Computer Speech & Language 65 (2021) 101135

Contents lists available at ScienceDirect

Computer Speech & Language

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

HSCJN: A holistic semantic constraint joint network for diverse response generation

Yiru Wang[a], Pengda Si[a], Zeyang Lei[a,c], Guangxu Xun[b], Yujiu Yang*[,a]

[a] Tsinghua Shenzhen International Graduate School, Tsinghua University, PR China
[b] Department of Computer Science, University of Virginia, USA
[c] Baidu Inc., Beijing, PR China

A R T I C L E I N F O

Article History:
Received 31 December 2019
Revised 2 July 2020
Accepted 13 July 2020
Available online 29 July 2020

Keywords:
Dialogue generation
Diversity
Holistic semantics
Word prediction
Entropy regularizer

A B S T R A C T

The sequence-to-sequence (Seq2Seq) model generates target words iteratively given the previously observed words during the decoding process, which results in the loss of the holistic semantics in the target response and the complete semantic relationship between responses and dialogue histories. In this paper, we propose a generic diversity-promoting joint network, called Holistic Semantic Constraint Joint Network (HSCJN), enhancing the global sentence information, and then regularizing the objective function with penalizing the low entropy output during the training stage. Our network introduces more target information to improve diversity and captures direct semantic information to better constrain relevance simultaneously. Moreover, the proposed method can be easily applied to any Seq2Seq structure. Extensive experiments on several dialogue corpuses show that our method effectively improves both semantic consistency and diversity of generated responses, and achieves better performance than other competitive methods.

1. Introduction

Recently, dialogue systems have attracted increasing attention in both academia and industry because of their potential applications and commercial values. Sequence-to-sequence (Seq2Seq) models form the cornerstone of popular dialogue generation models (Serban et al., 2016; Sordoni et al., 2015; Shang et al., 2015). However, neural dialogue systems based on Seq2Seq models tend to repeatedly generate universal and boring responses like “I don’t know.”, “Thank you.”. Although widely applied, conventional Maximum Likelihood Estimation (MLE) training could cause the low-diversity problem above (Li et al., 2016b). Since high-frequency words make up a big proportion of the training set, MLE encourages the model to excessively generate high-frequency words.

Moreover, when training a Seq2Seq model traditionally, we iteratively maximize the log predictive likelihood of each true token in the target sequence, given the previously decoded tokens. Therefore, the model can only see the previous information during learning, unable to grasp the holistic information of the target sequence when decoding tokens. This also leads to the loss of a complete semantic relationship between target sequences and source sequences.

As discussed, we argue that the current learning strategy heavily limits the Seq2Seq models to generate highly diverse responses, and the holistic semantic information of the target response, as well as the global semantic relationship between responses and dialog histories, are missing during the generation process. Most previous solutions simply rely on external

*Corresponding author.
E-mail addresses: wangyiru017@gmail.com (Y. Wang), spd18@mails.tsinghua.edu.cn (P. Si), zeyanglei@gmail.com (Z. Lei), gx5bt@virginia.edu (G. Xun), yang.yujiu@sz.tsinghua.edu.cn (Y. Yang).

https://doi.org/10.1016/j.csl.2020.101135
0885-2308/

information or post-processing models to mask the deficiencies of the Seq2Seq model, while the problem of the Seq2Seq model itself has not been addressed. Therefore, in this paper, we hope to fully exploit the learning potential of the Seq2Seq models without any external information to improve diversity while better to constrain the semantic relevance of generated responses simultaneously. We propose a Holistic Semantic Constraint Joint Network (HSCJN) for model training to predict the subsequent word set in each target utterance for direct supervision in decoding, which directly introduces more linguistic information from target utterances to increase diversity. More specifically, during the training stage, we require each hidden state in the decoder to predict the words in the target utterance which remain ungenerated, and the initial state of the decoder is required to predict all the words in the target utterance. Since the HSCJN enables the decoder network to see all words in the target utterance at every time step, our model also more likely captures direct semantic information such as keywords in target utterances to enhance relevance.

