English Review of
斉藤鉄也著『仮名字母の出現傾向を用いた池田本源氏物語の調査』

Evaluation 1 Paper for ASIAN 573, 2023 Spring

Ruixuan Tu
ruixuan.tu@wisc.edu
University of Wisconsin-Madison

21 February 2023

Introduction: In this paper [1], Saito has analyzed the kana 仮名 usage of Ikeda-bon 池田本 of *the Tale of Genji* 源氏物語 ("the tale'' in the following) by PCA, hieratical clustering, and n-gram statistics. Saito has discovered that for Ikeda-bon 池田本 (1) there could be two copyists (denoted as copyist A and copyist B); (2) there might be a correlation between the two chapters *Kiritsubo* 桐壺 and *Hananoen* 花宴 done by the copyist A. (p. 121)

Motivation: As I have majors in Computer Sciences, Data Science, Statistics, and Japanese, I am interested in this field of Japanese interpretation using statistics and machine learning, so I select this paper.

Methods: In the PCA setting, the author uses chapters with over 5000 characters as the dataset and analyzes 88 jibos 字母 of kana 仮名, by extracting the most significant projection of the vectors (i.e., jibos 字母 in chapters, of the dimension of the chapter length) on a 2D plane, and Saito has divided the points after projection to five groups by k-means (p. 122). In the hieratical clustering setting, Saito uses IR distance for comparing two chapter text vectors and the average linkage method (p. 123).

Discoveries and Interpretations: From the PCA (図 2, p. 122), I can see the two groups are close on the left side by copyist A, and there is one group containing almost all chapters by copyist B, showing the distinction between the copyists. From the hierarchical clustering (図 3, p. 123), I can find *Kiritsubo* 桐壺 and *Hananoen* 花宴 in the same cluster in the first level, standing out from the large cluster of all other chapters by copyist A. However, this paper fails to give the special jibos 字母 (i.e., features) of the two chapters from the n-gram analysis (p. 127).

Evaluation and Reflection:

The yields of this analysis are convincing to me, as I can see the clustering results from the graphs. Take the leading eight chapters as an example, only two (Murasaki 若紫 and Suetsumuhana 末摘花) are copied by copyist B, and these two chapters are both witty and regarding them as a whole, then this part of the tale could be seen as a relief for the readers from the heavy plot of the refusal from Utsusemi 空蝉 and the death of Yūgao 夕顔 in the previous two chapters, which should also suggest the difference.

As I was more into machine learning, this is the first time I know there could be multiple copyists for one book. But some of the methodologies in this paper using is outdated in machine learning, with the non-yielding n-gram detailed but not clearly explained in its results (p. 124-127). As n-gram only traces a few jibos 字母 in context, we might use further context in sentences to find the specialties of the two chapters, which can be archived by new methods like Transformer. In this paper, a single-phone kanji 漢字 is regarded as jibos 字母 in analysis (p. 125), which I think might not be a good choice for mixing the phonogram and logogram, as they are usually representing different things in language.

For the remaining question on the two chapters, there are some similarities in the plot: chapter *Kiritsubo* 桐壺 and chapter *Hananoen* 花宴 are both around Kiritsubo no kōi 桐壺更衣 and the family of Udaijin 右大臣 (minister of the right), but chapter *Kiritsubo* 桐壺 is much longer than chapter *Hananoen* 花宴 to include

almost every protagonist in the first part of the tale, and chapter *Hananoen* 花宴 has Oborozukiyo 朧月夜 and Murasaki 若紫 who are not introduced in chapter *Kiritsubo* 桐壺. Thus, I think the two chapters could also be a balanced workload of two standard chapters (not including these two short ones) for copyist A, from the perspective of workload, which can be used to explain the similarity, but not the specialty of the two chapters.

Values and Limitations: For researchers in machine learning or/and Japanese literature, this is an interdisciplinary paper that could be a fresh perspective and may give some unforeseen results. However, for literature researchers, this paper has unfamiliar quantitative analysis, and for machine learning researchers, they might not have read the tale, especially the two chapters (*Kiritsubo* 桐壺 and *Hananoen* 花宴), to understand the research topic.

Related Works: There are other papers [2] [3] [4] under the same author that researches the same topic on different copies of the tale.

