Augmenting Neural Response Generation with Context-Aware Topical Attention

Nouha Dziri Ehsan Kamalloo Kory W. Mathewson Osmar Zaiane

Department of Computing Science
University of Alberta
{dziri,kamalloo,korym,zaiane}@cs.ualberta.ca

Abstract

Sequence-to-Sequence (Seq2Seq) models have witnessed a notable success in generating natural conversational exchanges. Notwithstanding the syntactically well-formed responses generated by these neural network models, they are prone to be acontextual, short and generic. In this work, we introduce a Topical Hierarchical Recurrent Encoder Decoder (THRED), a novel, fully data-driven, multi-turn response generation system intended to produce contextual and topic-aware responses. Our model is built upon the basic Seq2Seq model by augmenting it with a hierarchical joint attention mechanism that incorporates topical concepts and previous interactions into the response generation. To train our model, we provide a clean and high-quality conversational dataset mined from Reddit comments. We evaluate THRED on two novel automated metrics, dubbed Semantic Similarity and Response Echo Index, as well as with human evaluation. Our experiments demonstrate that the proposed model is able to generate more diverse and contextually relevant responses compared to the strong baselines.

1 Introduction

With the recent success of deep neural networks in natural language processing tasks such as machine translation (Sutskever et al., 2014) and language modeling (Mikolov et al., 2010), there has been growing research interest in building data-driven dialogue systems. Fortunately, innovation in deep learning architectures and the availability of large public datasets have produced fertile ground for the data-driven approaches to become feasible and quite promising. In particular, the Sequence-to-Sequence (Seq2Seq) neural network model (Sutskever et al., 2014) has witnessed substantial breakthroughs in improving the performance of conversational agents. Such a model succeeds in learning the backbone of the conversation but lacks any aptitude for producing context-sensitive and diverse conversations. Instead, generated responses are dull, short and carry little information (Li et al., 2016a). Instinctively, humans tend to adapt conversations to their interlocutor not only by looking at the last utterance but also by considering information and concepts covered in the conversation history (Danescu-Niculescu-Mizil and Lee, 2011). Such adaptation increase the smoothness and engagement of the generated responses. We speculate that incorporating conversation history and topic information with our novel model and method will improve generated conversational responses. In this work, we introduce a novel, fully data-driven, multi-turn response generation system intended to produce context-aware and diverse responses. Our model builds upon the basic Seq2Seq model by combining conversational data and external knowledge information trained through a hierarchical joint attention neural model. We find that our method leads to both diverse and contextual responses compared to the literature strong baselines. We also introduce two novel quantitative metrics for dialogue model development, dubbed Semantic Similarity and Response Echo Index. While the former measures the capability of the model to be consistent with the context and to maintain the topic of the conversation, the latter assesses how much our approach is able to generate unique and plausible responses which are measurably distant from the input dataset. Used together, they provide a means to reduce burden of human evaluation and allow rapid testing of dialogue models. We show that such metrics correlate well with human judgment, making a step towards a good automatic evaluation procedure.

The key contributions of this work are:

Proceedings of the 1st Workshop on NLP for Conversational AI, pages 18–31
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

- We devise a fully data-driven neural conversational model that leverages conversation history and topic information in the response generation process through a hierarchical joint attention mechanism; making the dialogue more diverse and engaging.

- We introduce two novel automated metrics: Semantic Similarity and Response Echo Index and we show that they correlate well with human judgment.

- We collect, parse and clean a conversational dataset from Reddit comments[1].

2 Related Work

Neural generative models have been improved through several techniques. (Serban et al., 2016) built upon the Seq2Seq work by introducing a Hierarchical Recurrent Encoder-Decoder neural network (HRED) that accounts for the conversation history. (Li et al., 2016b) used deep reinforcement learning to generate highly-rewarded responses by considering three dialogue properties: ease of answering, informativeness and coherence. (Zhang et al., 2018) addressed the challenge of personalizing the chatbot by modeling human-like behaviour. They presented a persona-based model that aims to handle the speaker consistency by integrating a speaker profile vector representation into the the Seq2Seq model. (Xing et al., 2017) used a similar idea but added an extra probability value in the decoder to bias the overall distribution towards leveraging topic words in the generated responses. Their architecture does not focus on capturing conversation history. All of these improvements are motivated by the scarcity of diversity and informativeness of the responses. Our work follows on from these works with the additional aim of generating context-aware responses by using a hierarchical joint attention model. An important line of research that we also address in this work is automatically evaluating the quality of dialogue responses. In dialogue systems, automated metrics tend to be borrowed from other NLP tasks such as BLEU (Papineni et al., 2002) from machine translation and ROUGE (Lin, 2004) from text summarization. Yet, such metrics fail, mainly because they are focusing on the word-level overlap between the machine-generated answer and the human-generated answer, which can be inconsistent with what humans deem a plausible and interesting response. (Liu et al., 2016) have showed that these metrics correlate very weakly with human evaluation. Indeed, word-overlapping metrics achieve best results when the space of responses is small and lexically overlapping which is not the case for dialogue systems responses. Significant works have looked into this challenge. Examples include ADEM (Lowe et al., 2017), an evaluation model that learns to score responses from an annotated dataset of human responses scores. (Venkatesh et al., 2018) proposed a number of metrics based on user experience, coherence, and topical diversity and have showed that these metrics can be used as a proxy for human evaluation. However, engagement and coherence metrics are estimated via recruiting evaluators. In this work, we propose directly calculable approximations of human evaluation grounded in conversational theories of accommodation and affordance (Danescu-Niculescu-Mizil and Lee, 2011).

3 Topical Hierarchical Recurrent Encoder Decoder

Topical Hierarchical Recurrent Encoder Decoder (THRED) can be viewed as a hybrid model that conditions the response generation on conversation history captured from previous utterances and on topic words acquired from a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). The proposed approach extends the standard Seq2Seq model by leveraging topic words in the process of response generation and accounting for conversation history. Figure 1 illustrates our model. We detail below the components of our model.

