Computer Speech & Language 62 (2020) 101070

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Context and knowledge aware conversational model and system combination for grounded response generation

Ryota Tanaka*, Akihide Ozeki, Shugo Kato, Akinobu Lee

Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, Aichi, Japan

ARTICLE INFO

Article History:
Received 31 July 2019
Revised 26 December 2019
Accepted 16 January 2020
Available online 24 January 2020

Keywords:
DSTC
Dialogue system
Conversational AI
Sentence generation
Grounding knowledge

ABSTRACT

End-to-end neural-based dialogue systems can potentially generate tailored and coherent responses for user inputs. However, most of existing systems produce universal and non-informative responses, and they have not gone beyond chitchat yet. To tackle these problems, 7th Dialog System Technology Challenges (DSTC7-Track2) was developed to focus on building a dialogue system that produces informational responses that are grounded on external knowledge. In this study, we propose a Memory-augmented Hierarchical Recurrent Encoder-Decoder, called MHRED, that grounded on both multi-turn dialogue context and external knowledge. Furthermore, we apply a combination of multiple dialogue systems. Our final system is an ensemble that combines three modules: a generation-based module, a retrieval-based module, and a reranking module. First, responses are generated by MHRED, and retrieved from a pre-defined database focusing on informativeness. Next, the reranking module sorts these candidates using several hand-crafted features, and finally it selects a response with the highest score. Therefore, this system can return diverse and meaningful responses from various perspectives. Experimental results show that our proposed MHRED outperforms strong baseline models and combining multiple dialogue systems significantly improves the automatic evaluation and human evaluations.

1. Introduction

Generation-based dialogue systems using neural networks (Vinyals and Le, 2015; Serban et al., 2016; 2017b) have a potential to generate tailored and coherent responses for user inputs. However, there are still problems with the quality of responses. One is to suffer from non-informative responses that are different from real-world facts. Another problem is the low diversity of responses such as *I don't know*.

Several recent approaches have attempted to address these issues by incorporating external knowledge in the form of a knowledge base (Han et al., 2015; Xu et al., 2017), profile information as persona (Zhang et al., 2018), and review posts (Ghazvininejad et al., 2018). These were effective in supporting the lack of wide knowledge on dialogue data and in generating informative responses. However, most of the existing knowledge-based dialogue systems focused on single-turn conversation. This causes difficulties in understanding a complex dialogue context and in utilizing an appropriate knowledge, which leads to non-informative conversation.

In this study, to tackle challenges associated with multi-turn conversation, we propose a Memory-augmented Hierarchical Recurrent Encoder-Decoder (MHRED) model based on works by Serban et al. (2016) and Ghazvininejad et al. (2018). Our system generates responses grounded on both multi-turn dialogue context and external knowledge. In addition, we apply a system

* Corresponding author.
E-mail address: rtanaka@slp.nitech.ac.jp (R. Tanaka).

https://doi.org/10.1016/j.csl.2020.101070

combination technique (Serban et al., 2017a; Song et al., 2018; Pandey et al., 2018) to improve the overall performance. Compared to the generation-based model, the retrieval-based model, which searches for a corresponding utterance in a pre-defined database, can produce more fluent responses (Jr. et al., 2000; Ji et al., 2014). Thus, our final system consists of three modules including generation, retrieval, and reranking. First, MHRED and retrieval-based system generate and retrieve response candidates by feeding the dialogue context and facts extracted from information websites such as Wikipedia. To generate these candidates from MHRED with high diversity, we leverage a diverse beam search (Vijayakumar et al., 2018). Second, the reranking module sorts these candidates according to several features, taking into consideration appropriateness and informativeness. Finally, reranking module returns the final response that is the highest-ranked candidate.

We evaluated the performance using the knowledge grounded sentence generation task of the 7th Dialogue System Technology Challenge (DSTC7-Track2) (Yoshino et al., 2019). The challenge is devoted to focus on building end-to-end dialogue systems that can go beyond chitchat and produce both useful and appropriate responses. Automatic and human evaluations of the experimental results according to this challenge show that our final system performs better than competitive baseline models.

Our main contributions of this paper have two-fold : (1) we propose a knowledge grounded hierarchical encoder-decoder model that considers both the multi-turn conversation and external knowledge, and (2) we apply a combination of systems that reranks hypotheses produced from the context- and knowledge-aware generation-based and the retrieval-based models, which results in diverse and meaningful conversation.

2. DSTC7-Track2

As shown in Fig. 1, DSTC7-Track2 tries to push the data-driven conversation models beyond chitchat to produce both useful and appropriate responses like “ideal response”. In DSTC6-Track2 (Hori et al., 2019), the “The End-to-End Conversational Modeling” track was held to build a data-driven dialogue model trained on Twitter customer support data for a fully data-driven conversation in a goal-oriented scenario. Unlike the DSTC6-Track2, DSTC7-Track2 follows the previous work “A knowledge-grounded neural conversational model” (Ghazvininejad et al., 2018), and it takes the data-driven paradigm using not only dialogue data but also contextually relevant facts, which avoids hand-coding any linguistic, domain, or task-specific information (i.e., beyond chitchat). To create a grounded conversational dataset, multi-turn conversational threads on Reddit are used. On Reddit, user submissions typically consists of a submission title associated with a URL pointing to a website containing news or other background articles. Moreover, the Reddit conversation usually initiates a discussion about the contents of the article, and these can be seen as a form of contextually relevant facts. Therefore, this task uses the text snippets of articles as facts.

