An Ensemble Dialogue System for Facts-Based Sentence Generation

Ryota Tanaka, Akihide Ozeki, Shugo Kato, Akinobu Lee
Graduate School of Engineering, Nagoya Institute of Technology, Japan
{rtanaka, ozeki, shugo693, ri }@slp.nitech.ac.jp

Abstract

This study aims to generate responses based on real-world facts by conditioning context and external facts extracted from information websites. Our system is an ensemble system that combines three modules: generated-based module, retrieval-based module, and reranking module. Therefore, this system can return diverse and meaningful responses from various perspectives. The experiments and evaluations are conducted with the sentence generation task in Dialog System Technology Challenges 7 (DSTC7-Task2). As a result, the proposed system performed significantly better than sole modules, and worked fine at the DSTC7-Task2, specifically on the objective evaluation.

Introduction

The popularization of Social Networking Services (SNS) offers the advantage of reducing the burden of building large-scale open datasets. Therefore, recent works pertaining to dialogue systems have focused on end-to-end dialogue system using neural networks (Vinyals and Le 2015; Serban et al. 2016; 2017b). The end-to-end approach has a potential to generate tailored and coherent responses for user-input. However, there are still some problems with suffering from "safe response" phenomenon available to any utterance, such as the "*I'm sorry*"and "*I think so*," and generating words that have meanings different from real-world facts. This is because neural networks generally infer responses using only the collection of conversational transcriptions.

To tackle these problems, researchers have taken various approaches. (Ghazvininejad et al. 2018) proposed a knowledge-grounded dialogue system, conditioned on the context and facts extracted from online resources such as SNS posts utilizing location information. This easily and quickly enables to handle topics not appeared in training data and to adapt to a new domain. In the other approach, dialogue systems combining multiple dialogue models allow responses to be more diverse than those with a single model so that they can treat user-inputs from various viewpoints (Serban et al. 2017a; Song et al. 2018). We believe that combining these approaches is crucial to generate meaningful responses.

In this study, we propose an ensemble dialogue system conditioned on a previous context and external facts. This system consists of three modules including generation, retrieval and reranking. First, two modules generate and retrieve responses by feeding context and facts extracted from information websites such as Wikipedia. In generating candidates, we use the method extending Diverse Beam Search (DBS) (Vijayakumar et al. 2018) by enhancing the probability of words in facts data to treat low-frequency words such as proper nouns in external data adequately. Second, the reranking module sorts these candidates according to several features considering appropriateness and informativeness, and it finally returns the final response which is the highest-ranked candidate. Our main contributions of this paper has two-fold : (1) we propose a model for combining multiple hypotheses and injecting external facts, (2) we develop a method to decode diverse and informative words.

We evaluate the performance with the DSTC7-Task2 (Yoshino et al. 2018), which is devoted to building dialogue systems generating responses based on real-world facts. In this paper, we report our experimental results.

Problem Definition

The system outputs a response using a context $S = \{U_1, ..., U_M\}$ in M recent turns and N facts $F = \{f_1, ..., f_N\}$ relevant to the context, where F is a sentence, containing HTML tag, extracted information websites. Each utterance $U_m = \{x_{m,1}, ..., x_{m,n}\}$ is composed of n words.

Here, we categorize F as subject facts $F^{subj} = \{f_1^{subj}, ..., f_K^{subj}\}$ and description facts $F^{desc} = \{f_1^{desc}, ..., f_L^{desc}\}$ using the HTML tag rule. F^{subj} is a sequence enclosed by `<h>` tag or `<title>` tag, and F^{desc} is a sequence enclosed by `<p>` tag or not enclosed by any tag.

