

Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures

Hiroaki Sugiyama*, Toyomi Meguro*, Ryuichiro Higashinaka**, Yasuhiro Minami*

*NTT Communication Science Laboratories

2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, Japan

**NTT Media Intelligence Laboratories

1-1, Hikari-no-Oka, Yokosuka-shi, Kanagawa, Japan

{sugiyama.hiroaki,meguro.toyomi,higashinaka.ryuichiro,minami.yasuhiro}@lab.ntt.co.jp

Abstract

Even though open-domain conversational dialogue systems are required in many fields, their development is complicated because of the flexibility and variety of user utterances. To address this flexibility, previous research on conversational dialogue systems has selected system utterances from web articles based on surface cohesion and shallow semantic coherence; however, the generated utterances sometimes contain irrelevant sentences with respect to the input user utterance. We propose a template-based approach that fills templates with the most salient words in a user utterance and with related words that are extracted using web-scale dependency structures gathered from Twitter. Our open-domain conversational dialogue system outperforms retrieval-based conventional systems in chat experiments.

1 Introduction

The need for open-domain conversational dialogue systems continues to grow. Such systems are beginning to be actively investigated from their social and entertainment aspects (Shibata et al., 2009; Ritter et al., 2011; Wong et al., 2012); conversational dialogues also have potential for therapy purposes and for evoking a user's unconscious requests in task-oriented dialogues (Bickmore and Cassell, 2001). However, developing open-domain conversational dialogue systems is difficult, since the huge variety of user utterances makes it harder to build knowledge resources for generating appropriate system responses. To address this issue, previous research has selected system utterances from web articles or microblogs on the basis of surface cohesion and shallow semantic coherence (Shibata et al., 2009; Jafarpour and Burges, 2010; Wong et al., 2012); however, the selected utterances sometimes contain sentences irrelevant to the user utterance since they originally appeared in a different context.

To satisfy both web-scale topic coverage and suppression of irrelevant sentences, we propose a template-based approach that fills templates with words related to the topic of the user utterance and with words related to the topic-words. This approach enables us to generate a wide range of system responses when we properly extract related words. To obtain words related to topic-words, we analyzed the dependency structures of a huge number of sentences posted to such microblogs as Twitter, where a large number and variety of sentences are posted daily. This way, we can generate a variety of appropriate system responses despite wide variation in user utterances.

We develop a conversational dialogue system that generates system utterances with our proposed utterance generation approach and examine its effectiveness by chat experiments with real users.

2 Related Work

To generate system utterances for conversational dialogue systems, Ritter et al. (2011) proposed a statistical machine translation-based approach that considers source-reply tweet pairs as a bilingual corpus. They compared the following three approaches: IR-status, which retrieves reply tweets whose associated source tweets most resemble the user utterance (Jafarpour and Burges, 2010); IR-response, which retrieves reply tweets that are the most similar to the user utterance; and their proposed SMT-based approach, named MT-chat. They reported that MT-chat outperformed the other approaches and that IR-response was superior to IR-status. However, these approaches used only the words, and not the structures, of user utterances to generate system utterances.

Yoshino et al. (2011) proposed a QA system that answers questions about current events by retrieving, from news articles, descriptions containing similar dependency structures as those of the user's questions. Although this retrieval-based approach is effective for answering the user's factual questions, it is insufficient to generate subjective utterances for conversational dialogue systems since such systems are required to introduce

new topics or to respond with opinions related to user utterances.

3 Open-domain Utterance Generation

Open-domain conversational dialogue systems should be able to respond to any user utterance on any topic. To achieve this, we adopt a template-based approach that estimates the topic of the user utterance, extracts words related to the topic-words, and fills templates with these words. The template-based approach resembles previous rule-based approaches, but these dialogue systems had difficulty achieving coverage for template fillers. In contrast, our approach utilizes the dependency structures of sentences gathered from microblogs that have a wide range of topics, in order to extract the related words used in template-filling. The dependency parser we use is a state-of-the-art Japanese dependency parser that uses Conditional Random Fields trained on text and blog posts, and performs cascaded chunking until all dependencies are found. This parser achieved 84.59% dependency accuracy on a corpus of Japanese blog posts (Imamura et al., 2007).

Microblog posts do not typically contain formulaic utterances such as greetings or back-channels. Therefore, in addition to the template-filling approach, we adopt dialogue act based utterance generation for the formulaic utterances. Figure 1 illustrates the whole architecture of our system.

3.1 Topic-word-driven Template-based Utterance Generation

Our topic-word-driven template-based approach consists of the following three steps: topic estimation, related word extraction, and template-filling utterance generation.

3.1.1 Topic Estimation

We identify three types of potential topic in an input user utterance: proper nouns, common nouns, and predicates (verbs, adjectives, adjectival verbs, and verbal nouns).