In this way, the relationship of representation spaces between source sequences and target sequences, and the transition between different decoder states could be better constrained. In addition, we consider that the entropy of the output distribution is low if the model is over-confident about high-frequency words. Penalizing the low entropy output distribution can help regularize the model, optimize the predicted output distribution and alleviate the over-estimation of high-frequency words. Therefore, we devise a maximum entropy-based regularizer. Our learning framework can be used as a general joint training method with Seq2Seq models and requires no additional data or annotation. In general, our contributions are summarized as follows:

- We devise a joint training network to introduce future information during the decoding stage in the open-domain dialogue generation task, which can be applied to any Seq2Seq neural model. Our network introduces more linguistic information from target utterances to increase diversity, and likely captures key semantic information such as keywords in target utterances to enhance relevance in a direct manner.
- We regularize the model by penalizing low entropy output distribution at each time step in the decoder to alleviate the over-estimation of high-frequency words, which also enables the loss function to consider every word in the vocabulary to improve diversity.
- The experimental results on multi-turn dialogue datasets show the effectiveness of our method in terms of both diversity and relevance of generated responses.

2. Related work

The diversity of generated responses is an important issue of common concern. Serban et al. (2017) and Zhao et al. (2017) proposed to introduce variational auto-encoders (VAEs) to Seq2Seq models to bring informativeness by increasing variability. Some researches proposed several beam search based approaches (Li et al., 2017; Song et al., 2018; Vijayakumar et al., 2016). However, this kind of methods merely provide a criterion for reweighing response candidates, rather than producing more diverse responses in the first place. Besides, other previous works introduced additional information or knowledge such as the context (Serban et al., 2016; Tian et al., 2017; Yao et al., 2017), keyword (Serban et al., 2016; Xing et al., 2017; Yao et al., 2017) or knowledge-base (Young et al., 2018; Ghazvininejad et al., 2018) into the response generation process to produce informative content. Although effective, these approaches actually bypass the low-diversity problem by introducing the randomness of stochastic latent variables or additional information. The underlying Seq2Seq model remains sub-optimal in terms of diversity. Li et al. (2016a) proposed to use a Maximum Mutual Information (MMI) as an optimization objective to maximize the mutual information between messages and responses, but the MMI objective is used only during test time, and relies on many extra modules, like reverse models and beam search. Zhang et al. (2018) proposed the Adversarial Information Maximization (AIM) model which considers explicitly maximizing mutual information during training to generate informative responses, but it still needs to train an extra backward model generating source from the target, and it is implemented with complicated adversarial training strategy.

In other tasks, the word prediction technique has been applied in the neural machine translation (Weng et al., 2017; L'Hostis et al., 2016). Lin et al. (2019) added entropy to the loss function to make the sparse distribution more specifically concentrate on a small set of video segments in the VQA task. Unlike we consider entropy on the entire vocabulary, this work only considered the entropy on the current video segment.

3. Methodology

3.1. Task definition

Given a dialogue as a sequence of M utterances $\mathbb{U} = \{U_1, \ldots, U_M\}$, and each U_i as a sequence of N_i tokens $U_i = \{w_{i1}, \cdots, w_{iN_i}\}$, where w_{ij} represents the token at position j in utterance i from the vocabulary V, our task is to generate a response $Y = \{y_1, y_2, \ldots, y_m\}$ that is not only fluent and grammatical but also not repeated and trivial in content. Essentially, the goal is to estimate the conditional probability:

$$P(Y|\mathbb{U}) = \prod_{t=1}^{m} P\left(y_t \middle| \mathbb{U}, y_{<t}\right)$$
$$\mathbb{U} = \left\{ [w_{1k}]_{k=1}^{N_1}, \cdots, [w_{Mk}]_{k=1}^{N_M} \right\} \tag{1}$$

3.2. Model overview

Fig. 1 demonstrates the architecture of our model, in which we join a prediction network with the decoder network on each hidden state of the decoder. The encoder takes the word embedding sequences of context utterances as inputs and obtains the hidden representations of the context. The decoder starts the generation of the target sequence from the initial state s_0. Since the initial state is responsible for the generation of the whole target sequence, we optimize the initial state by making a prediction for all the target words to contain comprehensive target information. Similarly, at each time step in the decoder, we introduce the prediction network to predict the word set of the target subsequence that has not yet been generated. The response is generated from the decoder, and the HSCJN applies a constraint network to each hidden state in the decoder to introduce more direct language information from the Seq2Seq model. The joint prediction network only works for model training. During inference, the underlying encoder-decoder model is exactly the same. The HSCJN guides the underlying Seq2Seq model to capture the entire linguistic information and dialogue context from the sentence level during the learning phase, so that the model can exert a stronger ability to generate reasonable responses during the inference phase. There is a specific objective function for the HSCJN. In addition, we optimize the output distribution at each decoding step by adding a maximum entropy-based regularizer to the final objective function to generate the predictive response.

