Towards an open-domain conversational system fully based on natural language processing

Ryuichiro Higashinaka[1], **Kenji Imamura**[1], **Toyomi Meguro**[2], **Chiaki Miyazaki**[1]
Nozomi Kobayashi[1], **Hiroaki Sugiyama**[2], **Toru Hirano**[1]
Toshiro Makino[1], **Yoshihiro Matsuo**[1]
[1]NTT Media Intelligence Laboratories
[2]NTT Communication Science Laboratories
{higashinaka.ryuichiro, imamura.kenji, meguro.toyomi, miyazaki.chiaki,
kobayashi.nozomi, sugiyama.hiroaki, hirano.tohru,
makino.toshiro, matsuo.yoshihiro}@lab.ntt.co.jp

Abstract

This paper proposes an architecture for an open-domain conversational system and evaluates an implemented system. The proposed architecture is fully composed of modules based on natural language processing techniques. Experimental results using human subjects show that our architecture achieves significantly better naturalness than a retrieval-based baseline and that its naturalness is close to that of a rule-based system using 149K hand-crafted rules.

1 Introduction

Although task-oriented dialogue systems have been extensively researched over the decades (Walker et al., 2001; Williams et al., 2013), it is only recently that non-task-oriented dialogue, open-domain conversation, or chat has been attracting attention for its social and entertainment aspects (Bickmore and Picard, 2005; Ritter et al., 2011; Bessho et al., 2012). Creating an open-domain conversational system is a challenging problem. In task-oriented dialogue systems, it is possible to prepare knowledge for a domain and create understanding and generation modules for that domain (Nakano et al., 2000). However, for open-domain conversation, such preparation cannot be performed. Since it is difficult to handle users' open-domain utterances, to create workable systems, conventional approaches have used hand-crafted rules (Wallace, 2004). Although elaborate rules may work well, the problem with the rule-based approach is the high cost and the dependence on individual skills of developers, which hinders systematic development. Another problem with the rule-based approach is its low coverage; that is, the inability to handle unexpected utterances.

The recent increase of web data has propelled the development of approaches that use data retrieved from the web for open-domain conversation (Shibata et al., 2009; Ritter et al., 2011). The merit of such retrieval-based approaches is that, owing to the diversity of the web, systems can retrieve at least some responses for user input, which solves the coverage problem. However, this comes at the cost of utterance quality. Since the web, especially Twitter, is inherently noisy, it is, in many cases, difficult to sift out appropriate sentences from retrieval results.

In this paper, we propose an architecture for an open-domain conversational system. The proposed architecture is fully composed of modules based on natural language processing (NLP) techniques. Our stance is not just to hand-craft or to search the web for utterances, but to create a system that can fully understand and generate utterances. We want to show that it is possible to build an open-domain conversational system by combining NLP modules, which will open the way to a systematic development and improvement. We describe our open-domain conversational system based on our architecture and present results of an evaluation of its performance by human subjects. We compare our system with rule-based and retrieval-based systems, and show that our architecture is a promising direction. In this work, we regard the term open-domain conversation to be interchangeable with non-task-oriented dialogue, casual conversation (Eggins and Slade, 2005), chat, or social dialogue (Bickmore and Cassell, 2000). We use the term to denote that user input is not restricted in any way as in open-domain question answering

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 928–939, Dublin, Ireland, August 23-29 2014.

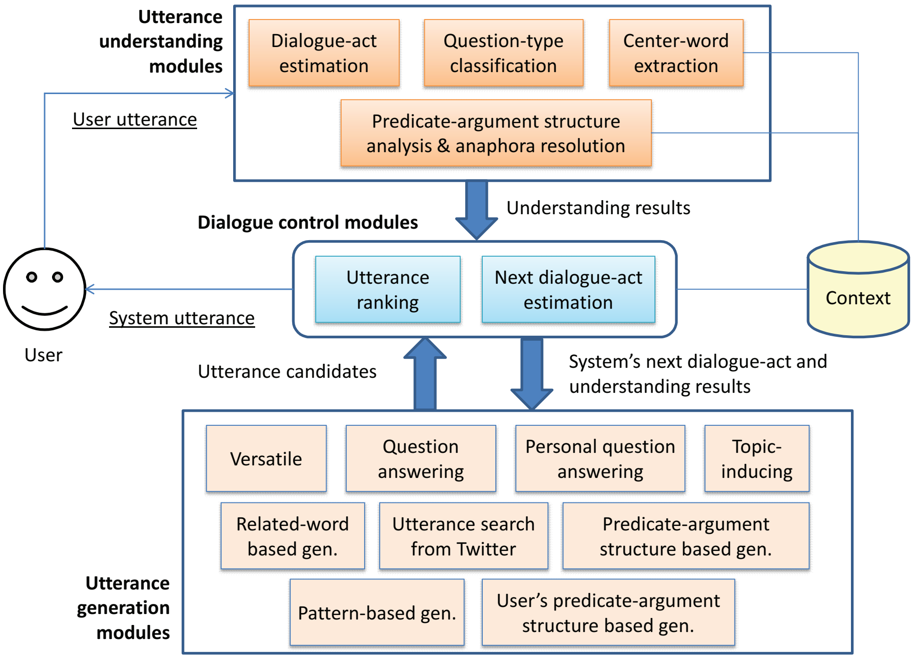
Utterance understanding modules
Dialogue-act estimation
Question-type classification
Center-word extraction
Predicate-argument structure analysis & anaphora resolution
User utterance
Understanding results
Dialogue control modules
Utterance ranking
Next dialogue-act estimation
Context
User
System utterance
Utterance candidates
System's next dialogue-act and understanding results
Versatile
Question answering
Personal question answering
Topic-inducing
Related-word based gen.
Utterance search from Twitter
Predicate-argument structure based gen.
Utterance generation modules
Pattern-based gen.
User's predicate-argument structure based gen.