3.1 Message Encoder

Let a sequence of N utterances within a dialogue $D = \{U_1, ..., U_N\}$. Every utterance $U_i = \{w_{i,1}, ..., w_{i,L_i}\}$ contains a random variable L_i of sequence of words where $w_{i,k}$ represents the word embedding vector at position k in the utterance U_i. The message encoder sequentially accepts the embedding of each word in the input message U_i and updates its hidden state at every time step t by a bidirectional GRU-RNN (Cho et al., 2014) according to:

$$h_{i,t} = GRU(h_{i,t-1}, w_{i,t}), \forall t \in \{1, \ldots, L_i\} \quad (1)$$

where $h_{i,t-1}$ represents the previous hidden state.

[1] The source code and the dataset are available at https://github.com/nouhadziri/THRED

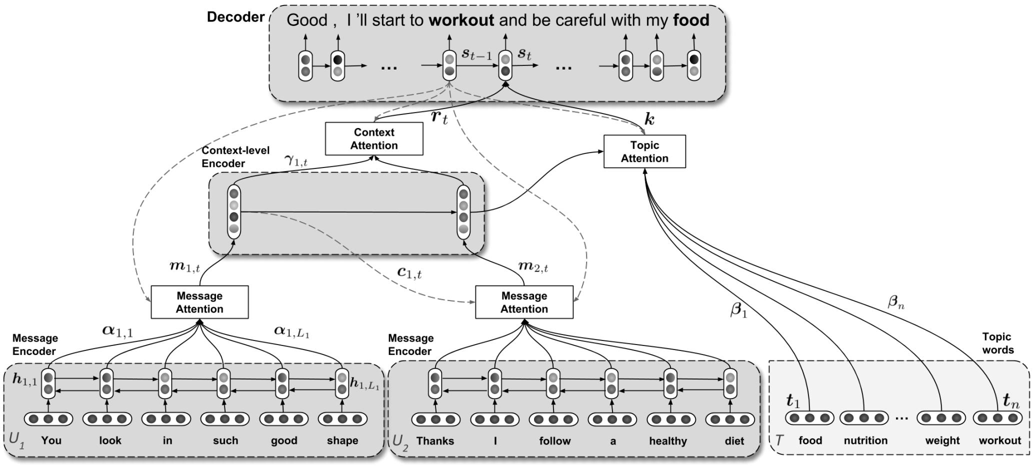


Figure 1: THRED model architecture in which we jointly model two specifications that presumably make the task of response generation successful: context-awareness (modeled by **Context Attention**) and diversity (modeled by **Topic Attention**).

3.2 Message Attention

Different parts of the conversation history have distinct levels of importance that may influence the response generation process. The message attention in THRED operates by putting more focus on the salient input words with regard to the output. It computes, at step t, a weight value $\alpha_{i,j,t}$ for every encoder hidden state $h_{i,j}$ and linearly combines them to form a vector $m_{i,t}$ according to Bahdanau attention mechanism (Bahdanau et al., 2015). Formally, $m_{i,t}$ is calculated as:

$$m_{i,t} = \sum_{j=1}^{L_i} \alpha_{i,j,t}\, h_{i,j},\ \forall i \in \{1, \ldots, N\} \qquad (2)$$

where $\alpha_{i,j,t}$ is computed as:

$$\alpha_{i,j,t} = \frac{\exp(e_{i,j,t})}{\sum_{k=1}^{L_i} \exp(e_{i,k,t})} \,;$$
$$e_{i,j,t} = \eta(s_{t-1}, h_{i,j}, c_{i-1,t})$$

where s_{t-1} represents the hidden state of the decoder (further details are provided later), $c_{i,t}$ delineates the hidden state of the context-level encoder (computed in Equation (3)) , η is a multi-layer perceptron having tanh as activation function. Unlike the Bahdanau attention mechanism, the attentional vector $m_{i,t}$ is based on both the hidden states of the decoder and the hidden states of the context-level encoder. We are motivated by the fact that $c_{i,t}$ may carry important information that could be missing in s_{t-1}. In summary, the attentional vector $m_{i,t}$ is an order-sensitive information of all the words in the sentence, attending to more important words in the input messages.

3.3 Context-Level Encoder

The context-level encoder takes as input each utterance representation ($m_{1,t}$, ..., $m_{N,t}$) and calculates the sequence of recurrent hidden states as shown in Equation (3):

$$c_{i,t} = GRU(c_{i-1,t}, m_{i,t}), \forall i \in \{1, \ldots, N\} \quad (3)$$

where $c_{i-1,t}$ delineates the previous hidden state of the context-level encoder and N represents the number of utterances in the conversation history. The resulted $c_{i,t}$ vector summarizes all past information that have been processed up to position i.

3.4 Context-Topic Joint Attention

Context Attention: On top of the context-level encoder, a context attention is added to attend to important utterances in the conversation history. Precisely, the context attention assigns weights $(\gamma_{1,t}, ..., \gamma_{N,t})$ to $(c_{1,t}, ..., c_{N,t})$ and forms a vector r_t as

$$r_t = \sum_{j=1}^{N} \gamma_{j,t} c_{j,t} \qquad (4)$$

where:

$$\gamma_{j,t} = \frac{\exp(e'_{j,t})}{\sum_{i=1}^{N} \exp(e'_{i,t})} \,; \qquad (5)$$
$$e'_{i,t} = \eta(s_{t-1}, c_{i,t})$$

