http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 2, December 2017: 13 - 22
e-ISSN: 2289-2192

ACQUISITION OF ANECDOTES FROM THE WEB FOR FREE CONVERSATION SYSTEMS

KIYOAKI SHIRAI
TOMOTAKA FUKUOKA

ABSTRACT

A free conversation system that can talk freely with users gives computers human-like qualities that can enrich our daily lives. This paper proposes a method to automatically acquire anecdotes from various people on the Internet as a source of knowledge to be accessed by free conversation systems. Firstly, passages that are related to a given person are retrieved by searching the relevant web pages. Secondly, each page is segmented and then, thirdly, each passage is judged as to whether it is an anecdote by applying several rules based on the linguistic features of the anecdotes. The experimental results showed that the precision of our method was increased by 25% when compared to the baseline.

Keywords: knowledge acquisition, Web, anecdote, free conversation system

INTRODUCTION

A free conversation system is a non-task-oriented dialog system that can talk freely with users. In recent years, the study of open-domain or free conversational systems has attracted research interest because it can be widely applicable for many applications such as robotic pets or nursing care robots (Libin & Libin 2004). In general, topics in free conversation are constantly changing and the initiative in conversation often passes between two or more participants (Kawachi 2003). In order to chat with users smoothly, the free conversation system should not only respond to a user's utterance simply but also be able to handle various topics and, sometimes, take the initiative in the conversation. To take the initiative, for example, the system is required to start a new topic and then continue to talk about it.

To generate an utterance in a free conversation system, an approach that retrieves sentences from a huge external knowledge source would be more appropriate than a rule based approach. This is because the system should respond to a user who may talk about a wide variety of topics. For example, several methods were proposed to generate the system's response to a user's inputs by using newspaper articles on the World Wide Web (WWW or 'the Web') or tweets posted on Twitter as knowledge sources (Mizuno et al. 2009; Shibata et al. 2009; Higashinaka et al. 2014a). However, they focused on replying to the user's utterance, where the user takes the initiative in the conversation and not the free conversation system.

In order to take the initiative, a system is required to produce a sequence of consistent utterance on the same topic. However, it is rather hard to automatically generate such coherent utterance. One of the possible approaches is to store a set of sentences regarding a certain topic (or a 'story') in advance and then generate the stored story when the system keeps the initiative. An anecdote from a famous person is a useful story to be stored in a free conversation system because the anecdote might be a good topic that the system can offer in conversation. Let us

suppose that the system has a database of people and their anecdotes. When a user mentions a certain person, the system can introduce his/her anecdote as a new topic, thereby taking the initiative of the conversation through telling the anecdote. Such transition of the initiative makes the conversation more natural, allowing a user to chat with the system for a long time without becoming bored.

The goal of this paper is to explore how to automatically acquire the anecdotes of any given person from the Web. The Web is the best source of such anecdotes, since many web sites describe interesting stories of famous people and celebrities. Note that only anecdotes written in Japanese will be retrieved in this study.

A simple way to acquire the anecdotes is to use Wikipedia however, not all entries contain anecdotes. When fourteen entries of famous people were checked in Japanese Wikipedia[1], only five entries were found to contain anecdotes. Therefore, it is essential to acquire the anecdotes from a variety of web sites.

RELATED WORK

Response generation in a free conversation system is difficult since the system needs to respond to a wide variety of users' utterances. Nevertheless, several attempts have been made on this issue.

Mizuno et al. (2009) presented a method to use news articles on the Web as a knowledge source for response generation in free conversation systems. For a given user's input, a relevant sentence was searched for based on any overlap of words between the system and the user's utterances, the position of it in the article and the length of it. Then the retrieved sentence was generated as the system's response. Three subjects evaluated the generated responses in terms of their fluency, relevance and interest (whether a user wanted to continue to talk). They concluded that news sites on the Web were a valuable source for the free conversation system.

