Improvements in the Utterance Database for Enhancing System Utterances in Chat-oriented Dialogue Systems

Yuiko Tsunomori\textsuperscript{†}, Ryuichiro Higashinaka\textsuperscript{††}, Takeshi Yoshimura\textsuperscript{†} and Yoshinori Isoda\textsuperscript{†}

In our commercial chat-oriented dialogue system, we have been using an utterance database created from a massive amount of predicate-argument structures extracted from the web for generating utterances. However, because the creation of this database involves several automated processes, the database often includes non-sentences (ungrammatical or uninterpretable sentences) and utterances with inappropriate topic information (called off-focus utterances). Additionally, utterances tend to be monotonous and uninformative because they are created from single predicate-argument structures. To resolve these problems, we propose methods for filtering non-sentences by using neural network-based methods and utterances inappropriate for their associated foci by using co-occurrence statistics. To reduce monotony, we also propose a method for concatenating automatically generated utterances so that the utterances can be longer and richer in content. Experimental results indicate that the non-sentence filter can successfully remove non-sentences with an accuracy of 95\% and that our focus filter can filter utterances inappropriate for their foci with high recall. We also examine the effectiveness of our filtering methods and concatenation method through an experiment involving human participants. The experimental results indicate that our methods significantly outperform a baseline in terms of understandability and that the concatenation of two utterances leads to higher familiarity and content richness while retaining understandability.

**Key Words:** Chat-oriented Dialogue System, Dialogue System, Utterance Database, Utterance Concatenation

1 Introduction

Chat-oriented dialogue systems (also known as non-task-oriented dialogue systems) have become increasingly popular (Onishi and Yoshimura 2014; Ritter, Cherry, and Dolan 2011; Vinyals and Le 2015; Yu, Xu, Black, and Rudnicky 2016; Banchs and Li 2012). Such systems need to generate a wide variety of utterances to manage the many topics in user utterances. Although rule-based methods have typically been used to generate system utterances, the topics that appear in chats are diverse, and it is extremely expensive to create rules with adequate

\textsuperscript{†} NTT DOCOMO, INC.  
\textsuperscript{††} NTT Corporation
coverage (Higashinaka, Meguro, Sugiyama, Makino, and Matsuo 2015b). Recent neural-based response-generation methods have also suffered from generating uninformative utterances (Li, Galley, Brockett, Gao, and Dolan 2016).

To overcome these weaknesses, Higashinaka et al. (Higashinaka, Imamura, Meguro, Miyazaki, Kobayashi, Sugiyama, Hirano, Makino, and Matsuno 2014) proposed a method of using a large volume of text data on the web to extract predicate-argument structures (PASs) and convert them into utterances. The result of this method is a database of utterances with their associated topics (called foci) (see Section 3 for details). We use the utterance database created with this method in our commercial chat-oriented dialogue system.\(^1\)

Although this method can generate utterances that correspond to a variety of foci by exploiting the richness of the web, system utterances have the following problems:

- Because of errors from automatic analysis of PASs and their automatic conversion into utterances, non-sentences (ungrammatical or uninterpretable sentences) and utterances inappropriate for their associated foci (called off-focus utterances) can sometimes be generated.
- System utterances tend to be monotonous and uninformative because they are created from a single PAS.

We propose methods for improving the quality of the utterance database created using Higashinaka et al.'s method (Higashinaka et al. 2014) and for reducing the monotony of system utterances. In particular, our methods filter non-sentences and off-focus utterances by using neural network-based methods and co-occurrence statistics. We also propose a method of reducing monotony by concatenating pairs of automatically generated utterances about the same focus so that the utterances can be longer and richer in content. We verified the effectiveness of our methods through an experiment involving human participants. Our contributions are as follows:

- We successfully created non-sentence and off-focus filters with our filtering methods that can greatly refine the utterance database created from PASs on the web. In terms of utterance quality, we observed significant improvements in familiarity, understandability, and content richness from subjective evaluations. By using these methods, the utterances of the database can be safely used for chat-oriented dialogue systems.
- We found that by concatenating two utterances about the same focus from the utterance database, we can create utterances that are significantly better in terms of familiarity and content richness. We confirmed that this effect occurs only when we use the utterance

\(^1\) https://www.katar.ai/
database refined using the non-sentence and off-focus filters. We believe our proposed methods will contribute to commercial chat-oriented dialogue systems in which the quality of utterances is critical.

The paper is structured as follows. In Section 2, we cover related work. In Section 3, we explain our PAS-based utterance database and examine the proportions of non-sentences and utterances inappropriate for their associated foci. In Section 4, we explain our proposed methods for filtering inappropriate utterances and our utterance-concatenation method. In Section 5, we explain our experiment involving human participants. Finally, we summarize the paper and discuss future work in Section 6.

2 Related Work

Various methods have been proposed to generate utterances in chat-oriented dialogue systems, such as rule-, retrieval-, and generation-based methods.

Rule-based methods generate system utterances on the basis of handcrafted rules. Representative systems that use such rules are ELIZA (Weizenbaum 1966) and A.L.I.C.E. (Wallace 2009). However, the topics that appear in chats are diverse, and handcrafting rules with wide coverage is extremely expensive (Higashinaka et al. 2015b).

Retrieval-based methods have been proposed to improve topic coverage. The recent increase in web data has propelled the development of methods that use data retrieved from the web for open-domain conversations (Bessho, Harada, and Kuniyoshi 2012; Shibata, Nishiguchi, and Tomiura 2009; Ritter et al. 2011). The advantage of such retrieval-based methods is that because of the diversity of the web, systems can retrieve at least some responses for user input, which can solve the coverage problem; however, this comes at the cost of utterance quality. Because the web is inherently noisy, it is, in many cases, difficult to extract appropriate sentences from retrieval results.