Proper Nouns We take the last proper noun that appears in the user’s utterance as a potential topic. Since general Japanese morphological analyzers cannot capture recent proper nouns, we complement the proper noun dictionary entries with Wikipedia entries¹.

Common Nouns To identify potential topics from common nouns, we calculate the inverse document frequency (IDF) of each common noun (all nouns except for proper, time-related, and verbal ones) in the user’s utterance. We use a corpus of microblog posts and treat each post as a document. We adopt the word with the highest IDF as a potential topic.

¹<https://github.com/nabokov/mecab-dic-overdrive>

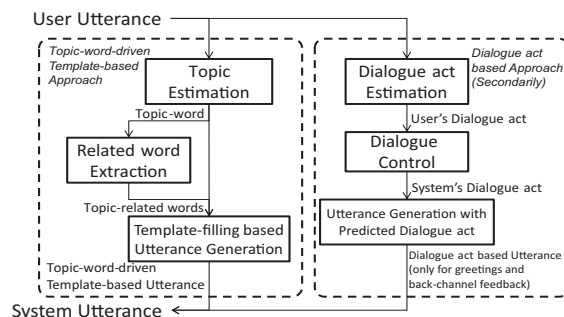


Figure 1: System Architecture

Predicates We take the predicate that composes a dependency in the highest layer of the dependency structure as a potential topic. For example, we adopt “ask”, but not “walk” from the utterance “I asked the man walking on the street”.

3.1.2 Related word Extraction

To obtain topic-related words, a thesaurus or topic model such as Latent Dirichlet Allocation are the most popular approaches (Blei et al., 2003). However, these approaches return semantically similar words to input query words, which do not effectively introduce new information into the system utterances. Therefore, we count the dependencies between words in a huge number of sentences gathered from microblogs, and utilize the most frequently dependent words. This approach enables us to extract adjectives related to proper noun topics; for example, the adjectives *beautiful*, *good*, *clear*, *white*, and *huge* are extracted for Mt. Fuji. Since microblogs contain a huge number of subjective posts, we expect the extracted words to be subjective and suitable for conversational dialogue systems. In this work, we extract adjectives for proper and common nouns, and nouns and their case frames for predicates. Examples of extracted words are shown in Table 2.

3.1.3 Template-filling Utterance Generation

We generate two types of system utterances using manually defined templates: subjective sentences with proper nouns and common nouns; and questions with predicates and their case frames.

Noun-driven Subjective Sentence Generation

We generate system utterances using the proper and common nouns and their related adjectives. Here, we adopt different templates for each word type; proper nouns have explicit meanings, so adjectives related to them are easily suited for any dialogue context. By contrast, since common nouns are used in various contexts in microblogs, adjectives related to common nouns may not fit the dialogue context. Thus, we use “suki” (“like” in English), or “nigate” (“don’t like” in English) in the templates based on the proportion of positive/negative adjectives in the set of related words for a common noun topic. Table 3 shows representative examples for each type. If the system gener-

ates subjective utterances as the system’s own impression of the dialogue topic, the user will expect the system to justify or explain its opinion; however, our system cannot answer that kind of question. Thus, we define the templates using hedges such as “I hear that...” to avoid such questions. The number of templates for proper nouns is eight, and for common nouns is four for each polarity.

Predicate-driven Question Sentence Generation We generate question sentences using predicates and their related nouns and case frames. To elicit user utterances on a particular topic, we generate How/What/Where/When types of questions as shown in Table 3. To select a question word, we use the predicate types and the classes of the related nouns. If the predicate type is adjective or adjectival noun, we select “how” for the question word. If the predicate type is verbal noun or verb and *location* class words appear in the related noun phrase, we select “where” for the question word; the *time* class induces the question word “when”. When no proper noun is found in the topic-word, we select “what”. The number of templates for proper nouns is three for each interrogative type.

3.2 Dialogue act based Utterance Generation

Our approach has difficulty generating appropriate responses to formulaic utterances such as greetings and back-channels. To address this weakness, we adopt dialogue act based utterance generation for these types of utterance. A dialogue act is an abstract expression of a speaker’s intention (Stolcke et al., 2000); we used the 33 dialogue acts defined in Meguro et al. (2010).

Our dialogue act based approach estimates the next dialogue act that the system should output based on the user’s utterance, and generates a system utterance based on the system’s predicted dialogue act if the dialogue act is greetings, sympathy, non-sympathy, filler, or confirmation.

3.2.1 User’s Dialogue act Estimation

We collected 1,259 conversational dialogues from 47 human subjects and labeled each sentence of the collected data using the 33 dialogue acts. 67,801 dialogue acts are contained in the corpus.