Figure 1: System architecture.

(QA) (Voorhees and Tice, 2000) and open information extraction (Etzioni et al., 2008). The application here in mind is one that can chat with users like chatbots. It should also be noted that we deal with Japanese text chat in this paper, although we believe the architecture to be largely language-independent and extendable with other modalities.

In Section 2, we describe the architecture and its underlying modules. In Section 3, we describe the rule-based and retrieval-based systems that we use for comparison. In Section 4, we describe the experiment we performed to evaluate our system. Section 5 summarizes the paper.

2 Architecture and System Description

Figure 1 shows the architecture we propose for an open-domain conversational system. The architecture has three main components: utterance understanding, dialogue control, and utterance generation. Following the literature on discourse theory (Grosz and Sidner, 1986), we regard *intention* (intentional structure), *topic* (attention state), and *content* (linguistic structure) as three important elements in conversation, and seek to create a system that can understand and generate on the basis of them in a general way. The dialogue control component works by ranking utterance candidates using a general coherence criterion (Hovy, 1991). Note that the overall architecture is roughly the same as conventional dialogue systems; however, the internal architecture is different so as to allow open-domain conversation. To give a rough idea of how the system works, Figure 2 shows an example dialogue between our system and a user (one of the subjects in our experiment). As this example shows, the system can handle various user utterances. Below, we describe how this is achieved.

2.1 Utterance Understanding Modules

We identify dialogue-act, question-type, center-word, and predicate-argument structure (PAS). Dialogue-act and question-type correspond to intention, center-word to topic, and PAS to content. We use PASs because they can represent an arbitrary sentence. For languages other than Japanese, instead of PASs, semantic role labeling (SRL) can be used (Palmer et al., 2010). Below, we describe each module.

Dialogue-act estimation: As a dialogue-act tag set, we use the one proposed by Meguro et al. (2013). Although their tag set is designed for annotating listening-oriented dialogue (LoD), since speakers in LoD are allowed to speak freely, the tag set can cover diverse utterances, making it suitable for open-domain conversation. There are 33 dialogue-acts in the tag set. See (Meguro et al., 2013) for details. We used 1259 LoDs annotated with dialogue-acts and trained a classifier using a support vector machine (SVM). The features used are word N-grams, semantic categories (obtained from a Japanese thesaurus Goi-Taikei (Ikehara et al., 1997)), and character N-grams. Here, unless otherwise noted, we use JTAG (Fuchi and Takagi, 1998) for morphological analysis in this work. When we use the LoD data for training and testing, by a ten-fold cross validation, the estimation accuracy is 45%, which is reasonable when considering that the inter-annotator agreement rate is 59%. For reference, the majority baseline, which estimates the dialogue-acts of all utterances to be information-provision, has 12% accuracy.

Question-type classification: We use the question taxonomy by Nagata et al. (2006) because it was derived by analyzing questions from the general public and therefore covers diverse questions. The taxonomy has 23 question types under five main categories: name, quantity, explanation, yes-no, and other. Since some types could be too specific, by merging similar ones, we shrinked the 23 types into 13: name-other, name-person, quantity-other, quantity-date, quantity-period, quantity-money, yes-no, explanation-reason, explanation-definition, explanation-method, explanation-reputation, explanation-association, and other. Using an in-house data set of about 48K questions annotated with the 13 types, we trained a logistic-regression-based classifier that achieves a classification accuracy of 92.5% by a five-fold cross validation. The majority baseline that always classifies to name-other has 39.5% accuracy.

Center-word extraction: We define a center-word as a noun phrase (NP) that denotes the topic of a conversation. We hypothesize that an utterance has at most one NP suitable for a center-word. To extract an NP from an utterance, we use conditional random fields (Lafferty et al., 2001); NPs are extracted directly from a sequence of words without creating a parse tree. For the training and testing, we prepared 10K sentences with center-word annotation. Here, the sentences were those randomly sampled from the open-domain conversation corpus (See Section 2.2). The feature template uses words, part-of-speech (POS) tags, and semantic categories of current and neighboring words. The extraction accuracy is 83.4% by a five-fold cross validation. This module has access to the context. When there are already center-words (represented by a stack) in the context, the center-word of the current utterance is placed at the top of the stack.

