Conversation Initiation by Diverse News Contents Introduction

Satoshi Akasaki*
The University of Tokyo
akasaki@tkl.iis.u-tokyo.ac.jp

Nobuhiro Kaji
Yahoo Japan Corporation
nkaji@yahoo-corp.jp

Abstract

In our everyday chit-chat, there is a conversation initiator, who proactively casts an initial utterance to start chatting. However, most existing conversation systems cannot play this role. Previous studies on conversation systems assume that the user always initiates conversation, and have placed emphasis on how to respond to the given user's utterance. As a result, existing conversation systems become passive. Namely they continue waiting until being spoken to by the users. In this paper, we consider the system as a conversation initiator and propose a novel task of generating the initial utterance in open-domain non-task-oriented conversation. Here, in order not to make users bored, it is necessary to generate diverse utterances to initiate conversation without relying on boilerplate utterances like greetings. To this end, we propose to generate initial utterance by summarizing and chatting about news articles, which provide fresh and various contents everyday. To address the lack of the training data for this task, we constructed a novel large-scale dataset through crowd-sourcing. We also analyzed the dataset in detail to examine how humans initiate conversations (the dataset will be released to facilitate future research activities). We present several approaches to conversation initiation including information retrieval based and generation based models. Experimental results showed that the proposed models trained on our dataset performed reasonably well and outperformed baselines that utilize automatically collected training data in both automatic and manual evaluation.

1 Introduction

Conversation[1] systems are becoming increasingly important as a means to facilitate human-computer communication. However, most of the studies on conversation systems have been based on the assumption that a human always initiates conversation. As a result, the systems are designed to be passive (Yan, 2018), meaning that they keep waiting until they are spoken to by the human and will never speak to the human proactively. For example, popular encoder-decoder models (Sutskever et al., 2014; Vinyals and Le, 2015) are designed to respond to input utterances provided by humans, and it is difficult for them to proactively initiate the conversation. Although some systems are able to initiate conversations, they basically adopt template-based generation methods and thus lack diversity.

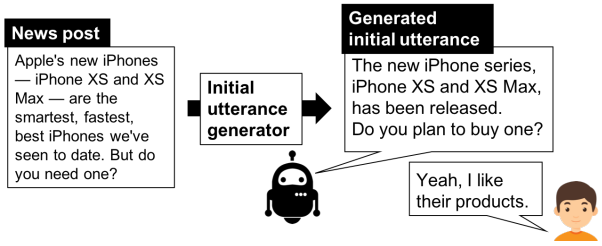

Figure 1: Conversation initiation task. The system in this example is given a news post about "iPhone" and generates an initial utterance for chatting about it.

This paper investigates generating the very first utterance in a conversation. We feel strongly that conversation systems should not always be passive; sometimes, they have to proactively initiate the conversation to enable more natural conversation. In addition, it is crucial to be able to initiate conversation in various ways in actual applications, since systems that initiate a conversation by always saying "*Let's talk about something*" or "*Hello*" are inherently boring.

We propose a task setting in which the system initiates a conversation by talking about a news topic. In this task, the system is provided with

*This work was done during research internship at Yahoo Japan Corporation.

[1]"Conversation" in this paper refers to open-domain non-task-oriented conversations and chit-chat.

Proceedings of NAACL-HLT 2019, pages 3988–3998
Minneapolis, Minnesota, June 2 - June 7, 2019.

a news post to talk about and uses it to generate the initial utterance of the conversation (Fig. 1). This task is referred to as *conversation initiation* in this paper. We have two primary reasons for using news posts. First, sharing and exchanging opinions about the latest news with friends is common in our daily conversations (Purcell et al., 2010) (*e.g,* asking something like *"What do you think about today's news on Trump?"*). Second, and more importantly, this task setting allows us to proactively generate diverse utterances to initiate conversations by simply using the latest news posts, which include a wide variety of content published daily.

We created a large-scale dataset for training and evaluating conversation initiation models through a crowd-sourcing service. The crowd-sourcing workers were presented with news posts collected from Twitter and asked to create utterances to initiate a conversation about the post. The resulting dataset will be released to facilitate future studies at the time of publication.

We developed several neural models, including retrieval-based and generation-based ones, to empirically compare their performances. We also compared the proposed models against baselines that utilize automatically constructed training dataset to investigate the effectiveness of our dataset. Both automatic and manual evaluation were used to assess not only the quality but also the diversity of the generated initial utterances. The results indicate that the proposed models successfully generated initial utterances for the given news posts, and significantly outperformed the baseline models.