Ensemble Dialogue System for Facts-Based Sentence Generation

We propose an ensemble dialogue system using external facts and context. As shown in Figure 1, it consists of the Memory-augmented Hierarchical Encoder-Decoder

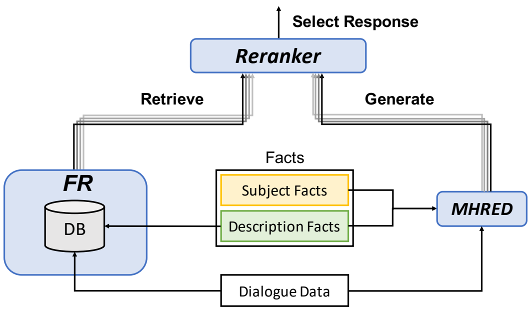


Figure 1: An overview of the proposed model

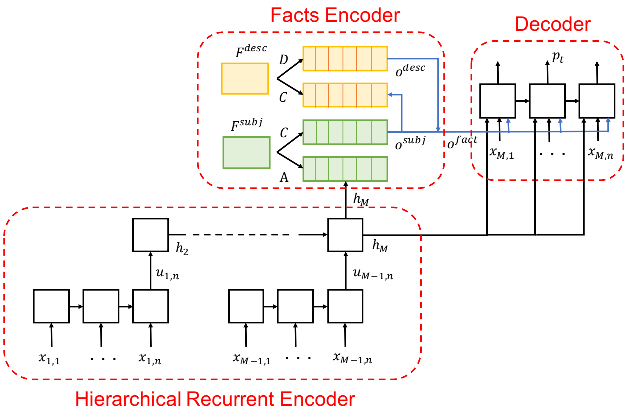


Figure 2: An overview of the MHRED

(MHRED), the sentence selection module with facts retrieval (FR), and the Reranker. This system has two processes: the generate-retrieval process and the reranking process. In the generate-retrieval process, the MHRED generates responses using context and external facts, and the FR retrieves the responses from a database containing important words extracted from the facts. In the reranking process, we use a binary classifier with various dialogue features to select the final response by feeding all the candidates from the MHRED and FR. In this section, each module of the proposed system is introduced in detail.

Memory-augmented Hierarchical Recurrent Encoder-Decoder

To inject facts into responses, a novel encoder-decoder model incorporating end-to-end memory networks (MemN2N) (Sukhbaatar et al. 2015) architecture into hierarchical recurrent encoder-decoder (HRED) (Serban et al. 2016) is proposed. We call this model as Memory-augmented HRED (MHRED). The overview of MHRED is shown in Figure 2.

Hierarchical Recurrent Encoder To encode the context, a Hierarchical Recurrent Encoder (HRE) is applied. Previous work has shown that hierarchical Recurrent Neural Networks (RNNs) have a higher ability to express the dialogue context than non-hierarchical RNNs (Tian et al. 2017). The HRE consists of two level encoders, one at the utterance level and the other at the context level, computed by the Gated Recurrent Unit (GRU) (Chung et al. 2014). An utterance encoder converts each utterance to an utterance vector. The utterance vector is the hidden state obtained after encoding the last word in each utterance. Let $w_{m,t}$ denote the word embedding of the t-th word in the m-th utterance. Then, utterance vector $u_{m,t}$ is computed as follows:

$$u_{m,t} = \mathrm{GRU}(u_{m,t-1}, w_{m,t}) \tag{1}$$

After processing each utterance, a context encoder outputs context vector h_m, which is a summary of the past utterances, as follows:

$$h_m = \mathrm{GRU}(h_{m-1}, u_{m-1,n}) \tag{2}$$

Facts Encoder A facts encoder is introduced to select facts that need to be injected in responses and map the facts to the continuous representation utilizing the concept of MemN2N architecture. F^{subj} contains many sentences, written headlines, and titles concerning facts, whereas F^{desc} mostly contains sentences explaining the headline and the title. To access the F^{desc} using F^{subj}, it is efficient to extract the detailed facts about the headlines and titles since they tend to contain vital information as a fact. Therefore, we extend the facts encoder proposed by (Ghazvininejad et al. 2018) and to store F^{subj} in the first memory (first hop), and F^{desc} in the last memory (second hop).