We estimated the 33 dialogue acts from user utterances using a logistic regression model and adopted 1- and 2-gram words and 3- and 4-gram characters as model features. We trained our model using 1,000 dialogues and evaluated it using 259 dialogues. The estimation accuracy was about 61%, whereas the human annotation agreement rate was about 59%.

3.2.2 Dialogue control Model and Utterance Generation with Predicted Dialogue act

We developed a dialogue control model that estimates the system’s next dialogue act based on the

user’s dialogue act. The model features are the user’s current dialogue act vector, the system’s last dialogue act vector, and the user’s last dialogue act vector. Each dialogue act vector consists of a 33-dimensional binary vector space. We used the dialogue corpus described above to train and evaluate our model, which we trained with 1,000 dialogues and evaluated using 259 dialogues. The estimation accuracy was 31%, whereas the dialogue act annotation agreement rate between humans is 60%. We exploited the fact that formulaic utterances can pre-define corresponding utterances regardless of the context. Table 4 shows example generated sentences for each dialogue act.

4 Experiment

4.1 Experiment Setting

We recruited ten native Japanese-speaking participants in their 20’s and 30’s (two males and eight females) from outside of the authors’ organization, who have experience using chat systems (not bots). Each participant chatted with the following systems, provided subjective evaluation scores for each system for each of the eight criteria shown in Table 1 (2)-(10) using 7-point Likert scales, and at the end ranked all the systems. We examined the effectiveness of our proposed approach by comparison with the following six systems.

We built the following proposed systems with about 150 M posts gathered from Twitter (excluding posts that contain “@”, “RT”, “http” and brackets, and posts that don’t contain any dependency pairs). At the beginning of a dialogue or the end of a conversation topic when the topic-based approach didn’t generate system utterances, the proposed approaches generated questions such as “What is your favorite movie?” to introduce the next conversation topic. These questions were gathered from utterances in the self-introduction phase (about the five initial utterances) of each dialogue in our dialogue corpus. We manually selected 109 questions that have no context from 179 questions gathered from our corpus, and chose a question at random to generate each topic-inductive question.

Proposed-All This approach used all found topics: proper and common nouns, and predicates. This approach is expected to be well-balanced since it generates both content-focused utterances and general WH-type questions.

Proposed-Nouns This approach used only proper and common nouns, not predicates.

Proposed-Predicates This approach used only predicates, not proper nor common nouns.

Retrieval-Self This approach resembles the IR-response method in Ritter et al. (2011). This approach chose the most similar posts to the user ut-

	Prop.-All	Prop.-Noun	Prop.-Pred.	Ret.-self	Ret.-reply	Human
(1) Number of superior prefs. vs. Prop.-All	-	4	3	0**	2*	9**
(2) Naturalness of dialogue flow	4.0	3.1**	3.5	2.2**	3.5	6.5**
(3) Grammatical correctness	4.0	3.7	4.4	4.1	3.9	6.4**
(4) Dialogue usefulness	3.7	2.9**	3.9	2.7**	3.5	6.1**
(5) Ease of considering next utterance	3.5	3.4	4.4**	2.4**	3.3	5.7**
(6) Variety of system utterances	4.3	4.0	4.2	2.9**	4.0	5.5
(7) User motivation	4.5	4.0*	4.7	3.7*	4.6	5.6**
(8) System motivation that the user feels	4.7	4.1*	4.3	3.5**	4.5	5.7*
(9) Desire to chat again	3.7	2.8**	3.3	2.0**	3.1	5.7**
(10) Averaged score of all evaluation items	4.05	3.50**	4.08	2.93**	3.8*	5.9**

Table 1: System preferences and evaluation scores on 7-point Likert scale (*: $p < .1$, **: $p < .05$)

terance from source posts using the Lucene² information retrieval library, which is an IDF-weighted vector-space similarity. We built about 55 M source-reply post pairs from Twitter.

Retrieval-Reply This approach is the same as the IR-status method in Ritter et al. (2011). It chooses a reply post whose associated source posts most resemble the user’s utterance.

Human As an upper-bound of these systems, the user chats with a human using the same chat interface used by the other systems.

Each dialogue took place over four minutes and was conducted through a text chat interface, and the orders of presentation of systems to participants was randomized. Since the humans have to type their utterances and the systems can generate utterances much faster than typing, we set the transition of the system utterances to about ten seconds to avoid different response intervals between the systems and the humans. Table 5 shows a dialogue example.

4.2 Results and Discussion

Table 1 shows that Proposed-All is ranked the highest of all the automatic systems (1), and achieves the best average evaluation scores (2)-(10). Statistical analyses were performed using the Binomial test for (1) and Welch’s t test for (2) to (10). Proposed-All was ranked higher than the retrieval-based approaches (10 of 10 participants ranked Proposed-All higher than Retrieval-Self, and 8 participants ranked Proposed-All higher than Retrieval-Reply), but none of our three proposed approaches was ranked significantly higher than the others.