Explanations of Terms in Paper:

Term	Explaination / Definition
kana, kanji, jibo [5]	kana 仮名, a Japanese phonetic writing system invented in the Heian period 平安時代, used to be written in scripting font of kanji 漢字, i.e., Chinese characters, incluing a set of Hentaigana 変体仮名 which is not currently used. Take the kana no の (in modern) as an example, the jibo 字母 (kanji base) of this form is 乃, and the kana was also written as 乃 (based on 乃), 能 (based on 能), and 農 (based on 農). Thus, there are many choices of scripts to use when one copies a text, which is a personal preference of copyist for research.
k-means clustering [6] (p. 6-9)	an algorithm to randomly pick k points as centroids, and generate clusters of closest points other than centeroids from the centeroids
hierarchical clustering [6] (p. 10-21)	an algorithm to generate $n-1$ clusters for n vectors, a linkage method combines the closest two clusters into one, so we can say a relation between any two chapters is closer if they are in a cluster at a lower level
PCA (Principal Components Analysis) [7] (p. 30-45)	a technique for extracting variance structure from high dimensional datasets as well as reducing dimensionality by orthogonal projection of the data recursively
IR distance between vectors $\mathbf{x}$ and $\mathbf{y}$ [8]	$d_{IR}(\mathbf{x},\mathbf{y}) = \sum_i \left(x_i \log\left(\frac{2x_i}{x_i+y_i}\right) + y_i \log\left(\frac{2y_i}{x_i+y_i}\right)\right)$
cosine distance between vectors $\mathbf{x}$ and $\mathbf{y}$ [8]	$d_{\cos}(\mathbf{x},\mathbf{y}) = 1 - s(\mathbf{x},\mathbf{y})$, where $s(\mathbf{x},\mathbf{y})$ is the cosine similarity between the two vectors, defined by $s(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$
n-gram [9] (p. 11)	limit context length to n words before the current sequence (e.g., word), and count the number of occurences
Transformer [10] (p. 12)	a sequence-to-sequence model based entirely on attention (i.e., long-sequence context)

References

[1] Saito Tetsuya 斉藤鉄也. "Kana jibo no shutsuken keikō o mochiita Ikeda-bon *Genji monogatari* no chōsa" 仮名字母の出現傾向を用いた池田本源氏物語の調査. Jinbun kagaku to konpūta shinpojiumu ronbunshu 人文科学とコンピュータシンポジウム論文集 (2020): 121-128. (https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_action_common_download&item_id=208687&item_no=1&attribute_id=1&file_no=1)

[2] Saito Tetsuya 斉藤鉄也. "Kana jibo no shutsuken keikō o mochiita Ōshima-bon *Genji monogatari* no chōsa" 仮名字母の出現傾向を用いた大島本源氏物語の調査. Jinbun kagaku to konpūta shinpojiumu ronbunshu 人文

科学とコンピュータシンポジウム論文集 (2019): 157-164. (https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_action_common_download&item_id=201088&item_no=1&attribute_id=1&file_no=1)

[3] Saito Tetsuya 斉藤鉄也. “Kana jibo no shutsuken keikō o mochiita Nidai sanjo nishike-bon *Genji monogatari* no chōsa” 仮名字母の出現傾向を用いた日大三条西家本源氏物語の調査. Jinbun kagaku to konpūta shinpojiumu ronbunshu 人文科学とコンピュータシンポジウム論文集 (2018): 59-66. (https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_action_common_download&item_id=192443&item_no=1&attribute_id=1&file_no=1)

[4] Saito Tetsuya 斉藤鉄也. “Kana jibo no shutsuken keikō o mochiita Bijūke kawachi-bon *Genji monogatari* no chōsa” 仮名字母の出現傾向を用いた尾州家河内本源氏物語関連写本の調査. 情報処理学会論文誌, Vol. 61, No. 2 (2020): 144-151. (https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_action_common_download&item_id=203133&item_no=1&attribute_id=1&file_no=1)

[5] Seishindo 誠心堂書店. “Hentaigana wo shiraberu rekishiteki kana shotai wo sagasu” 変体仮名を調べる歴史的仮名書体を探す. (http://www.book-seishindo.jp/kana/)

[6] Ilias Diakonikolas. Unsupervised Learning I. CS 760: Machine Learning, University of Wisconsin-Madison, Fall 2022. (http://www.iliasdiakonikolas.org/teaching/Fall22/slides/lec19.pdf)

[7] Ilias Diakonikolas. Unsupervised Learning II. CS 760: Machine Learning, University of Wisconsin-Madison, Fall 2022. (http://www.iliasdiakonikolas.org/teaching/Fall22/slides/lec20.pdf)

[8] Jin Minzhe 金明哲. “Tōkeiteki tekisuto kaiseki (13) – tekisuto no kurasutā bunseki” 統計的テキスト解析(13)～テキストのクラスター分析～. Furī sofuto niyoru dēta kaiseki & mainingu フリーソフトによるデータ解析・マイニング 67. (https://mjin.doshisha.ac.jp/R/68/68.html)

[9] Junjie Hu. Language Modeling. CS 769: Natural Language Processing, University of Wisconsin-Madison, Spring 2022. (https://junjiehu.github.io/cs769-spring22/assets/pdf/anlp-03-lm.pdf)

[10] Junjie Hu. Attention and Transformer. CS 769: Natural Language Processing, University of Wisconsin-Madison, Spring 2022. (https://junjiehu.github.io/cs769-spring22/assets/pdf/anlp-08-attention.pdf)