PAS analysis and anaphora resolution: In PAS analysis, predicates and their arguments are detected. A predicate can be a verb, adjective, or copular verb, and the arguments are NPs associated with cases in case grammar. As cases, we use standard cases `ga` (nominative), `wo` (accusative), `ni` (dative) as well as `de` (locative/instrumental), `to` (with), `kara` (source), `made` (goal). We use the PAS analyzer described in (Imamura et al., 2009; Imamura et al., 2014a; Imamura et al., 2014b). The analyzer works statistically by ranking NPs in the context using supervised learning with an obligatory case information dictionary and a large-scale word dependency language model. For the ranking, NPs in the current utterance and the center-words in the context are considered. Also considered are the special NPs ‘I’, ‘YOU’, ‘exog (exophora general)’, and ‘null’. Here, ‘exog’ means that there should be some argument for a case but it is missing in the context, and ‘null’ means that no argument is required for a case. When arguments are filled with NPs from the context, this means that zero pronoun resolution (a type of anaphora resolution) has been performed. The analyzer is reported to have an accuracy of 57-62% for `ga`, `wo`, and `ni` cases for dialogue data (Imamura et al., 2014b). When a zero pronoun is resolved, the NP used for the resolution is placed at the top of the center-word stack because such an elided NP is grounded among the conversational participants and is likely to be the focus of conversation (Walker et al., 1994).

2.2 Dialogue Control Modules

We have two modules for dialogue control. The next dialogue-act estimation module roughly decides the next user utterance on the intention level (dialogue-act level). The utterance ranking module ranks utterance candidates (See Section 2.3 for how utterance candidates are created) by a coherence criterion.

	Utterance (English translation by the authors)	DA	Gen. Module
SYS1	Doumo desu (Hi)	Greeting	Initial prompt
USR1	Doumo. Genki desuka? (Hi. How are you?)	Greeting	
SYS2	Saikin no oishii mono ni tsuite kikasete kudasai (Tell me about delicious food that you've had recently)	Q-Plan	Topic-Inducing
USR2	Karei ni hamatterunda! (I'm now into curry!)	S-Disc Fact	
SYS3	Aishou ga ii desuka (Does it go together well with you?)	Confirmation	PAS
USR3	Toku ni tai karei wa aishou ga yoimitai (Thai curry especially goes together well with me)	S-Disc Pref. Pos.	
SYS4	Guzai ga gorogoro shitete, natsukashii karei ga sukidesu (I like nostalgic curry with large chunks of ingredients)	S-Disc Pref. Pos.	Twitter
USR4	Yokosuka no karei nanka, soudayone (Curry in Yokosuka is like that, isn't it?)	Information	
SYS5	Watashi mo daisuki desu! (I like it very much!)	S-Disc Pref. Pos.	Versatile
USR5	Yokosuka, ikimasuka? (Do you go to Yokosuka?)	Q-Fact	
SYS6	Yokosuka sunderunde, nanka ureshii desu (I live in Yokosuka, so I'm flattered)	S-Disc Fact	Twitter

Figure 2: Example dialogue with our proposed system [Prop (tri); See Section 4.1]. DA denotes recognized dialogue-acts for user utterances and those chosen by the system for system utterances. Gen. Module indicates the generation module used. Underlined words denote newly recognized center-words.

Next dialogue-act estimation: Using the same approach as Sugiyama et al. (2013), this module takes into account three previous dialogue-acts in the context to estimate the system's next dialogue-act. We trained an SVM-based dialogue-act estimator using 1259 LoDs. Using 1,000 dialogues as training data and 259 as test data, the trained estimator can predict the next dialogue-act with the accuracy of 28% (NB. majority baseline has 15% accuracy). Although the accuracy is low, since the task is subjective and there is no definite answer, we consider the estimator to have sufficient ability to choose a reasonable next dialogue-act.

Utterance ranking: We adopt coherence as a general criterion for ranking utterances because it is a well-recognized measure of discourse and can be applied to arbitrary sentences (Lapata, 2003; Barzilay and Lapata, 2008). We hypothesize that an utterance that is the most cohesive to the current context should be chosen for the output of the system. To create the ranker, we first collected a data set of 3,680 open-domain conversations (hereafter, open-domain conversation corpus; 134K utterances) between humans, and from 3,496 of them (184 were held out for development), created dialogue snippets (excerpts) by taking N consecutive utterances. We use these dialogue snippets as references (positive examples). We also create counter-references (pseudo negative examples) by swapping the last utterance of each snippet with a randomly selected one from the dialogue from which the snippet was taken. This is similar to how Barzilay and Lapata (2008) created their training data for their coherence models. We then train a ranker by ranking SVM (Joachims, 2002) in the same manner as (Higashinaka and Isozaki, 2008). The ranker is trained so that references are ranked higher than counter-references. Following Lapata (2003), who used pairs of words for sentence ordering, we use, as features, the pairs of words, POS-tags, and semantic categories between the last utterance and each of the previous utterances. For example, when the last utterance U_l has k words and one of its previous utterance U_p has m words, we create $k \times m$ features by combining them. This feature generation is done also for POS-tags and semantic categories and is iterated over all previous utterances. We trained two rankers using 2 and 3 for N. When N is 2, we have 124,213 snippets. When N is 3, we have 120,717. By using four-fifths of the data for training and using the remaining one-fifth for testing, the rankers achieve 66.7% and 66.4% accuracies for N=2 and N=3, respectively. Here, the random baseline's accuracy is 50%. We only use 2 and 3 for N here since a larger N could lead to the explosion of features. By default, we use the ranker trained with N=3. The trained ranker ranks utterance candidates (generated by the modules in Section 2.3) and outputs the top one as a system utterance.

