# KIT Dialogue System for NTCIR-13 STC Japanese Subtask

Hiroshi Nakatani
Kyoto Institute of Technology, Japan
nakatani@ii.is.kit.ac.jp

Shigenori Nishiumi
Kyoto Institute of Technology, Japan
nishiumi@ii.is.kit.ac.jp

Takahiro Maeda
Kyoto Institute of Technology, Japan
maeda@ii.is.kit.ac.jp

Masahiro Araki
Kyoto Institute of Technology, Japan
araki@kit.ac.jp

## ABSTRACT

We introduce three methods for solving the NTCIR-13 STC Japanese Subtask. Method_1 is a retrieval-based method of scoring reply texts using TF-IDF, with relevance filtering using word2vec. Method_2 is a generation-based method using a seq2seq model. Method_3 is a retrieval-based method based on unsupervised clustering of dialogue acts. During the evaluation, Method_1 achieved the best results.

## Keywords

short text conversation, TF-IDF, word2vec, seq2seq, dialogue-act clustering

## Team Name

KIT16

## Subtasks

NTCIR-13 STC Japanese Subtask

## 1. INTRODUCTION

Recently, various types of dialogue systems have been developed. In task-oriented dialogue systems, the rule-based utterance generation method is predominant. However, this method requires the manual description of many rules for responding to the user's utterance; therefore, it is difficult to apply this rule-based method to the generation task of non-task oriented dialogue systems, such as chat bots. The NTCIR-13 STC Japanese Subtask is a challenge of response generation in non-task oriented dialogue systems [7]. In this task, two types of response generation methods are assumed; one is a retrieval-based method and the other is a generation-based method.

The information retrieval (IR) method, one of the retrieval-based methods, selects a reply utterance from a repository using IR technology. Previous work on response selection methods, such as IR-Status [1], returns a comment that is stored in the repository as a reply to an utterance similar to the given post. As a filter of this retrieval-based method, the filter based on the next utterance type can be useful. Such an utterance type cluster is acquired via an unsupervised learning method.

On the other hand, a generation method applies a machine translation framework to the response generation; it learns an internal representation of the dialogue using the Sequence-to-Sequence (seq2seq) model, and generates a response based on this model [9].

The task we address (NTCIR-13 STC Japanese Subtask) is to retrieve or generate appropriate replies to the input utterances using pairs of comments and replies from Yahoo! News comments data.

In the retrieval-based method, since it replies with sentences written by people, it is possible to prevent obvious grammar mistakes and inconsistency in sentences. Then it is possible to realize a natural dialogue if it can retrieve a coherent and cohesive sentence along the input data. Therefore, we tackle retrieval-based method with TF-IDF and word2vec (Method_1) with scoring added to meet the evaluation criteria. The generation method can handle utterances that the retrieval-based method cannot when the repository does not contain an appropriate utterance. In this task, we need to deal with various input utterances; therefore, we consider a generation-based method using seq2seq (Method_2). Furthermore, by clustering dialogue acts, we can use the characteristics of each utterance. Hence, we consider a retrieval-based method with topic-modeling, using the Chinese restaurant process (CRP) (Method_3).

## 2. RETRIEVAL-BASED METHOD WITH TF-IDF AND WORD2VEC

### 2.1 Overview

This method is mainly based on a Kyoto Institute of Technology (KIT) method [4] and on a Osaka Kyoiku University (OKSAT) method [6], from the previous NTCIR-12 STC Japanese Subtask. KIT ' s method uses the Latent Dirichlet Allocation (LDA) and the inverse document frequency (IDF) as a vector representation of queries, with a dialogue function filter acquired via CRP-clustering of the queries. OKSAT ' s method creates search terms by using proper nouns in a query and calculates the score as the number of common nouns for the queries in the repository, per frequency of appearance.

Both methods can be improved by content similarity, using a method based on the word frequency and their parts of speech. It is also possible to increase context-dependent appropriateness. Additionally, our proposed method uses a filter, based on TF-IDF and proper noun appearance, to maintain the topic coherence. We devise a method that can better evaluate the contents of sentences by scoring with a modifier.

