Developing a Collaborative Annotation System for Historical Documents by Multiple Humanities Researchers

Takafumi Sato, Makoto Goto, Fuminori Kimura, and Akira Maeda

***Abstract*—This paper describes a prototype Web-based system for collaboratively making annotations on historical documents by multiple humanities researchers who are distant from each other. The target document of this study is "Todaiji Yoroku", which was written in the 12th century in Japan, and the system is supposed to be used by humanities researchers who are actually making annotations to this document. We have implemented the system for humanities researchers and evaluated the effectiveness of the proposed method through experiments. The unique features of the proposed system are; 1) multiple users can make annotations to the same document simultaneously; 2) suggestion function of annotation, which highlights parts of the text that are likely to be annotated, using information such as existing annotation strings and their surrounding words in the text. In this paper, we especially focus on the feature of suggestion of annotation.**

***Index Terms*—Annotation suggestion, historical documents, web-based system.**

I. Introduction

This paper describes a prototype system for collaboratively making annotations on historical documents by multiple humanities researchers who are distant from each other. There are some existing systems for annotating historical documents such as SMART-GS [1], [2], which is a sophisticated annotation system for personal research that has a function to append information to historical iconographic documents, which enables discussions on the Web. However, this system is not capable of dealing with "multiple humanities researchers who are distant from each other" and "historical literature documents with a complex structure". Therefore, we propose a Web-based collaborative annotation system for historical documents with a complex structure that can be used by multiple humanities researchers simultaneously. To make it realized, the system is required to handle multiple users simultaneously, to be able to access the same document simultaneously by these users, and to handle multiple annotations to the same strings made by these users.

The target historical document of this study is "Todaiji Yoroku", which was written in the 12th century in Japan. The proposed system is intended to be used primarily by the group of humanities researchers who are conducting annotation task for this document. This paper focuses on the annotation suggestion function, which is a part of several core functions of the proposed annotation system.

Fig. 1 shows how the study group of "Todaiji Yoroku" annotates and discuses the document of "Todaiji Yoroku".

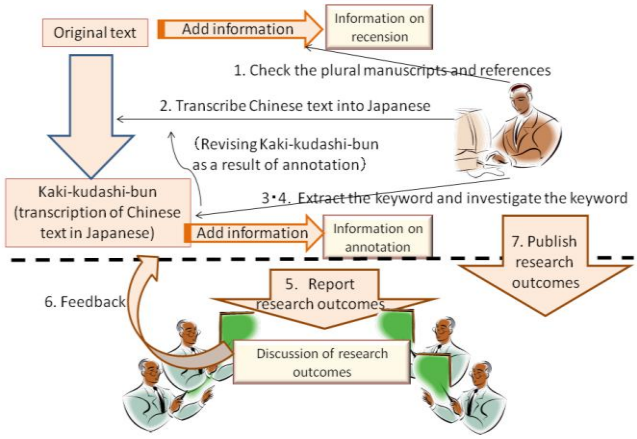

Fig. 1. How the study group of "Todaiji Yoroku" annotates and discuses the document of "Todaiji Yoroku".

II. Related Work

In this section, we describe existing works of digitization method for historical documents.

Nagasaki *et al.* developed a Web service for the text database of Buddhist scriptures called "Taisho Shinsho Daizokyo" [3], [4]. They have been published as a Web service called SAT2008 in 2008. The system is capable of searching full-text of the documents, and also the text can be displayed in Sanskrit and external characters, and the words are linked to the dictionary of Buddhist words. In Japan, such databases of historical documents opened to the public on the Web are still few.

Di Donato *et al*. developed a semantic annotation system called "Pundit" [5]. The system is capable of making annotations to documents in the form of RDF triples. Besides, it is capable of searching the Web for existing annotations, import the data, and make a link to it. Alexiev *et al*. have also developed a system called "Research Space" that can be used by researchers on the Web [6]. The purposes of this system are to help researchers study and to serve as a portal site among researchers. Munnelly *et al.* developed an annotation system for digital cultural heritage collections [7]. The system is capable of making annotations to images as well as texts, and an annotation can include links to external sources on the Web.

These systems have implemented a lot of functions in common with the system described in this paper. However, none of these systems have the function of suggesting

Manuscript received September 25, 2014; revised November 17, 2014.