2.3 Utterance Generation Modules

We prepared nine modules for generation. The versatile, QA, and personal QA modules generate on the basis of dialogue-acts and question-types (intention). The topic-inducing, related-word, Twitter, and

PAS modules generate on the basis of center-words (topic). The pattern and user PAS modules generate by using the surface string and PASs of user utterances (content). Note that, for all modules, the system's next dialogue-act is taken into account; that is, wherever necessary, the aforementioned dialogue-act estimation module is applied to generated utterances so that utterances whose estimated dialogue-acts match the system's next dialogue-act are returned.

Versatile: This module receives the system's next dialogue-act and returns utterances randomly chosen from the list of utterances for that dialogue-act. To create lists for dialogue-acts, we first extracted frequent utterances for each dialogue-act in the LoD corpus. Then, we selected context-independent utterances for the dialogue-act. We call such utterances "versatile utterances" because they can be used in various situations. For example, we have "I like it", "It is good", and "That's great" for S-Disc Pref. Pos. (a dialogue-act that discloses one's positive preference).

QA: When the user dialogue-act is a question and the question-type requires a named entity as an answer (i.e., when the question-type starts with a 'name' or 'quantity'), we call an off-the-shelf QA API that is publicly available[1]. The API returns top-N answers (NEs) for a natural language query (Uchida et al., 2013; Higashinaka et al., 2013). We refer to this API with the user input sentence as a query and obtain the top-five answers. This module returns these answers as utterance candidates.

Personal QA: When the user dialogue-act is a question, this module is called for answering personal questions. Answering such questions is important in chat (Batacharia et al., 1999) or even in task-oriented dialogue (Takeuchi et al., 2007). We use the same method as (Sugiyama et al., 2014b; Sugiyama et al., 2014a) and create a person database (PDB) of question-answer pairs for a persona. In the PDB, the questions are given category labels (e.g., favorite sport, whether the persona likes dogs, etc.) as well as question-types based on our taxonomy. Given a question, the answer is obtained by searching the PDB by the category label and the question-type for the question. To obtain the category label, a separately-trained logistic-regression-based classifier is used. We prepared a PDB for a persona 'Aiko' (a 29 year-old Japanese woman). The PDB contains 4,428 question-answer pairs. This module searches Aiko's PDB and returns obtained answers as utterance candidates.

Topic-inducing: When there is no center-word, this module returns utterances that introduce topics (e.g., "Let's talk about favorite foods!"). The utterances are chosen randomly from a list of utterances that we extracted from the dialogue-initiating utterances in the LoDs.

Related-word: The input to this module is the top center-word (C) and the next dialogue-act. For given C, we first get its related-words. Although we cannot describe details for lack of space, as related-words, we have attributes, question words, associative words, and category words. Such words are mined from blogs, Twitter, and Wikipedia by using lexico-syntactic patterns (Hearst, 1992). By combining related-words with a small number of templates, utterances are created. For example, we have a template "C wa ADJ desune (C is ADJ)" where ADJ is an adjectival attribute of C. The created utterances are returned as utterance candidates. This approach is similar to that used by (Higuchi et al., 2008) and (Sugiyama et al., 2013) in that words obtained from large text data are combined with templates for generation.

Twitter: We use the same approach as (Higashinaka et al., 2014), who created a database of Twitter sentences by word-level and syntactic-level filtering. The database is searched by a query expanded with its related-words so that tweets relevant to the query can be accurately retrieved. It has been reported that only 6% of the retrieved results are judged as inappropriate by subjective evaluation. Using the same database and method as (Higashinaka et al., 2014), this module returns the top-ten retrieved sentences from the database using the top center-word as a query. The database contains about 7M sentences.

[1] `https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_docs_id=6`

PAS: We created a database of PASs by processing more than three years' of blogs. For fear of noise, we only harvested PASs that have just a predicate and an argument for `ga` (nominative) with its topic (an NP) explicitly marked by a topic marker `wa`. From the blogs, we obtained 146K PASs for 50K topics. Given the top center-word and the next dialogue-act, this module looks for PASs whose topic matches the top center-word. Then, it converts the PASs into sentences so that they can convey the intention of the system's next dialogue-act. This conversion is automatic: we first convert the PASs into declarative sentences using a simple rule. Then, their sentence-end expressions (NB. In Japanese, modalities are mostly expressed by sentence-end expressions) are swapped with those matching the target dialogue-act. The sentence-end expressions used here are those automatically mined from dialogue-act annotated dialogue data. This module returns the converted sentences.