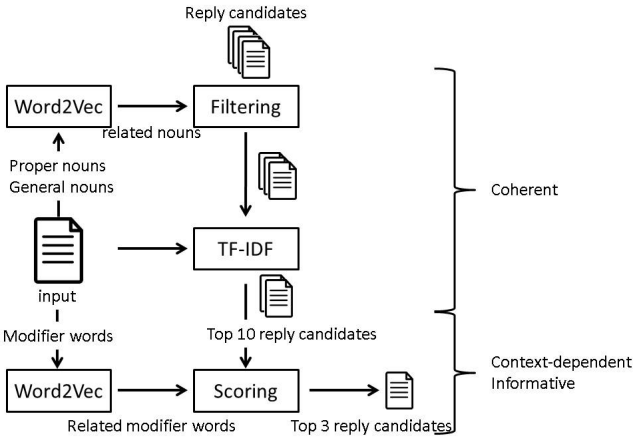


**Figure 1: Flow of Method_1.**

## 2.2 Implementation

This method selects reply candidates from the viewpoint of two types of evaluation criteria: "coherence" and "context-dependence and informative". Figure 1 shows the flow of this method. "Coherence" evaluation uses a candidate filtering and TF-IDF for scoring, and context-dependence and informative evaluation uses scoring based on OKSAT's method in the NTCIR-12 STC Japanese Subtask.

### *2.2.1 Coherence*

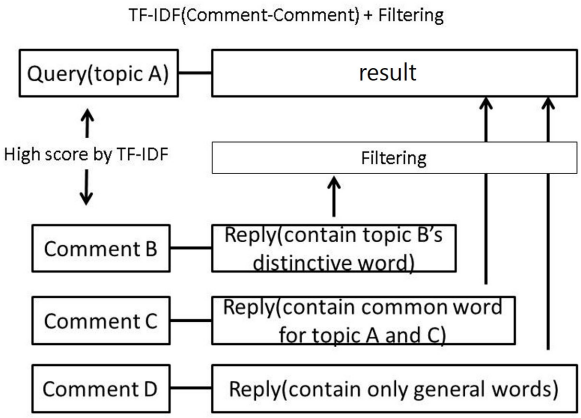


**Figure 2: Filtering model.**

First, we apply a filter, based on the proper noun, to all reply candidates. Next, a similarity calculation is done, based on TF-IDF.

By learning a repository using word2vec, we obtain meaningful words with high relevance by calculating the cosine distance between them. Based on the presence or absence of proper and general nouns, we divide queries into the following three cases.

**A proper noun is in the query**

Proper nouns, $q\_KM$, are extracted from the query, and five proper nouns, $w\_KM$, with cosine distances close to $q\_KM$, are obtained using word2vec.
Reply texts containing other proper nouns are then removed from reply candidates.

**A proper noun is not in the query**

General nouns, $q\_M$, are extracted from the query and five proper nouns, $w\_KM$, with cosine distances close to $q\_M$, are obtained using word2vec.
Reply texts, which have other proper nouns, are then removed from the reply candidates.
Nouns are important for topics in this case. Therefore, if the comment and reply texts have no $q\_M$ words, then the reply text is removed from the reply candidates.

**No nouns are in the query**

To prevent incoherent topics from being formed by careless noun usage, we remove reply texts having proper or general nouns from the reply candidates

*TF-IDF.*

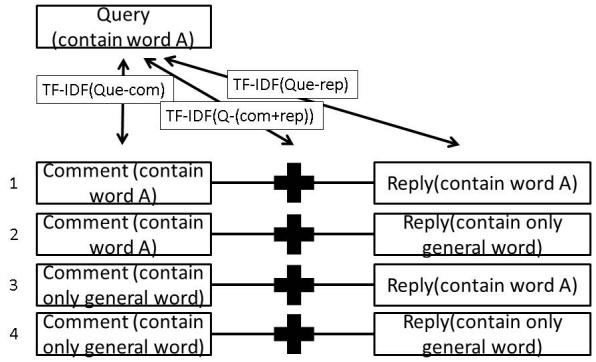


**Figure 3: Three methods of TF-IDF, this case uses TF-IDF(C-(com+rep)) and TF-IDF(Com-rep)**

We choose coherent candidates using TF-IDF with filtering result. TF-IDF is a good method for selecting related sentences, as seen in KIT's NTCIR-12 STC Japanese Subtask result. We use TF-IDF to calculate the cosine distance of "query" and "Comment text + Reply text". Thus, the ratio of words common to comment and reply texts is increased, compared to the case where only "Comment text" is used. Additionally, topic coherence is improved. The similarity of the query and the reply text is also necessary for the final reply candidate. Thus, the cosine distance of "query" and "Reply text" is obtained. These values are then multiplied to create coherence scores.