Shibata et al. (2009) proposed a method to retrieve a sentence from the Web on the basis of surface cohesion and shallow semantic coherence in order to enable the system to generate a response to the user. The surface cohesion followed the centering theory, while the shallow semantic coherence was calculated by the conditional distribution and inverse document frequency of content words. A proposed system that can converse about movies was evaluated. It was found that 66% of the system's utterance was appropriate in terms of fluency and meaningfulness.

Yoshino et al. (2011) developed a spoken dialog system that can answer a query given by the user or tell related news to the user if the exact answer was not found. The sentence for the system's response was retrieved from newspaper articles by the matching of a predicate-argument structure between the query and the sentence found in the news. Experimental results in the baseball domain showed that the proposed method outperformed a baseline only using Bag-of-Words by a 17% F-measure.

Sugiyama et al. (2013) proposed a method for generating a response to the user based on a template filling. The template was filled with the most salient word from the user's utterances, matched with a related word extracted beforehand from a collection of texts on Twitter. An example of the template was "I hear that [Noun] is [Adj], isn't it?" When the user mentioned "Mt. Fuji", it was filled at [Noun] and its related word "beautiful" was filled at [Adj]. They

[1] https://ja.wikipedia.org/

reported that the proposed system significantly outperformed the baselines (IR-response and IR-status as presented in (Ritter et al. 2011)) in a subjectivity test.

Twitter is also a valuable resource for response generation in dialog systems. In an open-domain conversational system developed by Higashinaka et al. (2014a) and Higashinaka et al. (2014b), tweets in Twitter were used as one of the sources of sentence generation. To choose an appropriate sentence from Twitter for a given topic word, methods of word-based filtering, syntactic filtering and content-based retrieval were proposed. The two filtering methods ensured grammatical correctness of the retrieved sentence, while the content-based retrieval can extract the sentences whose topics were similar to the given one. A dictionary of related words constructed from weblog articles was used for the content-based retrieval.

The previous studies, discussed above, focused on generating one sentence or message against a user's utterance. The free conversation system in this study will also react to a user's utterance but will try to generate a story (multiple sentences) in order to take the initiative in a conversation. This paper will present how to prepare a database of the stories (i.e. anecdotes) for such a dialog system.

Another related work is automatic generation of the story. Peinado & Gervas (2006) proposed a method based on the case-based reasoning with heuristic rules for producing a new story. Mcintyre & Lapata (2009) presented a novel data-driven approach with a sentence generator that operated predicate-argument structure. Imabuchi & Ogata (2012) proposed a method to generate a story in Japanese, which is based on Propp theory. Unlike these previous studies, we do not try to create a story but excerpt a complete story from the Web.

PROPOSED METHOD

In this study, an anecdote from a person is defined as text that fulfils three requirements, see FIGURE 1. Note that a story that only tells us a fact about a person (such as a career, prize, record and so on) is not regarded as an anecdote. In contrast, text can be defined as an anecdote if it contains a fact with an interesting story.

(a)	A story about a person which tells us about his/her character.
(b)	A story that is not well known by many people.
(c)	A story that is interesting or amusing, i.e. it attracts a user when it is told by a free conversation system.[2]

FIGURE 1. Definition of an anecdote

An example of an anecdote on a web page is shown in a solid box in FIGURE 2. The English translation is given on the left in italics.

FIGURE 3 illustrates an overview of the proposed method. Firstly, web pages including a name of a given person are retrieved via a Web search engine. Secondly, the retrieved web pages are segmented into passages. And then thirdly, a filtering is performed to judge whether each passage is an anecdote or not.

[2] Please note that the intensity of the interest is not considered in this paper. However, it is desirable to distinguish much interesting and a little interesting anecdote. It is our future work.