3.3. Joint network prediction

In the HSCJN, we require the hidden state at each time step in the decoder to predict the word set containing target words which remain ungenerated in the target utterance, where the order of words is not considered and we assume target words are independent with each other. In this way, at each time step, the decoder generates words not only conditioned on the previously generated subsequences within the original decoder network but also under consideration of the future words not yet seen in the target sequence through our HSCJN. That is, our joint network HSCJN can introduce and utilize the global sentence information in target utterances for every token generation, beneficial for both diversity and relevance. Specifically, for each time step j in the decoder, the hidden state s_j is required to predict the word collection of $Y_{j\sim m} \triangleq (y_j, y_{j+1}, \ldots, y_m)$. The conditional probability P_j of the prediction task in the HSCJN at hidden state s_j is defined as follow:

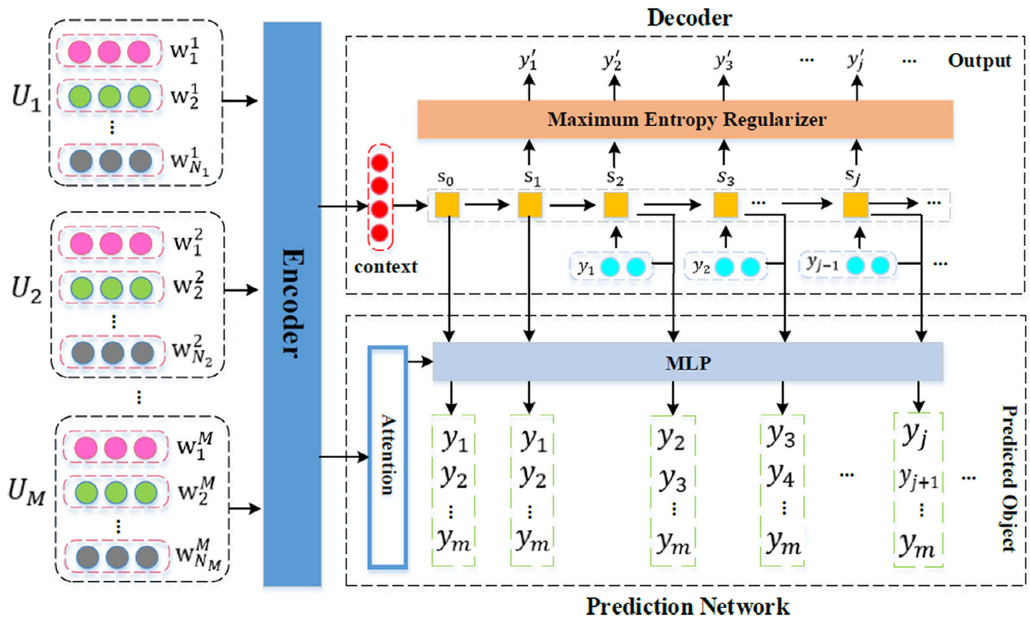

Fig. 1. The architecture of the proposed HSCJN for multi-turn dialogue generation. The prediction network only works during the training phase which helps improve the Seq2Seq model for inference.

$$P_j\left(Y_{j\sim m}\middle|\Psi\right)=\prod_{t=j}^{m}P(y_t|\Psi),\text{with }P(y_t|\Psi)=MLP\left(\left[e\left(y_{j-1}\right);s_j;c_j\right]\right) \tag{2}$$

where, $\Psi\triangleq\{y_{<j},U_1,\ldots,U_M\}$, and the set $Y_{j\sim m}$ is the word set of the future subsequence in target response Y at time step j. *MLP* is a multi-layer perceptron with two hidden layers using *tanh*(.) as an activation function, followed by one output layer with *sigmoid*(.) acting on each neuron. Here, we predict the target word set in a multi-classification way. $e(y_{j-1})$ is the embedding of the word y_{j-1}, and c_j is the context vector from the attention mechanism (Luong et al., 2015).

$$\begin{gathered} c_j=\sum_{i=1}^{T}a_{ji}h_i \\ a_{ji}=\frac{\exp\left(tanh\left(W_c\left[s_{j-1},h_i\right]\right)\right)}{\sum_{k=1}^{T}\exp\left(tanh\left(W_c\left[s_{j-1},h_k\right]\right)\right)} \end{gathered} \tag{3}$$

where W_c is the weight parameter, T is the input sequence length, h_i is the hidden state of the encoder RNN at time step i, and s_{j-1} is the hidden state of the decoder RNN at the previous time step $j-1$.