Topic Attention: In order to infuse the response with information relevant to the input messages, we enhance the model with topic information. We assign a topic T to the conversation context using a pre-trained LDA model (Hoffman et al., 2010). LDA is a probabilistic topic model that appoints multiple topics for the dialogue history. The LDA parameters were estimated using the collapsed Gibbs sampling algorithm (Zhao et al., 2011). We provide further details on how we train this model in the supplementary material. In our case, the conversation history is a short document, so we believe that the most probable topic will be sufficient to model the dialogue. After acquiring topic words for the entire history, we pick the n highest probable words under T (we choose $n = 100$ in our experiments). The topic words $\{t_1, \cdots, t_n\}$ are then linearly combined to form a fixed-length vector k. The weight values are calculated as the following:

$$\beta_{i,t} = \frac{\exp(\eta(s_{t-1}, t_i, c_{N,t}))}{\sum_{j=1}^{n} \exp(\eta(s_{t-1}, t_j, c_{N,t}))} \tag{6}$$

where $i \in \{1, \cdots, n\}$, $c_{N,t}$ is the last hidden state of the context-level encoder, and s_{t-1} is the $t-1^{th}$ hidden state in the decoder. The topic attention uses additionally the last hidden state of the context-level encoder $c_{N,t}$ in order to diminish the repercussion of impertinent topic words and feature the relevant ones to the message. Unlike (Xing et al., 2017), our model employs the final context-level encoder hidden state $c_{N,t}$ in order to account for conversation history in the generated response. In summary, the topic words are summarized as a topic vector k representing prior knowledge for response generation. The key idea of this approach is to affect the generation process by avoiding the need to learn the same conversational pattern for each utterance but instead enriching the responses with topics and words related to the subject of the message even if the words were never used before.

3.5 Decoder

The decoder is responsible for predicting the response utterance U_{m+1} given the previous utterances and the topic words. Following (Xing et al., 2017), we biased the generation probability towards generating the topic words in the response. In particular, we added an extra probability to the standard generation probability, enforcing the model to account for the topical tokens. Consequently, the generation probability is defined as the following:

$$p(w_i) = p_V(w_i) + p_K(w_i) \tag{7}$$

where K and V represent respectively topic vocabulary and response vocabulary; p_V and p_K are defined as follows:

$$p_V(w_i) = \frac{1}{M} \exp(\sigma_V(s_i, w_{i-1}))$$

$$p_K(w_i) = \frac{1}{M} \exp(\sigma_K(s_i, w_{i-1}, r_i))$$

where $s_i = GRU(w_{i-1}, s_{i-1}, r_i, k)$, σ is a tanh and M is calculated as follows:

$$M = \sum_{v \in V} \exp\left(\sigma_V(s_i, w_{i-1})\right) + \sum_{v' \in K} \exp\left(\sigma_K(s_i, w_{i-1}, r_i)\right)$$

4 Datasets

One of the main weaknesses of dialogue systems is caused by the paucity of high-quality conversational dataset. The well-known OpenSubtitles dataset (Tiedemann, 2012) lacks speaker annotations, thus making it more difficult to train conversation systems which demand high quality speaker and conversation level tags. Therefore, the assumption of treating consecutive utterances as turn exchanges uttered by two persons (Vinyals and Le, 2015) could not be viable. To enable the study of high-quality and large-scale dataset for dialogue modeling, we have collected a corpus of 35M conversations drawn from the Reddit data[2], where each dialogue is composed of three turn exchanges. The Reddit dataset is composed of posts and comments, where each comment is annotated with rich meta data (i.e., author, number of replies, user's comment karma, etc.)[3]. To harvest the dataset, we curated 95 English subreddits out of roughly 1.1M public subreddits[4]. Our choice was based on the top-ranked subreddits that discuss topics such as news, education, business, politics and sports. We processed Reddit for a 12 month-period ranging from December 2016 until December 2017. For each post, we retrieved all comments and we recursively followed the chain of replies of each comment to recover the entire conversation. Reddit dataset is often semantically well-structured and is not filled with spelling errors thanks to moderator's efforts. Therefore, we do not perform any spelling correction procedure. Due to resource limitations, we randomly sampled 6M dialogues as training data, 700K dialogues as development data, and 40K dialogues as test data.

[2] https://files.pushshift.io/reddit/

[3] https://github.com/reddit-archive/reddit/wiki/JSON

[4] As of February 2019

For OpenSubtitles, we trained the models on the same size of data as for Reddit.

5 Experiments

In this section, we focus on the task of evaluating the next utterance given the conversation history. We compare THRED against three open-source baselines, namely Standard Seq2Seq with attention mechanism (Bahdanau et al., 2015), HRED (Serban et al., 2016), and Topic-Aware (TA) Seq2Seq (Xing et al., 2017). As done in (Li et al., 2016b), for Standard Seq2Seq and TA-Seq2Seq, we concatenate the dialogue history to account for context in a multi-turn conversation. All experiments are conducted on two datasets (i.e., Reddit and OpenSubtitles). We report results on OpenSubtitles in the supplementary material.

5.1 Quantitative Evaluation

In the following subsections, we introduce two metrics that can impartially evaluate THRED and compare against the different baselines. These metrics were tested on 5000 dialogues randomly sampled from the test dataset. It is worth mentioning that we present word perplexity (PPL) on the test data in Table 4 (along with diversity metric). However, we do not believe that it represents a good measure for assessing the quality of responses (Serban et al., 2017). This is because perplexity captures how likely the responses are under a generation probability distribution, and does not measure the degree of diversity and engagingness in the responses.

5.2 Semantic Similarity

A good dialogue system should be capable of sustaining a coherent conversation with a human by staying on topic and by following a train of thoughts (Venkatesh et al., 2018). Semantic Similarity (SS) metric estimates the correspondence between the utterances in the context and the generated response. The intuition behind this metric is that plausible responses should be consistent with the context and should maintain the topic of the conversation. Our response generator THRED along with the baselines generate an utterance based on the two previous utterances in the dialogue (i.e., Utt1 and Utt2). We compute the cosine distance between the embedding vectors of the test utterances (Utt.1 and Utt.2) and the generated responses from the different models (i.e., THRED, TA-Seq2Seq, HRED and Seq2Seq). Therefore, a low score denotes a high coherence. More precisely, for each triple in the test dataset, we test two scenarios: (1) we compute the SS of each generated response with respect to the most recent utterance in the conversation (Utt.2) and (2) we compute the SS of each generated response with respect to the second most recent utterance (Utt.1). To render the semantic representation of an utterance, we leverage Universal Sentence Encoder (Cer et al., 2018) wherein a sentence is projected to a fixed dimensional embedding vector.