T. Sato is with the Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan (e-mail: is0069kh@ed.ritsumei.ac.jp).

M. Goto is with the Faculty of Letters, Hanazono University, Kyoto, Japan (e-mail: m-goto@hanazono.ac.jp).

F. Kimura is with Kinugasa Research Organization, Ritsumeikan University, Kyoto, Japan (e-mail: fkimura@is.ritsumei.ac.jp).

A. Maeda is with College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan (e-mail: amaeda@is.ritsumei.ac.jp).

possible annotations to the users, which we propose in this paper.

III. Overview of "Todaiji Yoroku" and the Details of Creating a Database of It

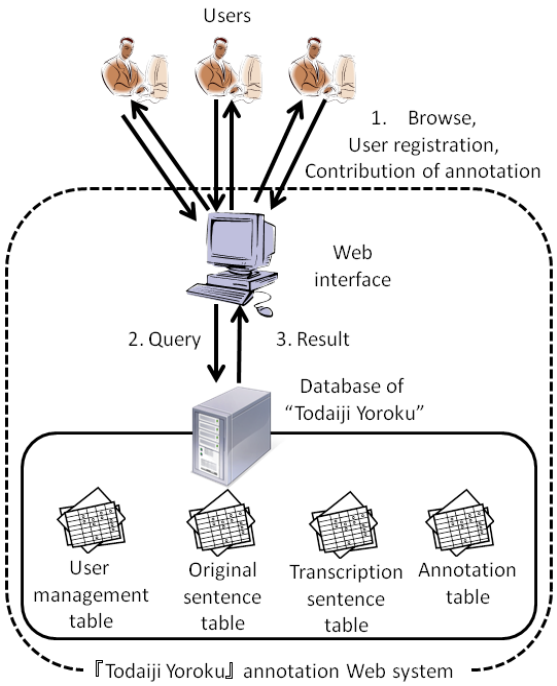

Fig. 2. Outline of the proposed system.

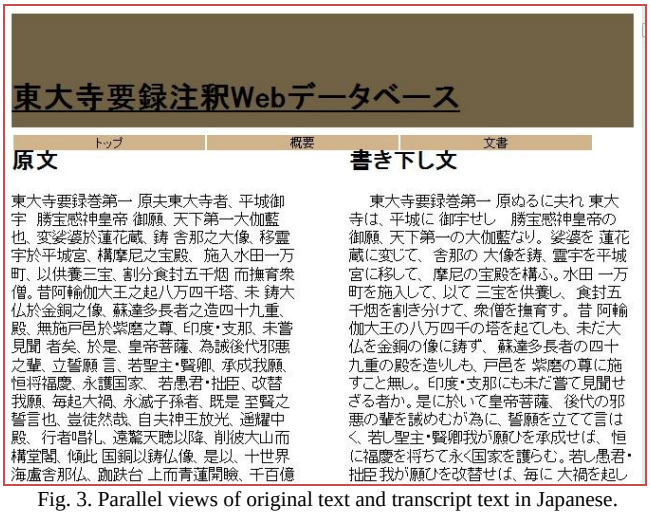

Fig. 3. Parallel views of original text and transcript text in Japanese.

"Todaiji Yoroku" was compiled by a monk of Todaiji Temple in 1106 in the hope of reviving the declined temple although actual name of the editor and the details are not known [8]. In later years Kangon augmented and revised it in 1134 and this is the version currently known. The original documents are scattered and lost. Some manuscripts are left in Todaiji Temple and Daigoji Temple. Todaiji Temple has ten volumes and Daigoji Temple has two volumes. There are two versions of printed book of the manuscripts, but both of them only put the original classical Chinese text into print, and no modern translation or detailed annotations are attached to them. Therefore, a research group was organized by the researchers of Japanese history, Buddhist studies, architecture and art history, and they are tackling annotation work. To help this work, we are developing a Web-based system for making annotations and revisions for this document by cooperating with this research group. In this paper, we propose a novel method for supporting researchers making new annotations to the document, which is a part of the functionalities required for the system.

Fig. 2 shows the outline of the proposed system. This system consists of four tables, "User management table", "Original sentence table", "Transcript sentence table" and "Annotation table".

Fig. 3 shows parallel views of original text and transcript text in Japanese.

IV. Requests from Humanities Researchers

The researchers, as actual users of the proposed system, made the following requests for our system: 1) The system should be available to users without stress, even if the user hardly has any skills in using computers, such as senior researchers of humanities field. 2) It should be able for several researchers to make different annotations or opinions to the same strings. 3) It should be able to store the history log of making annotation in order to check it or to rollback as necessary. 4) It should be able to make multiple annotations that are overlapped in text. 5) It should be able to request the verification of annotations among researchers.