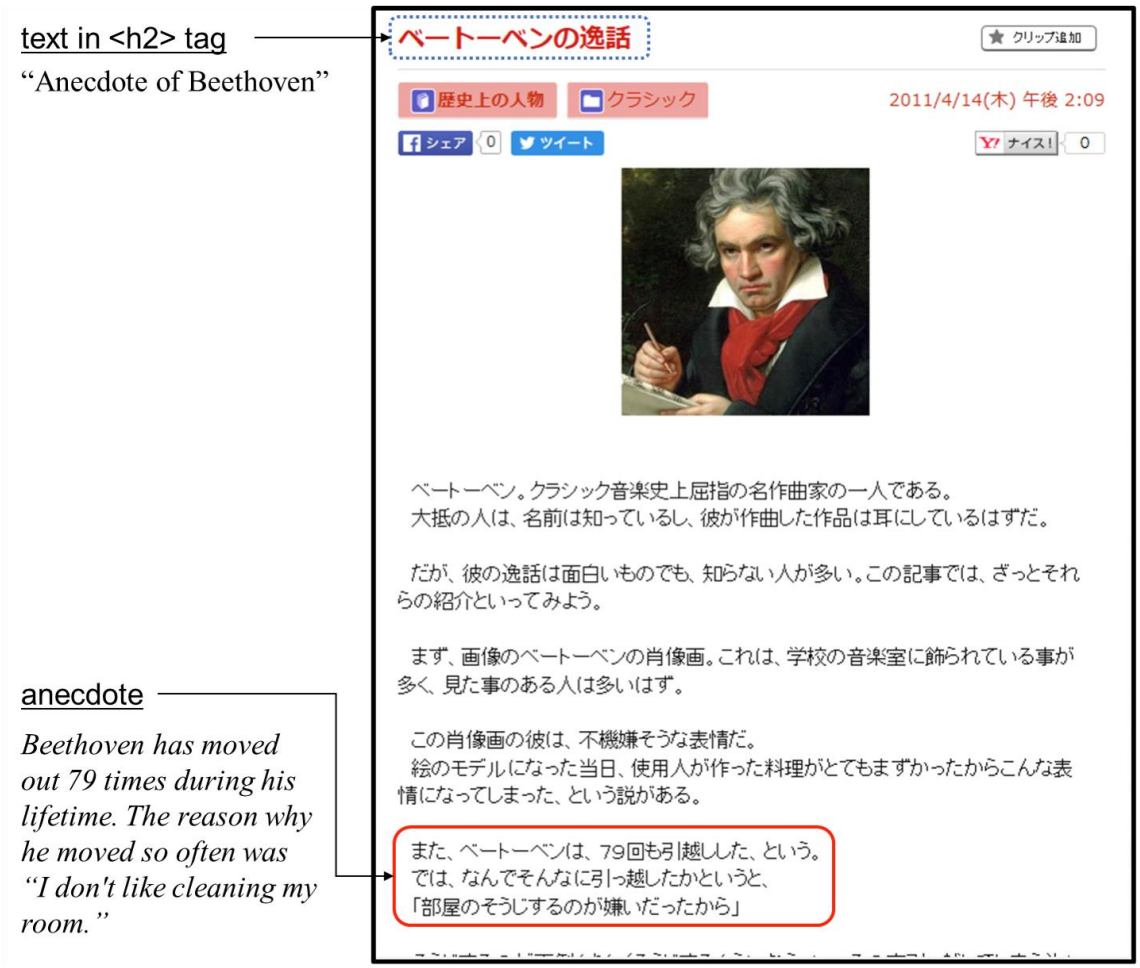

FIGURE 2. Example of an anecdote on a web page

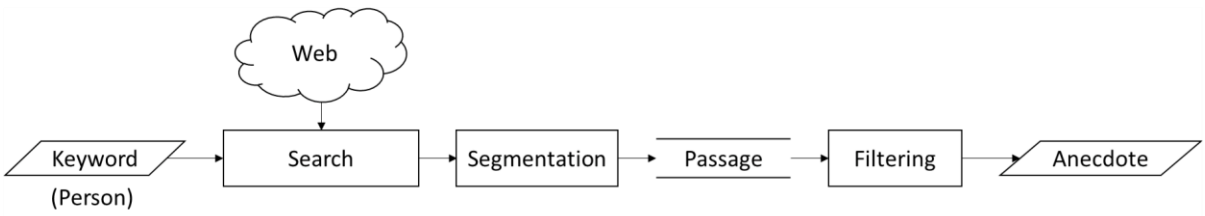

FIGURE 3. Overview of proposed method

PASSAGE RETRIEVAL

This procedure retrieves fragments of texts or passages from the Web that may prove to be candidates for anecdotes. Firstly we form a query in Equation (1), where [PERSON] is the given person.

[PERSON] & 逸話 (*itsuwa*;anecdote)[3] (1)

[3] In this paper, Japanese pronunciation and English translation of a Japanese word is written in parenthesis. The Japanese pronunciation is denoted in italics.

The query is entered into Bing API[4] and then the top 100 web pages are retrieved. Each web page is pre-processed by the Document Object Model (DOM). DOM represents an HTML source file as a tree structure, where each node corresponds to an HTML tag or text.

Next, each web page is segmented into passages. A passage is roughly defined as one or several paragraphs that denote the same topic. The HTML tags can be categorized into two groups: a block element (e.g. , <h1> and <p>) and an inline element (e.g. , <a> and <b>). We choose the smallest block elements (i.e. the nodes that are block elements and do not subordinate any other block elements) as their descendants in the DOM tree. We regard the text governed by the chosen node to be a passage.

Since the chosen DOM node sometimes covers blocks of text that are considered to be too large and may contain two or more genuine passages, we further divide the passages obtained by the above procedure into finer ones as follows. The text governed by the internal DOM node can be represented as $\{ t_1, \ldots, t_n \}$, where t_i is the text separated by the HTML tags. If t_i fulfils either of the following conditions, we segment the passage into smaller ones by using t_i as a separator so that $\{ t_1, \ldots t_{i-1} \}$ and $\{ t_{i+1}, \ldots, t_n \}$ become new passages.