Recently, generation-based methods based on neural networks have been extensively researched. However, these methods have generally tended to generate utterances with little content, despite the research on improving diversity in generated utterances (Li et al. 2016; Shao, Gouws, Britz, Goldie, Strope, and Kurzweil 2017). We acknowledge that the current neural network-based methods have been yielding promising results. However, we use an utterance database created from PASs on the web (Higashinaka et al. 2014) because it is guaranteed to output system utterances with content related to the focus of the conversation and because system utterances can be more controllable, which is particularly important for commercial applications.
The method by Higuchi et al. (Higuchi, Rzepka, and Araki 2008) generated system utterances by filling templates with keywords extracted from the web for user utterances. They evaluated the validity of the system utterances by using their frequency on the web. Our approach is similar to theirs in that we use the web to evaluate the validity of system utterances but is different in that system utterances are automatically generated from PASs.

The detection of inappropriate utterances including non-sentences is related to that of grammatical errors made by second-language learners. Imaeda et al. (Imaeda, Kawai, Ishikawa, Nagata, and Masui 2003) proposed a dictionary-based method for detecting case-particle errors by using a lexicon, Oyama et al. (Oyama, Matsumoto, Asahara, and Sakata 2008) proposed a support vector machine (SVM)-based method for detecting case-particle errors in documents created by non-native Japanese speakers, and Imamura et al. (Imamura, Saito, Sadamitsu, and Nishimura 2012) proposed a method for detecting all types of particle errors. However, these methods cannot be directly applied to the utterances of dialogue systems because the error tendency of automatically generated utterances differs from that of second-language learners. The detection of inappropriate utterances has also been tackled in dialogue breakdown detection challenges (DBDCs) (Higashinaka, Funakoshi, Kobayashi, and Inaba 2016; Higashinaka, Funakoshi, Inaba, Tsunomori, Takahashi, and Kaji 2017; Hori, Perez, Higashihara, Hori, Boureau, Inaba, Tsunomori, Takahashi, Yoshino, and Kim 2019). However, the main focus is on detecting inappropriate utterances in the context of dialogue, whereas we focus on refining an utterance database.

Inaba et al. (Inaba, Yoshino, and Takahashi 2016) proposed a monologue-generation method for non-task-oriented dialogue systems by concatenating sentences extracted from Twitter. Their method is similar to our concatenation method in that we concatenate utterances to reduce monotony, but their method differs from ours in that they focused on monologues rather than dialogues.

First, we describe the construction and details of the utterance database of our commercial chat-oriented dialogue system. Then, to illustrate the problems with the database, we examine the proportions of non-sentences and off-focus utterances.

3 PAS-based Utterance Database

3.1 Creation of Utterance Database

We use the utterance database created using the method described by Higashinaka et al. (Higashinaka et al. 2014). The method uses PAS analysis (Imamura, Saito, and Izumi 2009;
Tsutomori et al. Improvements in the Utterance Database

Imamura, Higashihara, and Izumi 2014) to extract PASs with their foci from a large amount of text data. To extract high-quality PASs and their foci, the method extracts predicates with two arguments explicitly marked with particles “wa” and “ga”. “Wa” is a topic marker and “ga” is a nominative case marker in Japanese. Therefore, a subject and predicate can be extracted as constituents of a PAS together with a focus. Here, “subject” is a noun phrase indicated by the case marker “ga”, and “focus” is a noun phrase indicated with the topic marker “wa” in a sentence. Hereafter, when we refer to “subject” and “focus”, they are used in this sense.

Because PASs cannot be uttered as they are, they need to be converted into utterances. Given a PAS and a dialogue-act type (we need this as input because utterances require underlying intentions; dialogue-act types are described at the end of this subsection), an utterance is automatically created. The PASs are first converted into declarative sentences using simple handcrafted rules. Then, their sentence-end expressions (NB. In Japanese, dialogue-act-related modalities are mostly expressed by sentence-end expressions) are swapped with those matching the target dialogue-act type. The sentence-end expressions we use are those automatically mined from dialogue-act-annotated dialogue data. The details of the method of obtaining and swapping sentence-end expressions are in Miyazaki et al. (Miyazaki, Hirano, Higashinaka, Makino, and Matsuo 2015).

We illustrate the process of creating an utterance with an example. First, when we have a PAS “[nominative = I, predicate = go, dative = England]” associated with a focus “England”, we create a declarative sentence “I go to England.”. Then, the sentence-end expression is converted by swapping it with that corresponding to a given dialogue-act type. When the dialogue-act type is “Self-disclosure (desire)”, the utterance is changed to “I want to go to England” by converting the modality in the sentence-end expression (i.e., “行く” to “行きたいです” in Japanese).

From the list of 32 dialogue-act types (Meguro, Higashinaka, and Dohsaka 2010), 21 of them mainly related to self-disclosure and questions, are used for conversion. From blog data (approximately three years’ of blog articles) and by combining the extracted PASs and the dialogue-act types, the resulting utterance database contains 7,116,597 utterances associated with 204,497 foci.

3.2 Quality of Utterance Database

Because the PASs are extracted and converted into utterances automatically, errors in the resulting utterances are inevitable, and affect the quality of the utterance database. We observe two types of erroneous utterances: non-sentence and off-focus utterances.

Non-sentence Sentences that we cannot understand because of grammatical errors or a strange
combination of words. Non-sentences are generated mainly in the conversion of sentence-end expressions; some propositions cannot be uttered with certain sentence-end expressions in Japanese (see (Higashinaka, Funakoshi, Arai, Tsukahara, Kobayashi, and Mizutani 2015a) for such examples).

**Off-focus utterances** Utterances inappropriate for their associated foci. Although the utterances in the database are created from PASs in which the focus and subject are explicitly marked by the topic marker and case marker, respectively, the focus and content of an utterance are often not closely associated. This phenomenon occurs when an error is in the PAS analysis or when the meaning of the focus is too broad or vague.

We investigate the current quality of the database in terms of how many non-sentences and off-focus utterances are contained. For this purpose, we conduct annotations regarding non-sentence and off-focus utterances, which are described below.