First, F^{subj} and F^{desc} are converted into memory vector $r^{subj} = \{r_1^{subj}, ..., r_K^{subj}\}, r^{desc} = \{r_1^{desc}, ..., r_L^{desc}\}$ by sum of word embeddings for each sentence. Then, context vector h_M, which is the last hidden state of HRE, is fed into the facts encoder in the first memory, and subject fact o^{subj} is obtained, as shown below:

$$m_i^{subj} = A r_i^{subj} \tag{3}$$
$$c_i^{subj} = C r_i^{subj} \tag{4}$$
$$p_i^{subj} = \mathrm{softmax}(h_M^{\mathrm{T}} m_i^{subj}) \tag{5}$$
$$o^{subj} = \sum_i^K p_i^{subj} c_i^{subj} \tag{6}$$

where $A, C \in \mathbb{R}^{d \times |V|}$ ($|V|$ denotes the vocabulary size) are trainable parameters. Moreover, h_M and o^{subj} are passed to the second memory, and we obtain the description fact o^{subj} as follows:

$$m_i^{desc} = C r_i^{desc} \tag{7}$$
$$c_i^{desc} = D r_i^{desc} \tag{8}$$
$$p_i^{desc} = \mathrm{softmax}\{(h_M + o^{subj})^{\mathrm{T}} m_i^{desc}\} \tag{9}$$
$$o^{desc} = \sum_i^L p_i^{desc} c_i^{desc} \tag{10}$$

where $C, D \in \mathbb{R}^{d \times |V|}$ are trainable parameters. Note that C denotes the shared weights between memories. Finally, vector concatenation across the rows on o^{subj}, o^{desc} is performed and facts vector $o^{fact} = [o^{subj}; o^{desc}]$ is obtained.

Decoder A decoder reads context vector h_M and facts vector o^{fact} and predicts the next utterance. Let the initial hidden state be $s_0 = h_M$. Then, the hidden state of decoder s_t is computed by GRU as follows:

$$s_t = \mathrm{GRU}(s_{t-1}, w_{M,t}) \tag{11}$$

In generating conversational responses such as "*I think*" and "*I know*" on the decoder, it is not always necessary to use facts relevant to the context at all time steps. Hence, the decoder should change the preference to whether facts or other information needs to be used. We use Maxout Networks (MN) (Goodfellow et al. 2013) to generate the response injecting facts. MN obtains the vector with the maximum value e, where e can be computed with linear transformation of input features. Its vector represents the most important features from among all features, and then enables the decoder to switch depending on whether only facts are required. The probability of generating the word p_t is calculated as follows:

$$z_t = W_z w_{M,t} + U_z h_M + V_z s_t + H_z o^{fact} \tag{12}$$
$$e_t = [\max\{z_{t,2j-1}, z_{t,2j}\}]^{\mathrm{T}} (j = 1, ..., d) \tag{13}$$
$$p_t = \mathrm{softmax}(W_e e_t) \tag{14}$$

where $W_z, U_z, V_z, H_z \in \mathbb{R}^{2d\times d}$, and $W_e \in \mathbb{R}^{|V|\times d}$ are trainable parameters.

Diverse Sentence Generation with Facts Most neural dialogue systems apply Beam Search (BS) to generate the optimal response (Vinyals and Le 2015; Serban et al. 2016; 2017b). However, BS does not guarantee diversity for the final response because word sequences within the beam width closely resemble each other. In addition, words such as proper nouns, which often appear in facts data, tend to be less selective than general words appearing in dialogue data.

Previous work extended BS to focus on alleviating the diversity problem. (Vijayakumar et al. 2018) proposed Diverse Beam Search (DBS), generating diverse word sequence alternatives to BS. Given a beam width B, groups G, and beam width in group $B' = B/G$, beam sets at time step t are divided into G subsets. The DBS selects the word Y_t^g in order of $g = 1, ..., G$ for these subsets as follows:

$$Y_t^g = \underset{y^g_{1,[t]},\dots,y^g_{B',[t]}}{\mathrm{argmax}} \sum_{b\in[B']} \Theta_t(y^g_{b,[t]}) + \lambda\Delta_{div} \tag{15}$$

where λ is the hyper-parameter, Θ is the log probability, and Δ_{div} is the penalty which is the hamming distance between the words selected in the other groups and $y^g_{b,[t]}$. Note that the DBS sets the penalty $\Delta_{div} = 0$ at $g = 1$.

