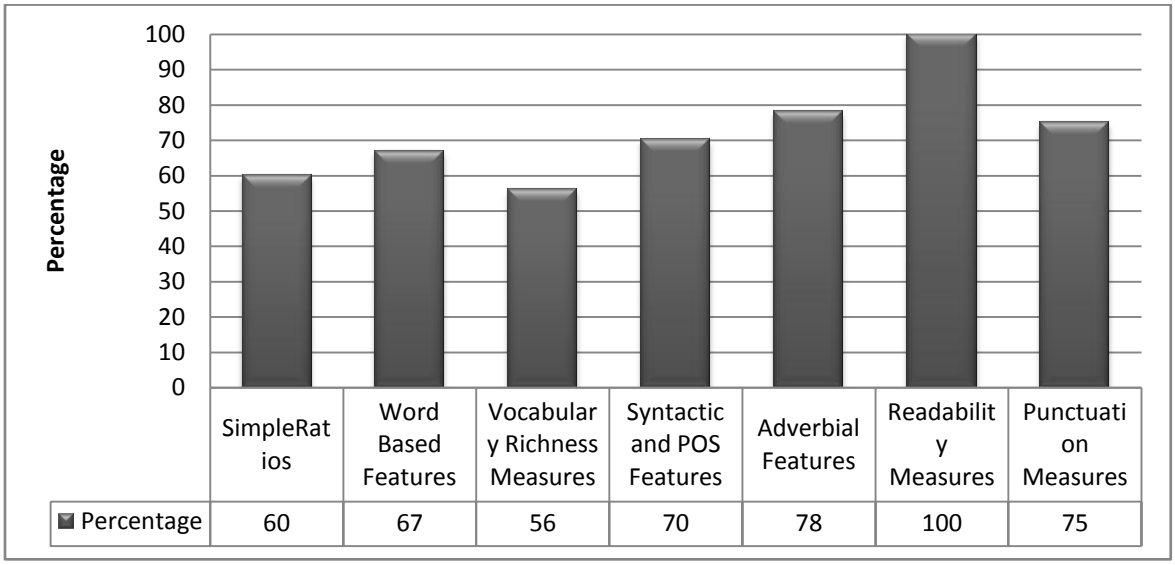

FIGURE 1: Usage of each feature category for clustering the documents.

The usage of readability measures get 100% and the newly introduced adverbial and punctuation measures also have played major roll. The minimum usage of 56% is presented by the traditional vocabulary richness measures. Generally, utilization of newly introduced Syntactic and part-of-speech features are in third place and it is higher than the traditional simple ratios, word based, and vocabulary richness style markers.

5. CONSLUSION & FUTURE WORK

The selections of features vary on the nature of the document. However, the four experiments cited above give evidence for using linguistic style markers which can be used more efficiently for authorship verification. Further, it is realized that SOM attempt to cluster the different theme of the same author. Because of the two books of the same author are in different context. It exhibits the used feature space capable to identify the shallow changes of the author's writing style accurately.

REFERENCES

1. R.M. Coulthard, Author identification, idiolect and linguistic uniqueness, Applied Linguistics Dumas B 'Reasonable doubt about reasonable doubt: assessing jury instruction adequacy in a capital case. in *J Cotterill (ed)*, pp. 246-259, 2002
2. R.M. Coulthard, Beginning the study of forensic texts: corpus, concordance, collocation, in M Hoey (ed.), *Data Description Discourse*, London: HarperCollins, PP 86-97, 1993.
3. M. Koppel, J. Schler, S.Argamon, Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, 60(1), pp 9 – 26, 2009.
4. D.I. Holmes, R.S. Forsyth, The Federalist Revisited: New Directions in Authorship Attribution, *Literary and Linguistic Computing* 10 (2), pp 111–127, 1995.
5. K. Marco, "Using Style Markers for Detecting Plagiarism in Natural Language Documents," Department of Computer Science, University of Sk¨ovde, Sk¨ovde, Theses 2003.
6. G. U. Yule, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to two Cases of Disputed Authorship," Biometrika, vol. 30 (3/4), pp. 363–390, 1939.
7. W. Fucks, "On Mathematical Analysis of Style," Biometrika, vol. 39 (1/2), pp. 122 – 129, 1952.
8. P. Clough, "Analyzing style - Readability," University of Sheffield, Technical report Available from: http://ir.shef.ac.uk/cloughie/papers/readability.pdf [Accessed 2010-11-02], 2000.
9. K. Johnson. (1998) Readability. [Online]. http://www.timetabler.com/readable.pdf
10. T. Honor´e, "Some simple measures of richness of vocabulary," Association for literary and linguistic computing bulletin, no. 7 (2), pp. 172–177, 1979.
11. S. Meyer zu Eissen and B. Stein, "Intrinsic plagiarism detection," in Lalmas et al. (Eds.): Advances in Information Retrieval Proceedings of the 28th European Conference on IR Research (ECIR), vol. 3936 of Lecture Notes in Computer Science, Landon, 2006 Springer, pp. 565-569.
12. N. Fakotakis, and G. Kokkinakis E. Stamatatos, "Automatic Text Categorization in Terms of Genre and Author," Computational Linguistics, vol. 26(4), pp. 471-495, December 2000.
13. K. Luyckx and W. Daelemans, "Authorship Attribution and Verification with Many Authors and Limited Data," in 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, 2008, pp. 513–520.

14. G. Hirst O. Feiguina, Authorship attribution for small texts: Literary and forensic experiments, 2007, SIGIR ’07, Amsterdam, Workshop on Plagiarism Detection, Authorship Identification and Near Duplicate Detection.
15. T. Kohonen, Self-organizing maps.: Springer-Verlag, 2001.
16. F. Tweedie, R. Baayen, How variable may a constant be? Measures of lexical richness in perspective, *Computers and the Humanities*, 32(5), pp 323–352, 1998.
17. E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic text categorization in terms of genre and author, *Computational Linguistics*, 26(4), pp 471–495, 2000.
18. S.Argamon, S.Levitan, Measuring the usefulness of function words for authorship attribution, In Proceedings of the *Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 2005.