Figure 3 shows combinations of TF-IDF$i.e., Q - (com + rep) \times Que - rep$ used in this method. It evaluates replies

containing word, $A$, as "high", and, if the comment also contains word, $A$, it evaluates it as "higher".

Therefore, the top ten candidates are obtained for each input. These candidates are then scored to further improve context-dependence and development as coherent candidates. To separately consider the coherence of two kinds of topics $e.g., context-dependence and informative$, we do not use this score for the next step.

### 2.2.2 Context-dependence and informative

We regard context-dependence to be the evaluation/impression of topics in the query and the focus of modifiers, particularly adjectives and adverbs with a higher frequency of appearance. If the comment text uses the same adjective or adverb as the query, then the reply text is appropriate in the query's context. Accordingly, if an adjective or adverb in the query is in the comment text, the score is increased, based on the number of occurrences of that word in the repository.

$$Pt_{adj} = \begin{cases} \log_2 \frac{1280}{t_{adj}} & (10 \leq t_{adj} \leq 1280) \\ 0 & (t_{adj} > 1280) \\ 7 & (t_{adj} < 10) \end{cases}$$

$t_{adj}$: Number of occurrences of an adjective in repository

$$Pt_{adv} = \begin{cases} \log_2 \frac{3200}{t_{adv}} & (100 \leq t_{adv} \leq 3200) \\ 0 & (t_{adv} > 3200) \\ 5 & (t_{adv} < 100) \end{cases}$$

$t_{adv}$:Number of occurrences of an adverb in repository

In part-of-speech judgement by MeCab, 1,230 adjectives exist in the repository, whereas 4,204 adverbs exist. The formula is based on OKSAT's method in the NTCIR-12 STC Japanese Subtask. To prevent a large difference between the adjective and adverb scores, the maximum value is set to "close".

When each adjective and adverb in the query is included in the reply candidate, $Pt$ is divided by 10c and added to the score. $c$ is the number of adjectives and adverbs in the reply candidate. Therefore, the context-dependence score, $Q\_con$, is given as the following expression.

$$Q_{con} = \sum \frac{Pt_{adj}}{10c_{adj}} + \sum \frac{Pt_{adv}}{10c_{adv}}$$

$Q_{con}$: Score of context-dependence
$c_{adj}$: Number of adjectives in the candidate
$c_{adv}$: Number of adverbs in the candidate

If there are nouns related to the noun in the query, the informative score is high. We get the next words by using the obtained vector, using word2vec from the repository, as follows.

- General nouns, $w\_M$, whose cosine distance is close to the general noun, $q\_M$,in the query.
- Proper nouns, $w\_KM$ whose cosine distance is close to the general noun, $q\_M$,in the query.
- Proper nouns, $w\_KM2$, whose cosine distance is close to the proper noun, $q\_KM$, in the query.

When $w\_M$, $w\_KM$, and $w\_KM2$ are in the reply candidate, the score is increased based on the number of the words in the repository.

$$Pt_{nou} = \begin{cases} \log_2 \frac{16384}{t_{nou}} & (10 \leq t_{nou} \leq 16384) \\ 0 & (t_{nou} > 16384) \\ 14 & (t_{nou} < 10) \end{cases}$$

$t_{nou}$: Number of general nouns in the repository

$$Pt_{pno} = \begin{cases} \log_2 \frac{102400}{t_{pno}} & (100 \leq t_{pno} \leq 102400) \\ 0 & (t_{pno} > 102400) \\ 10 & (t_{pno} < 100) \end{cases}$$

$t_{pno}$: Number of proper nouns in the repository

In the part-of-speech judgment by MeCab, the repository contains 182,356 nouns and 113,380 proper nouns. Resembling the context-dependence score, the informative score is given by the following formula.

$$Q_{inf} = \sum \frac{Pt_{nou}}{10c_{nou}} + \sum \frac{Pt_{pno}}{10c_{pno}}$$

$Q_{inf}$: Score of informative
$c_{nou}$: Number of general nouns in the candidate
$c_{pno}$: Number of proper nouns in the candidate

The final score is expressed by the following formula.

$$Score = Q_{con} + Q_{inf}$$

The top three scoring results are taken as the final candidates. When the results of the scoring are equivalent, the ranking in TF-IDF is used as is.

## 2.3 Evaluation

Experimental results related to this technique are shown below.