Furthermore, we extend the DBS to add a penalty with facts. In order to enhance the probability of generating the word sequence to contain words in facts data, we introduce a penalty Δ_{fact}, using the similarity between facts and the sequence of candidate words. Let γ be the hyper-parameter. The penalty term $\gamma\Delta_{fact}$ is added to the equation (15) when the word Y_t^g is selected. Here, Δ_{fact} is calculated as follows:

$$\Delta_{fact} = \frac{1}{K+L} \sum_{n=1}^{K+L} \mathrm{Sim}(\sum_{i=1}^{t} w_i^g, \sum_{j=1}^{|f_n|} w_j^{f_n}) \tag{16}$$

where w^g, w^f is the Y^g, f of word embeddings computed by Word2Vec (Mikolov et al. 2013) respectively, and $\mathrm{Sim}(\cdot,\cdot)$ denotes cosine similarity.

Sentence Selection with Facts Retrieval

In general, the raw human-human conversation is highly fluent and rich in variety, and often contains a considerable amount of information about a specific topic in itself. Thus, in this study a method that combines utterance selection based on facts is also proposed. Hence, Facts Retrieval (FR) is employed to output responses, including facts in responses and context. Let S be the context and R be the response. The database is constructed in the form of $< [S; R], R >$, where $[S; R]$ is a query, which is word sequence concatenation on S and R, and R is a system output. Note that the database is used from the training dialogue dataset.

For sentence selection, we extract important words Q from facts and feed them into the database. Here, Q denotes overlapping words in F^{subj}, which contains titles and headlines. In order to eliminate noises and improve the quality of retrieval, Q is restricted to word sequence that includes at least one noun, verb, adjective, and adverb. FR outputs R satisfying the relation $Q \in [S; R]$. If multiple sentences satisfy the relation, FR reranks sentences using the score produced by BM25F (Zaragoza et al. 2004) and outputs up to 10 sentences. Note that FR will not output sentences if the relation is unsatisfied.

Reranker

The outputs of the MHRED and FR modules may contain meaningless and non-fluent responses. Hence, these responses should be eliminated; the responses should be both appropriate and informative. The Reranker sorts candidates by feeding all of the results of the MHRED and FR, and the highest ranked candidate is returned to user as the final response. It classifies whether a candidate is "positive" or "negative" as a response, where the probability of being "positive" is computed as a confidence score from binary classification with XGBoost (Chen and Guestrin 2016). The features of the Reranker consist of three categories, "Candidate" (responses returned by both the FR and MHRED), "Pair" (a pair of a previous utterance and "Candidate"), and "Context" (a pair of a context and "Candidate"), as shown in Table 1. These categories enable evaluation of the quality of the responses.

Pairs of a context and a response from the dialogue dataset was used to build the training dataset. Contexts and responses with a high "response score" (over 100) on Reddit [1] was chosen as positive examples. Then, negative examples are generated on those contexts according to one of the following rules:

- A randomly selected response with a low "response score" (1 or less) from a dialogue on another topic.
- A response that swap words and eliminates some words randomly from a positive example.
- A response that matches both of above-mentioned descriptions.

Category	Features	About Features
Candidate	Length	Number of characters, and words
	Fluency	N-gram ($N = 2, 3$) language model
	POS	Number of nouns, verbs, adjectives, and adverbs
	Fact	Frequency of words appeared in F^{subj} and F^{desc} / number of words
Pair	Word sim	Cosine similarity between one-hot vectors of words
	N-gram sim	Cosine similarity between N-gram ($N = 2, 3$)
	Length sim	Similarity [2] of number of characters, and words
	Embedding sim	Cosine similarity between vectors computed as the averaged Word2Vec
	Sentimental sim	Similarity [3] of semantic orientations (Takamura, Inui, and Manabu 2005)
	POS sim	Cosine similarity between BoW (nouns, verbs, adjectives, and adverbs)
	Proper Noun sim	Cosine similarity between BoW of proper noun types, extracted by NLTK [4]
	Keyword sim	Cosine similarity between the averaged Word2Vec of keywords extracted by RAKE algorithm (Rose et al. 2010)
Context	Topic sim	Cosine similarity between topic vectors by feeding a context and candidate to LDA model (Blei, Ng, and Jordan 2003)

Table 1: Features used to select a response on the Reranker.