However, dull and generic responses such as "*i'm not sure*" tend to be semantically close to many utterances, hindering the effectiveness of the metric. To cope with this negative effect, we manually compiled a set of 60 dull responses and computed the SS score by multiplying the cosine distance with the following penalty factor (akin to length penalty in (Wu et al., 2016)):

$$P = 1 + \log \frac{2 + L'}{2 + L''}$$

where L' indicates the length of the response after dropping stop words and punctuation and L'' stands for the length of non-dull part of the response after dropping stop words. The intuition here is that the longer utterances, with nearly the same similarity, communicate the intention unequivocally since it takes more words to convey the same meaning.

The penalized Semantic Similarity score is therefore defined as:

$SS(utt_{i,j}\ ,\ resp_i) = P \times (1 - cos(\vec{utt_{i,j}}, \vec{resp_i}))$

where i represents the index for the dialogue in the test dataset and j denotes the index of the utterance in the conversation history. The results conducted on Reddit dataset are shown in Table 2. We can observe that THRED is able to generate responses which follow the topic and semantics of the input utterances. In particular, the responses generated by THRED tend to be closer to the context of the conversation (Utt.1 and Utt.2) compared to the responses generated from the baslines. To ensure the statistical significance of THRED, we conducted Student's t-test over the average values of SS metric. THRED outperforms all baselines ($p < 0.001$) especially when the comparison is made against the most recent utterance (Utt.2). On the other hand, THRED is level with compared models in semantic distance with respect to the second most recent utterance (Utt.1). This makes

CONTEXT	GENERATED RESPONSES
(Reddit) sanctions are an act of war → why do you think that ?	**THRED:** because it's really a theory that supports terrorism . and this has an effect on the idea of a regime that isn't the same as a government (*Excellent, Good, Excellent, Good, Excellent*) **HRED:** because the war is n't a war . it 's a war . (*Good, Poor, Poor, Poor, Poor*) **Seq2Seq:** because it 's an unpopular opinion , and that 's why it 's a bad thing to say . (*Good, Poor, Excellent, Good, Good*) **TA-Seq2Seq:** because it's a war . (*Good, Poor, Excellent, Poor, Good*)

Table 1: One cherry-picked dialogue out of 150 conversations along with the generated responses from all models. Human judgments are provided in the brackets. The blue arrow specifies a dialogue exchange and the highlighted words in red represent the topic words acquired from the pre-trained LDA model.

Stat.	**THRED**	**Seq2Seq**	**HRED**	**TA-Seq2Seq**
SS with respect to Utt.1				
μ	**0.680**	0.694	0.755	0.692
σ	0.200	0.236	0.283	0.252
SS with respect to Utt.2				
μ	**0.649****	0.672	0.720	0.702
σ	0.212	0.236	0.292	0.253

Table 2: Mean μ and standard deviation σ of SS scores for the responses generated from different models with respect to the most recent utterance (Utt.2) and the second most recent utterance (Utt.1) from conversation history on the Reddit test dataset (** indicates statistical significance over the second best method with p-value < 0.001).

sense because in a multi-turn dialogue, speakers are more likely to address the last utterance spoken by the interlocutor, which is why THRED tends to favour the most recent utterance over an older one. Additionally, the roughly similar distances for both utterances in Standard Seq2Seq and TA-Seq2Seq exhibit that by concatenating context as single input, these models cannot distinguish between early turns and late turns. Similarly, the results achieved on OpenSubtitles dataset (See Figure 4 in the supplementary material) illustrate that THRED succeeds in staying on topic and in accounting for contextual information.

5.2.1 Reliability Assurance

In order to ensure that the SS measurement is stable and void of random error, we investigate whether the SS metric is able to yield the same previous results regardless of a specific test dataset. Following (Papineni et al., 2002), the test dataset is randomly partitioned to 5 disjoint subsets (i.e., each one consists of 1000 test dialogues). Then, we compute standard deviation of SS over each dataset. The results, showcased in Table 3, indicate low standard deviation on the subdatasets, denoting that the SS metric is a consistent and reliable measure to compare different dialogue models.

Metric	**THRED**	**Seq2Seq**	**HRED**	**TA-Seq2Seq**
$\text{SS}_{\text{Utt.1}}$	0.008	0.009	0.001	0.006
$\text{SS}_{\text{Utt.2}}$	0.010	0.008	0.007	0.005

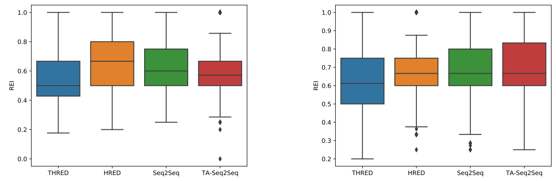
Table 3: Standard deviation of mean SS scores over the 5 different partitions of Reddit test dataset.

5.3 Response Echo Index

The goal of the Response Echo Index (REI) metric is to detect overfitting to the training dataset. More specifically, we want to measure the extent to which the responses generated by our model repeat the utterances appearing in the training data. Our approach is close to sampling and finding the nearest neighbour in image generative models (Theis et al., 2016). We randomly sampled 10% of the training data of both OpenSubtitles and Reddit. The nearest neighbour is determined via Jaccard similarity function. Each utterance is represented by lemmatized bag-of-words where stop words and punctuation marks are omitted. In effect, REI is defined as:

$$REI\left(resp_i\right) = \max_{utt_m \in \mathbb{T}_{0.1}} \mathcal{J}\left(\overline{resp_i}, \overline{utt_m}\right)$$

where $\bar{t}$ is the normalized form of text t, $\mathbb{T}_{0.1}$ denotes the sampled training data, and $\mathcal{J}$ represents Jaccard function. REI is expected to be low since the generated responses should be distant from the nearest neighbor. According to the results, presented in Figure 2, the REI scores of the responses generated from THRED are the lowest compared to the rest of the models. Such observation leads us to the conclusion that THRED is able to generate unique responses which appear to be drawn from the input distribution, while being measurably far from the input dataset. This strength in THRED is attributed to the topic attention and incorporating topic words in response generation.



(a) Reddit

(b) OpenSubtitles

Figure 2: Performance results of the generated responses from different models based on REI. From left to right, the labels in horizontal axis are THRED, HRED, Seq2Seq, TA-Seq2Seq.