1. KIT16-J-R4: Using only the TF-IDF from section 2.2.1.
2. KIT16-J-R1: Filtering + TF-IDF + context-dependence/informative scoring

Tables 1 and 2 shows the official STC Japanese Subtask results on the accuracy of execution.

From these results, KIT16-J-R1 shows better results than KIT16-J-R4. In Rule-2, because there is no difference from

**Table 1: Official results of the execution accuracy of Rule-1**

| | $Acc_{L2}$@1 | $Acc_{L2}$@2 | $Acc_{L1,L2}$@1 | $Acc_{L1,L2}$@2 |
|---|---|---|---|---|
| R1 | 0.1800 | 0.1690 | 0.8240 | 0.7980 |
| R4 | 0.1660 | 0.1610 | 0.8000 | 0.7700 |

**Table 2: Official results of the execution accuracy of Rule-2**

| | $Acc_{L2}$@1 | $Acc_{L2}$@2 | $Acc_{L1,L2}$@1 | $Acc_{L1,L2}$@2 |
|---|---|---|---|---|
| R1 | 0.1800 | 0.1690 | 0.6320 | 0.6050 |
| R4 | 0.1660 | 0.1610 | 0.6200 | 0.5900 |

Rule-1, context-dependence/informative scoring does not affect the desired function.

Moreover, in the case of executing only with TF-IDF, we assess how the results of executing only with TF-IDF + Scoring and TF-IDF + Filtering influence Filtering + TF-IDF + Scoring.

- If the result of TF-IDF, alone, remains in a result of the formal run,we infer the result is influenced by TF-IDF.
- If the result of TF-IDF + Scoring, without the result of TF-IDF, remains in the result of the formal run, we infer the result is influenced by Scoring.
- If the result of TF-IDF + Filtering, without the result of TF-IDF, remains in a result of the formal run, we infer the result is influenced by Filtering.
- If there is no result from TF-IDF, TF-IDF + Scoring, or TF-IDF + Filtering in the result of the formal run, we infer the result is influenced by Score + Filter.

**Table 3: Result of formal run top three candidates, and the method that influenced them.**

| TF-IDF | Score | Filter | Score+Filter | SUM |
|---|---|---|---|---|
| 153 | 13 | 89 | 45 | 300 |

From the above, half of the TF-IDF results are chosen, as they were in both top three and top one. Scoring has an influence of only about 10% when it is combined only with TF-IDF. However, it is effective when combined with Filtering. Filtering has a larger influence than Scoring, and there are 125 differences between TF-IDF-only results and TF-IDF + Filter results, 89 of which are final candidates.

As a result, because KIT16-J-R1 has a higher result than KIT16-J-R4, filtering for related terms using proper nouns and word2vec improves topic coherence. However, when not focusing on proper nouns, the evaluation concerns words with less involvement to topics. Even if there are characteristic words for other topics, and the similarity of other words is high, it will be selected as a candidate. To avoid this, we can improve topic coherence in advance by using proper nouns to remove inappropriate candidates.

**Table 4: Result of formal run top candidate and the method that influenced it.**

| TF-IDF | Score | Filter | Score+Filter | SUM |
|---|---|---|---|---|
| 56 | 9 | 15 | 20 | 100 |

Alternatively, scoring cannot improve context-dependence and informative score. For instance, only 44 of 100 input comment texts have candidates of a given value other than zero, and, more than half of the reply texts to input comment are were not influenced by this scoring. This is because most of the ten candidates, after maintaining topic consistency, lack target words to be scored. Additionally, many candidates have few words, as selected by TF-IDF, for scoring. When comparing the differences of $Acc_{L1,L2}$@1 of Rule-1 and Rule-2, 26 of 44 do not make a low evaluation with Rule-2 when there is a score, and 26 of 56, when there is no score. There are no significant differences in the number, and we conclude there is no contribution to context-dependence/informative by this scoring.

# 3. GENERATION-BASED METHOD WITH A SEQ2SEQ MODEL

## 3.1 Overview

Vinals et al proposed a method to apply the seq2seq model used for machine translation to dialogue. The seq2seq model can learn the relationship between sequence data end-to-end, and reduces the human cost required by the rule-based method. In addition, it shows that the performance has improved [9]. In this task, it is necessary to generate an appropriate reply for a very wide range of comments. In other words, it is necessary to generate replies flexibly and widely. The retrieval-based method replies only with the contents of a repository; so we use a generation-based method using the seq2seq model. In addition, in order to satisfy the evaluation measure "Coherent: The response keeps coherence with the topic of the news and the comment," we select a keyword that is talked about in comment text from the title in Yahoo! Topics and Theme and use it for training the model.

