Refinement of utterance database and concatenation of utterances for enhancing system utterances in chat-oriented dialogue systems

Yuiko Tsunomori[1], **Ryuichiro Higashinaka**[2], **Takeshi Yoshimura**[1]
[1] NTT DOCOMO
[2] NTT Corporation
[1] {yuiko.tsunomori.fc, yoshimurat}@nttdocomo.com,
[2] higashinaka.ryuichiro@lab.ntt.co.jp

Abstract

We have been using an utterance database created from a massive amount of predicate-argument structures extracted from the web for generating utterances of our commercial chat-orientated dialogue system. However, since 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). Also, utterances tend to be monotonous and uninformative because they are created from single predicate-argument structures. To tackle 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 our non-sentence filter can successfully remove non-sentences with an accuracy of 95% and that we can filter utterances inappropriate for their foci with high recall. We also examined the effectiveness of our filtering and concatenation methods through an experiment involving human participants. The experimental results show that our methods significantly outperformed the baseline in terms of understandability and that the concatenation of two utterances leads to higher familiarity and content richness while retaining understandability.

1 Introduction

Chat-oriented dialogue systems have become increasingly popular [1; 2; 3; 4; 5]. Such systems need to generate a wide variety of utterances to cope with the many topics contained 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 coverage [6].

To overcome this weakness, Higashinaka et al. [7] 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 are using the utterance database created by this method in our commercial chat-oriented dialogue system[1] [1].

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

- Because of errors resulting 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.
- The system utterances tend to be monotonous and uninformative because they are created from single PASs.

In this paper, we propose methods for improving the quality of the utterance database created by using Higashinaka et al.'s method [7] and for reducing the monotony of system utterances. In particular, our methods filter non-sentences and off-focus utterances 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 that can greatly refine the utterance database created from PASs on the web. In terms of the utterance quality, we observed significant improvements regarding familiarity, understandability, and content richness in subjective evaluations. By using our methods, the utterances of the database can be safely used by 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 is brought about only when we use the utterance

[1] https://dev.smt.docomo.ne.jp/

database refined by the non-sentence and off-focus filters.

We believe our proposed methods can especially 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 hand-crafted rules. Representative systems that use such rules are ELIZA [8] and A.L.I.C.E. [9]. However, the topics that appear in chat are diverse, and it is extremely expensive to hand-craft rules with wide coverage [6].

Retrieval-based methods have been proposed to improve coverage. The recent increase in web data has propelled the development of methods that use data retrieved from the web for open-domain conversation [10; 11; 2]. The advantage of such retrieval-based methods is that, owing to 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. Since the web is inherently noisy, it is, in many cases, difficult to sift out appropriate sentences from retrieval results.

Recently, generation-based methods based on neural networks have been extensively researched. However, these methods generally tend to generate utterances with little content, although there has been research on improving the diversity in generated utterances [12; 13]. We acknowledge that current neural-network-based methods are yielding promising results. However, we use an utterance database created from PASs on the web [7] 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 detection of inappropriate utterances including non-sentences is related to the detection of grammatical errors made by second-language learners. Imaeda et al. [14] proposed a dictionary-based method for detecting case particle errors by using a lexicon, Oyama et al. [15] 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. [16] proposed a method for detecting all types of particle errors. However, these methods cannot be directly applied to the utterances of dialogue systems since 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) [17; 18]. 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. [19] proposed a monologue-generation method for non-task-oriented dialogue systems by concatenating sentences extracted from Twitter. This is similar to our concatenation method in that it concatenates utterances to reduce monotony but different in that it targets monologues rather than dialogues.

3 PAS-based utterance database

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

3.1 Creation of the utterance database

We use the utterance database created by using the method described by Higashinaka et al. [7]. The method uses PAS analysis [16] 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 just two arguments explicitly marked with particles 'wa' and 'ga.' 'Wa' is a topic marker and 'ga' is a nominative case marker in Japanese. This way, a subject and a predicate can be extracted as constituents of a PAS together with a focus.

Since 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 below), an utterance is automatically created. The PASs are first converted 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 type. The sentence-end expressions used are those automatically mined from dialogue-act-annotated dialogue data. The details of the method of obtaining and swapping sentence-end expressions are given by Miyazaki et al. [20].