Due to the same reason, standard Seq2Seq and HRED fall short.

5.4 Degree of Diversity & Perplexity

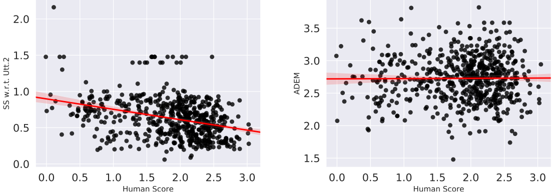
To account further for diversity in generated responses, following (Li et al., 2016a), we calculated $distinct$-1 and $distinct$-2 by counting unique unigrams and bigrams, normalized by the number of generated words. The results, given in Table 4, on Reddit indicate that THRED yields content rich and diverse responses, mainly ascribed to incorporating new topic words into response generation. Further, in perplexity, THRED performs slightly better.

5.5 Human Evaluation

Besides the quantitative measures, 4-scale and side-by-side human evaluation were carried out. Five human raters were recruited for the purpose of evaluating the quality of the responses. They were fluent, native English speakers and well-instructed for the judgment task to ensure quality rating. We showed every judge 300 conversations (150 dialogues from Reddit and 150 dialogues from OpenSubtitles) and two generated responses for each dialogue: one generated by THRED model and the other one generated by one of our baselines. The source models were unknown to the evaluators. The responses were ordered in a random way to avoid biasing the judges. Additionally, Fleiss' Kappa score is used to gauge the reliability of the agreement between human evaluators (Shao et al., 2017). An example of generated responses from the Reddit dataset are provided in Table 1 For the 4-scale human evaluation, judges were asked to judge the responses from Bad (0) to Excellent (3). Additional details are provided in the supplementary material. The results of this experiment, conducted on Reddit, are detailed in Table 5. The lablers with a high consensus degree rated 32.9% and 36.9% of the THRED responses in OpenSubtitles and Reddit respectively as Excellent, which is greatly larger than all baselines (up to 11.6% and 22.7% respectively). Apart from the 4-scale rating, we conducted the evaluations side-by-side to measure the gain in THRED over the strong baselines. Specific comparison instructions are included in the supplementary material. The results, illustrated in Table 5, suggest that THRED is substantially superior to all baselines in producing informative and plausible responses from human's perspective. The high Kappa scores imply that a major agreement prevails among the lablers. In particular, THRED beats the strong baselines in 52% of the test data in Reddit (the percentage is achieved by averaging the win ratio). However, for the rest of the cases, THRED is equally good with the baselines in 25% in Reddit (calculated similarly based on Table 5). Hence, the ratio of cases where THRED is better than or equal with the baselines in terms of quality is 77% in Reddit.

Method	PPL	*distinct-1*	*distinct-2*
Seq2Seq	62.12	0.0082	0.0222
HRED	63.00	0.0083	0.0182
TA-Seq2Seq	62.40	0.0098	0.0253
THRED	**61.73**	**0.0103**	**0.0347**

Table 4: Performance results of diversity and perplexity metrics of all the models on the Reddit test dataset. THRED surpasses all the baselines with a gain of 5% in $distinct$-1 and 37% in $distinct$-2 over TA-Seq2Seq (second best).



(a) SS w.r.t. Utt.2 ($\rho = -0.341$)

(b) ADEM ($\rho = 0.014$)

Figure 3: Scatter plots illustrating correlation between automated metrics and human judgment (Pearson correlation coefficient is reported in the brackets). In order to better visualize the density of the points, we added stochastic noise generated by Gaussian distribution $\mathcal{N}(0, 0.1)$ to the human ratings (i.e., horizontal axis) at the cost of lowering correlation, as done in (Lowe et al., 2017).

5.5.1 Automated metric vs. Human evaluation

We also carried out an analysis on the correlation between the human evaluator ratings and our

Side-by-Side	Wins	Losses	Equally Good	Equally Bad	Kappa
THRED vs Seq2Seq	**47.5%±4.4%**	19.1%±3.3%	28.5%±3.1%	4.9%±1.8%	0.80
THRED vs HRED	**51.7%±4.6%**	20.1%±3.4%	20.9%±3.1%	7.2%±2.3%	0.75
THRED vs TA-Seq2Seq	**55.7%±4.1%**	13.5%±2.6%	24.7%±3.0%	6.1%±1.8%	0.77
4-scale	**Excellent**	**Good**	**Poor**	**Bad**	**Kappa**
Seq2Seq	22.7%±2.6%	47.2%±3.5%	22.5%±3.5%	7.6%±2.7%	0.80
HRED	14.5%±2.8%	46.7%±3.8%	31.3%±3.8%	7.5%±2.5%	0.84
TA-Seq2Seq	17.1%±2.4%	44.8%±3.5%	30.1%±3.2%	8.0%±2.3%	0.72
THRED	**36.9%±3.0%**	51.1%±2.9%	10.3%±2.4%	1.7%±1.5%	0.84

Table 5: Side-by-side human evaluation along with 4-scale human evaluation of dialogue utterance prediction on Reddit dataset (mean preferences ±90% confidence intervals).

quantitative scores. The Semantic Similarity metric, which requires no pre-training, reaches a Pearson correlation of -0.341 with respect to the most recent utterance (Utt.2) on Reddit. A negative correlation is anticipated here since the higher human ratings correspond to the lower semantic distance. This compares with values of 0.351 for Automatic User Ratings (Venkatesh et al., 2018) and 0.436 for ADEM (Lowe et al., 2017) from recent models which required large amounts of training data and computation. The correlations are visualized as scatter plots in Figure 3. In addition, we assessed ADEM on our test datasets using the pre-trained weights[5], provided by the authors. ADEM achieves low correlation with human judgment ($\rho = 0.014$ on Reddit and $\rho = 0.034$ on OpenSubtitles) presumably since the quality of its predicted scores highly depends on the corpus on which the model is trained.

