

Personal Name Extraction from Ancient Japanese Texts

Mamoru Yoshimura
Graduate School of Information
Science and Engineering, Ritsumeikan
University, Japan
1-1-1 Noji-higashi, Kusatsu,
Shiga 525-8577, Japan
is046080@ed.ritsumei.ac.jp

Fuminori Kimura
Kinugasa Research Organization,
Ritsumeikan University, Japan
56-1 Toji-in Kita-machi, Kita-ku,
Kyoto, Kyoto 603-8577, Japan
fkimura@is.ritsumei.ac.jp

Akira Maeda
College of Information Science and
Engineering, Ritsumeikan University,
Japan
1-1-1 Noji-higashi, Kusatsu,
Shiga 525-8577, Japan
amaeda@is.ritsumei.ac.jp

ABSTRACT

Text analysis of ancient Japanese language is difficult due to the lack of language tools to segment a sentence into words. There exists some morphological analysis tools for ancient Japanese in a specific period, but there are no such tools that can be used for general purpose. Even if morphological analysis tools were not available, it would be beneficial for a certain kind of text analysis to be able to extract named entities, such as personal names, from ancient Japanese texts. In this paper, we propose a method of personal name extraction from ancient Japanese texts based on Support Vector Machine (SVM) using features of character appearance and probabilistic word segmentation information. Experimental results showed that our proposed method were able to extract personal names from ancient Japanese texts with approximately 4% better F-measure when utilizing our proposed word segmentation information.

Categories and Subject Descriptors

H.2.8 [Data mining]

General Terms

Algorithms, Experimentation.

Keywords

support vector machine, named entity extraction, chunking, ancient Japanese texts, personal name

1. INTRODUCTION

Ancient writings are increasingly being digitized in text form. This leads to a possibility of applying natural language processing techniques to digitized ancient writings. Natural language processing techniques for modern Japanese relies on morphological analysis tools in order to separate words from sentences, to identify the part of speech of a word, and so on. The situation is the same for ancient Japanese. However, it is usually impractical to use a morphological analyzer designed for modern language to analyze text written in an ancient language. It is also difficult to segment sentences into words, because there are no dictionaries that can be used for morphological analysis except for Japanese in some specific periods.

However, the following things become possible, if it is able to

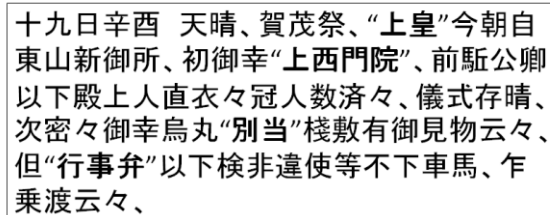
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '10, Month 1–2, 2010, City, State, Country.
Copyright 2010 ACM 1-58113-000-0/00/0010 ...\$15.00.

extract personal names from ancient Japanese texts. One is to utilize it for the construction of ancient Japanese dictionaries. Another is to utilize it for text analysis and text mining of ancient Japanese writings.

In this paper, we propose a method of personal name extraction from ancient Japanese texts. This method uses Support Vector Machine (SVM) in order to learn the rules for named entity extraction automatically. We used the term “personal name” to describe a person name, an alias of a person, and the official position name for a person.

We use three Japanese ancient writings, namely “Hyohanki”, “Azumakagami” and “Gyokuyo” as the corpora. These ancient writings were written between late Heian era and early Kamakura era (12th-13th century). These writings are written in the style of Kanbun (a style of ancient Japanese which is based on classical Chinese). We conduct experiments of personal name extraction from these Japanese ancient writings in order to verify the effectiveness of the proposed method.



十九日辛酉 天晴、賀茂祭、“上皇”今朝自
東山新御所、初御幸“上西門院”、前駟公卿
以下殿上人直衣々冠人数濟々、儀式存晴、
次密々御幸烏丸“別当”棧敷有御見物云々、
但“行事弁”以下檢非違使等不下車馬、乍
乘渡云々、

Figure 1. Excerpt of “Hyohanki”.

2. RELATED WORK

2.1 Support Vector Machine

Support Vector Machines (SVM) is one of the supervised learning models with associated learning algorithms that analyze data and recognize patterns. SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. We used “LIBLINEAR” [1] for implementing the proposed method. “LIBLINEAR” is a machine learning library that is specialized for linear prediction.