1) t_i only contains a space, a tab, or both.
2) The length of t_i is not more than 5. In this case, t_i may not be a complete sentence and can be a separator of the passage.

The passages obtained by this method are candidates for the anecdotes that can then be passed on to the next procedure: filtering.

FILTERING

We designed the following eleven rules to judge whether the retrieved passages fulfil the criteria to be anecdotes and filter out any non-anecdotal passages. A passage is communicated as an anecdote if it is not removed after all of the rules have been applied.

r_1: Heading

A heading in a web page is a strong indicator of its contents. If the heading includes the word "anecdote", then the text under it might well be an anecdote. This rule checks whether (1) the heading tag (e.g. <h1> and <h2>) is an ancestor or preceding sibling of the node of the passage in the DOM tree and (2) if this heading includes the person's name and "逸話(*itsuwa*;anecdote)" or "エピソード(*episôdo*;episode)". The passage is removed if the above conditions are not fulfilled. An example is shown in a dotted box in FIGURE 2. The heading "Anecdote of Beethoven" is marked up by the <h2> tag, that is an ancestor of the node of the anecdote passage.

r_2: Subject of sentence

This rule removes the passage if it contains no sentence in which the subject is the given person. It ensures that the person is the main topic of the passage. In Japanese, the subject of the sentence is followed by a nominative case marker "が(*ga*)" or topic case marker "は(*wa*)". Therefore, when the person's name is not followed by these case markers, the passage is filtered out.