5.6 Comparing Datasets

Finally, we investigate the impact of training datasets on the quality of the responses generated by THRED and all baselines. Table 6 has results which support that our cleaner, well-parsed Reddit dataset generates significantly improved responses over our metrics of interest. In particular, we contrast the two datasets in terms of human judgment and the automated metrics among all the models. Regarding human assessment, we took the mean evaluation rating (MER) per response in the test data to draw the comparison between the datasets. As demonstrated in Table 6 (see more details in Figure 6 in the Appendix), the human evaluators scored generated responses from the Reddit dataset higher than utterances generated from the OpenSubtitles dataset, which is true not only in THRED, but in all models. Consequently, the training data plays a crucial role in generating high-quality responses. Morever, in OpenSubtitles, the assumption of spotting a conversation, as stated in Section 4, tends to include extraneous utterances in the dialogue, impeding the response generation process. While such presumption may seem valid in dealing with two-turn dialogues, it can aggravate the quality of conversations in multi-turn dialogues.

Method	OpenSubtitles		Reddit	
	μ	σ	μ	σ
Human MER	1.681	0.639	**1.868**	0.624
SS w.r.t. Utt.1	0.642	0.167	**0.631**	0.270
SS w.r.t. Utt.2	0.662	0.209	**0.599****	0.262
REI	0.667	0.205	**0.546****	0.201

Table 6: Mean μ and standard deviation σ over metrics per dataset to fare Reddit against OpenSubtitles. (** indicates statistical significance with p-value < 0.001)

6 Conclusion

In this work, we introduce the Topical Hierarchical Recurrent Encoder Decoder (THRED) model for generating topically consistent responses in multi-turn open conversations. We demonstrate that THRED significantly outperforms current state-of-the-art systems on quantitative metrics and human judgment. Additionally, we evaluate our new model and existing models with two new metrics which prove to be good measures for automatically evaluating the quality of the responses. Finally, we present a parsed and cleaned dataset based on conversations from Reddit which improves generated responses. We expect more advanced work to be done in the area of chit-chat dialogue to improve the models, training data, and means of evaluation.

[5] https://github.com/mike-n-7/ADEM

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *international conference on learning representations*.

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

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

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 the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1116–1126.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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 *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 *AAAI*, pages 3295–3301.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *neural information processing systems*, pages 3104–3112.

Lucas Theis, Aron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. *international conference on learning representations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generatio. In *AAAI*, pages 3351–3357.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *ECIR 2011 Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611*, pages 338–349.

A Supplementary Material

A.1 Experimental Setup

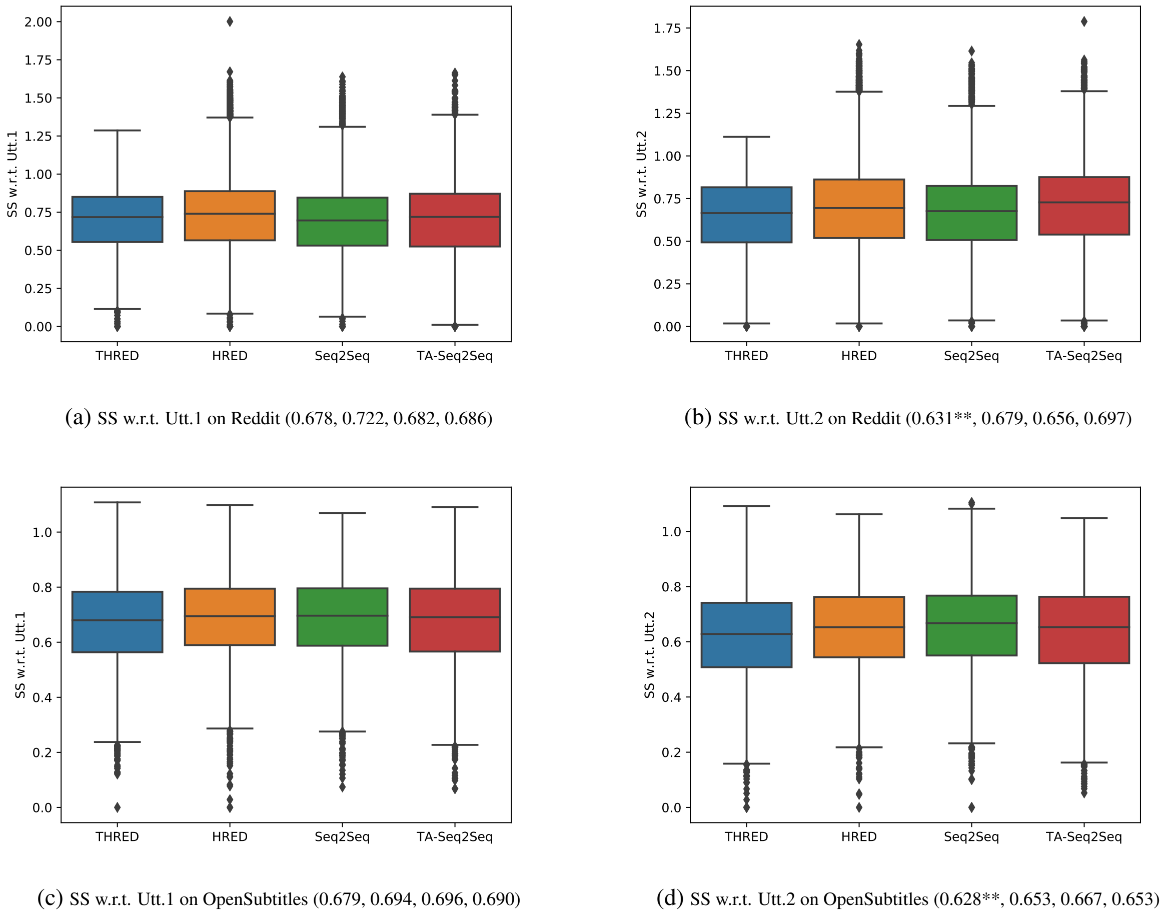
The model parameters are learned by optimizing the log-likelihood of the utterances via Adam optimizer with a learning rate of 0.0002; we followed (Luong et al., 2015) for decaying the learning rate. The dropout rate is set to 0.2 for both the encoder and the decoder to avoid overfitting. For all the baselines, we experimented hidden state units with the size of 1024. For our model, we tested with encoder and decoder hidden state units of size 800, the same for the context encoder. During inference, we experimented with the standard beam search with the beam width 5 and the length normalization $\alpha = 1$ (Wu et al., 2016). We noticed that applying the length normalization resulted in a more diverse and longer sentences but at the expense of the semantic coherence of the response in some cases.