Pattern: In everyday conversation, there are typical exchanges of utterances like adjacency pairs (Schegloff and Sacks, 1973). To obtain such exchanges, we mined Twitter. We first collected about 919M tweets. Then, by extracting tweets connected with an in-reply-to relationship, we created a Twitter conversation corpus (20M conversations containing 90M tweets). By taking two consecutive tweets in the corpus and retaining only the frequent ones by a cut-off threshold of ten occurrences, we obtained 22K utterance pairs. The input to this module is the user utterance string, and the module outputs utterances from matched utterance pairs.

User PAS: This module uses the PASs of the user utterance and the next dialogue-act. It performs the same operation as the PAS-based generation and returns the converted sentences. The merit of this module is that the system can use the user's content in its utterance, which has been found to be useful in casual conversation for showing understanding (Ivey et al., 2013) and entraining with users (Nenkova et al., 2008).

3 Rule-based and Retrieval-based Systems

For comparison, we prepared a rule-based system and a retrieval-based one. Since there is no off-the-shelf rule-based system in Japanese, we created one on our own. Because we wanted to compare our system with a state-of-the-art rule-based system, we put a great deal of effort in its development. Remember that creating rules is still the standard way of creating an open-domain conversational system. Last year's Loebner Prize (a chatbot contest) winner, Mitsuku, was based on rules written in artificial intelligence markup language (AIML) (Wallace, 2004). Beating a reasonably created rule-based system is a hard task. As for the retrieval-based system, we use a large Twitter database to replicate the current best retrieval-based system, which we regard as our baseline.

3.1 Rule-based System

We had a seasoned engineer, who specializes in text analysis, create rules in AIML for two and a half months. We used only one person to ensure consistency in the rules. He first created some initial rules by referring to the AIML rules of A.L.I.C.E. (Wallace, 2004). The rules here include common exchanges of utterances (e.g., compliments of the season) as well as those to evade weak topics, changes of topics, and innocuous utterances. Then, he used the open-domain conversation corpus (with 10% held out for evaluation as we describe later) to extract two consecutive utterances and coded them as rules, making the first utterance the matching pattern and the second utterance the output sentence (called a template in AIML). This process involved replacing certain words with asterisks (wildcard) to widen the coverage of patterns and modifying templates if necessary. To augment the rules, he also used the topic-based utterance corpus that we created. The corpus contains 25K utterances conceived by 20 individuals for 250 topic words. The instruction given to the individuals was to come up with an interesting utterance that you think you would like to hear for a given topic word. Here, the topic words are frequent NPs in the open-domain conversation corpus. The pairs of a topic word and its utterance were coded as rules. To enable the system to answer personal questions, the Aiko PDB was employed. Since the PDB comprises a set of question-answer pairs, the PDB was straightforwardly incorporated into AIML with minor modifications. The rules were loaded by ProgramD, an AIML interpreter (http://aitools.org/Program_D), and were tested

	Utterance (English translation by the authors)
SYS1	Doumo desu. (Hi)
USR1	Konnichiwa (Hello)
SYS2	Konnnichiwa. Ryokou wa suki desuka? (Hello. Do you like traveling?)
USR2	Hai, sukidesu. (Yes, I do)
SYS3	Kokunai ryokou ga ooi desuka, kaigai ryokou ga ooi desuka? (Do you travel in Japan or go overseas?)
USR3	Moppara kokunai desu nee. (Mainly in Japan)
SYS4	Watashi wa moppara kokunai ryokou desu. (I travel mainly in Japan)
USR4	Onaji desune (It's the same)
SYS5	Soudesu yone (It is indeed)
USR5	Hai (Yes)
SYS6	Shuugaku ryokou wa doko deshitaka? (Where did you go for the school trip?)
USR6	Nagano ken deshita. (We went to Nagano)
SYS7	Nagano no kikou, natsu wa hontoni iidesu yone. (The climate in Nagano, it is really great in summer)
USR7	A, eeto, shuugaku ryokou wa nagano no yukiyama deshita. (Well, actually, the school trip was in the snowy mountains in Nagano)

Figure 3: Example dialogue with the rule-based system.

	Utterance (English translation by the authors)
SYS1	Doumoo (Hi)
USR1	Konnnichiwa (Hello)
SYS2	Chuugaku jidai no uiuishii midiamuhea ga natsukashii (It brings back the memories about the medium hair in junior high)
USR2	Nani wo totsuzen, doushitan desuka! (Where did that come from? Are you OK?)
SYS3	Mou, kimochi iikara konomama nemashouyo suyaa (Well, since I feel very relaxed, let's sleep now)
USR3	Yaruki nai desunee (Are you even trying?)
SYS4	Baito ganbare! Yoru wa hanabi ja! (Good luck on your part-time job! We have fireworks at night!)
USR4	Hanabi ni ikun desuka? (Do you go to fireworks?)
SYS5	Ikou to omottoru. Iya, iku. (I'm planning to...No, I'm definitely going)
USR5	Dare to ikun desuka? (With who?)
SYS6	Kimetenai (I haven't decided yet)
USR6	Ima kara kimeruno? (Will you decide from now?)