As a result, the dataset contains 44,449 "positive" and "negative" examples respectively.

Experiments

Datasets

The experiment was performed according to the regulations of DSTC7-Task2. We crawled the dialogue dataset from 178 subreddits (subsidiary threads or categories on Reddit). Markdown and special symbols were eliminated from the crawled dialogue dataset, and for the same context, the context-response pair of the highest "response score" was selected. We crawled the facts dataset from 226 information sharing websites, such as Wikipedia. The facts were categorized into F^{subj} and F^{desc} as mentioned above, up to the top 10 sentences with the highest cosine similarity for each context. For calculating similarity, the average of the Word2Vec output with 256 dimension was used. Note that the Word2Vec model was trained only on the official training datasets according to DSTC-Task2 regulations. The pre-processing described above leads to the formation of the dialogue and facts datasets, as shown in Table 2.

Dialogue Dataset	train	dev	test
# Dialogues	832908	40932	13440
Avg. Turns	4.72	4.80	4.02
Avg. Tokens/Utterance	23.32	23.64	34.84
Facts Dataset			
Avg. Tokens/Sentence (s)	3.86	3.61	3.30
Avg. Tokens/Sentence (d)	17.11	16.67	15.63
# Topics (s)	27735	1152	3047
# Topics (d)	27645	1121	3063

Table 2: Statics of pre-processed dataset."s" and "d" denotes subject facts and description facts , respectively.

Evaluation Metrics

Automatic evaluation and human evaluation for responses were conducted in DSTC7-task2 organizers. For automatic evaluation, two types of metrics are used; one is word-overlap metrics, including BLEU (Papineni et al. 2002), NIST (Doddington 2002) and METEOR (Banerjee and Lavie 2005), and the other is the diversity metric using div (Li et al. 2015). In human evaluation, human evaluates responses rated with score 1 (Strong Disagree) to 5 (Strong Agree) for Appropriateness and Informativeness using crowdsourcing.

Models for Comparison

Several models are evaluated to show the effectiveness of the proposed model:

- `S2S`: Sequence-to-sequence (seq2seq) model (Vinyals and Le 2015).
- `HRED`: HRED model (Serban et al. 2016).
- `HRED-F`: Add the Δ_{fact} term to DBS (Vijayakumar et al. 2018), which generates the responses of HRED. $B' = 1 (B = 15, G = 15)$.
- `MHRED-F`: Add the Δ_{fact} term to DBS, which generates the responses of MHRED. $B' = 1$ $(B = 5, G = 5)$.
- `MHRED-F15-R, MHRED-F5-R`: Add the Δ_{fact} term to DBS, which generates the responses of MHRED. $B' = 1$ $(B = 15, G = 15)$ or $(B = 5, G = 5)$. Reranker selects the final response from candidates returned by `MHRED-F`.
- `Ensemble`: Reranker selects the final response from candidates returned by both `MHRED-F` and FR.

[1] https://www.reddit.com/

[2] Let $|U|$ and $|S|$ be the length of the previous utterance and candidate sentence normalized 0 to 1 respectively. Then, the similarity is calculated as $1.0 - |\,|U| - |S|\,|$.

[3] Let U_{sent} and S_{sent} be the average of the semantic orientations of the previous utterance and candidate. Then, the similarity is calculated as $1.0 - |(U_{sent} - S_{sent})/2|$.

[4] https://www.nltk.org/

Model	NIST4	BLEU4	METEOR	div1
`S2S`	0.023	0.34	3.92	0.026
`HRED`	0.730	0.58	5.65	0.049
`HRED-F`	0.766	0.68	5.61	0.049
`MHRED-F`	0.555	0.76	5.24	0.069
`MHRED-F15-R`	1.802	0.92	6.45	0.058
`MHRED-F5-R`	1.749	1.10	**6.74**	0.051
`Ensemble`	**2.047**	**1.35**	6.71	**0.094**

Table 3: Results of the automatic evaluation.