Training LDA model: We trained two LDA models[6]: one trained on OpenSubtitles and the other one trained on Reddit. Both of them were trained on 1M dialogues. We set the number of topics to 150, α to $\frac{1}{150}$ and γ to 0.01. We filtered out stop words and universal words. We also discarded the 1000 words with the highest frequency from the topic words.

A.2 Human Evaluation Procedure

For the 4-scale human evaluation, judges were asked to judge the responses from Bad (0) to Excellent (3). Excellent (score 3): The response is very appropriate, on topic, fluent, interesting and shows understanding of the context. Good (score 2): The response is coherent with the context but it is not diverse and informative. It may imply the answer. Poor (score 1): The response is interpretable and grammatically correct but completely off-topic. Bad (score 0): The response is grammatically broken and it does not provide an answer. Regarding the side-by-side evaluation, humans were asked to favor response 1 over response 2 if: (1) response 1 is relevant, logically consistent to the context, fluent and on topic; or (2) Both responses 1 and 2 are relevant, consistent and fluent but response 1 is more informative than response 2. If judges cannot tell which one is better, they can rate the responses as Equally good or Equally Bad.

[6] We used LDA model developed in Gensim library.



(a) SS w.r.t. Utt.1 on Reddit (0.678, 0.722, 0.682, 0.686)

(b) SS w.r.t. Utt.2 on Reddit (0.631**, 0.679, 0.656, 0.697)

(c) SS w.r.t. Utt.1 on OpenSubtitles (0.679, 0.694, 0.696, 0.690)

(d) SS w.r.t. Utt.2 on OpenSubtitles (0.628**, 0.653, 0.667, 0.653)

Figure 4: Box plots showcasing the performance of the generated responses from different models based on the Semantic Similarity metric with respect to Utt.1 and Utt.2 (complementary to Table 2). From left to right, the labels in horizontal axis are THRED, HRED, Seq2Seq and TA-Seq2Seq. THRED surpasses all baselines in similarity with Utt.2, and works mildly better in similarity with Utt.1. The median for every box plot is also reported between brackets. The order of the values is the same as the order of the methods.

CONTEXT	GENERATED RESPONSES
(OpenSubtitles) the giant observatory in hawaii are like time machines capturing the faint light that has travelled towards us through all of cosmic history → the deeper they look into space the farther back they see in time	**THRED:** as the stars move into space they see the rising sun and the stars rise above the clouds (*Excellent, Excellent, Good, Excellent, Poor*) **HRED:** this is the first time we have ever seen (*Poor, Poor, Bad, Good, Bad*) **Seq2Seq:**the stars and stars of the moon are the stars and stars of the stars and stars and stars and stars and stars and stars (*Poor, Bad, Bad, Bad, Poor*) **TA-Seq2Seq:** they have been searching for a long time for a long time (*Good, Poor, Good, Good, Good*)
(OpenSubtitles) and closer to home in baltimore prominent psychiatrist and the author of a newly released book on insanity disappeared mysteriously last night after a lecture at the university → former mental patient james cole is wanted for questioning regarding dr cole 's disappearance	**THRED:** authorities warn that cole has a history of violence and advise anyone to notify authorities (*Excellent, Excellent, Excellent, Excellent, Excellent*) **HRED:** they have been unable to identify him (*Good, Poor, Poor, Good, Poor*) **Seq2Seq:**cole cole is cole is cole(*Poor, Bad, Bad, Bad, Bad*) **TA-Seq2Seq:** authorities warn that cole has been kidnapped (*Poor, Good, Good, Good, Good*)

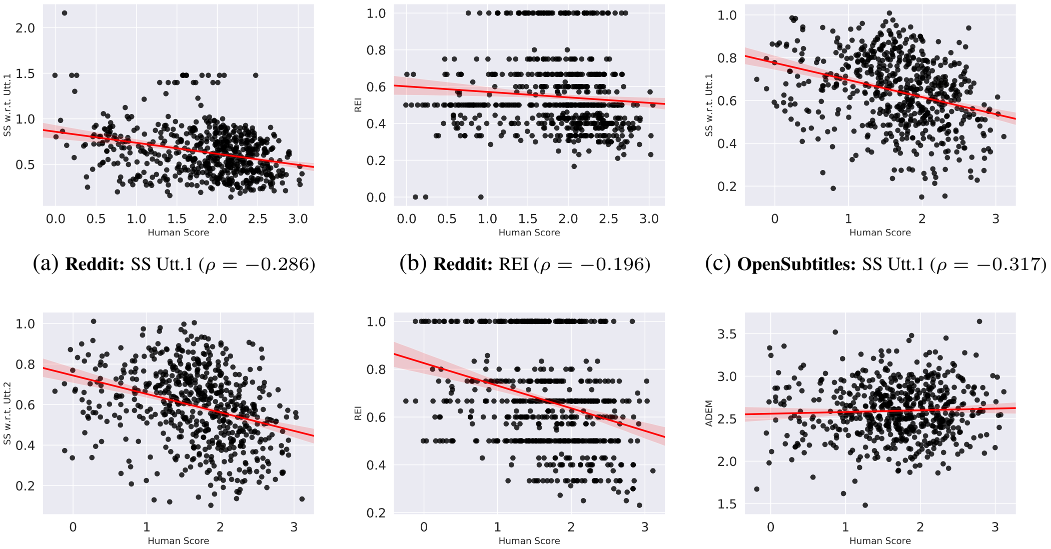
Table 7: One cherry-picked dialogues out of 150 conversations along with the generated responses from all models. Human judgments are provided in the brackets. The blue arrow specifies a dialogue exchange and the highlighted words in red represent the topic words acquired from the pre-trained LDA model.