Our contributions are the following:

- We investigate the task of conversation initiation, which has been largely overlooked in previous studies.
- We construct and release a large-scale dataset for conversation initiation.
- We develop several neural models and empirically compare their effectiveness on our dataset.

2 Related work

2.1 Non-task-oriented Conversation System

There are many existing studies on non-task-oriented conversation systems. Research started with rule-based methods (Weizenbaum, 1966; Wallace, 2009) and gradually shifted to statistical approaches (Ritter et al., 2011; Vinyals and Le, 2015), and many follow-up studies have since been undertaken to improve the quality of the generated responses (Hasegawa et al., 2013; Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016b; Serban et al., 2017).

However, the task of conversation initiation has been largely absent in these studies.

2.2 Data Grounded Conversation

There have also been efforts to develop systems that can chat with users about specific documents such as Wikipedia articles (Zhou et al., 2018) or reviews (Moghe et al., 2018). However, these studies did not investigate how to initiate such conversations, and as a result, their models assume that the initial utterance is always given by users. Also, their datasets are designed to be used to train models of multi-turn conversations about the given documents, rather than models of conversation initiation. For example, Moghe et al. (2018) utilized fixed templates to initiate conversations, and there are only a few (around 4k) utterances that can be used to train the model of conversation initiation in Zhou's dataset (2018).

In contrast, we focus on the conversation initiation task, which those studies have largely overlooked, and develop a large-scale dataset that includes 109,460 utterances for this task (see Section 3). Therefore, our work can be considered complementary to the previous studies.

In an approach that uses images rather than documents, (Mostafazadeh et al., 2016) proposed a method of generating questions about an image to initiate conversation. Although, like us, they explored initiating conversation, they focused only on generating questions. In contrast, we investigate generating other types of initial utterances than questions. Also, they investigated a task setting in which users can see the images along with the conversation, while we do not present the news posts to users. This difference makes our generation task a bit more complicated (see Section 3).

2.3 Proactive Conversation System

Some studies have attempted to make conversation systems more proactive rather than passively waiting for utterances from a user. (Li et al., 2016c) proposed a system that detects a stalemate in the conversation and then proactively

casts a specific response for breaking the stalemate. They use the history of the user's utterances to select response candidates. (Yan et al., 2017; Yan and Zhao, 2018) proposed a method of proactively suggesting the user's next utterance. Although these methods have been successfully used in proactive conversation systems, the conversation initiation has not been investigated.

2.4 Diverse Response Generation

A well-known problem of encoder-decoder-based conversational models is that they tend to generate generic responses such as "*I don't know*" (Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016). Such responses understandably bore users, so there has been much research focus on generating more diverse responses (Li et al., 2016a; Xu et al., 2018; Baheti et al., 2018).

We explore the problem of generating diverse initial utterances from a different perspective than other studies. In our problem setting, it is not obvious how to go beyond simple template-based systems, which cannot generate diverse utterances. We address this problem by generating initial utterances based on news posts, which feature various content and are updated every day.

This study is complementary to previous attempts at diversification. Our method exploits existing neural conversation models, which tend to generate generic responses, as a component. The previous diversification methods can be used to improve the initial utterances in our method.

2.5 Other Related Work

Question Answering (QA) tasks have long been studied in the research community (Rajpurkar et al., 2016, 2018). In recent years, conversational variants of this task such as visual QA (Antol et al., 2015; Das et al., 2017) and conversational QA (Reddy et al., 2018) have been proposed. All of these tasks differ from our conversation initiation task since they focus on how to respond to questions.

(Yoshino and Kawahara, 2014) proposed an information navigation system that presents users with the contents of news articles through conversation. Although this setting is similar to ours, their system always opens conversation by just presenting the news headline. Our study investigates initiating conversation in a more chatty way, and should contribute to making the systems more conversational and attractive.

(Qin et al., 2018) proposed the task of generating comments about given news articles. Although this task is similar to ours, it is not designed to converse with users. Our task focuses on conversation and tries to generate initial utterances using news articles (posts).

3 Conversation Initiation Dataset

In this section, we explain how we constructed the dataset for the task of conversation initiation. We then analyze the constructed dataset to provide insights into its effectiveness.