### 3.2.1 Non-sentence Annotation

We randomly sampled 200,000 utterances from the utterance database. The annotators labeled each utterance with the following instructions:

- If you think the utterance is a non-sentence, label it 0.
- If you do not think the utterance is a non-sentence (i.e., it is a valid-sentence), label it 1.

24 annotators participated; two annotators were randomly assigned to each utterance. Cohen’s $\kappa$ value, which assesses the agreement between the two annotators, was calculated as 0.56. This indicates an intermediate degree of agreement. Table 1 lists annotation examples, and Table 2 provides the annotation breakdown. Non-sentences accounted for 12% of the database. Hereafter, we call the non-sentence annotation data on which the annotators agreed “the non-sentence corpus” (containing $23,052 + 150,955 = 174,007$ utterances).

| Focus       | Utterance                                                                 | A1 | A2 |
|-------------|---------------------------------------------------------------------------|----|----|
| 秋冬 (Fall & winter) | どんな人が流行りますよね (What types are popular, aren’t they?)  
(NB. This sentence sounds odd because its subject is an interrogative while the sentence is declarative.) | 0  | 0  |
| 秋冬 (Fall & winter) | レギンス男子が増えてますねえ (Boys wearing leggings are increasing, aren’t they?) | 1  | 1  |
| 秋冬 (Fall & winter) | 空気が乾燥したりとかです (Air is dry and so on.) | 1  | 0  |

A1 and A2 indicate labels given by two different annotators. Utterances were originally in Japanese. English translations in parentheses were performed by authors.
3.2.2 Focus Annotation

By using the utterances annotated as valid sentences in the non-sentence corpus (i.e., 150,955 utterances), two annotators labeled whether the utterances were appropriate for their foci. The annotators were shown pairs of a focus and utterance and labeled each pair with the following instructions:

- If you feel that the combination of the utterance and focus is unnatural, label it 0 (off-focus).
- If you feel that the combination of the utterance and focus is natural, label it 1 (on-focus).

When the focus has multiple meanings, if there is at least one reasonable interpretation, label the combination 1.

24 annotators participated; pairs of annotators were randomly selected for labeling pairs of a focus and utterance. Cohen’s $\kappa$ value was 0.32, which indicates a rather low but reasonable degree of agreement when considering the subjective nature of judging naturalness. Table 3 shows an example of this annotation, and Table 4 provides the annotation breakdown. Utterances inappropriate for their associated foci accounted for 5% of the database. Hereafter, we call

| Focus            | Utterance                      | A1 | A2 |
|------------------|--------------------------------|----|----|
| 秋冬 (Fall & winter) | 単価が高いんですか？ (Is the unit price high?) | 0  | 0  |
| 秋冬 (Fall & winter) | ブーツが多いのでしょうか？ (Are there a lot of boots?) | 1  | 1  |
| 秋冬 (Fall & winter) | 空気が澄んでるんですかね？ (Is the air clear?) | 1  | 0  |

Annotators 1 and 2 provide labels by two different annotators (A1 and A2).

Table 2  Statistics of non-sentence annotation (0: non-sentence, 1: valid-sentence)

|                        | # of utterances | Percentage |
|------------------------|-----------------|------------|
| 2 annotators labeled 0 | 23,052          | 12%        |
| 2 annotators labeled 1 | 150,955         | 75%        |
| 1 annotator labeled 0, other labeled 1 | 25,993 | 13% |
| Total                  | 200,000         | 100%       |

Table 3  Examples of focus annotation (0: Off-focus, 1: On-focus)

| # of utterances | Percentage |
|-----------------|------------|
| 7,528           | 5%         |
| 121,511         | 80%        |
| 21,916          | 15%        |
| Total           | 150,955    | 100%       |
the focus annotation data on which the annotators agreed “the focus corpus” (containing 7,528 + 121,511 = 129,039 utterances).

4 Proposed Methods

We observe 12% non-sentences and 5% utterances inappropriate for their associated foci in our database. This finding means that the system utterances can often be erroneous; thus, we need to reduce these utterances to improve the quality of our database. We observe another problem: the utterances in our database are monotonous and uninformative because they were generated from single PASs.

We propose methods to filter non-sentences and off-focus utterances to refine the database. We also propose a method to concatenate pairs of utterances about the same focus to reduce the monotony of system utterances.

4.1 Method to Create a Non-sentence Filter

The detection of non-sentences can be regarded as a task of sentence classification. We use standard machine-learning methods for sentence classification, such as SVMs and neural network-based methods, which have been extensively used for that task in recent years. We used the following machine-learning methods to train our classifiers:

- **SVM** We train an SVM classifier with a linear kernel. The features are the averaged word vectors of words contained in an utterance. We use a pre-trained word vector provided by Suzuki et al. (Suzuki, Matsuda, Sekine, Okazaki, and Inui 2016), the dimensions of which are 200. We use the same pre-trained word vectors for the multilayer perceptron (MLP)-, convolutional neural network (CNN)-, and long short-term memory (LSTM)-based methods, which we describe as follows.

- **MLP** We train an MLP classifier. We have five layers: the input layer, three nonlinear layers (each layer having 200 units) with sigmoid activation, and the output layer. We use averaged word vectors as input. The output layer outputs a binary decision by a softmax function.

- **CNN** We train a classifier by using a CNN. We have an input layer, convolutional layer, pooling layer, and output layer. The model structure is the same as that in Kim (Kim 2014). A filter whose size is 200 × 3 is used for convolution. The stride is set to one. We use ReLU

---

2 We used scikit-learn (http://scikit-learn.org/) for the SVM-based method and Chainer (http://chainer.org/) for the MLP-, CNN-, and LSTM-based methods.
as an activation function. The max pooling layer uses a window size of three to output a fixed length vector. The output layer outputs a binary decision by a softmax function.

**LSTM** We train a classifier by LSTM. We have an input layer, LSTM layer, three hidden layers, and an output layer. The LSTM layer has 200 units. Each word is converted into an embedding, and the sequence of word embedding is converted into a hidden representation, corresponding to a sentence vector. This vector is then fed to three non-linear layers (each layer has 200 units) with sigmoid activation, the output of which is input to the output layer, making a binary decision by a softmax function.