Figure 4: Example dialogue with the retrieval-based system.

by an external human checker. The checker randomly extracted 100 utterances from the remaining 10% of the open-domain conversation corpus and put each utterance into ProgramD and subjectively evaluated the adequacy of each output. If more than 90% of the output was judged adequate, the rule creation process ended; otherwise, the rules had to be improved for another check. After several iterations, the 90% criterion was met, resulting in 149,300 rules. As this process illustrates, the rules were meticulously crafted. Our rule-based system uses ProgramD loaded with these rules. Figure 3 shows an example dialogue with the rule-based system (collected in our experiment; see Section 4.1).

3.2 Retrieval-based System

We used the Twitter conversation corpus (See Section 2.3) to create a database for retrieval. We extracted two consecutive utterances as input-output pairs and indexed them using the text search engine Lucene (http://lucene.apache.org/core/). For a given utterance as a query, the top-ten utterance pairs are retrieved on the basis of the similarity between the query and the input-part of the indexed pairs. Here, the similarity is the cosine similarity of TF-IDF weighted word vectors. Then, one of the retrieved pairs is randomly selected to produce the system's next utterance. Here, we adopt random selection so that the same utterance won't be uttered for the same input. Since the amount of indexed tweets is large (90M), we consider this to be a reasonable baseline. This system is our replication of IR-Status in (Ritter et al., 2011). Figure 4 shows an example dialogue with the retrieval-based system.

4 Experiment

We evaluated our proposed system in an experiment using human judges. We compared it with the rule-based and retrieval-based systems.

Questionnaire	(a) Rule	(b) Retrieval	(c) Prop (noTW)	(d) Prop (noPAS)	(e) Prop (bi)	(f) Prop (tri)
Q1 Naturalness	**3.88**^{bbe}	***2.68***	3.60^{bb}	3.48^{bb}	3.33^{bb}	3.44^{bb}
Q2 Generation	**4.40**^{bbe}	***2.80***	4.03^{bb}	3.98^{bb}	3.80^{bb}	3.92^{bb}
Q3 Understanding	**3.73**^{bb}	***2.61***	3.46^{bb}	3.33^{bb}	3.16^{b}	3.25^{bb}
Q4 Informativeness	**3.00**^{bb}	***2.24***	2.70	2.65	2.62	2.80^{b}
Q5 Diversity	**3.58**	3.44	***3.08***	3.17	3.27	3.38
Q6 Continuity	**3.87**^{bbf}	***2.63***	3.44^{bb}	3.38^{bb}	3.41^{bb}	3.23^{bb}
Q7 Willingness	**3.60**^{bb}	***2.64***	3.25^{b}	3.14^{b}	3.12^{b}	3.12^{b}
Q8 Satisfaction	**3.62**^{bb}	***2.72***	3.24^{b}	3.21^{b}	3.13	3.13

Table 1: Subjective evaluation results: ratings averaged over all dialogues for each system. Superscripts a–f next to the numbers indicate that the number is statistically better than systems (a)–(f), respectively. Double-letters (e.g., bb) mean $p < 0.01$; otherwise $p < 0.05$. For the statistical test, we used a Steel-Dwass multiple comparison test (Dwass, 1960). The largest and smallest numbers in a row are indicated by bold and bold italic font, respectively.

4.1 Experimental Procedure

We recruited 30 human subjects (14 males and 16 females, ages from 18 to 55). They were paid for their participation. Each participant took part in 24 dialogue sessions, talking four times to each of six different systems. The systems used were (a) the rule-based system, (b) the retrieval-based system, and four different configurations of our proposed system: (c) Prop (noTW), in which utterance search from Twitter is disabled; (d) Prop (noPAS), in which PAS-based generation is disabled; (e) Prop (bi), where N=2 is used for utterance ranking (See Section 2.2); and (f) Prop (tri), in which no module is disabled. All systems start a conversation with a greeting prompt. Each dialogue session lasted for two minutes. Two-minute interaction could be short, but we wanted to test the systems with different topics that can change dialogue-by-dialogue. The participants were instructed to enjoy the conversation with the systems. No dialogue topic was specified. No prior knowledge was provided about the systems, including the number of systems they were to talk to. The order of the systems was randomized. Since the rule-based and retrieval-based systems require less computation, four seconds of delay was inserted before their utterances. After each dialogue, each participant filled out a questionnaire comprising eight items (See the column Questionnaire in Table 1) asking for his/her subjective evaluation of the dialogue on a seven-point Likert scale, where 1 is the worst and 7 the best. We asked the participants not to take into account the delay of system responses for their evaluation. After all 24 sessions, each participant filled in a free-form opinion sheet to end the participation.

4.2 Results and Analyses

Table 1 shows the results of subjective evaluations. As can be seen from the table, the rule-based system performed the best and the retrieval-based system performed the worst. The retrieval-based system was the worst for all questionnaire items except Q5 (Diversity of system utterances); at least the large Twitter database produced diverse utterances. Our proposed systems placed between the rule-based and retrieval-based systems. The averaged scores and the results of statistical tests indicate that our systems are significantly better than the retrieval-based baseline and that our systems' performance is close to that of the rule-based system. The difference between the rule-based and our proposed systems is not statistically significant (except for a small number of cases). When we focus on Q1 (Naturalness of dialogue), Prop (noTW) attained a score of 3.6, which is close to that of the rule-based system (3.88). This indicates that our system has the ability to perform reasonably natural conversation and that it is possible to create a system of rule-based-level naturalness with our architecture. As for other questionnaire items, the difference between our systems and the rule-based system is a little wider in mean scores. Although further examination is needed, this is probably because user satisfaction is related to more sensitive issues such as politeness, linguistic style, consistency, and users' preferences.