Although almost all the facts contain HTML tags and tend to be noisy, the tags offer important hints to capture the high-level structure of the facts. We hypothesizes that the tags can be separated into two roles according to a tag rule. First one is a subject tag (<h> and <title>) representing topics of the article, and second one is a description tag (<p> and others) denoting the content of the article. Based on these categorized tags, we separated facts into subject facts $F^{subj} = \{f_1^{subj}, ..., f_K^{subj}\}$ and description facts $F^{desc} = \{f_1^{desc}, ..., f_L^{desc}\}$, where f_k^{subj} is a word sequence enclosed by <h> or <title> tag, and f_l^{desc} is a word sequence enclosed by <p> tag or not enclosed by any tag.

In summary, the task is as follows: given a multi-turn dialogue context $S = \{U_1, ..., U_M\}$ in M recent turns where each utterance $U_m = \{x_{m,1}, ..., x_{m,n}\}$ constitutes n tokens, and two types of facts $F^{subj} = \{f_1^{subj}, ..., f_K^{subj}\}$ and $F^{desc} = \{f_1^{desc}, ..., f_L^{desc}\}$, the model finally outputs a word sequence $Y = \{y_1, ..., y_T\}$ as a response.

3. Approach

Our approach is based on building both a generation-based model and a retrieval-based model grounded on external knowledge, and applying a system combination technique (Serban et al., 2017a; Song et al., 2018; Pandey et al., 2018). As shown in Fig. 2, our

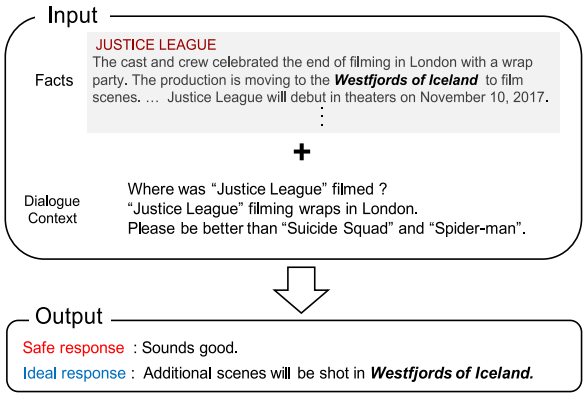

Fig. 1. A conversation example of DSTC7-Track2.

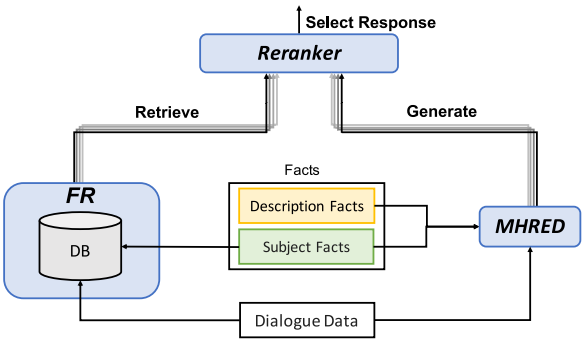

Fig. 2. An overall architecture of our approach.

proposed model consists of Memory-augmented Hierarchical Encoder-Decoder (MHRED), a retrieval-based module with Facts Retrieval (FR), and a reranking module. This system has two processes: the generation-retrieval process and the reranking process. In the generation-retrieval process, MHRED generates responses using dialogue context and contextually relevant facts, and FR retrieves responses containing important facts words from a pre-defined database. In the reranking process, a binary classifier with various dialogue features is used to select the final response by feeding all the candidates from MHRED and FR.

3.1. Memory-augmented hierarchical recurrent encoder-decoder

This model is focused on developing a sensitivity to both multi-turn dialogue context and relevant facts . In this study, we propose a hierarchical neural conversational model incorporating a MemN2N architecture into a hierarchical recurrent encoder-decoder (Serban et al., 2016) based on the idea of Ghazvininejad et al. (2018). As in Fig. 3, MHRED consists of three modules: Hierarchical Recurrent Encoder (3.1.1), Facts Encoder (3.1.2) and Decoder (3.2).

3.1.1. Hierarchical recurrent encoder

To encode the dialogue context, a hierarchical recurrent encoder is applied. A previous work has shown that hierarchical Recurrent Neural Networks (RNNs) have a higher ability to represent the dialogue context than non-hierarchical ones (Tian et al., 2017). The encoder consists of two-level encoders, one at the utterance-level and the other at the context-level, computed by a Gated Recurrent Unit (GRU) (Chung et al., 2014) with a single forward direction. An utterance encoder converts each utterance into an utterance vector. This vector is a hidden state obtained after encoding the last word of each utterance. Let $e(x_{m,t})$ be a word embedding of the tth word in the mth utterance. Then, the utterance vector $u_{m,t}$ is computed as follows:

$$u_{m,t} = \mathrm{GRU}\left(u_{m,t-1}, e(x_{m,t})\right). \tag{1}$$

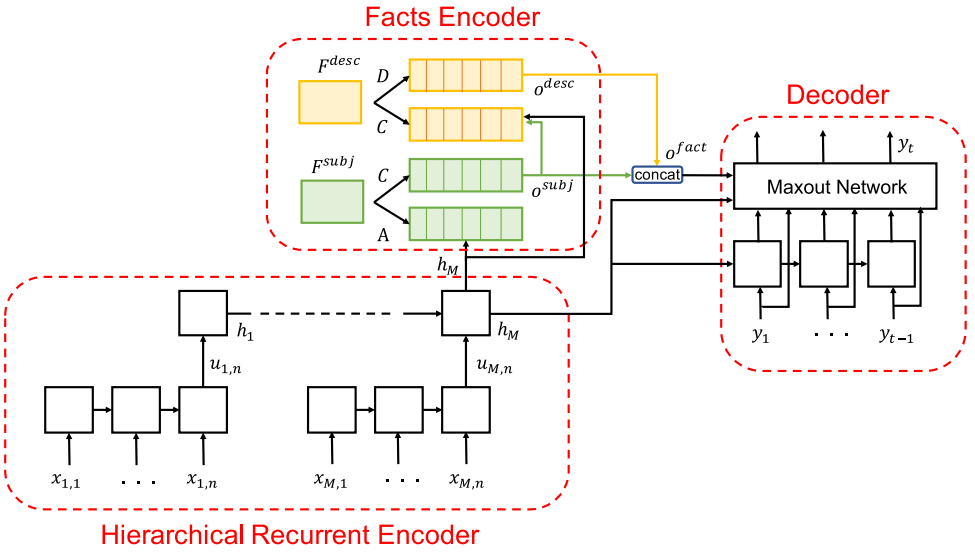

Fig. 3. An overview of Memory-augmented Hierarchical Recurrent Encoder-Decoder model.

After processing each utterance, a context encoder with the forward GRU finally outputs the context vector h_M that represents a summary of past utterances.

3.1.2. Facts encoder

The facts encoder is introduced to select facts that need to be injected in responses and map the facts into the continuous representation by utilizing the concept of the MemN2N architecture followed by Ghazvininejad et al. (2018). Our main idea is inspired from how humans read articles, where people often read the contents of articles after confirming main topics. Therefore, different from Ghazvininejad et al. (2018), instead of storing the facts into one memory, we store F^{subj} in the first memory and F^{desc} in the last memory.

We first convert F^{subj} and F^{desc} into memory vectors $r^{subj} = \{r_1^{subj}, ..., r_K^{subj}\}$ and $r^{desc} = \{r_1^{desc}, ..., r_L^{desc}\}$, respectively by sum of word embeddings. Subsequently, the context vector h_M computed from the hierarchical recurrent encoder is fed into the facts encoder in the first memory, and a subject facts vector o^{subj} is obtained in the following formulations:

$$m_k^{subj} = A r_k^{subj}, \tag{2}$$

$$c_k^{subj} = C r_k^{subj}, \tag{3}$$

$$p_k^{subj} = \mathrm{softmax}(h_M^{\mathrm{T}} m_k^{subj}), \tag{4}$$

$$o^{subj} = \sum_k^K p_k^{subj} c_k^{subj}, \tag{5}$$

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 a description facts vector o^{desc} is obtained as follows:

$$m_l^{desc} = C r_l^{desc}, \tag{6}$$

$$c_l^{desc} = D r_l^{desc}, \tag{7}$$

$$p_l^{desc} = \mathrm{softmax}\{(h_M + o^{subj})^{\mathrm{T}} m_l^{desc}\}, \tag{8}$$

$$o^{desc} = \sum_l^L p_l^{desc} c_l^{desc}, \tag{9}$$

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

3.2. Decoder

The decoder reads the results from the hierachical recurrent encoder and the facts encoder, and it predicts the next utterance word-by-word sequentially. The tth decoder hidden state s_t is computed by GRU where the initial hidden state s_0 is h_M and the initial word y_0 is $< \mathrm{SOS} >$ denoting the start of sentence.

$$s_t = \mathrm{GRU}\Big(s_{t-1}, e(y_{t-1})\Big). \tag{10}$$

For calculating a output word probability $p(y_t)$, we consider input features of the predictor as the context vector h_M, facts vector o^{fact}, previous predicted token embedding $e(y_{t-1})$, and current decoder hidden state s_t. These are thought to be useful, but it is not always necessary to leverage all information in each decoding step. For example, if the model generates a response like *Steve Jobs was CEO in Apple.*, general words (*was* and *in*) can be generated without facts. Zhou et al. (2017) proposed a selective encoder-decoder model that uses a maxout network (Goodfellow et al., 2013) to control the feature preference at each decoding step and performed good results in a text summarization task, and we also apply the maxout network to realize the function. The network outputs the vector with a maximum value v_t that can be computed with a linear transformation of input features. This vector represents the most salient feature among all features, and the decoder can switch the features at each decoding step. The probability $p(y_t)$ is estimated as follows:

$$z_t = W_e e(y_{t-1}) + W_h h_M + W_s s_t + W_o o^{fact}, \tag{11}$$

$$v_t = [\max\{z_{t,2j-1}, z_{t,2j}\}]^{\mathrm{T}} (j = 1, ..., d), \tag{12}$$

$$p(y_t) = \mathrm{softmax}(W_v v_t), \tag{13}$$

where $W_e, W_h, W_s, W_o \in \mathbb{R}^{2d \times d}$, and $W_v \in \mathbb{R}^{|V| \times d}$ are trainable parameters. For training the model, a negative log-likelihood is minimized with a teacher forcing.

3.2.1. Diverse sentence generation

Although most neural dialogue systems apply the beam search and the greedy search to generate the optimal response (Vinyals and Le, 2015; Serban et al., 2016; 2017b), many previous works have reported that thsese methods do not guarantee diversity for the final response (Vijayakumar et al., 2018; Fan et al., 2018; Holtzman et al., 2019). In order to enhance the diversity of responses, we apply a diverse beam search (Vijayakumar et al., 2018) that focuses on alleviating the low diversity problem. Let us consider a beam width B, groups G, and beam width in group $B' = B/G$; the beam sets at time step t are divided into G subsets. The word y_t^g is selected in order of $g = \{1, ..., G\}$ for these subsets as follows:

$$y_t^g = \underset{\{y_{1,t}^g, ..., y_{B',t}^g\}}{\operatorname{argmax}} \sum_{b \in B'} \Theta_t(y_{b,t}^g), + \lambda \Delta_{div} \tag{14}$$

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 $y_{b,t}^g$. Note that the penalty $\Delta_{div} = 0$ was set at $g = 1$.