## 3.2 Implementation

### *3.2.1 Filtering*

We used the following filtering for comment text and reply text in Yahoo! News comments data before training the model.

- Unify Katakana into full-width.
- Unify numbers and symbols into half-width.
- Delete the expressions referring to specific users.
- Delete Date and time of posting (e.g. "| 2016/12/28 17:27" etc.).
- Delete symbols that do not affect the contents of the text (e.g. ">" representing that this text is a reply

etc.).

- Remove unnecessary repetition of symbols (e.g. change "!!!" to "!" etc.).

These filtering exclude sentences unrelated to the utterance content and combine different expressions representing almost same meaning.

#### 3.2.2 Keyword selection

We select a keyword that is talked about in the comment text from the title in Yahoo! Topics and Theme, in order to get a good score in the evaluation measure 'Coherent'. First, we morphologically analyze the titles in Yahoo! Topics and Theme, using MeCab with NEologd [5]. Next, we scan titles in Yahoo! Topics and Theme, and when a proper noun is found, the word is taken as the keyword for the comment text. If both titles in Yahoo! Topics and Theme do not have a proper noun, we do not set the keyword.

#### 3.2.3 seq2seq model

The seq2seq model [8] is a neural network composed of the input layer (encoder) and output layer (decoder) of Long short-term memory networks (LSTM) [3]. This model can learn the relationship between input sequence data and output sequence data end-to-end.

For training the model, the sequence data needs to be separated into word. So, we separate comment texts and reply texts for each word with space as a delimiter, using MeCab with NEologd. We add the keyword picked up in section 3.2.2 to the end of each comment text, and train the model with it as in Figure 4. The "<eos>" is a symbol representing the end of sentence. Although it makes sentence ungrammatical, it is not a problem; this is because the model learns the relationship between comment text with keyword and reply grammatical text. The reason why we add the keyword at the end, not beginning, is that the last word of a sequence has more influence on the output than previous words.

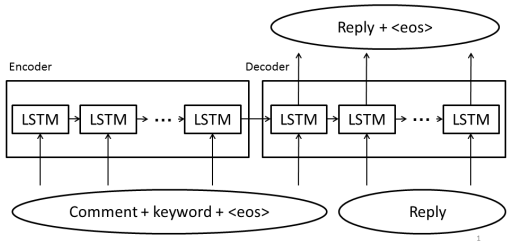


**Figure 4: Using the seq2seq framework for modeling**

#### 3.2.4 Supporting by Method_1

We trained a two-layered LSTM with 800 memory cells, and input data for development. As a result, we found some reply texts that agree or disagree only (e.g. "That's right!", "I think so.", "It is not." etc.). Such replies obtain a good score in evaluation measures 'Fluent' and 'Context-dependent', but obtain a bad score in 'Coherent' and 'Informative'. We added a text which Method_1 retrieved in section 2 to such replies, because it is difficult to solve this problem using Method_2.

### 3.3 Evaluation

As the official STC results (Table 5, 6 in section 5), Method_2 achieved lower scores for both Rule-1 and Rule-2 than Method_1. One of the major reasons for this is that in the NTCIR-13 STC Japanese Subtask, evaluation measures 'Context-dependent' and 'Informative' are not evaluated unless both evaluation measures 'Fluent' and 'Coherent' are satisfied; a reply generated by Method_2 tends to be inferior in evaluation measure 'Fluent' to the retrieval-based method.

## 4. RETRIEVAL-BASED METHOD WITH TOPIC-MODELING USING CRP

### 4.1 Overview

Higashinaka et al. [2], clustered documents by CRP and infinite HMM by the type of utterance (i.e. dialogue act). Following this research, Matsumoto et al [4] applied CRP at STC. In their method, it seemed difficult to cluster whole sentences. As in [2], good clustering results are achieved by clustering only sentences in a specific topic. The flow of Method_3 is shown in Figure 5. In Method_3, we perform clustering to estimate each topic and dialogue act. We made a cluster of the comment and reply texts by CRP and selected the response to the test data from the training data, using the dialogue act transition.