Model	Appropriateness	informativeness
baseline(constant)	2.60	2.32
baseline(random)	2.32	2.35
`Ensemble`	**2.69**	**2.58**

Table 4: Results of the human evaluation.

Moreover, the baseline models "baseline(random)" and "baseline(constant)" derived from the organizers are compared with the proposed model in human evaluation.

Note that only FR model should not be compared with other models since FR model is not able to output responses continuously.

Model Setup

We use a two-layer seq2seq, HRED and MHRED for training. All models are set to the word embedding dimension and hidden vector size of 256. Mini-batch training was employed with a batch size of 40. The models were trained with cross entropy loss function and adapted Adam optimization algorithm (Kingma and Ba 2014) with the initial learning rate of 0.0001. To alleviate over-fitting to the training dataset, a dropout rate of 0.2 was set for all models. Training was conducted for up to 20 epochs and the model with the lowest perplexity in the dev dataset was selected.

Hyper-parameters of DBS was set as $\lambda = 0.4$ and $\gamma = 10.0$ according to BLEU on dev dataset. Vocabulary size was set to 20k, which is shared between both the dialogue and facts data. In generating responses, the log probability of out-of-vocabulary (OOV) words was set to $-\infty$ so as not to generate the special symbol <unk>.

Results and Discussion

Table 3 shows results of the automatic evaluation. It can be seen that the proposed model `Ensemble` performs better than other models. This indicates that `Ensemble` enables to output more fluent responses similar to human and diverse responses.

Comparing the result of `MHRED-F` and `HRED-F`, notably at div1 score, it is apparent that the proposed MHRED architecture is superior to conventional models. It shows that MHRED can infer words and topics using facts that may be hard to handle only from conversation data and generate diverse responses on the new domain.

Comparison of `Ensemble` and `MHRED-F5-R` indicates that the FR module is effective. This is because the responses by the FR are parts of the conversation actually chatted by human and thus highly fluent. Thus, it is shown that the MHRED and the FR are useful in generating informative and appropriate responses.

Category	Feature	Difference: Category	Difference: Feature
Candidate	Length	$\mathbf{-0.1039}$	$+0.0004$
	Fluency		$\mathbf{-0.1053}$
	POS		-0.0038
	Facts		-0.0021
Pair	Word sim	-0.0155	-0.0021
	N-gram sim		$+0.0013$
	Length sim		-0.0027
	Embedding sim		-0.0021
	Sentimental sim		-0.0023
	POS sim		-0.0004
	Proper Noun sim		± 0.0000
	Keyword sim		$\mathbf{-0.0203}$
Context	Topic sim	-0.0015	-0.0015

Table 5: Differences of accuracy computed by Reranker when the target feature is excluded.

To analyze effectiveness of introducing the penalty Δ_{fact} and the Reranker, we compared `HRED-F`, `HRED`, `MHRED-F15-R` and `MHRED-F`. The model combining the Reranker (`MHRED-F15-R`) gives significantly higher performance than the model without the Reranker, even on a single model. It designates capturing diverse perspectives of dialogue with various features is important for response generation. Conversely, the model introducing the penalty Δ_{fact} (`HRED-F`) showed slight improvement on NIST4 and BLEU4. This indicates that adding the penalty of DBS has positive potential to generate responses similar to human-made.

Table 4 shows the results of human evaluation. Since our primary model beats official baseline models returning responses randomly and constantly, the proposed model is able to capture the context and generate responses fluently.

Case Study and Error Analysis

To validate the MHRED architecture, we looked into the details of the result with attention value p^{subj} and p^{desc} in the facts encoder. Figure 3 depicts an example of attention paid by the fact encoder. The F^{subj} captures "*seven wonders of the ancient world*", which refers to the topic of the context. Subsequently, The F^{desc} captures the facts containing "*pyramid*" considering both the context and F^{subj}. Finally, the MHRED generates a response including "*pyramid*". This indicates that this model enables to focus on the facts relevant to the context and generate responses injecting them.

