**SATO — Extracting**

<2 figures>

**Extracting Key Phrases for Suggesting Annotation Candidates from Japanese Historical Document**

Sato, T., Goto, M., Limura, F. and Maeda, A.

# Recent progress in digitizing historical documents has made it easier to work with the documents by enabling use of annotations and text analysis (Nagasaki et al., 2011). Di Donato et al. developed a system for annotating digitized historical documents (2013). Systems for analyzing historical documents focus on annotating not only text but also various kinds of objects such as images and figures and constructing an interface to facilitate annotating.

In the field of history and humanities, thorough annotation for a target document is carried out as research. However, efficiently annotating text in historical documents is not an easy task. Moreover, deciding which text should be annotated is challenging. Therefore, automatically suggesting annotations would be an invaluable resource for researchers.

We have been developing a system in which researchers can collaboratively annotate digitized Japanese historical documents. A team of Japanese historical researchers are currently annotating ‘Todaiji Yoroku’, a Japanese historical document written in the 12th century. However, progress has been slow because annotating requires specialized knowledge, and collaborative annotating involves a time-consuming process of face-to-face discussions and consensus among researchers. We aim to facilitate collaborative annotating through a web-based interface that enables annotation of the same document by multiple users simultaneously (Sato at al., 2014).

We developed a method for extracting key phrases from Japanese historical documents, and it will be used in our system for automatically suggesting annotations. We define a key phrase as a personal name, the name of a temple, or a Buddhist term.

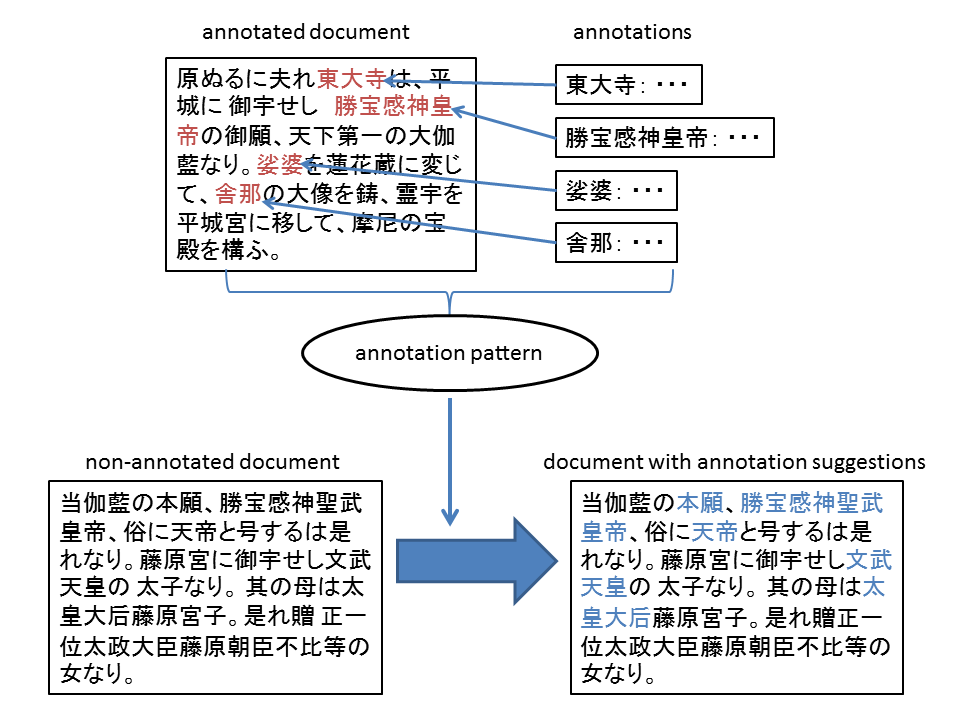


Figure 1. Outline of proposed method.

