Ghostwriter identification in Yasunari Kawabata's works in the 1960s

Hao Sun
sonnkou1985@gmail.com
Doshisha University, Japan

Mingzhe Jin
mjin@mail.doshisha.ac.jp
Doshisha University, Japan

Introduction

Yasunari Kawabata was a Japanese novelist who received the Nobel Prize for Literature in 1968. He was famous for his masterpieces such as *Snow Country*, *The Sound of the Mountain*, *The Old Capital*, *House of the Sleeping Beauties,* and so on. Kawabata had a shattered childhood. He was orphaned at five years old, and his other relatives including his grandfather, grandmother, and elder sister also passed away before he was fourteen. The successive deaths of his loved ones induced mental problems, which became worse in the 1960s. He became addicted to sleeping pills during those years. However, two of Kawabata's masterpieces *The Old Capital* and *House of the Sleeping Beauties* were published during his sleeping pills addiction period. These two novels were suspected as having been written by ghostwriters because it was hard to imagine that Kawabata could continue writing novels in his mental condition. There are already some pieces of evidence for the ghostwriter issue of *The Old Capital* and *House of the Sleeping Beauties*. Kawabata sent a letter to Sawano before the publication of *The Old Capital*. In the letter he wrote: "I have accepted to write a novel about Kyoto; the deadline is approaching but I even don't know how to start. Sawano was surprised when he received this letter. He went to Kyoto to meet Kawabata and gave advice on how to write *The Old Capital*. Itasaka mentions in his book that *The Old Capital* and *House of the Sleeping Beauties* were actually written by ghostwriters (Itasaka, 1997). *The Old Capital* may have been written by Kawabata's three disciples whose names are Hisao Sawano, Makoto Hokujyo, and Yukio Mishima. *House of the Sleeping Beauties* may have been written by Yukio Mishima. In this study, we show strong evidence suggesting the real author of *The Old Capital* and *House of the Sleeping Beauties* from a data analysis approach.

Method

The method of this study includes three main steps. Firstly, we digitized more than ten novels of both Kawabata and the three possible ghostwriters. Then, we extracted stylometric features from the novels, and all chapters of *The Old Capital* and *House of the Sleeping Beauties*. Finally, we applied the unsupervised and supervised methods to infer the possible author of *The Old Capital* and *House of the Sleeping Beauties*.

We used bigrams of characters and punctuation marks, part-of-speech bigrams, and phrase patterns as stylometric features, which have been proven useful in Japanese authorship attribution (Matsuura and Kanada, 2000; Jin, 2003, 2013).

Bigrams of characters and punctuation marks are pairs of two adjacent characters or punctuation marks extracted from plain text. Japanese texts should be tokenized previously for the extraction of part-of-speech features. We applied the Japanese morphological analyzer called MeCab to separate a Japanese sentence into morphemes. MeCab outputs the parts-of-speech of words in several layers. Deeper layers process more detailed information. In this study, we use information from the first layer. As an example, part-of-speech bigrams in the Japanese sentence "Ronbun wo kaku." are "Noun_Particle," "Particle_Verb," and "Verb_Period."

Phrase pattern is a powerful feature that can be extracted in terms of syntax. A Japanese parser (CaboCha) was introduced to separate sentences into phrases. Phrase pattern is defined as the smallest unit that divides the sentence into unnatural parts (Jin, 2013). It is a combination of two parts. One is the original form of the particles and punctuation marks, while the other is the parts-of-speech of the other materials, except for the particles and punctuation marks in the same phrase. The two phrase patterns in the sentence "Ronbun wo kaku." are "Noun_wo" and "Verb_Period."

We applied unsupervised methods and the integrated classification algorithm in this study. The idea of the integrated classification algorithm was to combine the results of several stylometric features and classifiers. It achieved highest classification accuracy in authorship attribution of literature (Jin, 2014). The integrated classification algorithm combines the results of stylometric features and classifiers to avoid the bias under a majority vote rule. AdaBoost (ADA), High-dimensional Discriminant Analysis (HDDA), Logistic Model Tree (LMT), Random Forest (RF), and

Support Vector Machine (SVM) were used as the base classifiers.

Results

The result of *House of the Sleeping Beauties* reveals that compared to Mishima, all chapters of *House of the Sleeping Beauties* are more likely to be written by Kawabata. The result in the classification between Kawabata and Hokujyo shows that the writing style in all chapters of *The Old Capital* is more like Kawabata's.

Bibliography

Itasaka G. (1997*). Gokusetsu Mishimayukio-Seppuku to furamenko.* Natsumesyobo Press.

Matsuura, T. and Kanada, Y. (2000). Identifying Authors of Sentences in Japanese Modern Novels via Distribution of N-grams. *Mathematical linguistics,* 22: 225-238.

Jin, M. (2003). Authorship Attribution and Feature Analysis Using Frequency of JOSHI with SOM. *Mathematical linguistics*, 23: 369-386.

Jin, M. (2013). Authorship Identification Based on Phrase Patterns*. The Japanese Journal of Behaviormetrics*, **40:** 17-28.

Jin, M. (2014). Using Integrated classification Algorithm to Identify a Text's Author. *The Japanese Journal of Behaviormetrics*, 41**:** 35-46.