CRP conventionally refers to clusters as tables, data as customers, and feature quantities as dishes. The probability of "table that customers are placed" is calculated as follows.

$$P(t_j|c_i) \propto \begin{cases} \frac{n(t_j)}{N+\alpha} \cdot P(c_i, t_j) & (if\ j \neq new) \\ \frac{\alpha}{N+\alpha} \cdot P(c_i, t_j) & (if\ j = new) \end{cases}$$

$$P(c_i|t_j) = \prod_{w \in W} P(w|t_j)^{count(c_i, w)},$$

$$P(w|t_j) = \frac{count(t_j, w) + \beta}{\sum_{w \in W} count(t_j, w) + |W| \cdot \beta}$$

$t_j$: Table

$c_i$: Customer

$n(t_j)$: Number of customers existing in the table

$N$: Total number of customers who have been seated so far

$\alpha$: Hyper parameter indicating the degree to which a customer is assigned to a new table

$\beta$: Hyper parameter to prevent the probability 0

$W$: Feature set

$count(*, w)$: Number of occurrences of feature, w, in customer or table

$t_{new}$ A new table, using uniform distribution.

### 4.2 Implementation

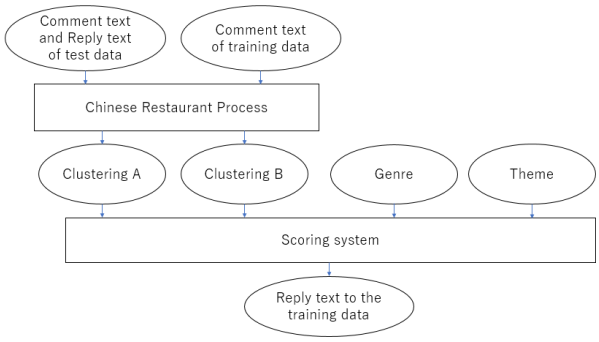


**Figure 5: Flow of Method_3**

In Method_3, we first separate words in Japanese with spaces using MeCab. We perform clustering with the CRP, using a bag of words composed of a group of nouns, adjectives or verbs as a feature. This is an estimate of the topic; let this be Clustering $A$.

Similarly, we perform clustering using a bag-of-words composed of a group of symbols, emotions, fillers, adverbs, particles, conjunctions, or adnominals as a feature. This is an estimate of the dialogue act; let this be Clustering $B$.

For each clustering, record the cluster with the highest probability of placing the test data comment text that separated words in Japanese with spaces; this is then added to the reply text of the cluster in which more reply text of the comment text is placed. Furthermore, if the characters match in each of the categories and theme of the article, the score is multiplied by 1.01. Let the reply statement with the highest score be a reply to the test data.

## 4.3 Evaluation

There are points that are considered important when using the coherency of dialogue act in Method_3 for retrieval.

In Clustering $A$, clusters containing much data correspond to topics, whereas clusters with fewer data do not. The value of $\alpha$ is set to 0.5, because we assume there are many clusters showing topics.

In Clustering $B$, clusters do not correspond to dialogue acts. There is a tendency in the cluster containing the comment text and the cluster containing the reply text corresponding to the comment text. The value of $\alpha$ is set to 0.01, because we assume there are few clusters showing dialogue act. We set $\beta$ it to 0.01, for both Clustering $A$ and $B$. For Clustering $A$, we predict that training data is in the same cluster, but only 87,777 out of 894,997 (about 10%) training data are in the same cluster.

Tables 5,and 6 show that. Clustering $A$ plays a role in replying the same topic reply text as comment text. As a result, $Acc_{L1,L2}$@1 is 0.530 in Rule-1. Clustering $B$ has the purpose of making the dialog action correspond to the comment text. However, the result of clustering is unsatisfactory. Besides, there is no system to deal with Informative. As a result, $Acc_{L2}$@1 is 0.086 in Rule-1.

# 5. RESULT

Table 7, 8, 9, 10 show some output examples for three methods and results of evaluation.

Method_1 select reply candidates by using similarity of words. This method is especially influenced by TF-IDF, so many replies include words which is included comment text. Many replies by using similarity of words are better than other methods, but sometimes select not coherent replies like Table 9. Furthermore, this method using TF-IDF can not select proper reply constructed by low similarity word like Method_3 of Table 8.

In Method_2, some replies have repetition of same sentence and tend to be ungrammatical like Table 9, others are good like Table 10. Method_2 generates replies using Title and Theme, reply in Table 9 is good reply that has "Sumo" from "Hakuho" (Hakuho is a Sumo wrestler), one in Table 8 is bad reply that has "Mr.Ishihara" from "Tokyo" (Mr.Ishihara is a former prefectural governor of Tokyo). The reply of Method_2 in Table 7 is supported by Method_1 (3.2.4). Thanks to it, we success to add information to the original reply, "It is not.", and get better score than the reply of Method_1.