3.3. Sentence selection with facts retrieval

In general, the raw human-human conversation tends to be highly fluent and rich in variety, and often contains a considerable amount of information about various topics. Inspired by information retrieval-based systems (Jr. et al., 2000; Ji et al., 2014), which search for a corresponding utterance that best matches the query in a pre-defined conversational repository, we develop a method to select responses focused on selecting informative responses based on the contextual relevant facts. As in Fig. 4, Facts Retrieval (FR) outputs responses using a query of overlapping facts words, and the responses contains informative words constantly.

First of all, the database is constructed in the form of a query-output pair $< [S; R], R >$, where $[S; R]$ is a concatenated sequence between a dialogue context S and a response R. As R is an unknown factor during the testing time, the database is constructed from only the training dialogue dataset given by DSTC7-Track2 and used during the testing time. For a sentence selection, a word-level matching method is applied. Specifically, we extract important words Q, which are overlapping words in F^{subj}, and feed them into the constructed database. In order to eliminate noises and improve the quality of retrieval, Q is restricted to certain Part Of Speech with noun, verb, adjective, and adverb. FR outputs R satisfying the relation $Q \in [S; R]$. Note that because we would like to keep responses high quality, FR will not output sentences continuously if the relation is unsatisfied. If multiple sentences satisfy this relation, the FR sorts responses using the score produced by BM25F (Zaragoza et al., 2004) and outputs at most 10 responses.

3.4. Reranker

The outputs of MHRED and FR may contain non-fluent and meaningless responses. Hence, these responses should be eliminated to improve the response quality. The Reranker sorts candidates by feeding all the response candidates from MHRED and FR, and the highest ranked candidate is returned as a final response. Specifically, it classifies whether a candidate is "positive" or "negative" as a response, and the probability of "positive" computed as a confidence score from a binary classification with XGBoost (Chen and Guestrin, 2016) is used to rerank. As shown in Table 1, the input features of Reranker can be summarized into three categories, "Response" (responses returned by both FR and MHRED), "Utterance" (pairs of previous utterance and response candidate), and "Context" (pairs of multi context and response candidate).

To reduce the spamming and offensive language while retaining the quality of responses, context-response pairs with a "response score" of 100 over were chosen as positive examples. The score is annotated by online users who can vote all

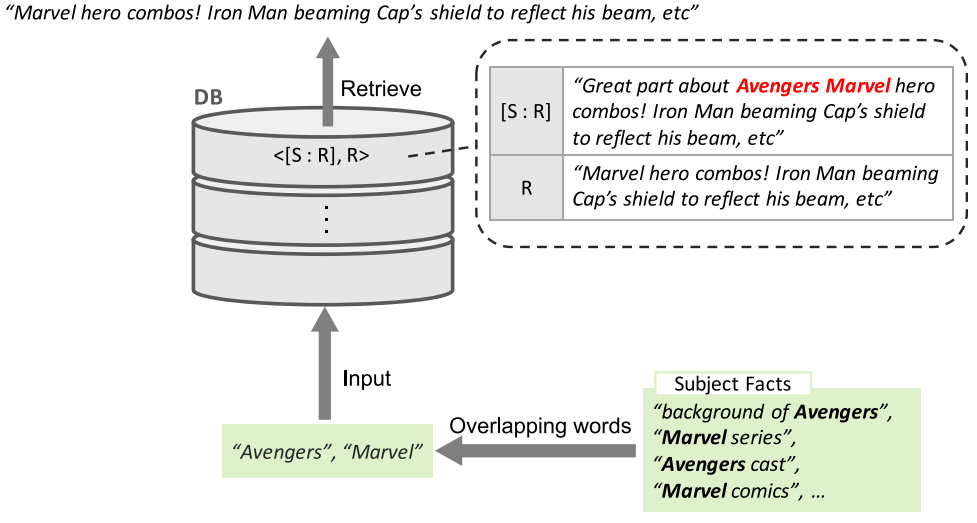

Fig. 4. An overview of facts retrieval.

Table 1
Features used to select a response on Reranker.

Categories	Features	About Features
Response	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
Utterance	Word sim	Cosine similarity between one-hot vectors of words
	N-gram sim	Cosine similarity between N-gram ($N=2,3$)
	Length sim	Similarity[a] of number of characters, and words
	Embedding sim	Cosine similarity between vectors computed as the averaged Word2Vec
	Sentimental sim	Similarity[b] of semantic orientations (Takamura et al., 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[c]
	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 context and response candidate into LDA model (Blei et al., 2003)

[a] 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 $1.0-||U|-|S||$.
[b] 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 $1.0-|(U_{sent}-S_{sent})/2|$.
[c] https://www.nltk.org/

conversation threads on reddit based on whether they liked or disliked the threads in the DSTC7 dataset. If our dialogue system generates artificial and non-fluent response candidates, the confidence score produced by the Reranker should be low. To obtain this behavior, we create negative examples based on both (1) the coherence between the context and response and (2) the quality of the response, and we randomly selected from one of the following rules:

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

The train and development context-response pairs are created from the DSTC7 dialogue dataset using the above rules contains 44449 and 5882 pairs, respectively.