### 4.2 Method for Creating Focus Filter

To filter out off-focus utterances, we use co-occurrence statistics, namely, pointwise mutual information (PMI) between the subject\(^3\) of the utterance and its focus. We use PMI because of its success in filtering out sentences unrelated to topics (Newman, Lau, Grieser, and Baldwin 2010). We calculate the PMI with the following equation:

$$PMI(S, F) = \log_2 \frac{\text{count}(S, F)/N}{\text{count}(S)/N \ast \text{count}(F)/N},$$

where \(S\) is a subject, \(F\) is a focus; `count` is a function that returns the number of documents containing \(S, F\), or both, and \(N\) is the total number of documents in a text database. We use a sentence as a document unit.

If the PMI value is below a certain threshold, we can filter the utterance because the association can be considered low. The threshold can be determined experimentally, that is, we find the threshold that produces the best accuracy by using training/development data. Note that the best accuracy depends on the objective. If we want the resulting database to be as clean as possible, we set a high threshold. If we do not want to lose much data, we set the threshold lower. In this study, we set the target recall for detecting off-focus utterances to 80% because we want most of the off-focus utterances removed. We determine the threshold that achieves this recall on the training/development data and use it to filter possible off-focus utterances.

Note that an appropriate text database must be selected to calculate the PMI. We use Wikipedia (containing approximately 8 M sentences) and blogs (one year of blogs that comprise approximately 2 B sentences); the former is smaller but provided more information, and the latter is larger but noisy and is a mixture of content of varying quality. We plan to verify which one is more useful in a later experiment, although we naturally assume that blog data are

\(^3\) “Subject” is a noun phrase indicated with a case marker “ga” in a sentence.
more suitable because they have more variety, a requirement for chat-oriented dialogue systems.

4.3 Utterance Concatenation

For one solution to reduce monotony, we propose a method of concatenating pairs of automatically generated utterances about the same focus so that the utterances can be longer and richer in content. More specifically, the method concatenates two random utterances that have the same focus, for example, when there are two randomly extracted utterances for the focus “dinner”, “彼女は懐かしのレストランでご飯を食べました (She/he says she/he wants to eat Japanese food for dinner.)” and “彼は特に夕食を楽しみにしています (It’d be good to have a hot pot for dinner.)”, we simply concatenate the two by a white space to create “彼女は懐かしのレストランでご飯を食べました  彼は特に夕食を楽しみにしています (She/he says she/he wants to eat Japanese food for dinner. It’d be good to have a hot pot for dinner.)”.

This approach may seem simplistic but can be effective because, at the very least, it increases the utterance length of a system. Note that the creation of a reasonable utterance by concatenating two utterances is not trivial. The literature has demonstrated that implicit discourse relations remain difficult to detect (Lin, Kan, and Ng 2009). Thus, utterances that are coherent in terms of discourse are difficult to accurately select. In addition, our simple concatenation method may be successful because the concatenated utterance will satisfy the local coherence (Barzilay and Lapata 2008) with the same underlying entity (i.e., the focus).

5 Evaluation

We first individually evaluated the performance of our non-sentence and focus filtering methods. Then, we conducted a subjective evaluation involving human participants on the filtered and concatenated utterances.

5.1 Evaluation of Our Non-sentence Filtering Method

We trained a non-sentence filter by using the non-sentence corpus (Section 3.2.1). We split the data into training, development, and test sets corresponding to 3,837, 500, and 500 foci, respectively.

We trained the classifiers by using the training data and evaluated the accuracy with the test data by using the highest F-measure model yielded from the development data.\textsuperscript{4} The classification

\textsuperscript{4} Note that for SVM, we used the training data for training and the test data for evaluation; we did not use the development data.
results are listed in Table 5. Our method successfully detected the non-sentences with high accuracy. The model that used LSTM had the highest accuracy (0.95) and F-measure (0.83). LSTM had the highest accuracy probably because the determination of non-sentences depends on the sequence of words that can best be captured with recurrent models.

5.2 Evaluation of Our Focus Filtering Method

We split the focus corpus (Section 3.2.2) into 80% training data and 20% test data. First, we calculated the PMI values between the subjects and foci for all utterances by using the training data. Then, we searched for the threshold of the PMI that achieved 80% recall for off-focus utterances through a grid search.

When we used Wikipedia as the data for PMI calculation, we obtained a threshold of 2.2, and when we used the blog data, the threshold was 2.8. Figures 1 and 2 present the changes in precision, recall, and F-measure when we changed the threshold by an interval of 0.1. Table

| Method | Accuracy | Precision | Recall | F-measure |
|--------|----------|-----------|--------|-----------|
| SVM    | 0.93     | 0.81      | 0.71   | 0.76      |
| MLP    | 0.90     | 0.63      | 0.84   | 0.72      |
| CNN    | 0.94     | 0.86      | 0.73   | 0.79      |
| LSTM   | **0.95** | **0.88** | 0.78   | **0.83** |

Bold font represents the top score for each evaluation criterion.

Fig. 1 Changes in precision, recall, and F-measure when we changed the PMI threshold by an interval of 0.1. Wikipedia was used for the PMI calculation.
Fig. 2 Changes in precision, recall, and F-measure when we changed the PMI threshold by an interval of 0.1. Blog data were used for PMI calculation.