3.1 Data Construction

We first collected 104,960 Japanese news posts from the Twitter account @YahooNewsTopics,[2] which delivers the latest news in the world every day. The data were collected between December 31, 2013 and October 31, 2017. Some example posts collected from this account are listed in the third column of Table 1.[3]

We investigate the task setting in which the system opens a conversation about a given news post. Here, we presume the post is not presented to the user during the conversation. Although letting users see the news posts would be possible, such a setting is not investigated here because our focus is a situation where users converse with the system only by voice. Such situations are growing more popular in recent years with the rise of voice-controlled conversation systems such as intelligent assistants (*e.g.,* Siri, Alexa, and Cortana) (Jiang et al., 2015; Sano et al., 2016; Akasaki and Kaji, 2017) and smart speakers (*e.g.,* Amazon Echo and Google Home).

Therefore, in our task setting, since the user does not always know about the news, it is preferred to first introduce the news summary so as to share the background knowledge before starting the conversation (see Fig. 1). In this sense, our task can be understood as a combination of summarization and chit-chat. Interestingly, the summarization subtask goes beyond the ordinary one in that we not only compress the content but also generate the text in a chit-chat-like style.

[2] https://twitter.com/yahoonewstopics

[3] Original news posts were written in Japanese. We have translated them for clarity.

dialogue acts	# examples	news posts (translated)	initial utterances (translated)
IMPRESSION	7,929	Yu Abiru announces her marriage on a broadcast. Her affiliation office and partner also commented.	It seems that Yu Abiru announced her marriage on a broadcast. *Congrats!*
		Major beer companies will increase beer production by about 10% this summer compared to the same period last year. Managerial resources have been shifted from other products due to a hot summer and tax cuts.	I heard that major beer companies are planning to increase beer production by about 10% this summer compared to last summer. *It makes me want to drink a cold beer on a hot day.*
URGING	273	The Korean Defense Department revealed that the North Korean army launched several "short-range projectiles" toward the Sea of Japan on the morning of the 3rd.	North Korean forces launched missiles toward the Sea of Japan. *Let's evacuate quickly!*
		Severe damage to the mind and body due to abuse cannot be healed easily. We followed the cases of women who suffered abuse as children.	Wounds by abuse are stored deeply in the body and mind. *Let's do something to help if child abuse is happening around you.*
QUESTION	1,028	Players of the national men's handball team smoked in a non-smoking area while staying at Ajinomoto national training center. They received an indefinite ban from JOC.	The national men's handball team players smoked in a non-smoking area and were expelled, I heard. *Have you ever seen a handball game?*
		An infant was caught between a bed guard and a mattress and subsequently died. In the US, 13 children were killed in the past 11 years due to such accidents. Japan Pediatric Society has called attention to this.	There seems to have been an accident caused by an infant's bed guard. *Do you think that bed guards are necessary?*

Table 1: Distribution over sampled dialogue acts and example initial utterances. *Italics* are chit-chat parts.

	# word	# sent.	vocab. size
News post	33.85	1.96	54,830
Initial utterance	31.50	2.03	49,211
*Summary part	22.27	1.00	45,850
*Chit-chat part	9.23	1.03	19,520

Table 2: Statistics of the dataset. First and second columns show the average numbers per utterance.

To construct the dataset, we had cloud workers create the initial utterance of a conversation on the basis of a given news post. We instructed workers to not only chat about the news post but also to provide its brief summary. The workers were asked to use colloquial expressions because users feel strange when spoken to in literary expressions. We obtained a total of 104,960 pairs of news post and initial utterance[4]. Note that we created only the initial utterances (same as (Mostafazadeh et al., 2016)) because our focus is how to initiate conversation[5].

[4]Some news posts (typically emergency news such as earthquake) were posted more than once, and as the consequence the dataset includes 102,844 unique news posts. In the experiment, we took care so that the training and test datasets do not include the same news posts.

[5]Of course it is necessary to continue successive conversation in an actual application, but we here leave this setting as a future work.

3.2 Data Analysis

Here we discuss our investigation of the 104,960 initial utterances. Some examples of the utterances are listed in Table 1. Most initial utterances first summarize the contents of the news post and then begin to chat about it, as we instructed. For subsequent analysis and model designing, we divided each initial utterance into sentences and then designated the one with the smallest edit distance from the input news post as "summary part" and the rest as "chit-chat parts". The rationale behind the use of this heuristic is that the summary part shares more words with the original news post than the chit-chat part and consists of just one sentence in most cases. The statistics of the dataset are shown in Table 2.

For the summary part, as seen in Tables 1 and 2, original news posts are compressed by 32.29% on average and are converted into a colloquial style. This indicates that the recruited cloud workers properly extracted the important contents from the input news posts and used them for making the summary part.