Method	OpenSubtitles			Reddit		
	PPL	*distinct-1*	*distinct-2*	PPL	*distinct-1*	*distinct-2*
Seq2Seq	74.37	0.0112	0.0258	62.12	0.0082	0.0222
HRED	74.65	0.0079	0.0219	63.00	0.0083	0.0182
TA-Seq2Seq	75.92	0.0121	0.0290	62.40	0.0098	0.0253
THRED	**73.61**	**0.0157** (+*30%*)	**0.0422** (+*45%*)	**61.73**	**0.0103** (+*5%*)	**0.0347** (+*37%*)

Table 8: Complete performance results of diversity and perplexity on Reddit test data and OpenSubtitles test data (complementary to Table 4). The numbers in the bracket indicate the gain of $distinct$-1 and $distinct$-2 over the second best method (i.e., TA-Seq2Seq).

Side-by-Side	Wins	Losses	Equally Good	Equally Bad	Kappa
THRED vs Seq2Seq	**54.0%±4.2%**	18.4%±3.4%	17.2%±3.0%	10.4%±2.3%	0.75
THRED vs HRED	**51.6%±4.4%**	19.5%±3.5%	18.4%±2.9%	10.5%±2.4%	0.72
THRED vs TA-Seq2Seq	**64.0%±4.3%**	14.4%±3.1%	14.1%±2.5%	7.5%±2.1%	0.90
4-scale Rating	**Excellent**	**Good**	**Poor**	**Bad**	**Kappa**
Seq2Seq	**8.4%±2.2%**	48.9%±3.9%	33.2%±3.7%	9.5%±3.1%	0.89
HRED	**11.6%±2.4%**	41.5%±3.4%	36.9%±3.9%	10.0%±2.8%	0.79
TA-Seq2Seq	**9.5%±2.1%**	42.3%±3.7%	34.7%±3.9%	13.6%±3.7%	0.92
THRED	**32.9%±3.6%**	49.2%±3.3%	16.8%±3.0%	1.1%±0.9%	0.83

Table 9: Side-by-side human evaluation along with 4-scale human evaluation of dialogue utterance prediction on OpenSubtitles dataset (mean preferences ±90% confidence intervals). Results on Reddit dataset are reported in Table 5.



(a) **Reddit:** SS Utt.1 ($\rho = -0.286$)

(b) **Reddit:** REI ($\rho = -0.196$)

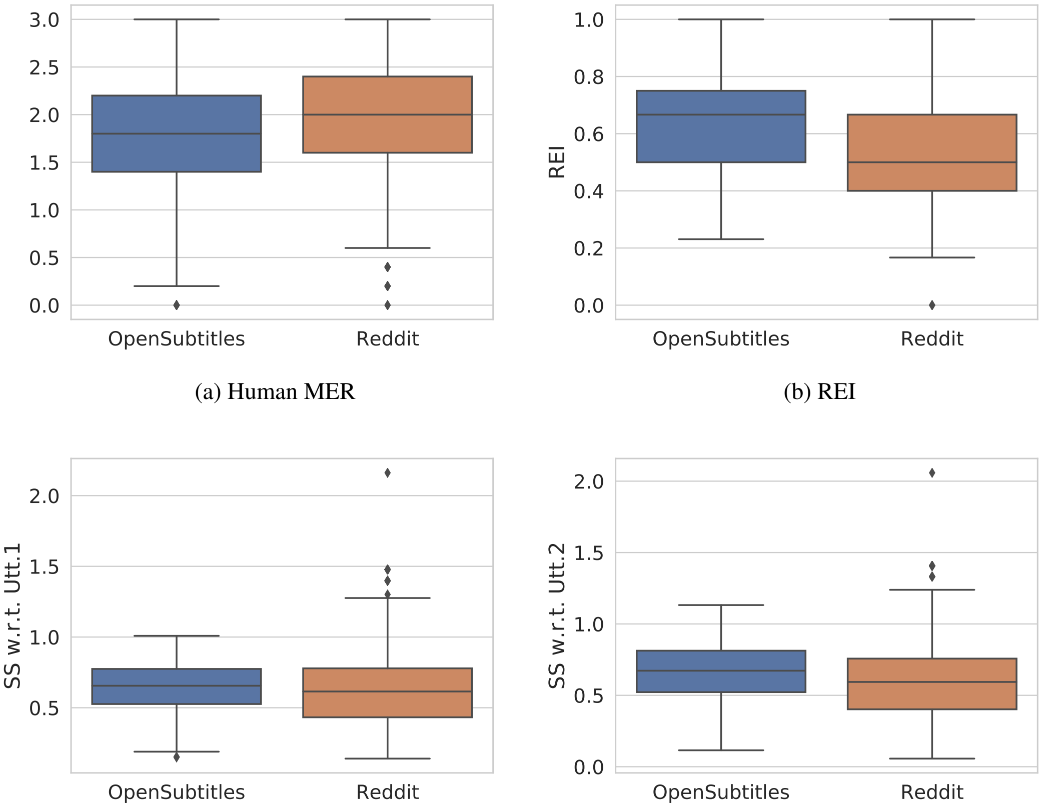
(c) **OpenSubtitles:** SS Utt.1 ($\rho = -0.317$)

(d) **OpenSubtitles:** SS Utt.2 ($\rho = -0.324$)

(e) **OpenSubtitles:** REI ($\rho = -0.295$)

(f) **OpenSubtitles:** ADEM ($\rho = 0.034$)

Figure 5: Scatter plots illustrating correlation between automated metrics and human judgment (Pearson correlation coefficient is reported in the brackets).



(a) Human MER

(b) REI

(c) Semantic Similarity w.r.t. Utt.1

(d) Semantic Similarity w.r.t. Utt.2

Figure 6: Box plots demonstrating the detailed comparison between OpenSubtitles and Reddit datasets. The metrics are calculated for all models in the cherry-picked data (150 samples for OpenSubtitles and 150 samples for Reddit). The results here complement what we found in Table 6 in which only mean and standard deviation are reported per metric.