Stylometric Studies based on Tone and Word Length Motifs

Hou Renkui
The Hong Kong Polytechnic University, Hong Kong
hourk0917@163.com

Huang Chu−Ren
The Hong Kong Polytechnic University, Hong Kong;
churen.huang@polyu.edu.hk

Abstract: We propose a new approach to stylometric analysis combining lexical and textual information, but without annotation or other pre-processing. In particular, our study makes use Chinese tones motifs and word length motifs automatically extracted from unannotated texts. The proposed approach is based on linked data in nature as tone and word-length information is extracted from a lexicon and mapped to the text. Support vector machine and random forest were used to establish the classification models for author differentiation. Based on comparative study of classification results of different models, we conclude that the combination of word-final tones motifs, segment-final motifs and word length motifs provides the best outcome and hence is the best model.

Keywords: Stylometric analysis, Tones motifs, Word length motif, Chinese prose

1 Introduction

Style refers to linguistic choices made by an author that distinguish his/her writing from those of other authors (Herdan 1966). Stylometric analysis, can distinguish texts written by different authors by measuring some stylistic features in text. It is assumed that quantitative authorship attribution is that the anonymous author of a text can be selected from a set of possible authors by comparing the values of textual measures in that text to their corresponding values in each possible author's writing samples (Grieve 2007). In fact, textual measurements are assumed to include conscious and unconscious aspects of the author's style. It would then be an asset to find the features of the unconscious aspect, since they can not be consciously manipulated by the author (García & Martin 2006). Stylometric analysis involves extracting style markers, i.e. stylometric features, and classifying the texts represented by those features according to authors (Stamatatos et al. 2000). These models can be seen as the text classification according to their authors.

The most effective features to discriminate between different authors, i.e. style markers, should be determined at first. A great variety of measures, including sentence length, word length, word frequencies, character frequencies and vocabulary richness had been proposed. Savoy (2012) compared the performance obtained when using word types or lemmas as text representations.

Koppel et al. (2009) compared the performances of several representative learning methods for authorship attribution and showed that the choice of the learning algorithm is no more important than the choice of the features by which the texts are to be represented.

This paper examines whether lexical information, such as tones motifs and word length motifs, can serve as effective stylometric features in authorship attribution. The motivation of such a study is both to find an effective model of stylometric study without annotation and processing, as well as to test the effectiveness of the linked data approach to stylometric studies.

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 56–63
Cebu City, Philippines, November 16-18, 2017

1.1 Literature review

Mosteller and Wallace's (1964) influential work in authorship attribution was based on Bayesian statistical analysis of the frequencies of a small set of common and topic-independent words (e.g., "and", "to", etc.) achieved productive and significant discrimination results between the candidate authors. Since then and until the late 1990s, research in stylometry was dominated by attempts to define features for quantifying writing style (Homes 1994, 1998), and to explore the new modeling methods.

Since the late 1990s, the study of authorship attribution have changed because of the vast amount of electronic texts available through Internet media. Koppel & Argamon (2009) considered a number of feature types that have been, or might be, used for the attribution problems. A number of earlier works that have surveyed and compared various types of feature sets, include Love (2002), Zheng et al. (2006), Abbasi and Chen (2008), and Juola (2008).

Most stylometric studies are lexically based, especially because it is the level of language where repetitions may be reliably used as a basis for measurement (Holmes 1994).

Grieve (2007) compared thirty-nine different types of textual measurements commonly used in attribution studies, in order to determine which are the best indicator of authorship. Stamatatos (2009) summarized the text representation features, style markers, and the computational requirements for measuring them. The most common words (articles, prepositions, pronouns, etc.) are found to be among the best features to discriminate between authors (Argamon & Levitan, 2005). Savoy (2015) found that some simple selection strategies (based on occurrence frequency or document frequency) may produce similar, and some times better, results compared with more complex ones. For example, García & Martin (2006) proposed that the function words prove to be more reliable identifiers of authorship attributions because of their higher frequencies.

There are also many researches for Chinese authorship attribution. Most of the researches focused on the distribution of character, word, lexical, syntax and semantic in the stylometric analysis. Wei (2002) examined the authorship attribution of the Chinese classical literary masterpiece, "The Dream of Red Mansion", using the distribution of common words. Ho (2015) thought Chinese auxiliary words, namely "的、地、得", can represent the writing style of different authors; Hence can be used as measurement to judge the author of literary texts. Xiao & Liu (2015) examined the stylistic difference between the literatures of Jinyong and Gulong using text clustering. He & Liu (2014) examined the difference of usage of rimes of a Chinese syllable in the prose of different Chinese authors based on text clustering. Other than this study, there were very few stylometric studies making use of the lexico-phonological characteristics of Chinese, and certainly not the unique tonal features.