Compared with the summary part, the number of words and vocabulary size for the chit-chat part are relatively small (Table 2). This is a natural phenomenon since the summary part uses more con-

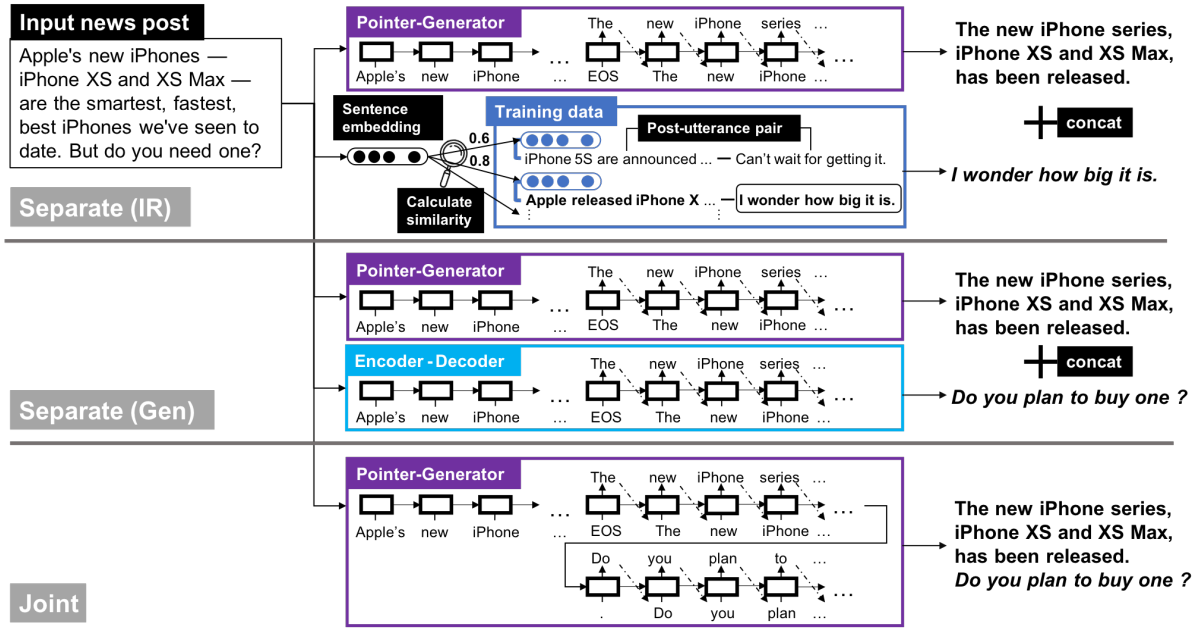

Figure 2: Overview of initial utterance generation by our proposed approaches. *Italics* are chit-chat parts.

tent words for summarization than the chit-chat part. To clarify how workers created these chit-chats, we randomly sampled 10,000 utterances and manually classified them according to their dialogue acts, as shown in Table 1. We found that the majority (92% = (7929 + 273 + 1082) / 10000) are classified into three dialogue acts (IMPRESSION, URGING, and QUESTION). The remaining 8% miscellaneous utterances that do not belong to any of the three dialog acts.

Most of the labeled initial utterances are the impressions and opinions of cloud workers about news posts (see the IMPRESSION act). Some of them are boilerplates (*e.g., "Congrats"*) while others show tremendous diversity (*e.g., "It makes me want to drink a cold beer on a hot day"*). It is interesting that some workers make an urging (*e.g., "Let's evacuate quickly"*) or ask a question (*e.g., "Have you ever seen a handball game?"*). These acts that attempt to solicit the user's response are important elements for conversation initiation.

4 Generation Method

As described in Section 2, most of the initial utterances in the dataset can be divided into a summary part and a chit-chat part. Because it is possible to generate these two parts by two separate models or by a single joint model, we investigate both approaches and compare their performance in an experiment. The overview of our proposed approaches is given in Fig. 2.

4.1 Separate Approach

The separate approach utilizes two different models to generate the summary part and the chit-chat part, respectively.

The summary part is generated by the pointer-generator model, which allows both copying words by pointing to the input sentence and generating words from a fixed vocabulary (See et al., 2017). This model is suitable for generating the summary part because it can appropriately select the contents of the input sentence while compressing them to a proper length.