2.2 Named Entity Extraction

In the methods of named entity extraction using SVM, it is popular to divide an input sentence into analysis units of proper sizes (tokens) and to group one or more analysis units into one token. The statuses of these grouped tokens are represented by a chunk tag set. “IOB2” chunk tag set is one of the best performed chunk tag set in previous studies. In “IOB2” chunk tag set, “B” tag is attached to the first token of a named entity, “I” tag is

attached to the following tokens in the named entity after “B”, and “O” tag is attached to tokens that are not a part of named entities.

In this paper, we use these tags as personal name tags. By doing so, personal name extraction is regarded as the learning of rules to classify each token in the input sentence into one of personal name tags. Table 1 shows an example of personal name tag attachment for a part of a sentence in “Hyohanki”. The character string “上皇” is a personal name (in this case, a official position that means a retired emperor). The first character “上” is attached the “B” tag, the following character “皇” is attached the “I” tag.

Table 1. An example of personal name tagging.

character	IOB2 tag
上	B
皇	I
今	O
朝	O
自	O

2.3 Named Entity Extraction using Support Vector Machine

Yamada et al. [2] proposed the grouping of tokens using SVM. This research reported that SVM is effective for the grouping of tokens. However, personal names shorter than a morpheme will not be able to be extracted if the result of morphological analysis is used as the analysis unit.

Asahara et al. [3] proposed a method that adopts a character as the analysis unit in order to solve this problem. This method can perform named entity extraction even if there is a difference in word boundary between the result of morphological analysis and named entities. However, it is impossible to attach the part of speech information to each character directly. For this purpose, this method also uses “Start/End” (SE) chunk tag set, which represents the position in a word. In this method, “B” tag is attached to the first character of a word, “E” tag to the last character of it, and “I” tag to the middle of it.

Table 2 shows an example tag attachment by SE chunk tag set for a sample sentence. This sample sentence is a part of the text in “Hyohanki”. This sample sentence is divided into words as shown

in figure 2 using our proposed word segmentation method, which is explained in the Section 3.

- 「上皇今朝自東山新御所、」
- ↓
- 上皇 今朝 自 東山 新 御所、

Figure 2. Excerpt of the segmentation result.

Table 2. An example of tagging for positions in a word.

character	SE tag with segmentation result
上	B-上皇
皇	E-上皇
今	B-今朝
朝	E-今朝
自	S-自

3. APPLYING THE METHOD TO ANCIENT JAPANESE TEXTS

Named entity extraction from modern Japanese texts usually utilizes part of speech and morpheme information obtained from a morphological analyzer and inputs them into SVM to conduct leaning and estimation. It also uses information of scripts, i.e. hiragana, katakana, and kanji.

In our proposed method, we target ancient Japanese writings written in Kanbun (a style of ancient Japanese which is based on classical Chinese). We cannot utilize the results of morphological analysis for the proposed method, because there is no morphological analyzer for ancient Japanese in Kanbun style. Therefore, we cannot obtain sufficient information from these texts.

In order to solve this problem, we use the word segmentation method that from our previous research [4]. We calculate the likelihood of character n-grams to be a word, and extract character n-grams with higher likelihood as ancient Japanese words.

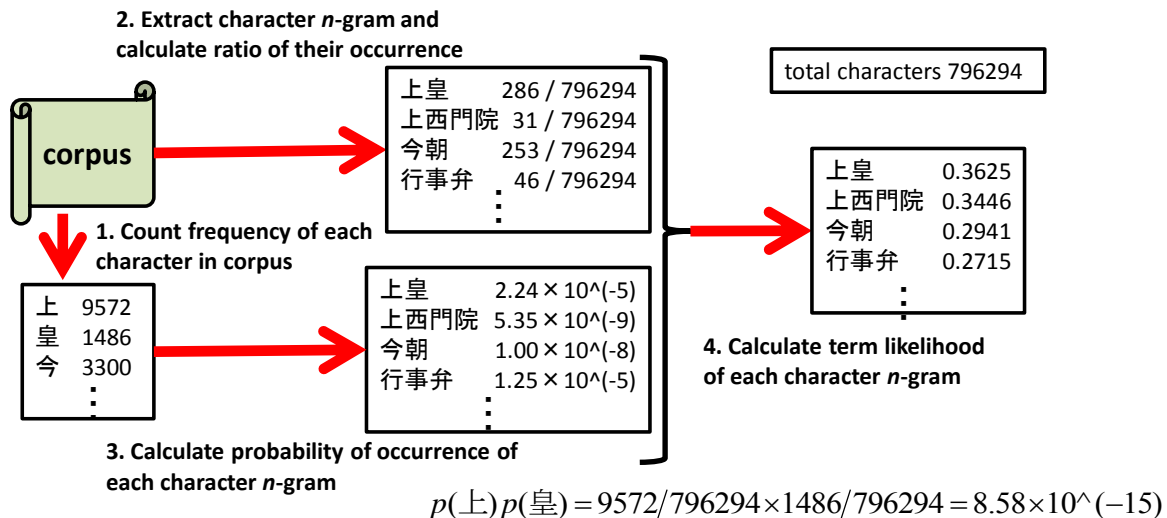


Figure 3. Processing flow of term likelihood calculation.