4. Experiment

4.1. Datasets

We trained and evaluated the models on the DSTC7-Track2 dataset. The dialogue dataset contains 2.8M conversation instances divided into train (years 2011−2016), development (Jan−Mar 2017) and test (the rest of 2017). To simplify the learning task of our model, markdown and special symbols were eliminated. For the train set, since many conversations have multiple references for a certain context, the condition may be difficult for the model to learn easily. To alleviate this problem, we restricted context-response pairs with the highest response score for the same context. For the test set, given the filtering criteria such as turn length, this yield a 5-reference test set of size 2208. For each instance, we set aside one of the 6 human responses to assess human performance on this task.

The dataset also contains facts extracted from 226 types of web-sites such as Wikipedia. Two types of facts F^{subj} and F^{desc} were prepared at most $K, L=10$ snippets for a certain conversation according to the highest cosine similarity between sentence vectors that obtained from the averaged Word2Vec. Note that Word2Vec was trained on only the official training datasets according to DSTC-Track2 regulations. The statistical information of the dataset after pre-processing is shown in Table 2.

Table 2
Statics of dataset after pre-processing.

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 (subject)	3.86	3.61	3.30
Avg. Tokens/Sentence (description)	17.11	16.67	15.63
# Topics (subject)	27735	1152	3047
# Topics (description)	27645	1121	3063

4.2. Evaluation metrics

Automatic evaluation and both online and offline human evaluations were conducted for the generated responses. For the automatic evaluation, two types of metrics were used; one is word-overlap metrics including BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and the other is the diversity metric using div (Li et al., 2016). For the online human evaluation, the evaluation was conducted by DSTC7-Track2 organizers. Workers recruited from Amazon Mechanical Turk[1] evaluated 1000 samples that randomly selected from the test set rated with score 1 (Strong Disagree) to 5 (Strong Agree) for "Appropriateness" and "Informativeness". More detail is described in Yoshino et al. (2019).

Since the online evaluation provided by the organizers was limited to only one system per participant, we also conducted the offline human evaluation by ourselves to evaluate our models on the same conditions. Each of five workers who have studied dialogue systems evaluated 50 samples. Given a dialogue context and responses generated from model A and B, workers evaluated which response is more adequate. In addition to the evaluation metrics ("Appropriateness" and "Informativeness") used in the online evaluation, "Fluency" was also used to measure whether the responses were fluent and could have been plausibly produced by humans. If the dialogue context couldn't be understood or if it could not be determined whether the response was good or not, the conversation sample was evaluated as Tie.

4.3. Models for comparison

Several models were selected to be evaluated.

- S2S: The normal sequence-to-sequence model that does not have access to the external knowledge is the same as the architecture of the official neural baseline model (Vinyals and Le, 2015). The experimental conditions of this model are the same as our models (not official models) which allows us to discuss the effectiveness of the model architecture.
- kgS2S: Knowledge grounded sequence-to-sequence model (Ghazvininejad et al., 2018). For the facts encoder, the two-hop MemN2N is used in a way similar to MHRED.
- HRED: HRED model that exploits the dialogue context for response generation without any grounding (Serban et al., 2016).
- FR: Retrieval-based dialogue system explained in Section 3.3. Each response is selected with the highest BM25F score. As this model could not generate responses constantly so as to maintain the response quality, all the test sets could not be evaluated.
- MHRED: Our proposed hierarchical conversational model described in Section 3.1. Different from HRED, this model generates responses grounded not only on dialogue context but also on facts. To verify the effectiveness of dividing facts into subject ones and description ones, we also compared MHRED-concat, where the facts are concatenated (not separated) and fed into two memories.
- Ensemble: Our final model described in Section 3. Reranker selects the final response from candidates returned by MHRED and FR. The accuracy of Reranker achieved 73.2% on the development dataset.

Moreover, official baseline models including (1) random that randomly selects responses from the train set, (2) constant that always responds "I don't know what you mean" and (3) human that generates responses by humans, derived from the task organizers were compared in our evaluation. The official neural baseline model, that is the sequence-to-sequence model with the greedy decoding, was also provided from organizers. In order to be fair with discussing all models in terms of highly influential experimental conditions such as decoding techniques, hyper-parameters and the way of pre-processing, the model was not compared with other models.

4.4. Model setup

The encoder and decoder had a two-layer GRU with 256 hidden states, but these did not share parameters. During training, the batch size was set to 40 and Adam optimization (Kingma and Ba, 2014) with the initial learning rate of 0.0001 was employed. To alleviate over-fitting to the training dataset, a dropout with the dropout rate of 0.2 was applied. All the generation-based models were trained at most 20 epochs, and the model with the lowest perplexity in the development dataset was selected.

For inference, we applied the diverse beam search where the beam size B was 15 and the group size G was 15. The hyper-parameter of decoding λ was set as 0.4 according to BLEU on the development dataset. The vocabulary was obtained from both the dialogue and facts data, and the size was set to 20k. In order to avoid generating the special symbol $<\text{unk}>$ that denotes out-of-vocabulary, the log probability of the symbol was set to $-\infty$.

5. Experimental results

5.1. Automatic evaluation

Table 3 and Table 4 show results of the automatic evaluation on all the test set and the restricted test set, respectively. It can be seen that our final model (Ensemble) performs better than other models on almost all the evaluation metrics. This indicates that

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

Table 3
Results of the automatic evaluation in all the test set. The highest score are shown in bold.