To generate the chit-chat part, both generation-based and information retrieval (IR)-based methods are investigated. We use a common encoder-decoder model (Vinyals and Le, 2015) as the generation-based method (see **Separate (Gen)** in Fig. 2). Since this model tends to generate generic sentences that lack diversity (Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016), we also adopt the MMI-antiLM method proposed by (Li et al., 2016a) to promote diversity. This method uses the following score function, instead of the commonly used log-likelihood, when decoding:

$$\log P(T|S) - \lambda \log U(T), \qquad (1)$$

where T is an initial utterance and S is a news post. $P(T|S)$ is the conditional likelihood of T

given S, and U is a language model. In decoding, output candidates are generated using beam search and are then reranked by Eq. 1. This model penalizes generic sentences by $U(T)$.

As the IR-based method, we utilize the embedding of an input news post to retrieve the closest news posts in the training data using cosine distance, and then extract the corresponding chit-chat part (Ritter et al., 2011) (see **Separate (IR)** in Fig. 2). We adopt Smooth Inverse Frequency (SIF)-based embedding (Arora et al., 2017) for inducing news post embeddings. This method first calculates a weighted average of word embeddings in a news post s as:

$$v_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + P(w)} v_w, \quad (2)$$

where a is a hyperparameter and $P(w)$ is the unigram probability calculated from the training data. Then, it reduces the influence of the first principal component by using the first singular vector u of the word vector matrix:

$$v_s = v_s - uu^{\mathrm{T}} v_s, \quad (3)$$

This method has demonstrated a competitive performance across various tasks (Arora et al., 2017).

4.2 Joint Approach

We concatenate the summary part and the chit-chat part of the training data and train only one pointer-generator model, as mentioned in Section 4.1 (see **Joint** in Fig. 2).

Unlike the separate approach, this method can be considered multi-task learning of the summary and the chit-chat part generation. Thus, we expect it can generate the initial utterance precisely by considering the coherence between the summary and the chit-chat part. We examine the effectiveness of this approach through experiments in the following section.

5 Experiments

We empirically evaluate the performance of the proposed methods on the constructed dataset.

5.1 Models

In addition to the proposed methods, we implemented baselines that do not use labor-intensive labeled data, since carefully preparing the dataset is one of our contributions. These baselines generate summary and chit-chat parts separately in the following way and concatenate them as output.

RNN type	Bi-LSTM
Layers	2
Hidden layer dim.	512
Embedding dim.	256
Dropout rate	0.2
Parameter init.	(-0.08, 0.08) (uniform)
Vocabulary size	50,000
Batch size	64
Epochs	30
Max. grad. norm.	1
Optimization	Adam
Learning rate	0.001
Beam size	5
λ (MMI)	0.3
γ (MMI)	0.15

Table 3: Hyperparameter settings for training encoder-decoder models.

We gathered tweets (news posts) of major news accounts from Twitter and their corresponding replies (regarded as chit-chats). Those tweet-reply pairs can be used as pseudo training data to generate the chit-chat part. Since we cannot automatically acquire training data for generating the summary part, we output the first sentence of the input news post as the summary part.

Overall, the following proposed and baseline methods were implemented for comparison:

Baseline Generate the summary part and the chit-chat part by separate models using the pseudo-training data collected from Twitter. There are three variants of this method for generating the chit-chat part. **Baseline (IR)** and **Baseline (Gen)** use the IR-based method and the generation-based method, respectively. **Baseline (Gen+MMI)** uses MMI-antiLM (Li et al., 2016a) for decoding.

Separate Generate the summary part and the chit-chat part separately using the approach described in Section 4.1 and the dataset described in Section 3.1. There are also three variants of this method, same as the baselines (**Separate (IR)**, **Separate (Gen),** and **Separate (Gen+MMI),** respectively).

Joint Generate the summary part and the chit-chat part jointly using the approach described

Model	R-1	R-2	R-L	D-1	D-2	D-S
Baseline	**70.2**	**59.1**	**67.5**	**17.7**	**60.1**	99.8
Separate	66.5	50.6	63.8	15.7	52.4	**99.8**
Joint	68.8	54.1	66.3	15.2	51.8	**99.8**

Table 4: Results of summary part generation.

Model	BLEU	D-1	D-2	D-S
Baseline (IR)	0.2	**21.8**	**65.0**	**90.3**
Baseline (Gen)	0.2	1.2	3.4	2.8
Baseline (Gen+MMI)	0.2	1.1	3.1	3.7
Separate (IR)	3.5	12.6	34.9	65.2
Separate (Gen)	6.3	1.5	2.9	3.2
Separate (Gen+MMI)	**9.6**	2.2	5.6	13.4
Joint	6.4	6.7	15.8	28.5

Table 5: Results of chit-chat part generation.

in Section 4.2 and the dataset described in Section 3.1.