In this method, we first calculate the likelihood of each character n-gram to be a word, and then select the n-grams with high likelihood. These extracted character n-grams can be considered as possible words. We call this likelihood “term likelihood”. It is assumed that a string that is a correct word will appear more frequently than the strings generated by randomly combining characters that constitute the word. Therefore, the n-grams with higher term likelihood can be considered to have higher possibilities to be correct words.

Figure 3 shows the processing flow of the term likelihood calculation. First, we count the frequency of each character in the corpus. Second, we extract character n-grams from the corpus and calculate the probabilities of their appearances in the corpus. Third, we calculate the joint probability of characters in each n-gram. Finally, we calculate the term likelihood of each n-gram.

This word segmentation method has not achieved sufficient precision of segmentation. Besides, this method has the problem for longer analysis unit. This word segmentation method can find correct word boundary in a high accuracy but tends to divide words into smaller character strings than a word. Therefore, we adopt the same approach as [3], which uses character as the analysis unit. The difference is that we use our proposed segmentation method instead of morphological analyzer. We consider the segmentation results of this method as words. These results are utilized for attaching “SE” tag to each character in the text.

4. PROPOSED METHOD

Our proposed method first learns the extraction rules of personal names from an annotated training corpus, and then extracts personal names from ancient Japanese texts by using SVM. We adopt a character as the analysis unit, and group them in order to extract personal names. The proposed method uses the IOB2 tag set as the personal name tag, which is reported to be effective in [2]. The proposed method classifies each character into one of three personal name tags in order to extract personal names. Besides, we also adopt the word segmentation method mentioned in section 3. The proposed method uses the results of this word segmentation as words, and attaches the SE tags according to these results.

The proposed method consists of the following three steps:

- 1) Word segmentation of input text (described in section 3)
- 2) Feature extraction for chunking by the method using the SE tag set (described in section 4.1)
- 3) Classification and grouping of tokens by the method using the IOB2 tag set (described in section 4.2)

4.1 Feature extraction for chunking

In this subsection, we explain the features used to learn the rules of personal name extraction.

Each character has three pieces of information; its own character, SE tag with segmentation result, and IOB2 tag. IOB2 tag is attached in the last step. In this step, the features of each character consists of information in two characters before and two characters after the character, and the character itself.

Table 3 shows an example of features. This example is the case for the focusing character “今” when the input text is “上皇今朝自”. The gray cells in table 3 are the information used for learning “今” in the next step.

Table 3. An example of features. This example is the case for a character “今”.

position	character	SE tag with segmentation result	IOB2 tag (attached in next step)
i-2	上	B-上皇	B
i-1	皇	E-上皇	I
i	今	B-今朝	O
i+1	朝	E-今朝	O
i+2	自	S-自	O

4.2 Classification and grouping of tokens by SVM

In this step, the proposed method conducts the classification and grouping of tokens by SVM. The features extracted in the previous step are entered to SVM and the method estimates the personal name tag for each character. The gray cells in table 3 are used for the features of the i-th character. However, the personal name tags at the i-2 and i-1 positions are known when learning but unknown when testing. Therefore, the personal name tag estimated at each position is used as the feature for the next character. Since the estimation is done one character by one character from the beginning of a sentence, we have estimated personal name tags for i-2 and i-1 positions when estimating the tag for i-th character. For example, when estimating the tag for the character “今” in the i-th position, the estimated tags “B” for “上” at the i-2 position, and “I” for “皇” at the i-1 position is used.

5. EXPERIMENTS

We conducted experiments of our proposed method of personal name extraction from ancient Japanese texts. For the experiments, we used three ancient Japanese writings, namely “Hyohanki”, “Azumakagami”, and “Gyokuyou”. For these documents, there are personal name indices that are manually compiled by scholars and are available in digital form. These indices were used as the correct answers for both training and test data. Table 4 shows the number of characters and the number of personal names contained in each document used in the experiments. For evaluation, we calculated precision, recall, and F-measure for each document by the 5-fold cross-validation.