From the list of 32 dialogue-act types [21], 21, which are mainly related to self-disclosure and question, are used for conversion. From blog data (about three years' worth of blog articles) and by the combination of 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

Since the PASs are extracted and converted into utterances automatically, errors in the resulting utterances are inevitable, affecting the quality of the utterance database. From our observation, there can be two types of erroneous utterances: non-sentence and off-focus utterances.

Non-sentence Sentences that we cannot understand due to 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 [22] for such examples).

Table 1: Examples of non-sentence annotation (0: non-sentence, 1: valid-sentence). A1 and A2 indicate the labels given by the two different annotators. Utterances were originally in Japanese. English translations in parentheses were done by authors.

Focus	Utterance	A1	A2
秋冬 (Fall & winter)	どんなんが流行りますよね (What kind of types is popular, isn't it?) (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ぇ (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

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%

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 occurs when there is an error in the PAS analysis or when the meaning of the focus is just too broad or vague.

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

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.

A total of 24 annotators participated; two annotators were randomly assigned to each utterance. Cohen's κ 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 gives 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 174,007 utterances).

Table 3: Examples of focus annotation (0: off-focus，1: on-focus). Annotators 1 and 2 give labels by two different annotators (A1 and A2).

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 air clear?)	1	0

Table 4: Statistics of focus annotation (0: Off-focus, 1: On-focus)

	# of utterances	Percentage
2 annotators labeled 0	7,528	5%
2 annotators labeled 1	121,511	80%
1 annotator labeled 0, other labeled 1	21,916	15%
Total	150,955	100%

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 to their foci. The annotators were shown pairs of a focus and utterance and labeled each pair with the following instructions:

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

A total of 24 annotators participated; pairs of annotators were randomly selected for labeling pairs of a focus and utterance. Cohen's κ value was 0.32, which indicates a reasonable degree of agreement when considering the subjective nature of judging naturalness. Table 3 shows an example of this annotation, and Table 4 gives the annotation breakdown. Utterances inappropriate for their associated foci accounted for 5% of the database. Hereafter, we call the focus annotation data on which the annotators agreed "**the focus corpus**" (containing 129,039 utterances).

4 Proposed methods

We found that there are 12% non-sentences and 5% utterances inappropriate for their associated foci in our database. Since this means the system utterances can often be erroneous, we need to reduce these utterances to improve the quality of our database. We also see it as a problem that the utterances in our database are monotonous and uninformative because they were generated from single PASs.

In this paper, we propose methods of filtering non-sentences and off-focus utterances for refining the database. We also propose a method to concatenate pairs of utterances about the same focus to reduce monotony of system utterances.

4.1 Method for creating non-sentence filter

Since the detection of non-sentences can be regarded as a task of sentence classification, we created a non-sentence filter by using machine-learning methods. We used standard machine-learning methods for sentence classification such as SVM and neural-network-based methods, which have been extensively used in recent years. We used the following machine-learning methods for training our classifiers[2]:

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 pretrained word vector provided by Suzuki et al. [23], the dimensions of which are 200. We use the same pretrained word vectors for MLP, CNN, and LSTM, which we describe below.

Multi-Layer Perceptron (MLP) We train a classifier by MLP. We have five layers: the input layer, three non-linear 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.

Convolutional Neural Network (CNN) We train a classifier by a CNN. We have an input layer, a convolutional layer, a pooling layer, and an output layer. The model structure is the same as that used by Kim [24]. A filter whose size is 200×3 is used for convolution. The stride is set to one. We used relu 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.

Long Short-Term Memory (LSTM) We train a classifier by LSTM. We have an input layer, an 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 embeddings is converted into a hidden representation, corresponding to a sentence vector. Then, this vector is fed to three non-linear layers (each layer having 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, point-wise mutual information (PMI) between the subject of the utterance and its focus. We use PMI because it has been successfully used to filter sentences unrelated to topics [25]. We calculate the PMI with the following equation:

$$\mathrm{PMI}(S, F) = log_2 \frac{\mathrm{count}(S, F)/N}{\mathrm{count}(S)/N * \mathrm{count}(F)/N}, \quad (1)$$

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 maximum number of documents in a text database. We use a sentence as a document unit.