5.2 Experimental Settings

We divided the 104,960 items of data (news post and initial utterance pairs) into 90,000, 10,000, and 4,960 for training data, development data, and test data, respectively. Input news posts that appear in the training data were removed from the test data. Consequently, 4,776 data were used as the final test data.

To train the baseline model, we collected 277,813 tweets and their corresponding replies from six major Japanese news accounts[6] on Twitter. We then divided those pairs into 260,000 and 17,813 for training data and development data for the baselines.

We performed tokenization using a Japanese morphological analyzer, MeCab,[7] with IPAdic dictionary,[8] and then removed usernames, URLs, and hashtags. We used OpenNMT-py (Klein et al., 2017)[9] for building the models described in Section 4. Their hyperparameter settings are given in Table 3. We used GloVe (Pennington et al., 2014)[10] to learn 300-dimensional word embeddings. We trained word embedding using a Japanese Wikipedia dump released on February 22nd, 2018. These embeddings were used for acquiring news post embeddings, as described in Section 4.1.

5.3 Automatic Evaluation

As discussed in Section 3.2, since the initial utterance can be divided into separate parts that have different properties, we evaluated each part separately to examine the generated initial utterances.

We automatically divided the generated sentences and reference sentences into summary parts and chit-chat parts, as explained in Section 3.2. We used ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) for evaluating the summary part (denoted as **R-1**, **R-2**, and **R-L**, respectively) and BLEU (Papineni et al., 2002) for evaluating the chit-chat part. We use different metrics for each part because ROUGE is often used for summarization tasks while BLEU is used for conversational tasks. Since these automatic metrics are insufficient for evaluation (Novikova et al., 2017), we also perform a manual evaluation in Section 5.4.

To evaluate diversity, we calculate the proportion of distinct unigrams, bigrams, and sentences (**D-1**, **D-2**, and **D-S**, respectively) in the generated initial utterances (Li et al., 2016a).

Table 4 lists the results of the summary part. The baseline method that outputs the first sentence of an input news post achieved higher ROUGE scores than the proposed methods. This does not necessarily mean that the proposed methods are poor because even the SOTA summarization system exceeds such a baseline by only a small margin (See et al., 2017). Also, our task has a requirement to convert sentences into colloquial expressions, and the ROUGE metric cannot capture such a subtle difference. We perform a deeper investigation into the quality of the generated initial utterance in the next section. Regarding the diversity, almost all of the generated initial utterances are distinct, as shown in Table 4.

Table 5 shows the result of the chit-chat part. The proposed methods outperformed the baselines in terms of BLEU score. Although the baselines use two times as much training data as the proposed methods, their scores were quite low. This demonstrates the quality of our dataset. The score of **Separate (IR)** was relatively low among the proposed methods, presumably because the chit-chat parts retrieved from the training data do not always match the content of the input news post.

[6] @YahooNewsTopics, @livedoornews, @asahi, @mainichi, @mainichi_jp, @nhk_news

[7] http://taku910.github.io/mecab/

[8] https://ja.osdn.net/projects/ipadic/

[9] https://github.com/OpenNMT/OpenNMT-py

[10] https://nlp.stanford.edu/projects/glove/

Model	Naturalness	Coherency
Human	3.31	3.39
Baseline (Gen+MMI)	2.26	2.24
Separate (IR)	2.32	2.21
Separate (Gen+MMI)	2.96	2.96
Joint	**3.07**	**3.06**

Table 6: Results of evaluating **Naturalness** and **Coherency** of the generated utterances by the manual evaluation (higer is better).

Model	Dullness
Human	2.18
Boilerplate	3.06
Separate (IR)	2.47
Separate (Gen+MMI)	2.39
Joint	**2.36**

Table 7: Results of evaluating **Dullness** of the generated utterances by the manual evaluation (lower is better). Boilerplate uses manually created boilerplate utterances.

We also see that all the BLEU scores of the models are quite lower than ROUGE scores in Table 4. In general, both summarization and chat generation tasks often use automatic evaluation metrics to evaluate generated sentences, their scores tend to be much lower in the chat generation task. This is because the answer sentences (utterances) of the chat generation task have more diverse candidates than other generation tasks such as machine translation and summarization (Li et al., 2016a,b; Baheti et al., 2018). We also examine the diversity of the chit-chat part in Table 5. Although the diversity of the IR-based methods was high, their BLEU scores deteriorated considerably. Among the generation-based methods, although **Separate (Gen+MMI)** achieved the highest BLEU score, it lacked diversity. In contrast, **Joint** achieved a reasonable BLEU score while maintaining diversity to some extent.