The evaluation scores also demonstrate the characteristics of each approach. Proposed-Nouns shows significantly low scores in dialogue flow (2), dialogue usefulness (4), and system motivation (9). Since this approach is overly affected by the nouns in the user utterances, users didn’t feel that the system was actually thinking. Proposed-Predicates shows a high score in ease of thinking about the next utterance (5) since it generates WH-type questions for which users can easily produce answer utterances.

²<http://lucene.apache.org>

For conventional retrieval-based approaches, contrary to Ritter et al. (2011), Retrieval-Self shows significantly lower scores in almost all the evaluation items, and Retrieval-Reply shows scores close to Proposed-All. These results reflect the retrieved corpus size, which is 40 times larger than that of Ritter et al. (2011). When the retrieval performance improves, Retrieval-Self returns posts that are too similar to user utterances, while Retrieval-Reply can find appropriate source posts. Retrieval-Reply shows almost the same scores as Proposed-All for each single evaluation metric, but Retrieval-Reply is inferior to Proposed-All in the averaged evaluation items (10). This is a reason why Retrieval-Reply is also inferior in (1).

None of the systems approached human performance. The users thought that the systems were not able to respond to user utterances that referred to the system itself, like personal questions; and that the systems didn’t understand user utterances since the systems sometimes generate a question that contains different but semantically similar words to those used by the user, due to the lack of thesaurus knowledge.

5 Conclusions

We proposed a novel open-domain utterance generation approach for a conversational dialogue system that generates system utterances using templates populated with topics and related words extracted from a huge number of dependency structures. Our chat experiments demonstrated that our template-based approach generated system utterances preferred over those produced with retrieval-based approaches, and that WH-type questions make it easy for users to produce their next utterance. Our work also indicated that template-based utterance generation, which is considered a legacy approach, has potential when the template-filling resource is huge. Future work includes improving the data-driven topic selection in the proposed approach, the aggregation of words with web-scale class structures like Tamagawa et al. (2012), response generation for utterances that describe the systems themselves, and exploitation of information about the user to generate system utterances.

References

Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 396–403.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Kenji Imamura, Genichiro Kikui, and Norihito Yasuda. 2007. Japanese Dependency Parsing Using Sequential Labeling for Semi-Spoken Language. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 225–228.

Sina Jafarpour and Christopher J.C. Burges. 2010. Filter, Rank, and Transfer the Knowledge: Learning to Chat. Technical Report MSR-TR-2010-93, Microsoft.

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 761–769.

Alan Ritter, Colin Cherry, and William.B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

Masahiro Shibata, Tomomi Nishiguchi, and Yoichi Tomiura. 2009. Dialog System for Open-Ended Conversation Using Web Documents. *Informatica*, 33:277–284.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Susumu Tamagawa, Takeshi Morita, and Takahira Yamaguchi. 2012. Extracting Property Semantics from Japanese Wikipedia. In *Proceedings of the 8th international conference on Active Media Technology*, pages 357–368.

Wilson Wong, Lawrence Cavedon, John Thangarajah, and Lin Padgham. 2012. Strategies for Mixed-Initiative Conversation Management using Question-Answer Pairs. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2821–2834.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 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*, pages 59–66.

Appendix

Topic-word	Extracted topic-related words
Mt.Fuji	beautiful, good, clear, white, huge (adj)
Bread	delicious, good, like, cute (adj)
Eat	I (“ga”: subject-case), rice (“wo”: target-case) together (“ni”: “with”; “to”)

Table 2: Extracted topic-related words (translated by authors)

Word type	Template example
Proper nouns	I hear that [Noun] is [adj], isn’t it? [Noun] ha [adj] deshitakke? (JP)
General nouns	Actually, I am weak in [Noun]... [Noun] ha chotto nigate desu... (JP)
Adjectives and adjectival nouns	How does it [Noun]? douiu huuni [Noun] nndesuka? (JP)
Verbs and verbal nouns	Where do you [Verb]? doko de [Verb] nndesuka? (JP)

Table 3: Templates for topic types (translated by authors)

Dialogue act	Example
Greetings	Hello.
Sympathy	Good./ Nice.
Non-sympathy	Not so good..
Filler	Uh.
Confirmation	Really?/Oh, [Noun].

Table 4: Generated utterances for dialogue acts (translated by authors)

Speaker	Utterance
User	Hello.
System	Hello. Do you like movies?
User	I like movies.
System	Good. I think movie is fun, isn’t it?
User	I like reading more than going to the movies.
System	Reading is also good.
User	I borrow books from my local library.
System	Nice. I hear that libraries are quiet.
User	Yes, they usually are.
	I used to study, write reports, and read books.

Table 5: Dialogues between user and Proposed-All (translated by authors)