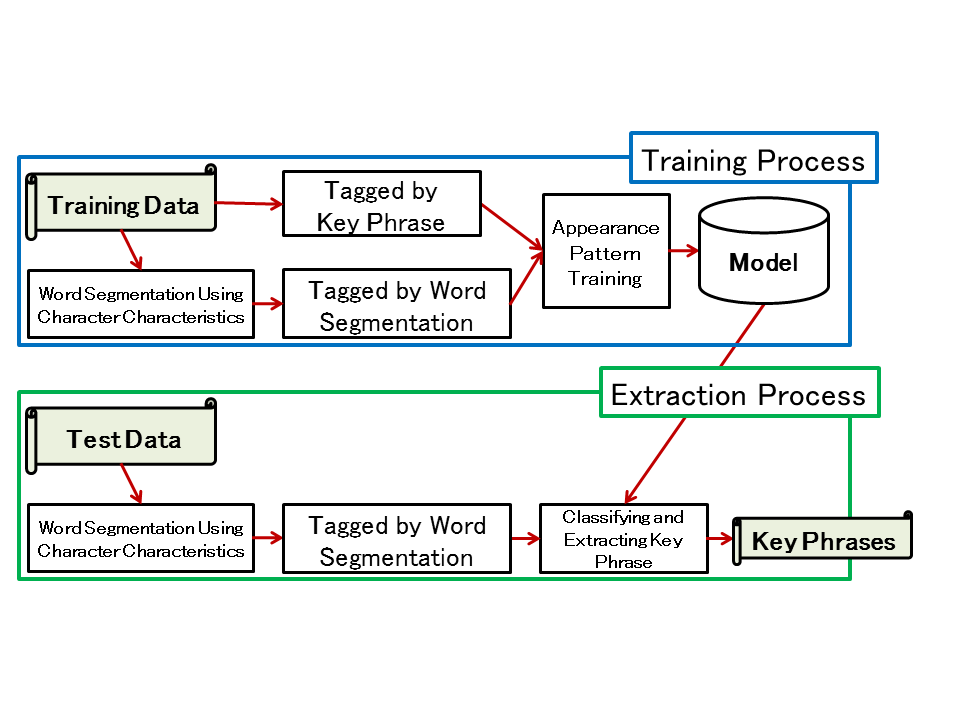


Figure 2. Flow of proposed method.

# **Proposed Method**

Figure 1 shows the outline of our proposed method, and Figure 2 illustrates the flow of the proposed method. Patterns of characters surrounding existing annotations are found in an annotated text by using support vector machine (SVM), and key phrases are extracted from a non-annotated text by utilizing these patterns. One of the unique characteristics of the proposed method is that the estimated word segmentation is used as a feature during the learning process. The ability to segment words is particularly important for analyzing ancient Japanese because the Japanese writing system does not use spaces between words, and automatic extraction of words from text in ancient Japanese is especially challenging.

## *Word Segmentation*

In the word segmentation process, sentences are divided into words. In Japanese, there are three kinds of scripts: cursive Japanese syllabary (hiragana), angular Japanese syllabary (katakana), and Chinese characters (kanji). ‘Todaiji Yoroku’ was written as a transcription of classical Chinese into classical Japanese. In the transcriptions from classical Chinese into classical Japanese, hiraganacharacters were inserted into original sentence written in classical Chinese in order to adapt to Japanese syntax. The cursive Japanese syllabary is often used, particularly for postpositional particle of auxiliary verbs. Therefore, we can assume that most nouns consist of only Chinese characters. The boundary between Chinese characters and hiragana characters can possibly be regarded as segmentation.

## *Key Phrase Extraction*

In the key phrase extraction process, patterns of characters surrounding key phrases are found by using SVM. SVM is applied to words that are found by the word segmentation process to learn where patterns of characters appear around key phrases and to attach tags to them. Key phrases are extracted by utilizing the results of SVM by referring to the attached tags and presented as suggested annotations.

# **Experiment**

We conducted an experiment to evaluate the effectiveness of the proposed method for extracting key phrases from the text of ‘Todaiji Yoroku’ by using existing annotations as the training data. We used 5-fold cross-validation. We achieved a 75% precision rate. Even though we have not yet closely examined whether the obtained results are appropriate as annotation suggestions, many of the extracted key phrases that were not judged as relevant were actually good annotation suggestions. This means that the proposed method should be an effective tool for suggesting annotations.

# **Conclusion**

We developed a method for extracting key phrases from digitized Japanese historical documents. We believe that automatically suggesting annotations will be an effective tool for annotating historical documents. We are planning to evaluate the appropriateness of key phrases extracted by the researchers using our method who are actually studying and annotating ‘Todaiji Yoroku’.

**References**

**Di Donato, F., Morbidoni, C. and Fonda, S.** (2013). Semantic Annotation with Pundit: A Case Study and a Practical Demonstration. In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies, and Techniques in the Digital Humanities* (*DH-CASE’13*), Florence, Italy, 10 September 2013.

**Nagasaki, K., Tomabechi, T. and Shimoda, M.** (2011). Toward a Digital Research Environment for Buddhist Studies. In *Book of Abstracts of Digital Humanities 2011*, Stanford, CA, 19–22 June 2011, pp. 342–43.

**Sato, T., Goto, M., Kimura, F. and Maeda, A.** (2014). Developing a Collaborative Annotation System for Historical Documents by Multiple Humanities Researchers. *Proceedings of the International Conference on Computer Science and Information Technology 2014 (ICCSIT2014)*, Barcelona, Spain, 22–24 December 2014.