	Model	NIST4	BLEU4	METEOR	div1	div2
Official	constant	0.184	2.87	7.48	0.000	0.000
	random	1.637	0.86	5.91	0.160	0.647
	human	2.650	3.13	8.31	0.167	0.670
Ours	S2S	0.023	0.34	3.92	0.026	0.161
	kgS2S	0.223	0.54	4.70	0.025	0.195
	HRED	0.730	0.58	5.65	0.049	0.309
	MHRED	0.645	0.60	5.48	0.069	**0.352**
	MHRED-concat	0.620	0.58	5.43	0.063	0.343
	Ensemble	**2.047**	**1.35**	**6.71**	**0.094**	0.334

Table 4
Results of the automatic evaluation in the part of test set (1148 pairs).

	Model	NIST4	BLEU4	METEOR	div1	div2
Official	constant	0.081	0.81	5.04	0.000	0.000
	random	–	–	–	–	–
	human	2.563	2.99	8.14	0.217	0.755
Ours	S2S	0.027	0.47	4.05	0.041	0.217
	kgS2S	0.215	0.67	4.77	0.038	0.250
	HRED	0.823	0.72	5.08	0.071	0.394
	MHRED	0.586	0.67	5.37	0.090	0.433
	MHRED-concat	0.599	0.58	5.41	0.083	0.418
	FR	1.847	1.23	5.72	0.115	0.300
	Ensemble	**2.263**	**1.48**	**6.90**	**0.146**	**0.467**

In order to be fair with the same condition of FR, we restricted and evaluated in test set pairs that FR could reply.
– indicates that outputs of the system were not be reproduced, and the evaluation could not be conducted.

Ensemble can output more fluent and diverse responses similar to humans. A comparison of the results from MHRED and kgS2S shows that our proposed architecture of MHRED outperforms the conventional model on the diversity score. This indicates that MHRED can more effectively capture topics using the hierarchical architecture. Moreover, MHRED also improves the performance in terms of all the diversity metrics on both the test sets when compared with the hierarchical neural model that does not access to the facts (HRED). Meanwhile, word-overlap metrics are not be improved consistently. Since the bottom line systems constant and random achieve high performance (particularly constant achieves the best BLEU4 among all competitive models in this task), they may not be reliable on evaluating dialogue systems as reported in the DSTC7 organizer (Yoshino et al., 2019). A comparison between the variants of MHRED on both datasets shows that MHRED achieves better performance on all the diversity score. This indicates that dividing the facts into two types is more effective than identifying the facts as one type.

The results of both MHRED and FR show that MHRED outperforms on div2 and FR performs well on all the word-overlap metrics. This indicates that each system has advantage of either response diversity or appropriateness. According to the results of Ensemble, combining MHRED and FR using Reranker results in making most of the advantages of the generation- and retrieval-based methods without any deterioration in performance.

5.2. Human evaluation

Table 5 shows the results of online human evaluation. Our proposed system Ensemble beats the official simple baseline models on all the evaluation metrics. However, in terms of the quality in generating responses, there is a huge gap between humans and our model, and more work is required to be carried out in future work. In order to compare with our models, we also conducted offline human evaluation by ourselves as described in Section 4.2. Table 6 shows that Ensemble has the strongest performance on all the metrics. Moreover, MHRED performs better than the conventional models including HRED and kgS2S. Most surprisingly, compared with kgS2S, MHRED significantly improves the performance on informativeness.

6. Analysis

To verify the effectiveness of the MHRED architecture, we looked into the attention weights p^{subj}, p^{desc} in the facts encoder. Fig. 5 depicts an example of attention paid by the fact encoder. F^{subj} captures "*seven wonders of the ancient world*" that refers to the main topic of conversation. Subsequently, F^{desc} captures the facts that contain "*pyramid*" considering both the dialogue

Table 5
Results of the online human evaluation.

	Model	Appropriateness	informativeness
	constant	2.60	2.32
Official	random	2.32	2.35
	human	3.61	3.49
Ours	Ensemble	2.69	2.58

Table 6
Results of the offline human evaluation.

			Fluency	Appropriateness	Informativeness
MHRED	vs	S2S	**55.6** ± **3.4**, 19.6 ± 3.2, 24.8 ± 2.4	**49.2** ± **4.1**, 31.6 ± 3.9, 19.2 ± 1.3	**56.0** ± **4.6**, 36.8 ± 4.34, 7.2 ± 0.7
MHRED	vs	HRED	**48.0** ± **6.0**, 26.0 ± 6.9, 26.0 ± 2.0	**44.4** ± **3.3**, 28.8 ± 2.0, 26.8 ± 4.3	**40.4** ± **0.9**, 32.0 ± 2.3, 27.6 ± 1.8
MHRED	vs	kgS2S	**44.8** ± **3.0**, 15.6 ± 4.1, 39.6 ± 3.2	**41.6** ± **2.2**, 29.2 ± 3.6, 29.2 ± 2.0	**47.6** ± **1.2**, 39.2 ± 3.8, 13.2 ± 2.6
MHRED	vs	FR	39.2 ± 2.6, 12.0 ± 2.9, **48.8** ± **2.3**	32.4 ± 2.5, 28.0 ± 2.7, **39.6** ± **3.1**	20.0 ± 1.8, 28.4 ± 4.5, **51.6** ± **2.9**
Ensemble	vs	FR	**40.0** ± **2.7**, 23.2 ± 4.6, 36.8 ± 2.8	**40.8** ± **2.7**, 30.8 ± 4.1, 28.4 ± 2.1	**36.8** ± **2.7**, 28.8 ± 5.8, 34.4 ± 3.6
Ensemble	vs	MHRED	**53.2** ± **3.9**, 16.0 ± 3.8, 30.8 ± 1.7	**44.4** ± **2.9**, 27.2 ± 3.6, 28.4 ± 2.8	**37.2** ± **3.1**, 45.2 ± 3.6, 17.6 ± 1.4

In each evaluation metric, it shows the percentage of Win, Tie and Lose to a compared model from left to right, together with 95% confidence intervals.