[4] https://datamarket.azure.com/dataset/bing/search

r_3: First person pronoun

This rule removes the passage if it includes the first person pronoun (私(*watashi*), 僕(*boku*), 俺(*ore*) and うち(*uchi*)). A sentence like "I love Beethoven" usually represents an opinion or sentiment of the writer of the web page and a passage including such a sentence may not be an anecdote. As an exceptional case, the passage is not removed when the first person pronoun appears in a quotation, such as "Beethoven said 'I love music'."

r_4: Subjective expression

This rule removes the passage if it contains the verb "思う(*omou*;think)", since a sentence like "I think Beethoven is a great composer" may express an opinion of the writer. Similar to **r_3**, the passage is not removed when "think" appears within a quotation.

r_5: Demand

The passage might not be an anecdote if it contains a writer's demand, such as "please tell me *something*." This rule removes the passage if "お願い(*onegai*;I have a favor to ask you)" or "教えて(*oshiete*;please tell me)" is a predicate of the sentence. Since Japanese is a head-final language, the predicate appears near the end of the sentence. We simply judge if these demand expressions are the predicates by checking the position of them. More precisely, we check if $p/l \geq 0.8$, where p is the position of the demand expression and l is the length of the sentence, respectively.

r_6: Introductory expression

This rule removes the passage if it includes "エピソード(*episôdo*;episode)" as well as one of the following keywords in the last two sentences:

まとめる(*matomeru*;summarize), 伝える(*tsutaeru*;tell), 紹介(*shôkai*;introduce), 披露(*hirô*;introduce).

It aims to identify an introductory expression, such as "We will introduce the episode of *someone* in this page." Although such a sentence indicates that the anecdote is written in the web page, the passage itself is not the anecdote.

r_7: Minimum length

If the passage is too short then it may be inappropriate as an anecdote because the dialog system should continue to produce several sentences in order to take the initiative in the conversation. This rule removes the passage if its length is less than 50 characters.

r_8: Maximum length

We observed that a long passage tended not to be an anecdote. Furthermore, if a story is too long then it is inadequate for the free conversation system. Therefore, we removed passages consisting of six or more sentences.

r_9: Ungrammatical sentence

This rule removes the passage if the last word is neither a period nor closing parenthesis (indicating the end of a quotation) as the last sentence of the passage would be ungrammatical.

r_{10}: Initial demonstrative

This rule removes the passage if a demonstrative "この(*kono*;this)" appears at the beginning of it. Since "this" refers to a noun outside of the passage, it may lack some information.

r_{11}: Link tag

This rule removes the passage if its current or ancestor DOM node is <a> tag as this would mean that the passage is within a hyperlink. This rule was constructed to remove advertisements.

A web page usually contains only one or a few anecdotes. However, we found that too many anecdotes are sometimes extracted from one web page by our method. It means that only a small number of the retrieved passages are actually anecdotes, which causes a serious decrease in precision. Therefore, we apply another filtering rule which considers the number of extracted anecdotes per web page. Let us suppose that *n* anecdotes are obtained from one web page. If *n* is greater than a threshold T_n, all anecdotes are discarded. It may wrongly eliminate some anecdotes however, the precision of anecdote extraction will be much improved. In this study, precision is preferred to recall, since it is not necessary to exhaustively acquire all possible anecdotes from the Web.

EVALUATION

EXPERIMENTAL SETUP

To evaluate the proposed method, the five men shown in TABLE 1 were chosen as the target people. For each person, the relevant 100 documents were retrieved as described earlier. The family and given names were concatenated by "or" and used as the query[5], which is shown in the second column in TABLE 1. The retrieved web pages were manually annotated with an anecdote tag that marked up the passage of relating to the anecdote[6]. TABLE 1 also shows the number of gold anecdotes. Note that some of the gold anecdotes were duplicated, i.e. they appeared more than once in the data set.

TABLE 1 Data set

Person	Query	Number of Anecdotes
Ichiro Suzuki*	Ichiro	60
Oda Nobunaga**	Oda **or** Nobunaga	85
Leonardo da Vinci	Leonardo **or** Vinci	104
Ludwig van Beethoven	Ludwig **or** Beethoven	104
Wolfgang Amadeus Mozart	Wolfgang **or** Amadeus **or** Mozart	65

* Japanese baseball player ** Famous Japanese military commander.