Specifically, for the initial state s_0, the HSCJN requires it to predict the word set containing all target words, so as to compress the overall information of the target sequence into the initial state. Therefore, the decoder can see the entire target sequence at the initial time step through the HSCJN. The conditional probability P_0 of the HSCJN at the initial state s_0 is defined as follow:

$$P_0(\tilde{Y}|\mathbb{U})=\prod_{t=1}^{m}P_0(y_t|\mathbb{U}),\ \text{ with }\ P_0(y_t|\mathbb{U})=MLP([s_0;c_0]) \tag{4}$$

where $\tilde{Y}$ is the word set containing all target words in the target response Y, and c_0 is the context vector from the attention mechanism for the initial state s_0.

To optimize the HSCJN network, we add an extra likelihood function L_{WP} into the training procedure:

$$L_{WP}=-\frac{1}{m}\log P_0-\sum_{j=1}^{m}\frac{1}{m-j+1}\log P_j \tag{5}$$

where P_0 and P_j are as previously defined, the coefficient of the logarithm is used to calculate the average probability of each prediction. This loss function is used to guide the HSCJN to accurately introduce the expected target semantic information.

3.4. Output distribution regularizer

When a dialogue model generates universal responses, the prediction of high-frequency words is too confident, that is, the entire output probability distribution is concentrated on high-frequency words. In result, the entropy of the output distribution is low. We consider that maximizing the entropy of the output distribution at each decoding step could help to regularize the model and produce more diverse responses. By this means, the token-level distribution $P(y_t|\mathbb{U},y_{<t})$ is better constrained to relieve over-estimation of high-frequency words. Therefore, we add a negative entropy term to the negative log-likelihood loss function during training. To minimize the overall loss function, the model encourages the maximization of entropy. Specifically, the loss term is expressed as follow:

$$\begin{gathered} L_{ME}=-\sum_{t=1}^{m}H\left(p\left(y_t|\tilde{\Psi}\right)\right) \\ H\left(p\left(y_t|\tilde{\Psi}\right)\right)=-\sum_{i=1}^{|V|}p\left(w_i|\tilde{\Psi}\right)*\log p\left(w_i|\tilde{\Psi}\right) \end{gathered} \tag{6}$$

where $\tilde{\Psi}\triangleq\{y_{<t},U_1,\ldots,U_M\}$, $H(\cdot)$ is the entropy of the output distribution at decoding step t, $|V|$ is the length of the vocabulary V, and w_i represents a word in the vocabulary.

This loss function not only penalizes the low entropy output distribution when predicting each token, but also considers the entropy over the entire vocabulary, so that the model likely takes into account more words to increase diversity.

3.5. Loss function

We add L_{WP} and L_{ME} to the original negative log-likelihood loss function. The final loss function for model training is as follow:

$$L=-\log P(Y|\mathbb{U})+\alpha L_{WP}+\beta L_{ME} \tag{7}$$

where α and $\beta\in(0,1]$ are weight coefficients, which control the strength of the joint prediction task and the output distribution regularizer, respectively. Our HSCJN builds a training objective at sentence level instead of the traditional token-level transition, considering the complete linguistic information in target utterances for every token generation.

4. Experiment

4.1. Data preparation

We evaluate our proposed method on two multi-turn dialogue datasets, DailyDialog (Li et al., 2017) and OpenSubtitles (Tiedemann, 2012).

DailyDialog: It is a high-quality and less noisy dataset, which contains 13,118 multi-turn dialogues, separated into training/validation/test sets with 11,118/1,000/1,000 conversations. For the computational efficiency, we remove the dialogues with more than 300 tokens, which only makes up a small proportion of the whole dataset, and finally our training/validation/test sets of DailyDialog dataset contain 10,712 / 976 / 960 conversations, respectively.