The proposed system should satisfy these requests. Making annotations to historical documents is the fundamental process for humanities researchers to analyze them. Therefore, it is important to be able to make annotations to documents as freely as possible. In addition, it is necessary to verify whether the attached annotations are appropriate. Thus, these requirements have to be met.

In order to help the researchers making annotations, we propose a method to support making new annotations. This function is focused on building an interface that can be operated intuitively and easily even if the users are researchers of humanities field who are not necessarily proficient with computers.

The current Web-based proposed system can do the following three types of operations. 1) To show the original Chinese text and the transcript of the original text in Japanese. 2) To view existing annotations. 3) To post a new annotation. In this paper, we propose an additional function to the current system, in which existing annotations are suggested when a user is making an annotation to the same thing at the different part of the text. We call this function "suggestion function". We also propose a method for ranking the results of the suggestion function, and this is the main topic of this paper.

Generally, "suggestion" means a dynamic query expansion method that presents a user with the related words when entering keywords into a Web search engine. Note that "suggestion function" in this paper is different from suggestion in Web search engines.

V. Proposed Method

In this section, we explain the method of annotation suggestion function in detail.

A. Overview of the Method

In this method, the input is an annotation made by one user. The output is annotations that are already attached to the document and are similar to the input annotation. The document of "Todaiji Yoroku" has the original text and the transcript of Chinese text in Japanese. The original text is written in the style of classical Chinese. The transcript of Chinese text in Japanese is close to the contemporary writing of Japanese. Our method currently focuses on annotations of the transcript of Chinese text in Japanese.

B. Selecting a New Annotation String

The user can drag the mouse cursor on the string that he/she wants to make a new annotation within the text that is displayed on the Web browser, and the system highlights the string to be annotated. Since senior humanities scholars are generally not very familiar with the operation of a computer, we chose a simple mouse input method rather than typing input from a keyboard. Fig. 4 shows an example of selecting a new annotation string.

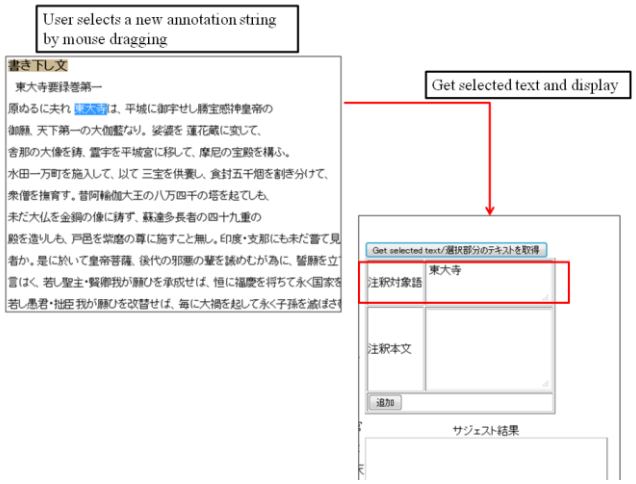

Fig. 4. An example of selecting a new annotation string.

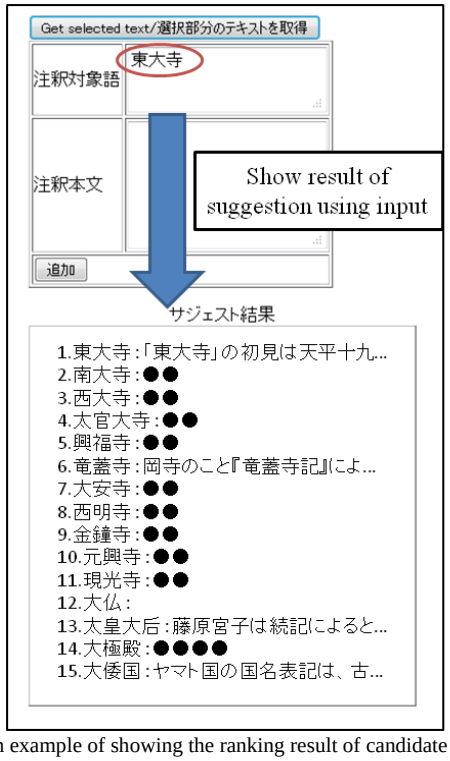

Fig. 5. An example of showing the ranking result of candidate strings.