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

Fig. 5. Attention weights produced by the facts encoder. Sentences painted in darker shades represent greater attention.

context and F^{subj}. Finally, MHRED generates a response including "*pyramids*". This indicates that our model can focus on the facts relevant to the dialogue context and generate responses injecting them.

Table 7 shows how the accuracy of Reranker is affected when one of the target features is excluded. The category "Response" shows the largest decrease among all the categories. Moreover, in the features, "Fluency" shows the largest decrease, followed by "Keyword sim". These indicate that Reranker has a tendency to select the final response that focuses on fluency and contextually informativeness in the dialogue. One of the most likely causes of this tendency is the way of creating the training dataset for Reranker. Specifically, negative examples were created using hand-crafted rules such as swapping and eliminating words, resulting in sensitivity to "Fluency" and "Keyword sim".

Table 8 shows examples of responses. As can be observed from the table, Ensemble outputs informative words related to the dialogue context such as "*lake sperior*" (#1) and "*george*" (#2). Table 9 shows examples of reranking by Reranker. In example #1, MHRED is explicitly designed for the previous context, and Reranker selects the most meaningful response. In example #2, the response returned by FR has high fluency and many content words. Conversely, the response is not suitable for the dialogue

Table 7
Differences of accuracy computed by Reranker when the target feature is excluded in the development dataset.

Category	Feature	Difference	
		Category	Feature
Response	Length		+0.0004
	Fluency		**−0.1053**
	POS		−0.0038
	Facts	**−0.1039**	−0.0021
Utterance	Word sim		−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	−0.0155	± 0.0000
	Keyword sim		**−0.0203**
Context	Topic sim	−0.0015	−0.0015

The lower score is, the more important feature is.

Table 8
Example of a response.

#	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.	HRED: we are there. i don't really count, but its not. kgS2S: That makes sense, and i know. 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 seinfeld wasn't actually a "show about nothing," it was a show about how comedians get their material. → they never really had an episode where someone read a book to themselves for 22 minutes. that, would be a show "about nothing"	HRED: and then had one episode and it is that was it was the show. kgS2S: thats what a great ! Ensemble(MHRED): and georges show up episode aired in a show.

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

Table 9
Example of reranking by 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

The token → indicates a change of turn. CS denotes the confidence score produced by XGBoost.

context in terms of the topic. This indicates, as above mentioned, that Reranker tends to focus on "Response" strongly because of the way of creating examples for Reranker. We expect that creating examples from various perspectives will improve the performance further, and this would be the future work.

7. Related work

The popularization of social networking services offers the advantage of reducing the burden of building large-scale open datasets, and end-to-end conversational models have been the popular approaches. These models can be further classified into two broad categories; generation-based models and retrieval-based models.

Generation-based models are developed based on sequence to sequence learning problem (Vinyals and Le, 2015; Serban et al., 2016; 2017b). Various methods proposed to improve the content of diversity promotion by introducing Maximum Mutual

Information as the objective function (Li et al., 2016) and unknown words handling (Gu et al., 2016; See et al., 2017). However, there are still problems in tending to generate non-informative responses since they rely on only dialogue pairs as training set. Incorporating external information sources based on real-world facts has shown to be effective for solving this problem. Several studies incorporated structured formats such as knowledge-graph (Zhou et al., 2018) and knowledge-base (Han et al., 2015; Xu et al., 2017), and unstructured text such as web sites (Yoshino et al., 2019), posts (Ghazvininejad et al., 2018), and crowdsourcing (Zhang et al., 2018). However, most of existing systems have focused on single-turn conversation and there have difficulties of understanding the dialogue context and extracting appropriate knowledge, which causes a non-informative conversation. Our approach tackles this problem to model both muti-turn context and facts based on the idea of (Ghazvininejad et al., 2018) and (Serban et al., 2016).

Retrieval-based models respond by selecting the utterance that best matches an input from a pre-defined response set (Jr. et al., 2000; Ji et al., 2014), and the resulting responses tend to be highly fluent.

Making most of the advantages of the two methods (generation-based and retrieval-based) is hypothesized to improve the response quality (Serban et al., 2017a; Song et al., 2018; Pandey et al., 2018). In contrast to these previous works, we attempt to combine generation-based and retrieval-based models that are aware of both dialogue context and knowledge.

8. Conclusion

In this study, we proposed a context- and knowledge-aware neural conversational model (MHRED) for grounded response generation. Furthermore, we applied a system combination technique for improving the response quality. Our final system consists of three modules: MHRED, a retrieval-based module with Facts Retrieval (FR), and a reranking module. When the response candidates are generated from MHRED, a diverse beam search is used to promote the diversity of responses. The experimental results on DSTC7-Track2 dataset showed that MHRED can generate more diverse and informative responses than generation-based baselines. Moreover, we confirmed that FR's responses have the most informativeness among all single models, and the combination of multiple modules significantly improved overall automatic and human evaluation metrics. Future work will be focused on introducing an end-to-end learning for multiple systems simultaneously.

References

Banerjee, S., Lavie, A., 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL 2015 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. https://aclanthology.info/papers/W05-0909/w05-0909.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022. http://portal.acm.org/citation.cfm?id=944937.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the SIGKDD 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR https://arxiv.org/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. Morgan Kaufmann Publishers Inc., pp. 138–145. http://dl.acm.org/citation.cfm?id=1289189.1289273.

