Objective measurement of the relationship between variants in classical literature

Yuuki Tachioka
The College of Humanities and Sciences, The Nihon University
Chiyoda-ku, Tokyo, Japan
Email: yuuki_tachioka@yahoo.co.jp

***Abstract*—Stylometrics is a method of analyzing the style of a text using metric features. To apply this to classical literature, it is required that the diversity of variants of the same work is sufficiently smaller than that between different works because for the same work there are many variants, which have been changed from the original form. In this paper, this prerequisite will be confirmed, and the use of the Chinese character ratio, which is affected by the original form, Levenshtein distance, and perplexity is introduced. The experimental results show that the Chinese character ratio is effective for discriminating series of variants and that the Levenshtein distance and perplexity are also effective in addition to principal component analysis of features, which is general in Stylometrics. Especially, by using perplexity, the diversity between variants can be quantitatively compared in different works.**

***Keywords*-Stylometrics; variants of text; principal component analysis; Levenshtein distance; perplexity**

I. Introduction

Stylometrics is a method for analyzing the style of a text using statistical methods [1]. For Japanese these types of research are more difficult than for English because analyses of agglutinative languages, such as Japanese, need part-of-speech (POS) tagging, which divides sentences into words and gives them the POS of the words; but development of the POS tagger enables mechanical POS tagging, and currently the amount of this type of research (e.g., [2]) is increasing.

There are some studies dealing with classical literature but these types of research use reprinted texts. However, to deal with classical literature quantitatively, it is essential to consider variants. There are few original manuscripts written by the authors of classical literature. Most of the texts have been preserved by being transcribed repeatedly, and errors, reformations, and additions change the original contents. The metric analyses are used to discriminate between works, but it is required that a diversity of variants of the same work is sufficiently smaller than that between works. This prerequisite of the research has not been confirmed. This paper focuses on the Japanese classics, but the same problems occur in other languages. The objects of this research are four variants of "Izumishikibu Nikki" with comparison to another work ("Sarashina Nikki"). This paper introduces the use of the Chinese character ratio, which reflects the original usage, Levenshtein distance, and perplexity, and shows that these reflect the degree of diversity of the variants or works.

II. PCA of text features for style analysis

Style analysis uses some features that can discriminate between works effectively. Though general features are shown in IV, here the Chinese character ratio (the ratio of Chinese characters to all characters) is introduced to discriminate between different variants of the same work. Since Japanese sentences consist of Chinese characters and hiragana (Japanese inherent characters), the usage of Chinese characters is affected by the original usage and this ratio characterizes the series of variants. Features' dimensions are reduced by the principal component analysis (PCA) [3].

III. Objective measurement of variants using Levenshtein distance and perplexity

To calculate the diversity of variants, Levenshtein distance is introduced. An arbitrary sentence can be converted to any by three procedures: substitution, insertion, and deletion. The Levenshtein distance is defined as the minimum steps required (i.e., cost) to convert one sentence to another. Dynamic programming (DP) enables fast calculation.

The basic model that analyzes the sentences is N-gram, which is a chain probability of words or characters. The N-gram model is more flexible than Levenshtein distance because, if contents are totally different, texts can be compared quantitatively by perplexity defined as $P(w_1,\dots,w_N)^{\frac{1}{N}}$ where P represents a probability of observed N words sequence $w_1,\dots,w_N$ and perplexity is an inverse of its geometrical mean. Perplexity is related to the number of the next word candidates assuming that the emerging probability is the same. For texts which can be easily estimated by N-gram models, perplexity is low. Using this property, similarity of texts can be quantitatively evaluated.

It is inefficient to deal with variants by respective databases because the variants have a similarity. Here, a data structure which can deal with different texts in a single database is proposed as $\backslash\mathrm{d}\{[\mathrm{t}_i]\mathrm{text}_i@\mathrm{text}_0\}$ where text_i is a different part compared with the base text (text_0). For example, there are four different sentences (AAD, ABD, ABD, ACD) compared with the base sentence (ACD). This part is integrated into the form: $\mathrm{A}\backslash\mathrm{d}\{[\mathrm{t}_1]\mathrm{A}[\mathrm{t}_2\mathrm{t}_3]\mathrm{B}@\mathrm{C}\}\mathrm{D}$. This form also enables fast computation of the Levenshtein distance because the DP alignments are easily arranged.

978-0-7695-5047-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CultureComputing.2013.65

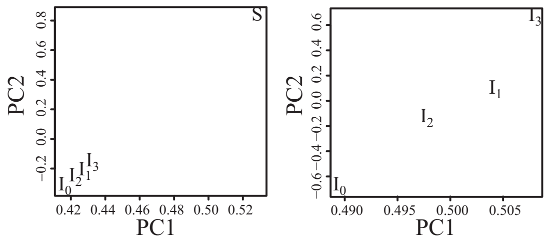

Figure 1. Principle components mapped down from 16 metric features where I_0–I_3 are variants of “Izumishikibu Nikki” and S is “Sarashina Nikki”. (left: PCA among I_0–I_3 and S, right: PCA among I_0–I_3)

Table I
LEVENSHTEIN DISTANCE BETWEEN FOUR VARIANTS.