In Method_3, a reply text, including a word associated with one included in the comment text, is selected ,in contract to other methods. The reply "Okinawa" is made against the topic, "Osprey". If genre and theme is the same as the comment text, the score is multiplied. Therefore sometimes "boxing" responses are chosen for comment texts whose topics are "sumo" as in Table 9; and, responses related to "smartphone" are chosen for comment texts related to "smartphone" as in Table 10.

Table 5, 6 show the STC-2 Japanese Subtask official results for our systems. From the results, Method_1 got the best score in both Rule-1 and Rule-2.

**Table 5: Official STC results. (Rule-1)**

| | $Acc_{L2}$@1 | $Acc_{L2}$@2 | $Acc_{L1,L2}$@1 | $Acc_{L1,L2}$@2 |
|---|---|---|---|---|
| Method_1 | 0.1800 | 0.1690 | 0.8240 | 0.7980 |
| Method_2 | 0.0960 | 0.0960 | 0.6320 | 0.6320 |
| Method_3 | 0.0860 | 0.0860 | 0.5300 | 0.5300 |

**Table 6: Official STC results. (Rule-2)**

| | $Acc_{L2}$@1 | $Acc_{L2}$@2 | $Acc_{L1,L2}$@1 | $Acc_{L1,L2}$@2 |
|---|---|---|---|---|
| Method_1 | 0.1800 | 0.1690 | 0.6320 | 0.6050 |
| Method_2 | 0.0960 | 0.0960 | 0.4680 | 0.4680 |
| Method_3 | 0.0860 | 0.0860 | 0.3840 | 0.3840 |

# 6. CONCLUSION

We addressed on the task of retrieving or generating an appropriate reply for a comment using Yahoo! News comments data using three methods.

Method_1 uses the similarity of sentences using TF-IDF and filters and scores by obtaining words associated with words in the query using word2vec. As a result, regarding the consistency of topics, we could obtain higher accuracy than only TF-IDF by filtering, but the scoring approach in terms of context-dependence and informative scores was not effective. In the future, we will identify elements contributing to context-dependence and informative scores in the document and apply them to the selection of candidates.

Method_2 is a generation-based method using a seq2seq model. In order to maintain the consistency of the topic, we extract keywords and use them for training the seq2seq

**Table 7: An example post and its three candidate comments with human annotation.,Title:"Osprey is already mentioned about danger in 1996", Genre:Domestic, Theme:Osprey | Futenma base relocation problem | F-35 (fighter plane) | Aegis ship**

| Post | 20 年間、オスプレイの性能は向上してないっていうのか？<br>Hasn't Osprey's performance improved for 20 years? | Rule-1<br>$Acc_{L1,L2}$@1 | Rule-2<br>$Acc_{L1,L2}$@1 |
|---|---|---|---|
| Method_1 | 頑張れ！オスプレイ！<br>Hang in there! Osprey! | 1.0000 | 0.6000 |
| Method_2 | それはない。頑張れ！オスプレイ！<br>It has not. Hang in there! Osprey! | 1.0000 | 1.0000 |
| Method_3 | アリランを沖縄で踊ってるんでしょうねw<br>I guess they dancing Arirang in Okinawa. lol. | 0.2000 | 0.0000 |

**Table 8: An example post and its three candidate comments with human annotation.,Title:"Postponement of Toyosu's hot spring facilities construction", Genre:Region, Theme:Toyosu | Hot Springs | Koto-ku, Tokyo | New Toyosu Market**

| Post | ベンゼン、シアン、ヒ素などなど、そんなの染み出てるところの温泉なんて誰も入りたくないです。<br>Nobody wants to take a hot spring bath where it seeps out benzene, cyanide, arsenic and so on. | Rule-1<br>$Acc_{L1,L2}$@1 | Rule-2<br>$Acc_{L1,L2}$@1 |
|---|---|---|---|
| Method_1 | ベンゼンもシアンも食べたくないからね。<br>Nobody does not want to eat benzene nor cyan. | 1.0000 | 1.0000 |
| Method_2 | 石原さん<br>Mr.Ishihara | 0.2000 | 0.0000 |
| Method_3 | そもそも今までの測定で測定値がゼロばかりだったてのは … 。<br>Why had measured values were 0 in the past ... ? | 1.0000 | 1.0000 |