Table 6 Precision, recall, and F-measure for off-focus/on-focus utterances for training and test data when thresholds of 2.2 and 2.8 were used for Wikipedia and blog data, respectively

|            | Precision | Recall | F-measure |
|------------|-----------|--------|-----------|
| **Wikipedia** |           |        |           |
| train      |           |        |           |
| off-focus  | 0.09      | 0.82   | 0.16      |
| on-focus   | 0.98      | 0.49   | 0.65      |
| test       |           |        |           |
| off-focus  | 0.09      | 0.80   | 0.16      |
| on-focus   | 0.97      | 0.42   | 0.59      |
| **Blog data** |         |        |           |
| train      |           |        |           |
| off-focus  | 0.12      | 0.81   | 0.20      |
| on-focus   | 0.98      | 0.62   | 0.76      |
| test       |           |        |           |
| off-focus  | 0.13      | 0.81   | 0.23      |
| on-focus   | 0.98      | 0.64   | 0.77      |

Table 6 shows the precision, recall, and F-measure for off-focus/on-focus utterances for the training and test data when the thresholds of 2.2 and 2.8 were used for Wikipedia and the blog data, respectively. As we expected, the use of blog data yielded much better results, resulting in higher precision/recall for on-focus utterances at the point of 80% recall for off-focus utterances. The results indicate that our off-focus filter can successfully filter utterances not associated with their foci (off-focus utterances).
5.3 Subjective Evaluation

We conducted a subjective evaluation involving human participants to verify the effectiveness of our non-sentence and focus filtering methods as well as our concatenation method (Section 4.3).

5.3.1 Evaluation Procedure

Four participants who specialized in text annotation participated in the evaluation. We made each of the eight methods for comparison (Section 5.3.2) generate utterances for 100 randomly selected foci, resulting in 800 utterances (8 × 100 foci) for use in the experiment. The utterances were randomly shuffled and presented to the participants. The system utterances together with their foci were shown for evaluation, that is, the participants evaluated the utterances with regard to the dialogue context indicated by the focus. Each participant rated the 800 utterances in terms of familiarity, understandability, and content richness (Section 5.3.3).

5.3.2 Methods for Comparison

We compared the following eight methods (a)–(h). Note that, for non-sentence filtering, we used the LSTM model, which had the best performance in our experiment. For focus filtering, we used the PMI threshold of 2.8 calculated using the blog data.

(a) Random (Single): Baseline
We randomly select a single utterance from the utterance database.

(b) Random (Pair): Proposed
We randomly select two utterances associated with the same focus from the utterance database and concatenate them to create a system utterance.

(c) NS-filtered (Single): Proposed
We randomly select one utterance from the test data of the non-sentence corpus that was classified as a valid-sentence with non-sentence filtering.

(d) NS-filtered (Pair): Proposed
We randomly select two utterances associated with the same focus from the test data of the non-sentence corpus that were classified as valid sentences with non-sentence filtering.
We then concatenate these utterances to create a system utterance.

(e) NS+F-filtered (Single): Proposed
We randomly select one utterance from the test data of the non-sentence corpus that was classified as a valid sentence with non-sentence filtering and as on-focus with focus filtering.
(f) **NS+F-filtered (Pair): Proposed**

We randomly select two utterances associated with the same focus from the test data of the non-sentence corpus that were classified as valid sentences with non-sentence filtering and as on-focus with focus filtering. Then, we concatenate these utterances to create a system utterance.

(g) **Gold NS (Single)**

We randomly select one utterance annotated as a valid sentence in the test data of the non-sentence corpus.

(h) **Gold F (Single)**

We randomly select one utterance annotated as on-focus in the test data of the focus corpus.

Random (Single) is the baseline, which is our current method of merely using a single utterance for a given focus from the utterance database. Table 7 lists the example utterances generated with these eight methods.

| Method                      | Focus          | Utterance                                                                 |
|-----------------------------|----------------|---------------------------------------------------------------------------|
| (a) Random (Single): Baseline | Carbonara      | カルボナーラはパスタがいいですか？ (Does carbonara want to say pasta?) (NB. This is a non-sentence; the inanimate subject carbonara cannot be the subject of “say.”) |
| (b) Random (Pair): Proposed  | Eye sight      | 視力は出ないってことがわかりますねえ 視力は右が下がります？？ (We understand that your eyesight is not good. Has the sight of your right eye decreased?) |
| (c) (NS)-filtered (Single): Proposed | Scarf          | マフラーやバーバリーマフラーや欲しいですね (I want a Burberry scarf.) |
| (d) NS-filtered (Pair): Proposed | Wednesday     | 水曜は授業が終わってますか？水曜は授業が入ってないんです (Has Wednesday’s class ended? There is no class on Wednesday.) |
| (e) NS+F-filtered (Single): Proposed | Banana        | バナナはおいしいのが多いですね (Bananas are generally delicious, aren’t they?) |
| (f) NS+F-filtered (Pair)     | Dinner         | 夕食は和食が食べたいんですって 夕食は鍋がいいですね (She/he says she/he wants to eat Japanese food for dinner. It’d be good to have a hot pot for dinner.) |
| (g) Gold NS (Single)         | Doggy          | ワンコは耳がいいですよね (Doggies have good ears, don’t they?) |
| (h) Gold F (Single)          | Tourist        | 観光客は欧米人が多いですか？ (Are there many tourists from Europe and the US?) |

Underlines indicate the subjects of the utterances. In the English translations, phrases corresponding to the subjects are underlined. “subject” is a noun phrase indicated by the case marker “ga” in a sentence.
5.3.3 Evaluation Criteria

Sugiyama et al. (Sugiyama, Meguro, and Higashinaka 2014) used the semantic differential (SD) method to derive the dimensions to evaluate utterances in chat-oriented dialogue systems. They identified three dimensions, and we used them in our evaluation. The evaluation criteria and the statements used in the evaluation were as follows:

- Familiarity: You feel familiar with the system and that you want to talk more.
- Content Richness: You feel that the utterance is interesting and informative.
- Understandability: You feel that the utterance is natural and easy to understand.

Each participant rated their level of agreement to the aforementioned statements by using a Likert scale between 1 and 5, where 5 indicated the highest agreement. Note that our evaluation setting differed from that in (Sugiyama et al. 2014), that is, we showed system utterances and their foci as a dialogue context to the participants, whereas Sugiyama et al. (Sugiyama et al. 2014) showed system utterances and previous user utterances.