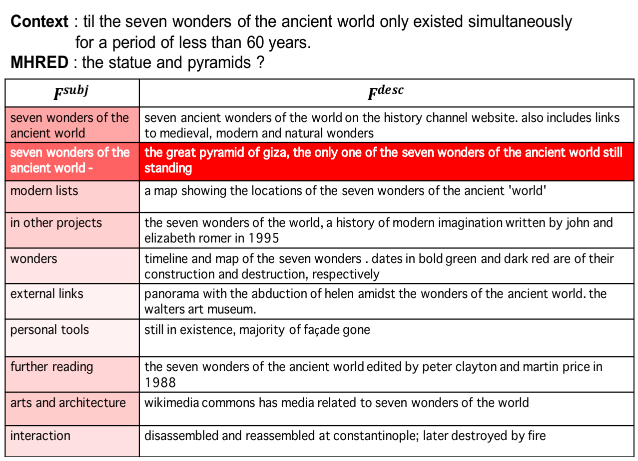
Table 5 shows how the accuracy of the Reranker changes when one of the target features is excluded, per feature or per category. A negative value implies that the corresponding feature is important. The category "Candidate" showed significant decrease of all categories, and the feature "Fluency" showed the biggest decrease by -0.1053, followed by "Keyword sim" by -0.0203. Thus, the Reranker has a tendency to select the final response focusing on fluency and contextually informativeness in dialogue. This tendency is

#	Context	Response
1	til there is enough water in lake superior to flood the entire landmasses of north and south america to a depth of 1 foot. it contains over 3 quadrillion gallons of fresh water. → the world's largest by volume and deepest lake is located in southern russia. lake baikal.	`S2S`: i'm not true. `HRED`: we are there. i don't really count, but its not.. `HRED-F`: also, and i have been posted *lake baikal* are on. the water lake . `Ensemble (FR)`: *lake baikal* volume. so the answer is yes a bit more than the all the great lakes, and a little under twice the volume of *lake superior*.
2	til that president jimmy carter placed his peanut farm in a blind trust when he was elected president. after his term ended, he found that it had been badly mismanaged and he was $1 million in debt. → god bless jimmy carter america's last honest president.	`S2S`: what do you think, i think of. `HRED`: not sure. the best really liked by far. `HRED-F`: yeah. but that is a president *obama*. `Ensemble (MHRED)`: *obama* was also a bit of course he was president *obama*. *democrat*.

Table 6: Example of a response. The token → indicates a change of turn. `Ensemble (·)` denotes the output source (either MHRED or FR) selected by the Reranker.

#	Context	Response	CS	Rank
1	til that kyoto, the former capital of japan, just means "capital city" and tokyo means "eastern capital" → i only just noticed that tokyo and kyoto are anagrams.	`MHRED`: i think tokyo godzilla, but as well and kyoto.	0.9602	1
		`MHRED`: they also have been a lot of tokyo as the tokyo are they have the same as well. the kyoto is the only one.	0.9342	2
		`FR`: villages arent cities.	0.1817	worst
2	til german animal protection law prohibits killing of vertebrates without proper reason. because of this ruling, all german animal shelters are no-kill shelters. → i am german. til that there are kill shelters.	`FR`: wow! i didnt know there was a tv show about for pets/animal shelters. thats pretty cool! do you know if that sort of advertising caused a lot more people to adopt animals?	0.8144	1
		`MHRED`: its a good thing about cats are occupying breeds cats.	0.7297	2
		`FR`: use its hide as shelter.	0.7234	3

Table 7: Example of reranking by the Reranker. The token → indicates a change of turn. CS represents the confidence score produced by XGBoost.


Context : til the seven wonders of the ancient world only existed simultaneously for a period of less than 60 years.
MHRED : the statue and pyramids ?