In order to verify how the information of term segmentation influences the extraction accuracy, we conducted comparable experiments of using/not using the feature of term segmentation results shown in Table 3. The results of these experiments are shown in Tables 5 and 6.

Table 4. The number of characters and the number of personal names in each document used in the experiments.

	Number of characters	Number of personal names
“Hyohanki”	796,294	22,488
“Azumakagami”	787,250	39,909
“Gyokuyou”	1,934,754	22,823

Table 5. The experimental results of not using the feature of term segmentation.

	Precision	Recall	F-measure
“Hyohanki”	0.6149	0.5272	0.5676
“Azumakagami”	0.6829	0.6117	0.6454
“Gyokuyou”	0.6697	0.5973	0.6314

Table 6. The experimental results of using the feature of term segmentation.

	Precision	Recall	F-measure
“Hyohanki”	0.6699	0.5956	0.6086
“Azumakagami”	0.7151	0.6697	0.6917
“Gyokuyou”	0.6857	0.6507	0.6678

We also conducted an experiment for examining the variation of accuracy when varying the amount of text used for training and testing. We used “Hyohanki” for this experiment. We calculated F-measure for several different portions of “Hyohanki” text by the 5-fold cross-validation. The experimental conditions are the same as the previous experiments except for the amount of text used. The result of this experiment is shown in Figure 4.

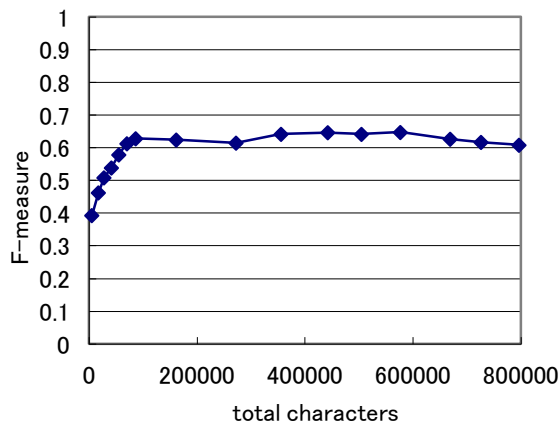


Figure 4. The relations of accuracy and the number of text used.

6. DISCUSSION

Table 5 and 6 show the experiment results for “Hyohanki”, “Azumakagami” and “Gyokuyo”. Table 5 is the case of not using the feature of term segmentation, and table 6 is the case of using it. These results indicate that the case of using the feature of term segmentation increased extraction accuracy than the case of not using the feature of term segmentation in each three ancient writings.

Figure 4 shows the relationship between total number of characters in corpus and accuracy of person name notation. The result of figure 4 shows that the F-measure increase in proportion to the number of character in range of 0 to about 90,000 characters, and it stabilizes in the range of over 90,000 characters. This result indicates that the proposed method with more 100,000 characters in corpus performs sufficiently. There is a possibility

that sufficient amount of data in ancient writings is not prepared because they are limited for amount of data. Therefore, it is important to perform with fewer data amount in corpus.

7. CONCLUSION

In this paper, we proposed a method of named entity extraction from ancient Japanese digitized text based on Support Vector Machine using features of character appearance and probabilistic word segmentation information. We improved accuracy of person name extraction by using the result of our proposed word segmentation method as feature for SVM. We verified that the proposed method is effective for extracting personal names in ancient Japanese texts on which morphological analyzers are not available.

In our future work, we should improve the accuracy of the proposed method. We will focus on notations appearing before and after the personal name. Besides, we are planning to apply our proposed method to extract place names as well as personal names, which will make our method more useful for various kinds of text analysis of ancient Japanese writings.

8. ACKNOWLEDGMENTS

This work was supported in part by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities “Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials” from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, MEXT Grant-in-Aid for Young Scientists (B) 23700302 “Information Extraction and Visualizing from Archaic Documents”, and JSPS Grant-in-Aid for Scientific Research (C) 24500300 “Research on Integrated Information Retrieval from Multilingual Digital Archives” from Japan Society for the Promotion of Science (JSPS).

9. REFERENCES

- [1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008), pp. 1871-1874.
- [2] Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machines. In *Proceedings of The Second Meeting of the North American Chapter of the Association for Computational Linguistics for Computational Linguistics on Language technologies (NAACL2001)* (2001), pp. 1-8.
- [3] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language (NAACL2003)* (2003), pp. 8-15.
- [4] Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Word Segmentation for Text in Japanese Ancient Writings Based on Probability of Character N-grams. In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL2012)* (Nov.2012), pp. 313-316.