C. Obtaining Candidates for Suggestion of Annotation from the Annotation Database

We obtain candidate strings to be displayed for the suggestion function from exiting annotations in the annotation database. In this method, we do not take into account the contents of annotations. The candidate strings are ranked according to the method described in the next section. Fig. 5 shows the image of showing the ranked result of candidate strings.

D. Ranking

We do the ranking of candidates when displaying the results of the candidate suggestion function. As explained in the previous section, we obtain the candidates from the annotation database in the order stored in the database, which is the same as the order of appearance in the text. However, we consider that displaying candidates in such order is not necessarily useful for researchers. For example, when a user is editing the annotation at the bottom of the document on the Web browser, and the candidates are displayed in the order of appearance in the text, appropriate candidates might be displayed at the bottom, and it makes the system difficult to use. Besides, if we use ranking, we only need to show several candidate strings at the top rank rather than showing all candidates, thus the operation of the Web browser will be faster and smoother.

1) Levenshtein distance

When measuring the similarity of two strings, edit distance is often used. The Levenshtein distance, a kind of edit distance, is a numerical value that indicates to which extent two strings are different [9]. Specifically, it is the minimum number of operations, i.e., insertion, deletion or substitution, required to transform one string into the other. The Levenshtein distance is useful for measuring the similarity of two strings. For example, the minimum number of operations required for transforming the string "kitten" into "sitting" is three as shown below, thus the Levenshtein distance of these strings is 3.

kitten

1) sitten (substitute "k" with "s")
2) sittin (substitute "e" with "i")
3) sitting (insert "g" and finish)

In the above example, the cost of each operation (insertion, deletion, and substitution) is set to 1. It is possible to set different values for different operation. For example, to avoid substitution, we can set the cost of 2 for substitution and the cost of 1 for insertion and deletion.

2) Calculation of the Levenshtein distance

The proposed method calculates the Levenshtein distance between the new annotation string and each of the candidate strings obtained in section V.C. We use three types of operations, i.e., insertion, deletion, and substitution, for calculating the Levenshtein distance, and we set the cost of all the operations to 1. The Levenshtein distance will be smaller when two strings are more similar. Therefore, we sort the candidate strings in ascending order of Levenshtein distance.

E. Displaying the Ranked List of Candidates for Annotation

The system displays a box to show the new annotation

string selected by the user on a Web browser. The results of the annotation suggestions obtained by the method described in the previous sections will be displayed in the box below. The results will be displayed in ascending order of the Levenshtein distance. By sorting the candidate strings in ascending order of the Levenshtein distance, it becomes possible to display the annotation candidates that are expected by the researchers at the top of the results, thus improve the efficiency of annotation task.

TABLE I: THE RESULT OF THE BASELINE METHOD

	new input strings		
Rank	天皇(Emperor)	天智天皇(Emperor Tenji)	東大寺(Todaiji Temple)
1	天皇(Emperor)	天璽国押開豊…(Posthumous title of Emperor Shomu)	東大寺(Todaiji Temple)
2	天皇(Emperor)	天皇(Emperor)	太官大寺(Daikandaiji Temple)
3	天璽国押開豊…(Posthumous title of Emperor Shomu)	氷高内親王(Emperor Gensho)	南大寺(Nandaiji Temple)
4	氷高内親王(Emperor Gensho)	太上天皇(Abdicate Emperor)	西大寺(Saidaiji Temple)
5	太上天皇(Abdicate Emperor)	又勅所司、縁…	阿輸伽大王(Ashoka the Great)
6	又勅所司、縁…	天皇御中宮(empress)	迦維羅の大士(Bodhisattva of Karura)
7	天皇御中宮(empress)	天皇(Emperor)	其の母は 太皇…(The mother is empress…)
8	高野天皇(Emperor Koken)	天智天王(Emperor Tenji)	太皇大后(Grand empress dowager)
9	太上天皇(Abdicate Emperor)	高野天皇(Emperor Koken)	正一位太政大…(part of "Prime Minister of the Imperial Court")
10	勝宝感神皇帝(honorary name of Emperor Shomu)	太上天皇(Abdicate Emperor)	文武天皇治天…(Emperor Monmu governs)
11	天帝(honorific name of Emperor Shomu)	勝宝感神皇帝(honorary name of Emperor Shomu)	大宝元年(The Daiho first year (701))
12	太皇大后(Grand empress dowager)	天帝(To name Emperor Shomu)	大極殿(Council hall in the imperial palace)
13	文武天皇治天…(Emperor Monmu…)	太皇大后(Grand empress dowager)	東方に慶雲(auspicious cloud to the east)
14	皇太子始めて…(The Crown Prince for the first time…)	文武天皇治天…(Emperor Monmu governs…)	天下に大赦す…(grant an amnesty to the whole country…)
15	天下に大赦す…(grant an amnesty to the whole country…)	皇太子始めて…(The prince for the first time…)	大嘗会(播磨／… (first ceremonial offering of rice by newly-enthroned Emperor…)
Average Lavenshtein distance	7.14	7.53	7.4