5.3.4 Results

Table 8 lists the results of the evaluation and statistical test. By comparing (a) Random (Single) to (c) NS-filtered (Single), we observe that understandability and familiarity improved by using non-sentence filtering. The reason that non-sentence filtering improved familiarity is probably related to the cooperative behavior of the system; generating unclear utterances would indicate the system’s uncooperative behavior, leading to a decrease in familiarity. By comparing (c) NS-filtered (Single) to (e) NS+F-filtered (Single), although no significant difference was observed, we observed that understandability further improved. Both (c) NS-filtered (Single)

| Method          | Familiarity | Understandability | Content richness |
|-----------------|-------------|-------------------|-----------------|
| Baseline        |             |                   |                 |
| (a) Random (Single) | 3.52        | 3.37              | 3.25            |
| (b) Random (Pair)  | 3.60        | 2.87              |                 |
| (c) NS-filtered (Single) | 3.75        | 3.73              | 3.42            |
| (d) NS-filtered (Pair) | 3.76        | 3.17              |                 |
| (e) NS+F-filtered (Single) | 3.75        | 3.87              | 3.53            |
| (f) NS+F-filtered (Pair) |           |                   |                 |
| Gold            |             |                   |                 |
| (g) Gold NS (Single) | 3.63        | 3.69              | 3.40            |
| (h) Gold F (Single) | 3.88        | 4.21              | 3.64            |

Superscripts a–h next to the numbers indicate methods with which that value was statistically better. Double-letters (e.g., aa) indicate p < .01; otherwise, p < .05. For a statistical test, we used the Steel-Dwass multiple comparison test. Bold font represents the top three scores for each evaluation criterion.
and (e) NS+F-filtered (Single) significantly outperformed the baseline, verifying the effectiveness of our filters. By comparing (g) Gold NS (Single) to (h) Gold F (Single), we also confirmed that utterances need to be appropriate for their associated foci. The results indicate that our filters contribute greatly to the understandability of the utterances in the utterance database. In addition, we also observed improvements in familiarity and content richness.

By comparing (a) Random (Single) to (b) Random (Pair), we observe that understandability significantly decreased with our concatenation method. The reason that understandability decreased is probably because the meaning of the utterance became unclear because the semantic relation of the utterance pairs is not considered by our proposed methods. However, familiarity and content richness improved. As a hypothesis, the reason that familiarity improved by concatenating pairs of utterances is because the system speaking longer may indicate its active participation in the dialogue, hence, cooperative behavior. By comparing (a) Random (Single) to (d) NS-filtered (Pair), we observe that our concatenation method improved familiarity and content richness while maintaining understandability. By comparing (d) NS-filtered (Pair) to (f) NS+F-filtered (Pair), we observe further improvements in content richness and understandability. Although it does not seem to be a good idea to concatenate possibly low-quality utterances, it is beneficial to concatenate valid and on-focus utterances. Because content richness improved without a loss of understandability, we posit that our concatenation method can reduce the monotony and generate richer utterances.

5.4 Error Analysis

We conducted an error analysis of system utterances for which two or more evaluators rated the understandability as two or lower (i.e., 233 such utterances out of 800 utterances used in the evaluation). The reason we focused on understandability was that an utterance needs to be understandable for consideration of familiarity or content richness.

One of the authors annotated error types to the 233 utterances following the taxonomy of errors in chat-oriented dialogue systems (Higashinaka, Araki, Tsukahara, and Mizutani 2018). We used eight error types under the main categories of “Utterance” and “Environment” of the taxonomy, which are related to single utterances. Table 9 shows the taxonomy, and Figure 3 shows the decision flow for the annotation of error types (Higashinaka et al. 2018).

Figure 4 shows the annotation results. We observe that the syntactic error type has the largest number of instances; in particular, many instances from (b) Random (Pair). Because we observed many instances from (a) Random (Single) in this error type as well, the syntactic ill-formedness of (a) Random (Single) was perceived more problematic by the concatenation of
Table 9  Taxonomy of errors (abridged) in chat-oriented dialogue systems

| Main category | Subcategory | Description                      |
|---------------|-------------|----------------------------------|
| Utterance     | Syntactic error | Grammatically invalid utterance  |
|               | Semantic error | Semantically invalid utterance    |
|               | Uninterpretable | Not understandable               |
|               | Others (Utterance) | Other utterance-level error       |
| Environment   | Lack of common ground | Utterance has no factual grounding |
|               | Lack of common sense | Utterance lacks common sense      |
|               | Lack of sociality | Offensive utterance               |
|               | Others (Environment) | Other environment-level error     |

Fig. 3  Decision flow for annotating error types

Fig. 4  Results of error-type annotation
sentences. However, when we examined (c)–(f), this type of error greatly decreased by using our filtering methods.

The uninterpretable error type had the second largest number of instances. Many instances were observed from (b) Random (Pair) and (d) NS-filtered (Pair). However, a small number were observed from (f) NS+F filtered (Pair), suggesting that this error type can be reduced by using our off-focus filter. Comparing (a) Random (Single), (c) NS-Filtered (Single), and (e) NS+F-filtered (Single), we observed a steady decrease in erroneous instances, indicating the effectiveness of our filters.

We observed many instances of the semantic error. Particularly, many instances of this error type were from (f) NS+F-filtered (Pair). Because we observed a small number from (e) NS+filtered (Single), this indicates our inability to concatenate reasonable sentences to form a pair; we must consider methods to consider the relations between sentences when considering the concatenation of sentences.

Although many errors were observed in Others (Environment), this finding is mainly due to the consequence of the decision flow, which forces the utterances to be classified somewhere in the taxonomy. This error type has many instances that are difficult to classify into clear error types. A more fine-grained taxonomy would be necessary to further categorize such instances.