1.2 Research question and methodology

The information of character features is easily available for any natural languages and corpus, and they have been proven to be quite useful to quantify the writing style (Grieve 2007). The tones are the important and essential components and play an important role in Chinese language to determine the meaning of different words and characters. While there are few studies to examine whether tones can be used as stylometric in authorship attribution in Chinese language.

There are four tones which are high and level tones (阴平 *YinPing*), rising tones (阳平 *YangPing*), falling-rising tones (上声 *ShangSheng*), falling tones (去声 *QuSheng*). Except these

four tones, there is also a light tone.

This study hypothesizes that different authors tend to have different characteristic pattern of tone motifs and word length motifs usage. We selected the tone motifs and word length motifs in the different specific positions in the sentences as the characteristics to classify the texts according to their authors.

Support vector machine (SVM) algorithm and Random Forest were selected to establish the classification model. 5-fold cross-validation was used to measure the generalization accuracy. In order to avoid the contingency, the 5-fold cross-validation was run 30 times repeatedly. The average value of identification error rate (Stamatatos & Fakotakis 2000, Tan et al. 2006), i.e., erroneously classified texts/total texts, was used to validate the classification result.

We use the open source programming language and environment R (R Core Team 2016) to realize the classification experiments. The function of *ksvm* in R package *kernlab* and the function *randomForest* of R package *randomForest* were used to classify the texts from different authors.

2 Corpus

In the studies of stylometric analysis, an important problem is that the distribution of the training corpus over the different authors is uneven. For example, it is not unusual to have multiple training texts for some authors and very few training texts for other authors.

Another important question is the size of one text sample per authors. The text samples should be long enough to adequately extract the style of them which can be used as text representation features. Different from the existing researches of authorship attribution, this study focuses on the stylometric analysis of Chinese literary texts of different authors and explores whether the tones motifs and word length moitfs of Chinese language can be used as stylometric properties. It isn't rigorous than the authorship attribution in the data collection of this study. So we selected the similar number of texts of different authors and the similar size of every texts to establish the corpus for this study.

In this study, the proses of four Chinese writers were selected to build the corpus, as shown in Table 1. They are Congwen Shen, Zengqi Wang, Qiuyu Yu and Ziqing Zhu.

Table 1: Corpus scale using this study

	Text number	Word type	Word token
Congwen Shen	40	11551	101670
Zengqi Wang	38	14289	111589
Qiuyu Yu	38	11294	90132
Ziqing Zhu	38	13011	123674

Chinese language texts are written Chinese character by character. We try to resolve the question of multi-sound characters by segmenting the texts from character sequences to word sequences using the Chinese lexical analysis system created by Institute of Computing Technology of Chinese Academy of Science (ICTCLAS). Most of multi-sound characters have one pronunciation in a word.

Then we establish a system for extracting the tones of the characters based on the grammatical knowledge-base of contemporary Chinese.

3 Experiments results

Firstly, Chinese sentence should be defined in this study because the sentence-initial and

sentence-final characters will be considered. A sentence in Chinese text, however, is not easily defined for the lack of reliable convention to mark end-of-sentence, and because of frequent omission of sentential components including subjects and predicates (Huang and Shi 2016). Consequently, Chinese sentences are often defined in terms of characteristics of speech, rather than text (Lu 1993; Huang & Shi, 2016). Chao (1968) and Zhu (1982) offer similar definitions that rely on pauses and intonation changes at the boundaries of sentences.

According to the approach of many Chinese Treebank (e.g. Chen et al. 1996 for Sinica TreeBank, Huang and Chen 2017) and the analysis of sentence length distribution in quantitative linguistics (Hou et al. 2017) all segments between commas, semicolons, colon, periods, exclamation marks, and question marks expressing pauses in utterances are marked as sentences. Actually, the sentences by this definition are the clauses and conform to the sentence definitions relying on pauses and intonation changes in the utterances. In Wang & Qin (2013) and Chen (1994), the sentence by this operational definition is called sentence segment (hereinafter segment). Wang & Qin (2013) considered that sentence segment length is more relevant to language use in Chinese. So the sentence segments are used as the unit for extracting the sentence-initial and sentence-final characters.

There are often unique rhythms when the different proses are read. This unique rhythm is an inherent characteristic of a prose. Wang et al. (2011) proposed that there are different rhythms between the texts from different authors whilst there are similar rhythms between the texts of an author.