OpenSubtitles: It is a collection of movie subtitles. Following previous work (Xu et al., 2018), we treat each turn in the dataset as the target text and the two previous sentences as the source text. We randomly sample 200,000 / 50,000 / 10,000 dialogues for training, validation, and testing, respectively. Similarly, we also remove dialogues with more than 300 tokens, and finally our training/validation/test sets of OpenSubtitles dataset contain 199,992/ 49,995 / 9984 dialogues, respectively.

4.2. Baselines

AttnSeq2Seq: A vanilla Seq2Seq model with attention mechanism (Bahdanau et al., 2014). The encoder and decoder are both recurrent neural networks (RNN) with LSTM as the basic cell, and the encoder RNN is bidirectional.

HRED: HRED (Serban et al., 2016) considers dialogue histories in multi-turn dialogue generation at two levels: a sequence of words for each utterance and a sequence of utterances, and models this hierarchy of conversations accordingly.

VHRED: VHRED (Serban et al., 2017) augments the HRED model with a stochastic latent variable at the decoder, trained by maximizing a variational lower-bound on the log-likelihood. The latent variable helps facilitate the generation of long utterances with more informative content.

4.3. Model settings

Our proposed method is generic since it can be combined with any Seq2Seq model. In our experiments, we use HRED as the basis of our learning framework. We initialize the recurrent parameter matrices as orthogonal matrices while all the bias vectors are set to **0**. Other parameters are initialized by sampling from the Gaussian stochastic distribution $\mathcal{N}(0, 0.01)$. The vocabularies are limited to the most frequent 25K and 30K words for DailyDialog dataset and OpenSubtitles dataset, respectively. We apply GRU with 500 hidden states and GRU with 1000 hidden states to the encoders at word-level and utterance-level respectively, and LSTM with 500 hidden states to the decoder. The dimension of word embedding is set to 300. We use the Adam optimizer (Kingma and Ba, 2014) to update the parameters, with a batch size of 8. The learning rate is 0.0002 and the dropout rate is 0.75. Meanwhile, we set α to 1 and β to 0.13. For decoding during test time, we simply decode until the end-of-utterance symbol *eou* occurs, using a beam search with a beam width of 5. All baseline models are implemented with the same settings.

4.4. Automatic evaluation

We adopt BLEU (Chen and Cherry, 2014; Papineni et al., 2002), Distinct-1, Distinct-2 and Distinct-3 (Li et al., 2016a) to evaluate the models at the quality and diversity level. The higher BLEU values demonstrate the responses are closer to the ground truth. Distinct-1, Distinct-2 and Distinct-3 are the proportion and number of distinct unigrams, bigrams, and trigrams in all the generated tokens, respectively. Higher Distinct-*n* values are better for the overall diversity.

Table 1 shows the experimental results of 1-turn response generation on DailyDialog corpus and OpenSubtitles corpus. It is obvious that our HSCJN generates remarkably more distinct unigrams, bigrams and trigrams than all the baselines on both two datasets. Besides, our model achieves the highest BLEU-2/3/4 values on both two datasets, compared with all the baseline models. Confirmed by our experiments, our model achieves an excellent performance in terms of both quality and diversity, regardless of the scale of datasets.

Furthermore, we conduct experiments on 2-turns dialogue generation. Given dialogue histories as input, we require models to generate the next two consecutive utterances. Since dialogues in the OpenSubtitles dataset contain only three turns, we conduct 2-turns dialogue generation on the DailyDialog corpus. The results of 2-turns dialogue generation by our model and the other two multi-turn dialogue generation models are shown in Table 2. From the results, our model exceeds all baseline models with diversity improvement in multi-turn dialogue generation, and even higher BLEU-3 and BLEU-4.

To verify the effectiveness of our model in optimizing the output distribution, we perform word segmentation and word frequency statistics on the generated responses. Fig. 2 draws the distribution of ten most frequent words in responses generated by corresponding models on the OpenSubtitles dataset, excluding punctuations. "Natural" represents the natural distribution of the ground truth. The horizontal axis represents the rank of word frequency, and the vertical axis is word frequency values. The curves above columns fit the distribution trends of the word frequencies for different models, whose colors are consistent with the columns. It shows that the frequencies of words generated by our model are not as high as the baseline models, and the frequency distribution is flatter. Moreover, the distribution trends of our model and VHRED are basically consistent with the natural distribution, and our model is closer to the natural distribution in word frequency values than VHRED.