Fan, A., Lewis, M., Dauphin, Y.N., 2018. Hierarchical neural story generation. In: Proceedings of the ACL 2018, pp. 889–898. https://doi.org/10.18653/v1/P18-1082. https://www.aclweb.org/anthology/P18-1082/.

Ghazvininejad, M., Brockett, C., Chang, M., Dolan, B., Gao, J., Yih, W., Galley, M., 2018. A knowledge-grounded neural conversation model. In: Proceedings of the AAAI 2018, pp. 5110–5117. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710.

Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A.C., Bengio, Y., 2013. Maxout networks. In: Proceedings of the ICML 2013, 23, pp. 1319–1327. http://proceedings.mlr.press/v28/goodfellow13.html.

Gu, J., Lu, Z., Li, H., Li, V.O.K., 2016. Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the ACL 2016, pp. 1631–1640. http://aclweb.org/anthology/P/P16/P16-1154.pdf.

Han, S., Bang, J., Ryu, S., Lee, G.G., 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In: Proceedings of the SIGDIAL 2015, pp. 129–133. http://aclweb.org/anthology/W/W15/W15-4616.pdf.

Holtzman, A., Buys, J., Forbes, M., Choi, Y., 2019. The curious case of neural text degeneration. CoRR. 1904.09751.

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., Kim, S., 2019. Overview of the sixth dialog system technology challenge: DSTC6. Comput. Speech Lang. 55, 1–25. https://doi.org/10.1016/j.csl.2018.09.004.

Ji, Z., Lu, Z., Li, H., 2014. An information retrieval approach to short text conversation. CoRR https://arxiv.org/abs/1408.6988.

Jr., C.L.I., Kearns, M.J., Kormann, D.P., Singh, S.P., Stone, P., 2000. Cobot in lambdamoo: a social statistics agent. In: Proceedings of the AAAI 2000, pp. 36–41. http://www.aaai.org/Library/AAAI/2000/aaai00-006.php.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR https://arxiv.org/abs/1412.6980.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2016. A diversity-promoting objective function for neural conversation models. In: Proceedings of the NAACL-HLT 2016, pp. 110–119. http://aclweb.org/anthology/N/N16/N16-1014.pdf.

Pandey, G., Contractor, D., Kumar, V., Joshi, S., 2018. Exemplar encoder-decoder for neural conversation generation. In: Proceedings of the ACL 2018, pp. 1329–1338. https://aclanthology.info/papers/P18-1123/p18-1123.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the ACL 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135.

Rose, S., Engel, D., Cramer, N., Cowley, W., 2010. Automatic keyword extraction from individual documents. Text Min. Appl. Theory 1–20. https://doi.org/10.1002/9780470689646.ch1.

See, A., Liu, P.J., Manning, C.D., 2017. Get to the point: Summarization with pointer-generator networks. ACL 2017, pp. 1073–1083. https://doi.org/10.18653/v1/P17-1099. https://www.aclweb.org/anthology/P17-1099.

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., Bengio, Y., 2017. A deep reinforcement learning chatbot. CoRR https://arxiv.org/abs/1709.02349.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J., 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the AAAI 2016, pp. 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957.

Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A.C., Bengio, Y., 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the AAAI 2017, pp. 3295–3301. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567.

Song, Y., Li, C., Nie, J., Zhang, M., Zhao, D., Yan, R., 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. In: Proceedings of the IJCAI 2018, pp. 4382–4388. https://doi.org/10.24963/ijcai.2018/609.

Takamura, H., Inui, T., Manabu, O., 2005. Extracting semantic orientations of words using spin model. In: Proceedings of the ACL 2005, pp. 133–140. http://aclweb.org/anthology/P/P05/P05-1017.pdf.

Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D., 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In: Proceedings of the ACL 2017, pp. 231–236. https://doi.org/10.18653/v1/P17-2036.

Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D.J., Batra, D., 2018. Diverse beam search for improved description of complex scenes. In: Proceedings of the AAAI 2018, pp. 7371–7379. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17329.

Vinyals, O., Le, Q., 2015. A neual conversational model. In: Proceedings of the ICML, Deep Learning Workshop.

Xu, Z., Liu, B., Wang, B., Sun, C., Wang, X., 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In: Proceedings of the IJCNN 2017, pp. 3506–3513. https://doi.org/10.1109/IJCNN.2017.7966297. https://ieeexplore.ieee.org/document/7966297.

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., Batra, D., 2019. The 7th dialog system technology challenge. CoRR https://arxiv.org/abs/1901.03461.

Zaragoza, H., Craswell, N., Taylor, M.J., Saria, S., Robertson, S.E., 2004. Microsoft cambridge at TREC 13: web and hard tracks. In: Proceedings of the TREC 2004. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.965.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., 2018. Personalizing dialogue agents: i have a dog, do you have pets too? In: Proceedings of the ACL 2018, pp. 2204–2213. https://doi.org/10.18653/v1/P18-1205. https://www.aclweb.org/anthology/P18-1205.

Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X., 2018. Commonsense knowledge aware conversation generation with graph attention. In: Proceedings of the IJCAI 2018, pp. 4623–4629. https://doi.org/10.24963/ijcai.2018/643.

Zhou, Q., Yang, N., Wei, F., Zhou, M., 2017. Selective encoding for abstractive sentence summarization. In: Proceedings of the ACL 2017, pp. 1095–1104. https://doi.org/10.18653/v1/P17-1101.