The motif was inspired by the F-motiv for musical “texts” (Boroda 1982) and continued in linguistics by Köhler (2006, 2008) who used the concept of L-motifs, i.e. length motifs. Boroda defined the “F-Motiv” with respect to the duration of the notes of a musical piece because units common in musicology were not usable for his purpose.

According to Köhler & Naumann (2010) and Köhler (2015), linguistic motif is defined as:

The longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit. Thus *a L-motif is a continuous series of equal or increasing length values.*

Following the definition, any text or discourse can be segmented in an objective, unambiguous, and exhaustive way, i.e. it guaranties that no rest will remain (Köhler 2008).

In addition, motifs can be defined for any linguistic unit and for any linguistic property.

And motifs have an appropriate granularity, with respect to which motifs are scalable.

Word length is an important indicator for stylometric analysis and has significances in prosodic linguistics. L-Motif of word was defined as a maximal sequence of monotonically equal and increasing numbers which represent the length of the adjacent words in a sentence segment. According to this definition, a given text can be segmented some paragraphs which are represented by an uninterrupted sequence of L-segments of word. For example, in the following paper, the word L-moitf is (2), (1, 2), (1, 2, 2), (1, 1, 1, 1, 2), (1, 2, 2), (1, 2).

白河 到 沅陵 与 沅水 汇流 后 ， 便 略 显 浑浊 ， 有 出山 泉水 的 意思 。

Tone is the category variable, we defined the tone-motif as the longest continuous sequence of equal tones.

This part will examine whether tone motifs, word length motifs and their combination can be used as stylistic characteristics of the different authors. The segment-initial and segment-final tone motif, the word-final tone motif were considered. The paragraph was considered as a unit to compute the segment-initial and segment-final tone motif. The segment was considered as a unit to

compute the word-final tone motif and word length motif.

Table 2: The classification results using the tone motif, word length motif and their combination as characteristics

		Identification error rate	
	Stylometric markers	SVM	RF
1	word-final tone motifs	27.77%	26.01%
2	segment-final tone motifs	47.85%	50.91%
3	word-final tone motifs + segment-final tone motifs	24.15%	20.7%
4	bigrams of word-final tone motifs	34.35%	36.1%
5	word-final tone motifs + their bigrams	30.75%	26.85%
6	word length motifs	35.16%	33.83%
7	word-final tone motifs + word length motifs	20.07%	19.07%
8	word-final tone motifs + segment-final tone motifs + word length motifs	14.02%	14.62%

The texts from different authors were represented by the motifs and classified according to their authors. SVM and random forest were used to establish the classification model and the 5-fold cross validation was used to validate the classification results, as shown in Table 2 and Figure 1.