[2]We used scikit-learn (http://scikit-learn.org/) for SVM and Chainer (http://chainer.org/) for MLP, CNN, and LSTM.

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 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 can set a high threshold. If we do not want to lose much data, the threshold can be set lower. In this study, we set the target recall for detecting off-focus utterances to 80% because we want most off-focus utterances removed. We determine the threshold that achieves this recall on the training/development data and use it for filtering possible off-focus utterances.

Note that an appropriate text database must be chosen for calculating the PMI. We consider using Wikipedia (containing roughly 8M sentences) and blogs (we use one year's worth of blogs containing about 2B sentences). The former is smaller but more informational. The latter is larger but noisy and is a mixture of contents of varying quality. We will verify which one is more useful in a later experiment, although we naturally assume that blog data are more suitable because they have more variety, which is 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, we propose concatenating two random utterances that have the same focus.

Although this approach may seem simplistic, it can be effective because, at the very least, it increases the utterance length of a system. Note that it is not trivial to create a reasonable utterance by concatenating two utterances. It has been shown that implicit discourse relations are still hard to detect [26]. This means that utterances that will be coherent in terms of discourse are difficult to accurately select. In addition, we believe our simple concatenation method may just work because the concatenated utterance will satisfy the local coherence [27] 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 and then conducted a subjective evaluation involving human participants on the filtered and concatenated utterances.

5.1 Evaluation of our non-sentence filtering methods

We trained a non-sentence filter by using the non-sentence corpus (see Section 3.2). We split the data into training, development, and test sets corresponding to 3837, 500, and 500 foci, respectively.

We trained the classifiers using the training data and evaluated the accuracy with the test data by using the highest

Table 5: Precision, recall, and F-measure for the detection of non-sentences. Bold font represents top score for each evaluation criterion.

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**

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 are 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

F-measure model yielded from the development data[3]. The classification results are listed in Table 5. We can see that our method successfully detected the non-sentences with high accuracy. The model that uses LSTM had the highest accuracy (0.95) and F-measure (0.83). LSTM has 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 (see Section 3.2) into 80% training data and 20% test data. We first calculated the PMI values between the subjects and foci for all utterances by using the training data. Then, we looked 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. See Figures 1 and 2 for the changes in precision, recall, and F-measure when we changed the threshold by an interval of 0.1. Table 6 shows the precision, recall, and F-measure for off-focus/on-focus utterances for training and test data when the thresholds of 2.2 and 2.8 are used for Wikipedia and the blog data, respectively. As 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 that are not associated with their foci (off-focus utterances).

[3]Note that for SVM, we used the training data for training and the test data for evaluation; we did not use the development data.

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 (see Section 4.3).

Evaluation procedure

Four participants took part in the evaluation. We made each of eight methods for comparison (see the following subsection for details) 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. Each participant rated the 800 utterances in terms of familiarity, understandability, and content richness (we describe these criteria later).

Methods for comparison

We compared the following eight methods (a)–(h). Note that for non-sentence filtering, we use the LSTM model, which showed the best performance in our experiment. For focus filtering, we use the PMI threshold of 2.8 calculated by 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 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 from the test data of the non-sentence corpus that were classified as valid-sentences with non-sentence filtering. Then we 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 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-sentences corpus.

(h) Gold F (Single)
We randomly select one utterance annotated as on-focus in the test data of the focus corpus.

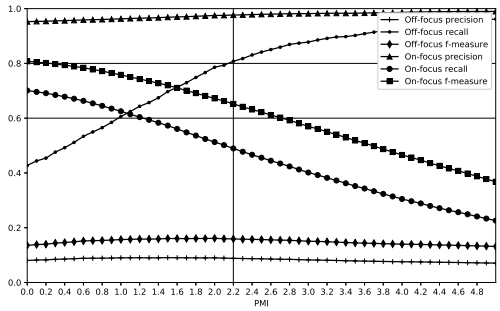

Figure 1: Changes in precision, recall, and F-measure when we changed PMI threshold by interval of 0.1. This is when Wikipedia is used for PMI calculation.