When we look at the difference in the four configurations, we see that Prop (noTW) is consistently

System		# Uniq utt	# Uniq word	# Utt	# Word	# Word/Utt	Perplexity
(a) Rule	USR	915	956	1049	5838	5.565	59.81
	SYS	***353***	803	1169	9565	**8.182**	***23.46***
	ALL	1263	1333	2218	**15403**	6.945	60.47
(b) Retrieval	USR	937	995	1067	5007	***4.693***	61.22
	SYS	1016	2043	1186	7744	6.530	80.30
	ALL	**1936**	**2449**	**2253**	12751	5.660	**100.48**
(c) Prop (noTW)	USR	750	889	879	4875	5.546	58.76
	SYS	613	***698***	999	5820	5.826	34.17
	ALL	1345	1187	1878	10695	5.695	57.76
(d) Prop (noPAS)	USR	744	852	***865***	***4823***	5.576	54.82
	SYS	551	807	985	6394	6.491	45.50
	ALL	1279	1242	1850	11217	6.063	67.89

Table 2: The number of unique utterances, unique words, utterances, words, words per utterance, and perplexity for systems (a)–(d). The results for systems (e) and (f) are omitted because they are between those for (c) and (d). USR, SYS, and ALL indicate rows for user, system, and all utterances, respectively. For perplexity calculation, half the data were used to train a trigram language model to be tested with the other half. Bold and bold italic font indicates max and min in each column.

Module	(c) Prop (noTW)	(d) Prop (noPAS)	(e) Prop (bi)	(f) Prop (tri)
Versatile	0.281 (247)	0.274 (237)	0.198 (174)	0.205 (173)
QA	***0.008*** (7)	***0.003*** (3)	***0.008*** (7)	***0.005*** (4)
Personal QA	0.047 (41)	0.054 (47)	0.033 (29)	0.046 (39)
Topic-Inducing	0.143 (126)	0.142 (123)	0.129 (113)	0.157 (132)
Related-word	0.060 (53)	0.090 (78)	0.043 (38)	0.027 (23)
Twitter	N/A	**0.358** (310)	0.187 (164)	0.253 (213)
PAS	**0.383** (337)	N/A	**0.280** (246)	**0.265** (223)
Pattern	0.032 (28)	0.031 (27)	0.081 (71)	0.021 (18)
User PAS	0.046 (40)	0.046 (40)	0.042 (37)	0.021 (18)

Table 3: Selected ratios (raw counts in parentheses) for the generation modules. Bold and bold italic font indicate max and min in each column.

better than the others except for Q4 and Q5. Since the main difference is whether Twitter sentences are used, this is probably the cause. The reason could be the inconsistency of linguistic styles in Twitter or the noise that could not be suppressed by the filtering. Since Twitter sentences surely augment diversity, we would like to consider ways to make better use of them, for example, by normalizing the linguistic style and applying stricter filters. There is a slight tendency for Prop (tri) to be preferred to Prop (bi), which is reasonable because it uses more context for deciding the next utterance. In the future, we would like to pursue methods that can exploit longer context, such as entity grids (Barzilay and Lapata, 2008) and co-reference structures (Swanson and Gordon, 2012).

We performed a brief analysis of the collected dialogues. Table 2 shows, for each system, the number of unique utterances, unique words, utterances, words, words per utterance, and perplexity. It can be seen that the utterances of the rule-based system are very rigid: the perplexity is very low (23.46) and there are only 353 unique utterances, which is about half of that of the other systems. It is interesting that, despite this fact, the rule-based system was perceived to produce the most diverse utterances by questionnaire. Since the rule-based system produced much longer utterances (8.182), this probably had a positive effect for the perceived diversity. In terms of natural interaction, it is not desirable for one participant to contribute more than the other. In this respect, our proposed systems seem appropriate because the users and the systems exchange a similar number of words per utterance.

Table 3 shows the selected ratios for the generation modules. It can be seen that all modules contributed to conversation. The most frequent ones were Twitter and PAS-based generation, followed by the versatile and topic-inducing modules. Although QA and personal QA were not used as frequently for output, when we examined the logs, we found that there were many cases where these modules could not obtain any answer from the QA API or the PDB. Since answering questions is a basic function in conversation, this needs to be improved. Similarly, we also want to evaluate the contribution of each module quantitatively, for example, by associating the behavior of each module with user subjective evaluations in a framework similar to PARADISE (Walker et al., 2000). Enabling this kind of analysis is a clear benefit of having an architecture such as the one we proposed.