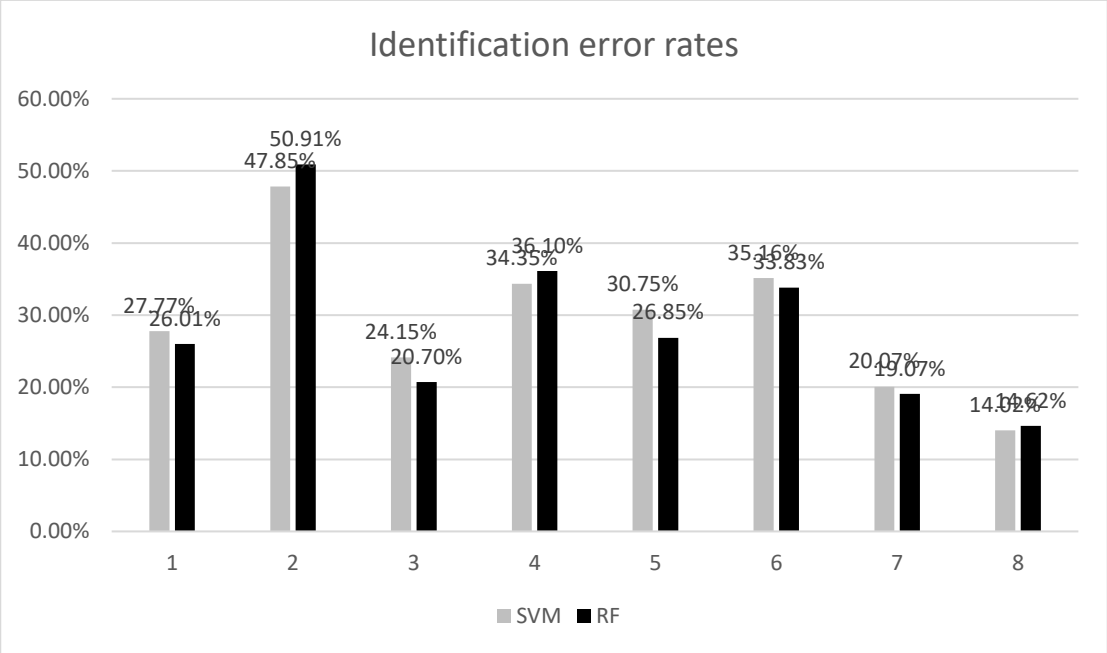

Figure 1: Classification results using the tone motif, word length motif and their combination as characteristics (1-8 on the horizon level represent the characteristics as shown in table 2)

From textural characteristics 1 and 5 in Table 2, we can see that the bigrams of word-final tone motif can't improve the classification result when they combines word-final tone motifs. So we can say that the bigrams of word-final tone motifs can't help to differ different authors and influence the classification results. Maybe this is because the bigrams of word-final motif are sparse.

Although, the identification error rate of classification result is very high when only the segment-

final tone motifs were used as the textual measurement, the combination of them and word-final tone motifs can reduce the identification error rate. This is the unexpected and interesting result. Compared with SVM, the classification model established using random forest can obtain the good classification result.

Combination of word length motifs and word-final tone motifs can make a relative low identification error rate.

From Table 2 and Figure 1, we can see that the classification result is well when the combination of segment-final and word-final tone motifs and word-length motifs is selected to represent the different texts from different authors.

Classification And Regression Trees (CART) was selected to establish a classification tree, as shown in Figure 2, using combination of word-final tone motifs and segment-final tone motifs and word length motifs as text characteristics. The tree outlines a decision procedure for determining the author of the texts.

In Figure 2, the leaf nodes specify a partition of the data, i.e. a division of the data set into a series of non-overlapping subsets that jointly comprise the full data set (Baayen 2008). For any node, the most useful predictor was selected to split it, for example word length motif (1-2-2, represented by x47). From the classification tree, we can see that the few predictors can roughly determine the authors of the texts. This conforms to the classification results using SVM and random forest establish the classification model.

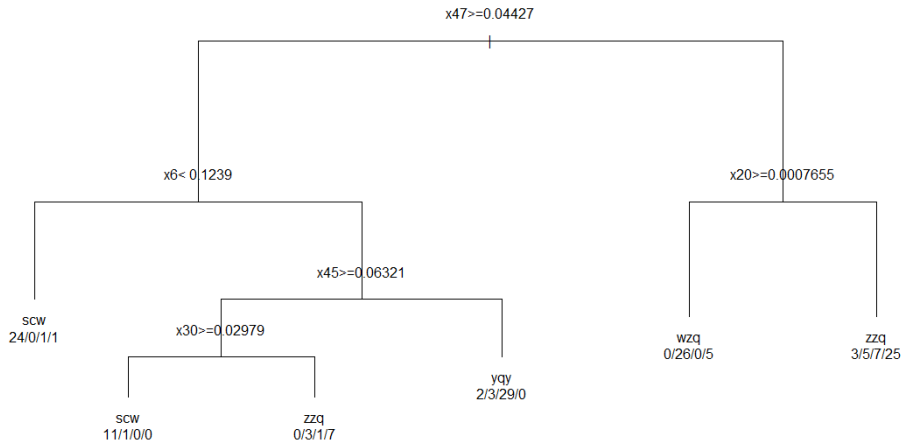

Figure 2: CART tree for the identification of authors

4 Conclusion

Chinese is a tonal language where tones, just like other lexical features, differentiate meanings. Most of previous studies in the Chinese stylometric analysis selected features at the words level or higher level as the textual measurement to identify the authors of the texts. Some examples of the selected features included words and syntactic information of the texts. Very few studies select the sub-lexical features mark the writing style of an author. In this study, we examine whether the Chinese tones motifs and word length motifs can be used as the stylometric characteristics. The tone motifs and word length motifs are both lexical feature that can be linked from other lexical resources and do not required annotated texts.

After comparing the classification results when using all the mentioned linguistic characteristics represent texts respectively, the experiments show that the combination of word-final tones motifs and segment-final tones motifs and word length motifs can effectively differentiate texts from these selected four authors.