**Table 9: An example post and its three candidate comments with human annotation.,Title:"Hakuho praised Kisenosato obediently who won the victory", Genre:Sports, Theme:(None)**

| Post | 今日の直接対決も、壁にはなれなさそう。<br>In today's battle, he can't to be a wall. | Rule-1<br>$Acc_{L1,L2}$@1 | Rule-2<br>$Acc_{L1,L2}$@1 |
|---|---|---|---|
| Method_1 | 壁、壁、壁<br>Wall, wall, wall | 0.6000 | 0.0000 |
| Method_2 | 白鵬は、相撲を見て、横綱になって、横綱になって、と思う。<br>Hakuho, learning Sumo, become Yokozuna, become Yokozuna, I think. | 0.6000 | 0.6000 |
| Method_3 | まあ、ボクシングにラッキーパンチなんてないけどな それでも技術の差は歴然だし今のままじゃ何回やっても井岡に勝てないだろうね<br>Well, there is no lucky punch in boxing but the difference in technique is still evident and it will not be possible to win against Ioka no matter how many times he challenged. | 0.0000 | 0.0000 |

**Table 10: An example post and its three candidate comments with human annotation.,Title:"New iOS have a function of finding AirPods", Genre:Computer, Theme:iOS\|Apple Inc.\|iPhone\|iPad**

| Post | 純正イヤホンより音が悪くて値段が高くてオマケに気を使うって どんな罰ゲームだよ<br>It's sound worse and higher price than genuine earphone and must take care about it...It is like a punishment game. | Rule-1<br>$Acc_{L1,L2}$@1 | Rule-2<br>$Acc_{L1,L2}$@1 |
|---|---|---|---|
| Method_1 | 純正使わずに他の会社のを使ってるわ<br>I use other company's one which is not genuine. | 1.0000 | 1.0000 |
| Method_2 | それは、あなたの価値観の問題ですよ。<br>That is a matter of your sense of values. | 1.0000 | 1.0000 |
| Method_3 | どうしたら文鎮化しますか？ 詳しく教えて下さい<br>What causes it is bricked? Please tell me more. | 0.2000 | 0.2000 |

model. As a result, extraction of keywords was mostly successful, but we could not generate an appropriate reply. Therefore, it is necessary to review the parameters and network structure and to consider a better usage method of keywords.

Method_3 retrieved an appropriate reply by clustering comments using the CRP. It was difficult to specify the dialogue act in clustering with words whose part of speech is either a symbol, an emotion, a filler, an adverb, a particle, a conjunction, or an adnominal. Furthermore, words that are not used as features for CRP cannot be the topic. As a future task, we will consider a method of extracting words that can successfully cluster dialogue acts.

Thus, Method_1: Retrieval-based method with TF-IDF and word2vec showed the best results in both Rule-1 and Rule-2 in the NTCIR-13 STC Japanese Subtask.

## 7. ACKNOWLEDGMENTS

Our thanks to TIS Inc. for supporting us and to the NTCIR-13 STC Japanese Subtask organizer and Yahoo Japan Corporation for preparing Yahoo! News comments data.

## 8. REFERENCES

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

[2] R. Higashinaka, N. Kawamae, K. Sadamitsu, Y. Minami, T. Meguro, K. Dohsaka, and H. Inagaki. Unsupervised clustering of utterances using non-parametric Bayesian methods. In *Proceedings of Interspeech 2011*, pages 2081–2084, 2011.

[3] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780, 1997.

[4] S. Matsumoto and M. Araki. Scoring of response based on suitability of dialogue-act and content similarity. In *Proceedings of the NTCIR 12 Conference*, 2016.

[5] T. Sato. Neologism dictionary based on the language resources on the Web for Mecab, 2015.

[6] T. Sato, Y. Morishita, and S. Shibukawa. Oksat at ntcir-12 short text conversation task priority to short comments, filtering by characteristic words and topic classification. In *Proceedings of the NTCIR 12 Conference*, 2016.

[7] L. Shang, T. Sakai, H. Li, R. Higashinaka, Y. Miyao, Y. Arase, and M. Nomoto. Overview of the NTCIR-13 short text conversation task. In *Proceedings of NTCIR-13*, 2017.

[8] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[9] O. Vinyals and Q. V. Le. A Neural Conversational Model. In *Proceedings of ICML Deep Learning Workshop*, 2015.