5 Summary

This paper proposed an architecture for an open-domain conversational system and evaluated an implemented system. The results indicate that our architecture enables better dialogue than a retrieval-based baseline using a large Twitter database. Although our system could not reach the level of a carefully crafted rule-based system and still has a number of limitations, our architecture can achieve naturalness close to that of the rule-based system. The contributions of this paper are that we introduced a viable architecture for an open-domain conversational system and experimentally verified its effectiveness. Rather than creating rules on the basis of developers' intuition, our architecture will enable module-by-module development, which will lead to rapid improvement in open-domain conversational systems in the future.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

B Batacharia, D Levy, R Catizone, A Krotov, and Y Wilks. 1999. CONVERSE: a conversational companion. In *Machine conversations*, pages 205–215. Springer.

Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 2012. Dialog system using real-time crowdsourcing and Twitter large-scale corpus. In *Proc. SIGDIAL*, pages 227–231.

Timothy Bickmore and Justine Cassell. 2000. How about this weather? social dialogue with embodied conversational agents. In *Proc. AAAI Fall Symposium on Socially Intelligent Agents*.

Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Meyer Dwass. 1960. Some k-sample rank-order tests. *Contributions to probability and statistics*, pages 198–202.

Suzanne Eggins and Diana Slade. 2005. *Analysing Casual Conversation*. Equinox Publishing Ltd.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Takeshi Fuchi and Shinichiro Takagi. 1998. Japanese morphological analyzer using word co-occurrence: JTAG. In *Proc. COLING*, volume 1, pages 409–413.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. COLING*, volume 2, pages 539–545.

Ryuichiro Higashinaka and Hideki Isozaki. 2008. Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 7(2):6.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, and Nozomi Kobayashi. 2013. Question answering technology for pinpointing answers to a wide range of questions. *NTT Technical Review*, 11(7).

Ryuichiro Higashinaka, Nozomi Kobayashi, Toru Hirano, Chiaki Miyazaki, Toyomi Meguro, Toshiro Makino, and Yoshihiro Matsuo. 2014. Syntactic filtering and content-based retrieval of Twitter sentences for the generation of system utterances in dialogue systems. In *Proc. IWSDS*, pages 113–123.

Shinsuke Higuchi, Rafal Rzepka, and Kenji Araki. 2008. A casual conversation system using modality and word associations retrieved from the web. In *Proc. EMNLP*, pages 382–390.

Eduard H Hovy. 1991. *Approaches to the planning of coherent text.* Springer.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. Goi-Taikei—A Japanese lexicon.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proc. ACL-IJCNLP (Short Papers)*, pages 85–88.

Kenji Imamura, Ryuichiro Higashinaka, and Tomoko Izumi. 2014a. Adaptaion of predicate-argument structure analysis with zero-anaphora resolution to dialogues. In *Proc. Annual Meeting of the Association for Natural Language Processing*, pages 709–712. (In Japanese).

Kenji Imamura, Ryuichiro Higashinaka, and Tomoko Izumi. 2014b. Predicate-argument structure analysis with zero-anaphora resolution for dialogue systems. In *Proc. COLING*.

Allen Ivey, Mary Ivey, and Carlos Zalaquett. 2013. *Intentional interviewing and counseling: Facilitating client development in a multicultural society*. Cengage Learning.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. KDD*, pages 133–142.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proc. ACL*, volume 1, pages 545–552.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2013. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):15.

Masaaki Nagata, Kuniko Saito, and Yoshihiro Matsuo. 2006. Japanese natural language search system: Web Answers. *Proc. Annual Meeting of the Association for Natural Language Processing*, pages 320–323. (In Japanese).

Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa. 2000. WIT: A toolkit for building robust and real-time spoken dialogue systems. In *Proc. SIGDIAL*, pages 150–159.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proc. ACL-HLT (Short Papers)*, pages 169–172.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proc. EMNLP*, pages 583–593.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Masahiro Shibata, Tomomi Nishiguchi, and Yoichi Tomiura. 2009. Dialog system for open-ended conversation using web documents. *Informatica (Slovenia)*, 33(3):277–284.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2013. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proc. SIGDIAL*, pages 334–338.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014a. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *Proc. IVA*. (to appear).

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014b. Response generation for questions about dialogue system's personality. In *JSAI Technical Report (SIG-SLUD-B303)*, pages 33–38. (In Japanese).

Reid Swanson and Andrew S Gordon. 2012. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):16.

Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano. 2007. Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proc. COCOSDA*.

Wataru Uchida, Chiaki Morita, and Takeshi Yoshimura. 2013. Knowledge Q&A: Direct answers to natural questions. *NTT DOCOMO Technical Journal*, 14(4):4–9.

Ellen M Voorhees and DM Tice. 2000. Overview of the TREC-9 question answering track. In *Proc. TREC*.

Marilyn Walker, Sharon Cote, and Masayo Iida. 1994. Japanese discourse and the process of centering. *Computational linguistics*, 20(2):193–232.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3&4):363–377.

Marilyn Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proc. ACL*, pages 515–522.

Richard S. Wallace. 2004. *The Anatomy of A.L.I.C.E.* A.L.I.C.E. Artificial Intelligence Foundation, Inc.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proc. SIGDIAL*, pages 404–413.