6 Summary and Future Work

To refine our utterance database and generate non-monotonous utterances, we proposed methods of filtering non-sentences and utterances inappropriate for their associated foci by using neural network-based methods and co-occurrence statistics. To reduce monotony, we also proposed a simple but powerful method of concatenating two utterances related to the same focus so that the utterances could be longer and richer in content. Experimental results indicated that our non-sentence filter successfully removed non-sentences with an accuracy of 95% and that we could filter utterances inappropriate for their foci with high recall. We also examined the effectiveness of our filtering methods and concatenation method through an experiment involving human participants. The experimental results indicated that our automatic methods significantly outperformed the current single-utterance baseline. The experimental results also indicated that the concatenation of two utterances led to higher familiarity and content richness while maintaining understandability. We posit that our proposed methods contribute to commercial chat-oriented dialogue systems, in which the quality of utterances is critical.
For future work, we plan to update the utterance database of our current commercial chat-oriented dialogue system with our filtering methods and concatenation method. We also plan to consider methods of concatenating two utterances more appropriately, for example, by considering discourse relations (Otsuka, Hirano, Miyazaki, Higashinaka, Makino, and Matsuo 2017; Lin et al. 2009). The goal of utterance concatenation is to demonstrate active participation in the dialogue and cooperative behavior of the system and to convey more complicated meaning that arises from the relationship between two sentences (e.g., elaboration and contrast as in rhetorical structure theory). One of the limitations of this method is that control of the meaning of utterances is difficult because the unit of concatenation is coarse. For example, in “島根は温泉が多いですよ 島根は魚がうまくいますよね (Shimane has a lot of hot springs, doesn’t it? Shimane has good fish, doesn’t it?)”, the latter part of the utterance should be “島根は魚もうまくいますよね (Shimane has good fish too.)”; however, modification of words in the utterance is currently impossible. Additionally, in “彼は食べたいんですって 彼は鍋がいいですね (She/he says she/he wants to eat Japanese food for dinner. It’d be good to have a hot pot for dinner.)”, generating an utterance aggregated in a natural manner, such as “彼は食べたいんです となったら 彼は鍋がいいですね (She/he says she/he wants to eat Japanese food for dinner, so It’d be good to have a hot pot for dinner.)” is currently impossible. We plan to pursue methods to make system utterances more controllable.

Regarding the lengths of utterances made with our concatenation method, the average length of an utterance in the chat dialogue corpus (Higashinaka et al. 2014), which recorded the human-human chat dialogue, was 15.35 characters (standard deviation: 9.74). Therefore, the length of a system utterance represented by this number is probably appropriate to use as the length of system utterances. By contrast, the average length of utterances in our utterance database is 14.94 characters (standard deviation: 2.81), indicating similar lengths on average but less variation in lengths. Although long utterances can be realized by our concatenation approach, combining our method with other methods to generate utterances with human-like length variations is important. We plan to consider methods to achieve this objective in the future.

Acknowledgement

Part of this study was presented at the fourth Linguistic and Cognitive Approaches To Dialog Agents Workshop co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (Tsunomori, Higashinaka, and Yoshimura 2018). The authors would like to thank the Journal of Natural Language Processing...
for English writing check for this article. The authors also thank the anonymous reviewers for their insightful comments.

Reference

Banchs, R. E. and Li, H. (2012). “IRIS: A Chat-oriented Dialogue System Based on the Vector Space Model.” In Proceedings of the ACL 2012 System Demonstrations, pp. 37–42.

Barzilay, R. and Lapata, M. (2008). “Modeling Local Coherence: An Entity-based Approach.” Computational Linguistics, 34 (1), pp. 1–34.

Bessho, F., Harada, T., and Kuniyoshi, Y. (2012). “Dialog System Using Real-time Crowdsourcing and Twitter Large-scale Corpus.” In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 227–231.

Higashinaka, R., Araki, N., Tsukahara, H., and Mizutani, M. (2018). “Improving Taxonomy of Errors in Chat-oriented Dialogue Systems.” In Proceedings of the International Workshop on Spoken Dialogue Systems Technology, pp. 331–343.

Higashinaka, R., Funakoshi, K., Arai, M., Tsukahara, H., Kobayashi, Y., and Mizutani, M. (2015a). “Towards Taxonomy of Errors in Chat-oriented Dialogue Systems.” In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 87–95.

Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., and Kaji, N. (2017). “Overview of Dialogue Breakdown Detection Challenge 3.” In Proceedings of Dialog System Technology Challenges Workshop (DSTC6).

Higashinaka, R., Funakoshi, K., Kobayashi, Y., and Inaba, M. (2016). “The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics.” In Proceedings of the 2016 Language Resources and Evaluation Conference, pp. 3146–3150.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). “Towards an Open-domain Conversational System Fully Based on Natural Language Processing.” In Proceedings of the 25th International Conference on Computational Linguistics, pp. 928–939.

Higashinaka, R., Meguro, T., Sugiyama, H., Makino, T., and Matsuo, Y. (2015b). “On the Difficulty of Improving Hand-crafted Rules in Chat-oriented Dialogue Systems.” In Proceedings of the Signal and Information Processing Association Annual Summit and Conference, pp. 1014–1018.

Higuchi, S., Rzepka, R., and Araki, K. (2008). “A Casual Conversation System Using Modality
Tsunomori et al. Improvements in the Utterance Database

and Word Associations Retrieved from the Web.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 382–390.

Hori, C., Perez, J., Higashihara, R., Hori, T., Boureau, Y. L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., and Kim, S. (2019). “Overview of the Sixth Dialog System Technology Challenge: DSTC6.” Computer Speech & Language, 55, pp. 1–25.

Imaeda, K., Kawai, A., Ishikawa, Y., Nagata, R., and Masui, F. (2003). “Error Detection and Correction of Case Particles in Japanese Learner’s Composition.” In Proceedings of the Information Processing Society of Japan SIG, No. 13 (2002-CE-068), pp. 39–46. (In Japanese).

Imamura, K., Higashihara, R., and Izumi, T. (2014). “Predicate-argument Structure Analysis with Zero-anaphora Resolution for Dialogue Systems.” In Proceedings of the 25th International Conference on Computational Linguistics, pp. 806–815.