	Sanjo (I_0)	Kangen (I_1)	Ouei (I_2)	Konsei (I_3)
Sanjo	0	4916	5480	5751
Kangen	4916	0	5412	5148
Ouei	5480	5412	0	3305
Konsei	5751	5148	3305	0

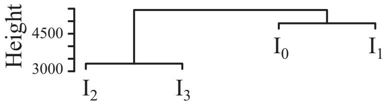

Figure 2. Dendrogram based on the Levenshtein distance.

IV. RELATIONSHIP BETWEEN VARIANTS OF THE SAME WORK WITH COMPARISON TO ANOTHER WORK

To clarify the relationship between variants, this paper focused variants of “Izumishikibu Nikki” and “Sarashina Nikki”, which are of similar length. “Izumishikibu Nikki” has four series of variants: Sanjonishi-ke (Sanjo), Kangen, Ouei, and Konsei. Sanjo is the oldest and the most popular. Three other variants were compared with Sanjo after constructing an aforementioned form database with reference to [4]. The base text of “Sarashina Nikki” was Teika-bon, which is the most common. For POS tagging, [5] was used. The tagging errors (up to 5%) were manually modified.

The number of features is 16 as follows: The first four are the sentence length, the quotation ratio (Japanese poem and conversation), the ratio of representing one’s feelings directly[1] and additionally the Chinese character ratio; after POS tagging, twelve other features are added to analyze a style of text: modifier-verb ratio[2] and other POS ratios (the ratio of independent word, pronoun, adjective, nominal adjective, adverb, noun, verb, Japanese-origin word, Chinese origin-word, mixed-origin word, and proper noun). According to the text style analyses, the Chinese character ratio (I_0: 7.2%, I_1: 8.3%, I_2: 8.6%, I_3: 10.7%, S: 9.1%) is effective for discriminating between variants where I_0, I_1, I_2, and I_3 denote Sanjo, Kangen, Ouei, and Konsei, respectively, whereas S denotes Sarashina. On the other hand, the quotation ratio (47.5%, 46.6%, 47.4%, 46.4%, 33.2%), the ratio of parts representing one’s feelings (12.2%, 12.3%, 12.4%, 12.8%, 4.1%), and the noun ratio (37.1%, 37.3%, 36.6%, 36.7%, 46.6%) are effective for discriminating between works. Principle components obtained by principal component analysis depict the relationship as shown in Fig. 1. Attribution ratio is 100% by these two principal components. PCA can distinguish I_0–I_3 from S, but the relationship among I_0–I_3 is not clear quantitatively because the value range is different at each feature and the Euclid distance between points on the graph is meaningless. This shows the limitation of PCA.

Table II
PERPLEXITY USING WORD TRI-GRAM OF FOUR VARIANTS OF “IZUMISHIKIBU NIKKI” COMPARED WITH “SARASHINA NIKKI”.

train \ eval	Sanjo	Kangen	Ouei	Konsei	Sarashina
Sanjo	8.9	35.2	35.6	41.4	198.5
Kangen	36.4	9.0	40.0	40.2	205.7
Ouei	37.2	40.3	9.3	26.7	195.3
Konsei	44.2	41.1	26.6	9.5	204.1
Sarashina	242.6	241.6	244.3	235.9	9.0

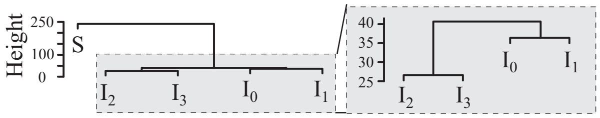

Figure 3. Dendrogram based on perplexity.

Table I shows the Levenshtein distance between four variants of “Izumishikibu Nikki”. Fig. 2 shows the relationship between variants. Levenshtein distance clarifies the relationship between variants quantitatively, but this technique cannot be applied to other works (e.g., “Sarashina Nikki”). Prior alignment should be arranged to some extent manually (e.g., per sentence) before DP. If not, mis-alignment would decrease accuracy. On the other hand, Table II and Fig. 3 show the perplexity and its dendrogram using a tri-gram model between variants with the different work. Perplexity calculation only needs a tri-gram model after “correct” POS tags are obtained. One text is selected for training and tri-gram models are constructed by [6], and the other texts are evaluated in terms of perplexity. Dendrograms obtained by using Levenshtein distance and perplexity are similar.

V. CONCLUSION

This paper aims to clarify the relationship between variants of texts. Experiments show that the Chinese character ratio is effective for discriminating between variants and that the Levenshtein distance and perplexity are also effective. Especially, using perplexity, the relationship between variants can be quantitatively compared to different works.

ACKNOWLEDGMENT

The author would like to thank Prof. Ogino at Nihon Univ.

REFERENCES

[1] A. Kenny, *The computation of style*. Pergamon Press, 1982.

[2] M. Jin. and M. Murakami, “Authorship identification using random forests,” *Proc of the Inst of Stat Math (J)*, vol. 55, pp. 255–268, 2007.

[3] D. Kaplan, “A computational approach to style in American poetry,” *in Proc Int Conf on Data Mining*, pp. 553–558, 2007.

[4] T. Ito, *Izumishikibu Nikki*. Izumi Shoin, 1991.

[5] “http://www2.ninjal.ac.jp/lrc/”.

[6] “http://www.speech.sri.com/projects/srilm/”.

[1] These two features show whether description is direct/indirect-oriented.

[2] The ratio of the number of modifiers (adjective, nominal adjective, adverb, and pronoun adjectival) to the number of verbs: A higher value shows static-oriented and a lower value shows dynamic-oriented.