The acquired anecdotes were evaluated by measuring their precision and recall. In the task of anecdote extraction for the free conversation system, precision is more important than recall. Boundaries of the extracted and gold anecdotes were not always the same. If the extracted anecdote covered the whole of the gold anecdote, it was regarded as being correct. If the

[5] We only used the given name for "Ichiro" because he is often called by that name alone.
[6] Since one author did it, we cannot show an inter-annotator agreement.

anecdote covered the whole of two or more gold anecdotes, it was judged that only one anecdote had been correctly retrieved when calculating the recall. On the other hand, if the extracted anecdote contained only a part of a gold anecdote, it was regarded as being faulty.

RESULTS

TABLE 2 shows the precision and recall of two methods. The baseline was a method that extracted a passage if it contained both the person's name and the word "逸話(*itsuwa*;anecdote)", while the proposed method retrieved the anecdote after applying the eleven rules $\mathbf{r_1}$ ~ $\mathbf{r_{11}}$. The baseline seems too naive. However, it is difficult to compare our method with other strong baselines, since no previous work focused on the exactly same task. Comparing the micro average of five people, the precision of the proposed method was shown to be better than the baseline by 11% with a small loss of recall.

TABLE 2 Results of anecdote acquisition

Person	Baseline		Proposed method	
	Precision	Recall	Precision	Recall
Ichiro Suzuki	0.179	0.083	0.429	0.056
Oda Nobunaga	0.049	0.035	0.375	0.141
Leonardo da Vinci	0.692	0.173	0.238	0.048
Ludwig van Beethoven	0.125	0.029	1.000	0.029
Wolfgang Amadeus Mozart	0.345	0.154	0.167	0.015
Micro average	0.232	0.093	0.348	0.057

Although our method outperformed the baseline, the precision and recall were still low (34.8% and 5.7% respectively). The reason for the low precision was that sentences containing interesting stories could not be distinguished from sentences containing facts about the person. For example, a sentence "Ichiro is a major league baseball player" mentions the target person "Ichiro" but it is not an anecdote, it is stating a fact. The passages containing such sentences (only containing facts) were often wrongly retrieved. Discrimination between anecdotal and non-anecdotal passages is difficult, especially when considering novel or amusing passages (see conditions (b) and (c) in FIGURE 1). Deep understanding of the text is required to precisely filter out non-anecdotal passages. The recall was also low and one of the major reasons was that multiple anecdotes were extracted as one passage due to segmentation error. Such long passages were wrongly filtered out by the rule of maximum length $\mathbf{r_8}$. Therefore, the method for passage segmentation should be improved. Although difficulties arise because web pages are written in a wide variety of styles, it is necessary to explore a general method that can precisely identify the boundaries of the passages.

The performance of the anecdote acquisition method was very sensitive to the target people. For example, the precision for "Beethoven" was 100%, while that for "Mozart" was 17%. Furthermore, the performance of the proposed method was worse than the baseline for "Leonardo" and "Mozart". It could indicate that the writing styles of the anecdotes obtained might be different for those people.

Next, we evaluated a heuristic rule by limiting the number of extracted anecdotes, i.e. the number of the extracted anecdotes per web page was limited to the threshold T_n. TABLE 3 reveals the precision when T_n was set to 1, 2, 3 or unset (no limitation was applied). On average,

the precision was improved by between 8.4% and 13.6%. This indicates that the proposed method was effective in improving precision, although this is based on a simple heuristic rule.