The most important feature of our proposed methodology is the linked data approach without any dependence on annotated data or complex text processing, such as PoS tagging or parsing. Note complex processing introduces errors that can be propagated and that requirement of annotated data often lead to data sparseness problems. Our proposed methodology can apply to any plain text, as long as a link to existing lexicon for tonal and word-length information. The tonal and word-length information are inherent information carried by the word, the basic textual elements; hence the methodology is applicable to unannotated big data and will have wide applications in nearly all forms of big data as well as literary texts.

Funding

Reference

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. New York: Cambridge University Press.

Boroda, Moisei (1982): Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Jurij K./Boroda, Moisei G./Nadarejšvili, Isabela Š.[eds.]: Sprache, Text, Kunst. Quantitative Analysen. Bochum: Brockmeyer, 231-262.

Chao, Yuen Ren. (1968). A Grammar of Spoken Chinese. Berkeley and Los Angeles: University of California Press.

Chen, H. H. (1994). The contextual analysis of Chinese sentences with punctuation marks. Literary and linguistic computing, 9(4), 281-289

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp. 167-176.

García, A. M., & Martin, J. C. (2006). Function words in authorship attribution studies. Literary and Linguistic Computing. Vol. 22, No. 1, 49-66.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and linguistic computing, 22(3), 251-270.

Herdan, G. (1966). The advanced theory of language as choice and chance. New York: Springer-Verlag.

Ho, James. (2015). From the Use of Three Functional Words “的, 地, 得” Examining Author’s Unique Writing Style – And on Dream of Red Chamber Author Issues. BIBLID. 120:1, 119-150.

Holmes, D. I. (1994). Authorship attribution. Computers and the Humanities, 28(2): 87-106.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing. 13(3), 111-117.

Hou, R., Huang, C. & Liu, H. (2017). A study on Chinese register characteristics based on regression analysis and text clustering. Corpus Linguistics and Linguistic Theory, 0(0),

https://doi:10.1515/cllt-2016-0062.
Huang, Chu-Ren and Dingxu Shi. 2016. A Reference Grammar of Chinese. Cambridge: Cambridge University Press.
Huang, C.-R. & K.-J. Chen. (2017). Sinica Treebank. In N. Ide and J. Pustejovsky (eds), Handbook of Linguistic Annotation. Berlin & Heidelberg: Springer.
Juola, P. (2008). Author attribution, Foundations and Trends in Information Retrieval, 1(3), 233–334.
Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1), 9-26.
Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: J. Genzor and M. Bucková [Eds.]: Favete linguis. Studies in honour of Victor Krupa. Slovak Academic Press, Bratislava, 145-152.
Köhler, R. and S. Naumann (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Burkhardt, Schmidt-Thieme, Decker [Hrsg.]: Data Analysis, Machine Learning and Applications. Berlin, Heidelberg: Springer, S. 637-646.
Köhler, Reinhard; Naumann, Sven (2010). A syntagmatic approach to automatic text clas sification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, Peter; Kelih, Emmerich; Mačutek, Ján (eds.), Text and Language: 81-89. Wien: Prae sens.
Köhler, R. (2015). Linguistic Motifs. Sequences in Language and Text, 69, 89
Love, H. (2002). Attributing authorship: An introduction. Cambridge University Press.
Lu, Jianming. (1993). The features of Chinese sentences. Chinese Language Learning. No.1, 1-6.
Mosteller, F., and D.L. Wallace. Inference and Disputed Authorship: The Federalist. Reading. Reading, Mass: Addison Wesley, 1964.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. https://www.R-project.org
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational linguistics, 26(4), 471-495.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology. 60(3), 538-556.
Tan, Pang-Ning, Michael Steinbach & Vipin Kumar (Translated by Ming Fan & Hongjian Fan). (2006). Introduction to Data Mining. P115. Beijing, China: Posts & Telecom Press
Wang, K., & Qin, H. (2014). What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. Corpus Linguistics and Linguistic Theory, 10(1), 57-77.
Wang, Shao-kang, Dong Ke-jun & Yan Bao-ping. (2011). Research on Authorship Identification Based on Sentence Rhythm Feature. Computer Engineering. Vol.37, No.9. 4-5 +8.
Wei, Peiquan. (2002). From the distribution of common words examining the author issue of Dream of Red Chamber Author. Memorial Li Fanggui's 100th Anniversary International Symposium on Chinese History. Seattle: University of Washington.
Xiao, Tianjiu & Liu, Ying. (2015). A stylistic analysis of Jin Yong's and Gu Long's fictions based on text clustering and classification. No. 5: 167-177.
Zhu, Dexi. (1982). Lectures on Grammar. Beijing, China: Commercial Press.