5.4 Manual Evaluation

Although diversity of utterances can be quantified automatically, ROUGE and BLEU scores do not always follow human intuition (Novikova et al., 2017; Lowe et al., 2017). Therefore, we evaluate the generated initial utterances manually. We picked the three proposed models with good performance in the automatic evaluation along with one baseline for this manual evaluation. 300 posts were sampled as the input news posts, and the outputs of the four methods were manually evaluated from two perspectives: 1) **Naturalness**: *Does the utterance naturally initiate conversation?* and 2) **Coherency**: *Is the content of the utterance coherent with the given news post?* We recruited crowd workers to score each utterance on a 4-point scale (Agree, Slightly Agree, Slightly Disagree, Disagree).

Table 6 show the results of the manual evaluation for **Naturalness** and **Coherency** of the generated initial utterances. The proposed methods excluding **Separate (IR)** outperformed **Baseline (Gen+MMI)** in both perspectives and achieved reasonable scores compared to human upper-bound. The scores of **Separate (IR)** are quite low because the retrieval result does not follow the input news post in many cases. This reveals that although those sentences have high diversity, their quality is poor as initial utterances. Although **Baseline (Gen+MMI)** achieved high ROUGE scores in Table 4, its style is not colloquial. Thus, workers felt odd and lowered their scores. In conclusion, it is better to use the generation-based methods for conversation initiation.

We also evaluated **Dullness**: *Is the given utterance dull or boring?* We used 15 manually created boilerplate utterances (*e.g.,* Hello., How are you?, Let's talk with me.) rather than **Baseline (Gen+MMI)** to confirm the effectiveness of utilizing news contents as the initial utterances. Table 7 show the results of the manual evaluation for **Dullness** of the generated initial utterances. We see that compared to our proposed methods, the score of the boilerplate baseline is quite high. This indicates that using boilerplate utterances for conversation initiation often bores users and possibly leads to early abandonment of the conversation.

To determine the statistical significance of our results, we performed Wilcoxon signed-rank tests with Bonferroni correction (Wilcoxon, 1945). In Table 6, for all combinations except **Baseline (Gen+MMI)** vs. **Separate (IR)** and **Separate (Gen+MMI)** vs. **Joint**, there were significant differences (p-value < 0.005 (corrected)) in both perspectives. Similarly, in Table 7, there were statistically significant differences for all combinations except **Separate (IR)** vs. **Separate (Gen+MMI)** and **Separate (Gen+MMI)** vs. **Joint**.

5.5 Examples

Finally, we investigated the initial utterances generated by **Separate (Gen+MMI)** and **Joint**. Ex-

news posts	initial utterances
A parade for the Rio Olympics and Paralympic medalists will be held in October. Approximately 500,000 people gathered at the time of the London Olympics.	**Separate (Gen+MMI):** I heard that a parade for the Rio Olympics and Paralympic medalists will be held in October. *That's amazing.*
	Joint: I heard that a parade for the Rio Olympics and Paralympic medalists will be held in October. *I would like to see what parade it is.*
A Chinese captain who was poaching a coral in the offshore of Kagoshima was arrested. The number of poaching boats has sharply declined in Ogasawara.	**Separate (Gen+MMI):** I heard that the coral was arrested in Kagoshima prefecture offshore because of poaching. *Get it together.*
	Joint: I heard that a Chinese captain who poached a coral in the offshore of Kagoshima was arrested. *Do not do poaching!*
On a suicide bombing that happened at a concert in England, the homeless action around the scene attracted praise. The UK raised the terrorist threat level to the "highest."	**Separate (Gen+MMI):** I heard that the homeless action around the scene of a suicide bombing that happened in England was praised. *That's scary.*
	Joint: I heard that the homeless action on a suicide bombing in England was praised. *Do you take measures against terrorism?*
The Meteorological Agency announced that Typhoon 11 will approach Tohoku in the early morning of 21st. Typhoon 9 is expected to approach Tokai on 21st and Typhoon 10 will approach Tokai or Kanto.	**Separate (Gen+MMI):** It seems that Typhoon 11 will approach Tohoku on 21st. *Let's watch out!*
	Joint: I heard that The Meteorological Agency announced that Typhoon 11 will approach Tohoku in the early morning of 21st. *We should pay attention to the future movement.*

Table 8: Examples of generated initial utterances. *Italics* are chit-chat parts.

amples of these utterances are provided in Table 8.