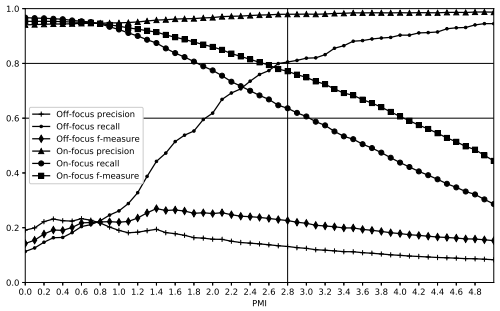

Figure 2: Changes in precision, recall, and F-measure when we changed PMI threshold by interval of 0.1. This is when blog data are used for PMI calculation.

Table 7: Example utterances generated by eight methods used in subjective evaluation

Method	Focus	Utterance
(a) Random (Single)	カルボナーラ (Carbonara)	カルボナーラはパスタがいいたいですか？ (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)	視力 (Eye sight)	視力は出ないってことがわかりますねぇ 視力は右が下がります?? (We understand that your eye sight is not good. Has the sight of your right eye decreased?)
(c) NS-filtered (Single)	マフラー (Scarf)	マフラーはバーバリーマフラーが欲しいですね (I want a Burberry scarf.)
(d) NS-filtered (Pair)	水曜 (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)	バナナ (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)	夕食 (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 (Somebody wants to have Japanese for dinner. Japanese stew should be fine.)
(g) Gold NS (Single)	ワンコ (Doggy)	ワンコは耳がいいですよね (Doggies have good ears, don't they?)
(h) Gold F (Single)	観光客 (Tourist)	観光客は欧米人が多いとかですか？ (Are there many tourists from Europe and the US?)

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

Evaluation criteria

Sugiyama et al. [29] 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 together with 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 above statements using a Likert scale between 1 and 5, where 5 indicates the highest agreement.

Results

Table 8 lists the evaluation results. By comparing (a) Random (Single) to (c) NS-filtered (Single), we can see that understandability and familiarity were improved by using non-sentence filtering. By comparing (c) NS-filtered (Single) to (e) NS+F-filtered (Single), although there was no significant difference, we can see that understandability further improved. Since both (c) NS-filtered (Single) and (e) NS+F-filtered (Single) significantly outperform the baseline, this verifies the effectiveness of our filters. In addition, by comparing (g) Gold NS (Single) to (h) Gold F (Single), we can confirm that utterances need to be appropriate for their associated foci. The results here indicate that our filters contribute greatly to the understandability of the utterances in the utterance database. In addition, we surprisingly also see improvements in familiarity and content richness.

By comparing (a) Random (Single) to (b) Random (Pair), although content richness improved, we can see that understandability significantly decreased. This means that just

Table 8: Subjective evaluation results (5 is high). Superscripts a–h next to numbers indicate methods with which that value was statistically better. Double-letters (e.g., aa) mean $p < .01$; otherwise, $p < .05$. For statistical test, we used Steel-Dwass multiple comparison test [28]. Bold font represents top three scores for each evaluation criterion.

		Familiarity	Understandability	Content richness
Baseline	(a) Random (Single)	3.52	3.37^{bb}	3.25
Proposed	(b) Random (Pair)	3.60	2.87	$\mathbf{3.74}^{aacceegg}$
	(c) NS-filtered (Single)	3.75^{aa}	$\mathbf{3.73}^{aabbddf}$	3.42
	(d) NS-filtered (Pair)	$\mathbf{3.76}^{aa}$	3.17^{b}	$\mathbf{3.89}^{aacceeggh h}$
	(e) NS+F-filtered (Single)	3.75^{a}	$\mathbf{3.87}^{aabbddf f}$	3.53^{aa}
	(f) NS+F-filtered (Pair)	$\mathbf{3.90}^{aabbgg}$	3.49^{bbdd}	$\mathbf{4.12}^{aabbccdeegghh}$
Gold	(g) Gold NS (Single)	3.63	3.69^{abbdd}	3.40
	(h) Gold F (Single)	$\mathbf{3.88}^{aabbgg}$	$\mathbf{4.21}^{aabbccddeeffgg}$	3.64^{aag}

randomly concatenating utterances in the current utterance database for the same focus does not lead to good utterances. However, by comparing (a) Random (Single) to (d) NS-filtered (Pair), we can see 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 can see 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 a good idea to concatenate valid and on-focus utterances. Because content richness has improved without loss of understandability, we can safely say that our concatenation method can reduce the monotony and generate richer utterances.