TABLE 3. Precision for different T_n

Person	T_n=1	T_n =2	T_n =3	unset T_n
Ichiro Suzuki	0.400	0.500	0.429	0.429
Oda Nobunaga	0.375	0.429	0.412	0.375
Leonardo da Vinci	1.000	1.000	0.250	0.238
Ludwig van Beethoven	1.000	1.000	1.000	1.000
Wolfgang Amadeus Mozart	0.500	0.167	0.167	0.167
Micro average	0.471	0.484	0.432	0.348

CONCLUSION

This paper proposed a method for automatically acquiring a person's anecdotes from the Web. The extracted anecdotes can provide a useful knowledge base for the dialog system that could then take the initiative in a free conversation. Web pages that included the person's name and the keyword "anecdote" were segmented into passages by analyzing the DOM tree, then the non-anecdotal passages were filtered out by applying the several rules that we have proposed. The experimental results indicated that our method outperformed the baseline by 11% with respect to precision. In addition, our proposed method of limiting the number of the extracted anecdotes per page could further improve precision by 14%. Through error analysis, we found that discrimination between interesting stories and facts was difficult. Although a deep understanding of the text is required to accurately filter out non-anecdotal passages, this paper has reported the initial results of applying relatively simple heuristic rules. The segmentation of passages also needs to be improved, especially when the anecdotes are extracted from a great number of web sites.

In future, we will refine the filtering rules using the syntactic and semantic parsing of the sentences. Syntactic and semantic patterns that frequently appear in the anecdotes will be explored and incorporated into the rules. A machine-learning approach will also be considered in order to determine whether a passage is an anecdote. As for passage segmentation, web pages will be classified into several types according to their writing styles, by the analysis of the structure of the DOM trees. The segmentation methods that are appropriate for each individual type will then be explored.

REFERENCES

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. 2014a. Towards an Open-Domain Conversational System Fully based on Natural Language Processing. *In Proceedings of the 25th International Conference on Computational Linguistics*, 928-939.

Higashinaka, R., Kobayashi, N., Hirano, T., Miyazaki, C., Meguro, T. Makino, T., and Matsuo, Y. 2014b. Syntactic Filtering and Content-based Retrieval of Twitter Sentences for the Generation of System Utterances in Dialogue Systems. *In Proceedings of the 5th International Workshop on Spoken Dialog Systems*, 113-123.

Imabuchi, S. and Ogata, T. 2012. A Story Generation System based on Propp Theory: As a Mechanism in an Integrated Narrative Generation System. *Advances in Natural Language Processing: the 8th International Conference on NLP, JapTAL*, 312-321.

Kawachi, A. 2003. An Analysis of Patterns of Topic-Shift in Japanese Conversations (in Japanese). *Waseda Journal of Japanese Applied Linguistics*, 3, 41-55.

Libin, A. V. and Libin, E. V. 2004. Person-Robot Interactions from the Robopsychologists Point of View: The Robotic Psychology and Robotherapy Approach. *Proceedings of the IEEE*, 92(11), 1789-1803.

McIntyre, N. and Lapata, M. 2009. Learning to tell tales: A data-driven approach to story generation. *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pp. 217-225.

Mizuno, J., Inui, K., and Matsumoto, Y. 2009. Human-computer Interaction System using Web News (in Japanese). *In Proceedings of Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD), The Japanese Society for Artificial Intelligence*, 55, 1-6.

Peinado, F. and Gervas, P. 2006. Evaluation of automatic generation of basic stories. *New Generation Computing*, 24(3):289-302.

Ritter, A., Cherry, C., and Dolan, W. B. 2011. Data-driven response generation in social media. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 583-593.

Shibata, M., Tomiura, Y., and Nishiguchi, T. 2009. Method for Selecting Appropriate Sentence from Documents on The WWW for the Open-Ended Conversation Dialog System (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 24(6), 507-519.

Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y. 2013. Open-Domain Utterance Generation for Conversational Dialogue Systems Using Web-Scale Dependency Structures. *In Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 334-338.

Yoshino, K., Mori, S., and Kawahara, T. 2011. Spoken Dialogue System Based on Information Extraction Using Similarity of Predicate Argument Structures. *In Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 59-66.

Kiyoaki Shirai
Japan Advanced Institute of Science and Technology, JAPAN
kshirai@jaist.ac.jp

Tomotaka Fukuoka
Nextremer Co., Ltd.
tomotaka.fukuoka@nextremer.com

Received: 30 May 2017
Accepted: 30 November 2017