F^{subj}	F^{desc}
seven wonders of the ancient world	seven ancient wonders of the world on the history channel website. also includes links to medieval, modern and natural wonders
seven wonders of the ancient world -	the great pyramid of giza, the only one of the seven wonders of the ancient world still standing
modern lists	a map showing the locations of the seven wonders of the ancient 'world'
in other projects	the seven wonders of the world, a history of modern imagination written by john and elizabeth romer in 1995
wonders	timeline and map of the seven wonders . dates in bold green and dark red are of their construction and destruction, respectively
external links	panorama with the abduction of helen amidst the wonders of the ancient world. the walters art museum.
personal tools	still in existence, majority of façade gone
further reading	the seven wonders of the ancient world edited by peter clayton and martin price in 1988
arts and architecture	wikimedia commons has media related to seven wonders of the world
interaction	disassembled and reassembled at constantinople; later destroyed by fire

Figure 3: Attention paid by the facts encoder. Sentences painted in darker shades of red represent greater attention.

probably due to making training dataset for the Reranker. Negative examples are generated using hand-crafted rules such as swapping and eliminating words, thus resulted in the tendency to select more higher "Fluency" and "Keyword sim" sentences preferentially.

Table 6 shows examples of responses predicted by the models. As can be observed from the table, `HRED-F` and `Ensemble` output more informative words related to the context such as "*lake bikal*" (#1) or "*obama*" (#2). Table 7 presents examples of reranking by the Reranker. In example #1, the MHRED is explicitly designed for the previous context, and the Reranker selects the most meaningful response. In example #2, the response returned by the FR has high fluency and many content words. Conversely, the response is not suitable for the context in terms of the topic. This indicates, as above mentioned, that the Reranker tends to focus on "Candidate" strongly due to the way of making examples for the Reranker. However, we expects making examples from the various perspective will improve the performance more.

Conclusion and Future Work

In this paper, we proposed an ensemble dialogue system using external facts for DSTC7-Task2. The proposed system is a combination of three modules: the MHRED, a neural dialogue system which incorporates external facts into the procedure of response generation, the FR, and the Reranker. In generation, we extend the DBS to generate more meaningful words containing facts data. The experimental results showed that the MHRED especially improved the diversity of the response sentence over the baseline model. Moreover, we confirmed that the combination of multiple modules improved overall automatic metrics and generates more informative responses. In future work, we plan to extend the proposed model to introduce an end-to-end learning for multiple systems simultaneously for improvements.

References

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145. Morgan Kaufmann Publishers Inc.

Ghazvininejad, M.; Brockett, C.; Chang, M.; Dolan, B.; Gao, J.; Yih, W.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018*, 5110–5117.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A. C.; and Bengio, Y. 2013. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, 1319–1327.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *CoRR* abs/1510.03055.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26, 2013*, 3111–3119.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory* 1–20.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 3776–3784.

Serban, I. V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N. R.; Mudumba, S.; de Brébisson, A.; Sotelo, J.; Suhubdy, D.; Michalski, V.; Nguyen, A.; Pineau, J.; and Bengio, Y. 2017a. A deep reinforcement learning chatbot. *CoRR* abs/1709.02349.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, , 2017*, 3295–3301.

Song, Y.; Li, C.; Nie, J.; Zhang, M.; Zhao, D.; and Yan, R. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, 2018*, 4382–4388.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28 , 2015*, 2440–2448.

Takamura, H.; Inui, T.; and Manabu, O. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 133–140.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao, D. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 231–236.

Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D. J.; and Batra, D. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018*, 7371–7379.

Vinyals, O., and Le, Q. 2015. A neual conversational model. In *ICML, Deep Learning Workshop*.

Yoshino, K.; Hori, C.; Perez, J.; D'Haro, L. F.; Polymenakos, L.; Gunasekara, C.; Lasecki, W. S.; Kummerfeld, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B.; Gao, S.; Marks, T. K.; Parikh, D.; and Batra, D. 2018. The 7th dialog system technology challenge. *arXiv preprint*.

Zaragoza, H.; Craswell, N.; Taylor, M. J.; Saria, S.; and Robertson, S. E. 2004. Microsoft cambridge at TREC 13: Web and hard tracks. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*.