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 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 can be longer and richer in content. Experimental results show that our non-sentence filter can successfully remove non-sentences with an accuracy of 95% and that we can filter utterances inappropriate for their foci with high recall. Also, we examined the effectiveness of our filtering methods and concatenation method through an experiment involving human participants. Experimental results show that our automatic methods of incorporating non-sentence and focus filtering significantly outperformed the current single-utterance baseline. The experimental results also indicate that the concatenation of two utterances leads to higher familiarity and content richness while maintaining understandability. We believe our proposed methods can especially 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 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 taking discourse relations [30; 26] into account.

References

[1] Kanako Onishi and Takeshi Yoshimura. Casual conversation technology achieving natural dialog with computers. *NTT DOCOMO Technical Jouranl*, Vol. 15, No. 4, pp. 16–21, 2014.

[2] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 583–593, 2011.

[3] Oriol Vinyals and Quoc Le. A neural conversational model. In *Proceedings of the 32nd International Conference on Machine Learning Deep Learning Workshop*, 2015.

[4] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander I Rudnicky. 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, 2016.

[5] Rafael E Banchs and Haizhou Li. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 37–42, 2012.

[6] Ryuichiro Higashinaka, Toyomi Meguro, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 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, 2015.

[7] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 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, 2014.

[8] Joseph Weizenbaum. ELIZA— a computer program for the study of natural language communication between

man and machine. *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45, 1966.

[9] Richard S Wallace. The Anatomy of A.L.I.C.E. In *Parsing the Turing Test*, pp. 181–210. Springer, 2009.

[10] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 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, 2012.

[11] Masahiro Shibata, Tomomi Nishiguchi, and Yoichi Tomiura. Dialog system for open-ended conversation using web documents. *Informatica (Slovenia)*, Vol. 33, No. 3, pp. 277–284, 2009.

[12] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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*, 2016.

[13] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 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, 2017.

[14] Koji Imaeda, Atsuo Kawai, Yuji Ishikawa, Ryo Nagata, and Fumito Masui. 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, 2003. (In Japanese).

[15] Hiromi Oyama, Yuji Matsumoto, Masayuki Asahara, and Kosuke Sakata. Construction of an error information tagged corpus of Japanese language learners and automatic error detection. In *Proceedings of the Computer Assisted Language Instruction Consortium*, 2008.

[16] Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 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, 2012.

[17] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the 2016 Language Resources and Evaluation Conference*, pp. 3146–3150, 2016.

[18] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. Overview of dialogue breakdown detection challenge 3. In *Proceedings of Dialog System Technology Challenges Workshop (DSTC6)*, 2017.

[19] Michimasa Inaba, Yuka Yoshino, and Kenichi Takahashi. Open domain monologue generation for speaking-oriented dialogue systems. *Journal of Japanese Society for Artificial Intelligence*, Vol. 31, No. 1, pp. DSF–F_1, 2016. (In Japanese).

[20] Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. Automatic conversion 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, 2015.

[21] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. Controlling listening-oriented dialogue using partially observable Markov decision processes. In *Proceedings of the 23rd international conference on computational linguistics*, pp. 761–769, 2010.

[22] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 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, 2015.

[23] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 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, 2016.

[24] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing*, pp. 1746–1751, 2014.

[25] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 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, 2010.

[26] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 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, 2009.

[27] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34, No. 1, pp. 1–34, 2008.

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

[29] Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 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, 2014. (In Japanese).

[30] Atsushi Otsuka, Toru Hirano, Chiaki Miyazaki, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. Utterance selection using discourse relation filter for chat-oriented dialogue systems. In *Dialogues with Social Robots*, pp. 355–365. Springer, 2017.