TABLE II: THE RESULT OF USING THE PROPOSE METHOD

	new input strings		
Rank	天皇(Emperor)	天智天皇(Emperor Tenji)	東大寺(Todaiji Temple)
1	天皇(Emperor)	天智天王(Emperor Tenji)	東大寺(Todaiji Temple)
2	天皇(Emperor)	天皇(Emperor)	南大寺(Nandaiji Temple)
3	天帝(honorific name of Emperor Shomu)	太上天皇(Abdicate Emperor)	西大寺(Saidaiji Temple)
4	天(Part of character of Emperor)	高野天皇(Emperor Koken)	太官大寺(Daikandaiji Temple)
5	太上天皇(Abdicate Emperor)	太上天皇(Abdicate Emperor)	興福寺(Kofukuji Temple)
6	高野天皇(Emperor Koken)	氷高内親王(Emperor Gensho)	竜蓋寺(Ryugaiji Temple)
7	太上天皇(Abdicate Emperor)	天帝(To name Emperor Shomu)	大安寺(Taianji Temple)
8	四天王(the Four Devas)	四天王(the Four Devas)	西明寺(Saimyoji Temple)
9	皇子(imperial prince)	智奴王(Grandchild of Emperor Tenmu)	金鐘寺(Kinshoji Temple)
10	氷高内親王(Emperor Gensho)	智行(Training to purchase wisdom)	元興寺(Gangoji Temple)
11	天皇御中宮(empress)	天(Part of character of Emperor)	現光寺(Genkoji Temple)
12	太皇大后(Grand empress dowager)	天皇御中宮(empress)	大仏(Great statue of Buddha)
13	天智天王(Emperor Tenji)	太皇大后(Grand empress dowager)	太皇太后(Grand empress dowager)
14	四天王寺(Shitennoji Temple)	皇子(Imperial prince)	大極殿(Council hall in the imperial palace)
15	皇后宮(Empress' palace)	日並智王子(Another name of Kusakabe Prince)	大倭国(Another name of Japan)
Average of distance	2.13	3.13	1.93

VI. Experiment

We conducted an experiment of the proposed ranking method mentioned in Section IV in order to verify its effectiveness. In this experiment, we used three strings "Emperor", "Emperor Tenji" and "Todaiji Temple" as new input strings. We show the top 15 existing annotations suggested by our proposed method and the baseline method for each input string.

A. Baseline Method

The baseline method uses exact and partial string matching between the new input string and existing annotations. The result of the exact and partial string matching is used as the suggested annotations. In this experiment, we used prefix matching, suffix matching and substring matching as partial matching. In addition, we considered that the longer the partial string matches, the higher the suggested annotation is ranked.

B. Calculation of the Levenshtein Distance

In the proposed method, the system calculates the Levenshtein distance between the input string and each of the existing annotations stored in the database. The system counts the number of operations of insertion, deletion and substitution of a character in order to estimate how similar each existing annotation is to the input string. In this experiment, we set the cost of insertion, deletion and substitution to 1.

We calculated the Levenshtein distance between suggested annotations and each of three input strings in order to evaluate how appropriate suggested annotations are.

C. Result of Experiment

Table I shows the result of the baseline method, and Table II shows the result of the proposed method. In both tables, suggested annotations in colored cells indicate appropriate suggestions. In addition, we show the average of the Levenshtein distance in both tables.

As shown in these tables, the averages of the Levenshtein distance of our method were much lower than the baseline for all three input strings. It indicates that the suggestions of our method might be more useful than the baseline method.

VII. Future Work

The proposed system is currently at the stage of obtaining comments from the study group of "Todaiji Yoroku", since the group is still in progress in creating annotations. Therefore, we still have to constantly improve the system through feedbacks from them. However, we consider that our proposed system enabled substantial collaboration with the research group from varieties of fields, such as history, religion, architecture and art through the Web-based system. One of our future tasks will be to consider how to brush-up this database to better meet the demands of the users. Another future task will be to make the text conforming to the TEI Guidelines [10]. Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. It has been used since 1994. By conforming to TEI, the text data of "Todaiji Yoroku" including the annotation data will be able to be standardized and shared with others.