We found that **Separate (Gen+MMI)** tended to generate generic utterances (*e.g., "That's amazing", "Get it together"*) as the chit-chat part that fit any context, even though it uses a diversity-promoting function when decoding. In contrast, **Joint** could generate more diverse chit-chat parts by utilizing contents words such as *"parade"* and *"poaching"*. One possible reason for this phenomena is that the generated summary part acts like an additional condition of $P(T|S)$ at the time of decoding the chit-chat part. This does not happen with **Separate (Gen+MMI)**, which simply concatenates the outputs of separate models.

Interestingly, we found that there are some utterances giving a *question* (third example of **Joint** in Table 8) or making an *urging* (fourth example of **Separate (Gen+MMI)** in Table 8). Controlling utterances of the model by such dialogue acts (Wen et al., 2015; Zhao et al., 2017) can make the conversation initiation more diverse and attractive. We leave them as the future work at this time.

We should note that although it is a problem common to all the generation-models, there is a possibility of transmitting false news contents (as in the second example of **Separate (Gen+MMI)** in Table 8) or ethically inappropriate contents to the users. Therefore, when adopting our method into an actual conversation application, we have to pay close attention to this problem.

6 Conclusion

In this paper, we proposed the new task of conversation initiation. To generate diverse initial utterances that can improve user engagement, we utilized news articles that provide fresh and varied information every day and constructed a large-scale dataset using crowd workers. To perform the conversation initiation, we designed separate and joint approaches including both IR-based and generation-based methods. Empirical experiments showed that the proposed methods outperformed the baselines in both automatic and manual evaluation, and can generate diverse initial utterances that template-based methods cannot make. These results demonstrate the quality of our constructed dataset, that will be released for future studies[11].

As a natural next step, we plan to develop a more sophisticated conversation model, which can not only generate initial utterances but also continue the conversation for the given news contents (Yoshino and Kawahara, 2014). In that case, depending on the user's interest, the model needs to determine whether to do a usual chat or talk about the news contents. We also plan to improve the proposed method so that it can generate even better initial utterances. Since our task has two elements, summarization and chit-chat, the focus of our future work will be a more sophisticated multi-task model that considers these relations.

Acknowledgments

We thank Manabu Sassano for fruitful discussions and comments. We also thank the anonymous reviewers.

[11] https://research-lab.yahoo.co.jp/en/software/

References

Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. In *Proceedings of ACL*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of ICCV*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceesdings of ICLR*.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of EMNLP*, pages 3970–3980.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of CVPR*, pages 1080–1089. IEEE.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of ACL*, pages 964–972.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of WWW*, pages 506–516.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *In Proceedings of ACL*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of ACL*, pages 994–1003.

Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. 2016c. Stalematebreaker: a proactive content-introducing approach to automatic human-computer conversation. In *Proceedings of IJCAI*, pages 2845–2851.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop*, pages 74–81.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of ACL*, volume 1, pages 1116–1126.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of EMNLP*, pages 2322–2332.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of ACL*, pages 1802–1813.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of EMNLP*, pages 2241–2252.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *In Proceedings of EMNLP*, pages 1532–1543.

Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenny Olmstead. 2010. Understanding the participatory news consumer. In *Pew Internet and American Life Project*.

Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *Proceedings of ACL*, pages 151–156.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of ACL*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

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

Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of prospective user engagement with intelligent assistants. In *Proceedings of ACL*, pages 1203–1212.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL*, pages 1073–1083.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of AAAI*, pages 3295–3301.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL*, pages 196–205.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of ICML Workshop*.

Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of EMNLP*, pages 1711–1721.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of EMNLP*, pages 3981–3991.

Rui Yan. 2018. "chitty-chitty-chat bot": Deep learning for conversational ai. In *Proceedings of IJCAI*, pages 5520–5526.

Rui Yan and Dongyan Zhao. 2018. Smarter response with proactive suggestion: A new generative neural conversation paradigm. In *Proceedings of IJCAI*, pages 4525–4531.

Rui Yan, Dongyan Zhao, et al. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of SIGIR*, pages 685–694. ACM.

Koichiro Yoshino and Tatsuya Kawahara. 2014. Information navigation system based on pomdp that tracks user focus. In *Proceedings of SIGDIAL*, pages 32–40.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of ACL*, volume 1, pages 654–664.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of EMNLP*, pages 708–713.