Imamura, K., Saito, K., and Izumi, T. (2009). “Discriminative Approach to Predicate-argument Structure Analysis with Zero-anaphora Resolution.” In Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, pp. 85–88.

Imamura, K., Saito, K., Sadamitsu, K., and Nishimura, H. (2012). “Grammar Error Correction Using Pseudo-error Sentences and Domain Adaptation.” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 388–392.

Inaba, M., Yoshino, Y., and Takahashi, K. (2016). “Open Domain Monologue Generation for Speaking-oriented Dialogue Systems.” Journal of Japanese Society for Artificial Intelligence, 31 (1), pp. DSF–F_1. (In Japanese).

Kim, Y. (2014). “Convolutional Neural Networks for Sentence Classification.” In Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing, pp. 1746–1751.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). “A Diversity-promoting Objective Function for Neural Conversation Models.” In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 110–119.

Lin, Z., Kan, M. Y., and Ng, H. T. (2009). “Recognizing Implicit Discourse Relations in the Penn Discourse Treebank.” In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 343–351.

Meguro, T., Higashinaka, Minami, Y., and Dohsaka, K. (2010). “Controlling Listening-oriented Dialogue Using Partially Observable Markov Decision Processes.” In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 761–769.

Miyazaki, C., Hirano, T., Higashinaka, R., Makino, T., and Matsuo, Y. (2015). “Automatic Con-
version of Sentence-end Expressions for Utterance Characterization of Dialogue Systems.” In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 307–314.

Newman, D., Lau, J., Grieser, K., and Baldwin, T. (2010). “Automatic Evaluation of Topic Coherence.” In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108.

Onishi, K. and Yoshimura, T. (2014). “Casual Conversation Technology Achieving Natural Dialog with Computers.” NTT DOCOMO Technical Journal, 15 (4), pp. 16–21.

Otsuka, A., Hirano, T., Miyazaki, C., Higashinaka, R., Makino, T., and Matsuo, Y. (2017). “Utterance Selection using Discourse Relation Filter for Chat-oriented Dialogue Systems.” Dialogues with Social Robots, pp. 355–365.

Oyama, H., Matsumoto, Y., Asahara, M., and Sakata, K. (2008). “Construction of an Error Information Tagged Corpus of Japanese Language Learners and Automatic Error Detection.” In Proceedings of the Computer Assisted Language Instruction Consortium.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). “Data-driven Response Generation in Social Media.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 583–593.

Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R. (2017). “Generating High-quality and Informative Conversation Responses with Sequence-to-sequence Models.” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2210–2219.

Shibata, M., Nishiguchi, T., and Tomiura, R. (2009). “Dialog System for Open-ended Conversation Using Web Documents.” Informatica (Slovenia), 33 (3), pp. 277–284.

Sugiyama, H., Meguro, T., and Higashinaka, R. (2014). “Multi-aspect Evaluation for Utterances in Chat Dialogues.” In Proceedings of the Special Interest Group on Spoken Language Understanding and Dialogue Processing, Vol. 4, pp. 31–36. (In Japanese).

Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., and Inui, K. (2016). “Neural Joint Learning for Classifying Wikipedia Articles into Fine-grained Named Entity Types.” In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, pp. 535–544.

Tsunomori, Y., Higashinaka, R., and Yoshimura, T. (2018). “Refinement of Utterance Database and Concatenation of Utterances for Enhancing System Utterances in Chat-oriented Dialogue System.” In Proceedings of the Linguistic and Cognitive Approaches To Dialog Agents Workshop, pp. 44–51.

Vinyals, O. and Le, Q. (2015). “A Neural Conversational Model.” In Proceedings of the 32nd
Improvements in the Utterance Database

International Conference on Machine Learning Deep Learning Workshop.

Wallace, R. S. (2009). “The Anatomy of A.L.I.C.E.” In Parsing the Turing Test, pp. 181–210. Springer.

Weizenbaum, J. (1966). “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine.” Communications of the ACM, 9 (1), pp. 36–45.

Yu, Z., Xu, Z., Black, A. W., and Rudnicky, A. (2016). “Strategy and Policy Learning for Non-task-oriented Conversational Systems.” In Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue, pp. 404–412.

Yuiko Tsunomori: She received her B.S. in information science from Nara Women’s University in 2013 and M.E. in information science from Nara Institute of Science and Technology in 2015. She joined NTT DOCOMO, Inc. in 2015. She is currently a research engineer of the Service Innovation Department. Her research interests include dialogue systems and natural language processing. She is a member of the Japanese Society for Artificial Intelligence.

Ryuichiro Higashinaka: He received his B.A. in environmental information, Masters of Media and Governance, and Ph.D. from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. He is currently a distinguished senior researcher at NTT Media Intelligence Laboratories. His research interests include building question-answering systems and spoken-dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield in the UK. He received the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan, and the Association for Natural Language Processing.

Takeshi Yoshimura: He received his B.E., M.E., and Ph.D. degrees from the University of Tokyo in 1997, 1999, and 2008, respectively. He joined NTT DOCOMO, Inc. in 1999. He has expertise in artificial intelligence (AI), including natural language processing, image recognition, machine learning, etc. He is currently a senior manager of the Service Innovation Department, NTT DOCOMO, Inc. He received the IPSJ Kiyasu Special Industrial Achievement.
Award in 2012 and the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of the Institute of Electronics, Information and Communication Engineers and the Information Processing Society of Japan.

**Yoshinori Isoda:** He received his B.E. and M.E. degrees from the University of Osaka in 1991 and 1993. He joined NTT, Inc. in 1993 and transferred to NTT DOCOMO, Inc. in 1999. He received his Ph.D. degree from the University of Osaka in 2011. He has expertise in artificial intelligence, including natural language processing, speech recognition, machine learning, etc. He is currently a senior manager of the Service Innovation Department, NTT DOCOMO, Inc. He received the Hisoka Maejima Award in 2014 and the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of the Robotics Society of Japan and the Information Processing Society of Japan.

(Received July 1, 2019)
(Revised October 1, 2019)
(Accepted November 7, 2019)