VIII. Conclusion

In this paper, we proposed a collaborative annotation system for historical documents, which has a function of suggesting possible annotation candidates derived from existing annotations. In particular, we describe a method for ranking the annotation candidates based on the similarities of newly input string and existing annotation strings in the database. Our aim is to support the annotation work in the study group of "Todaiji Yoroku".

The result of the experiment showed that the effectiveness of the proposed method is promising. However, in this paper we only tested the Levenshtein Distance for similarity ranking. In future work, we are considering testing different string similarity metrics. Besides, several problems such as insufficient amount of existing annotation data have been pointed out. We need to more closely collaborate with the researchers in the study group in order to encourage them to use our system and thus increase the annotation data.

Acknowledgment

Development of this system has been promoted by opinions from the members of the study group of "Todaiji Yoroku". In addition, we have received the cooperation of Prof. Towao Sakaehara, Osaka City University Professor Emeritus, and the director of the study group of "Todaiji Yoroku".

References

[1] S. Hayashi, K. Aihara, M. Kukita, and M. Ohura, "SMART-GS system: a software for historians by historians," presented at the JADH, Tokyo, Japan, September 15-17, 2012.
[2] Y. Hashimoto, "SMART-GS Web: A HTML5-Powered, Collaborative Manuscript Transcription Platform," presented at the JADH, Ibaraki, Japan, September 19-21, 2014.
[3] K. Nagasaki, T. Tomabechi, and M. Shimoda, "Toward a digital research environment for Buddhist studies," *Book of Abstracts of Digital Humanities*, California, pp. 342-343, 2011.
[4] K. Nagasaki, A. C. Muller, and M. Shimoda, "Aspects of the interoperability in the digital humanities: A case study in Buddhist studies," *iBook of Abstracts of Digital Humanities*, Maryland, pp. 375-377, 2009.
[5] F. D. Donato, C. Morbidoni, and S. Fonda, "Semantic annotation with Pundit: a case study and a practical demonstration," in *Proc. the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, Florence, Italy, September 10, 2013.
[6] V. Alexiev, S. Kostadinov, and J. Parvanova, "RDF Data and Image Annotations in ResearchSpace," in *Proc. the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities*, Florence, Italy, September 10, 2013.
[7] G. Munnelly, C. Hampson, N. Ferro, and O. Conlan, "The FAST-CAT: Empowering Cultural Heritage Annotations" *Book of Abstracts of Digital Humanities*, Nebraska, pp. 320-322, 2013.
[8] E. Tsutsui, *Interpretation of Todaiji Yoroku*, *Todaiji Yoroku*, Tokyo, Japan: Kokushokankokai, 1971.
[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710, 1966.
[10] TEI: P5 Guidelines. [Online]. Available: http://www.tei-c.org/Guidelines/P5/

Takafumi Sato is a master's student at the Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. He is interested in constructing support system for annotation.

Makoto Goto is an associate professor of digital archives at The National Institutes for the Humanities (NIHU). He obtained the Ph.D. degree in literature from Osaka City University in 2007. He has served as the chair of the Special Interest Group of Computers and Humanities under the auspices of the Information Processing Society of Japan from 2009 to 2010. His research interests are primarily Japanese ancient history and digital humanities research.

Fuminori Kimura is a senior researcher in Art Research Center, Ritsumeikan University. He obtained the Ph.D. degree in engineering from Nara Institute of Science and Technology in 2007. His research interests include information retrieval, and text mining and multilingual information processing.

Akira Maeda is a professor at the Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University. He received the B.A. and M.A. degrees in library and information science from University of Library and Information Science (ULIS) in 1995 and 1997, respectively, and received the Ph.D. degree in engineering from Nara Institute of Science and Technology (NAIST) in 2000. He has visited Virginia Polytechnic Institute and State University (Virginia Tech) from October 2000 through March 2001 as a postdoctoral visiting scholar. He has worked as a postdoctoral researcher of the CREST program, Japan Science and Technology Corporation (JST) from April 2001 through March 2002. He was a visiting professor at King's College London from September 2011 through September 2012. His research interests include digital libraries, digital humanities, information retrieval, and multilingual information processing.