Table 1
The results of 1-turn response generation on DailyDialog and OpenSubtitles datasets.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Distinct-1	Distinct-2	Distinct-3
DailyDialog Corpus							
Attnseq2seq	7.17	0.15	0.03	0.02	0.031/247	0.092/523	0.162/754
HRED	9.44	1.85	0.83	0.36	0.049/276	0.145/516	0.252/651
VHRED	11.28	2.30	1.05	0.44	0.068/418	0.197/778	0.342/1007
HRED+PN	**11.48**	2.55	1.28	**0.69**	**0.084/536**	**0.276/1149**	**0.475/1481**
HRED+ME	10.57	2.30	1.13	0.58	0.063/386	0.200/788	0.357/1043
HSCJN	11.05	**2.60**	**1.29**	0.56	0.075/463	0.242/954	0.424/1235
OpenSubtitles Corpus							
Attnseq2seq	13.27	0.88	0.28	0.12	0.006/530	0.022/1623	0.043/2741
HRED	12.50	1.25	0.47	0.16	0.007/459	0.027/1391	0.050/2039
VHRED	**14.64**	1.28	0.36	0.13	0.007/707	0.034/2627	0.073/4855
HRED+PN	13.15	**1.48**	0.54	0.23	0.011/880	0.056/3411	0.114/5773
HRED+ME	12.61	1.44	0.51	0.25	0.008/621	0.039/2231	0.078/3650
HSCJN	13.29	1.45	**0.55**	**0.26**	**0.012/883**	**0.063/3550**	**0.131/6056**

4.5. Ablation study

We conduct the ablation study to examine the effectiveness of each mechanism, and the results are shown in Table 1. PN and ME represent the prediction sub-network and the maximum entropy regularizer, respectively. On the DailyDialog dataset, HRED +PN generates the most distinct unigrams, bigrams and trigrams among all compared models, and also surpasses all the baseline models on Distinct-*n* metrics on the OpenSubtitles dataset, indicating that our joint network can obviously generate more diverse responses. HRED+PN also achieves higher BLEU scores, which proves that our joint network can directly capture semantic information to enhance relevance. Since HRED+ME only adds a regularization item to the loss function in the training process of HRED, the comparison with the HRED model is sufficient for verifying the performance of the maximum entropy regularizer. From the results, HRED+ME achieves obvious improvement in both of the quality and diversity performance compared with HRED, demonstrating the effectiveness of the maximum entropy regularization.

Based on the Distinct-*n* metrics, it appears that the addition of PN improves the diversity more than the addition of ME regularization. This maybe because that the prediction network works by bringing in richer semantics and expands the flow of information. At the same time, the introduced complete semantic information is helpful for modeling the semantic relationship between conversation contexts and responses, and helps the model better understand the conversation content. Differently, the maximum entropy regularizer does not increase the information flow in the encoder-decoder framework, but optimizes the prediction distribution and changes the word frequency of the output, so the scope of its action is relatively limited. Overall, both the prediction network and maximum entropy regularizer contribute to the improvement of diversity and quality in response generation.

4.6. Manual evaluation

Since automatic metrics for open-domain generative models may not be consistent with human perceptions, the quality scores from the human annotations are more reliable. Therefore, we further recruit human annotators to evaluate the quality of the generated responses. We randomly select 100 testing dialogues with responses generated by different models for each dataset and for both 1-turn and 2-turns generation. Responses generated by different models are randomly shuffled for each annotator. 5 annotators with linguistics experience are recruited to refer to the test dialogue histories and judge the quality of the responses of all compared models according to the following criteria:

0: The response cannot be used as a response to the conversation context. It is semantically unrelated or disfluent.

+1: The response can be used as a reply to the message, but it is too universal like “Yes, I see.”, “Thank you.” and “I don’t know.”.

Table 2
2-turns dialog generation results on the DailyDialog corpus.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Distinct-1	Distinct-2	Distinct-3
HRED	10.23	1.78	0.72	0.33	0.037/437	0.107/827	0.192/1104
VHRED	**12.24**	**2.26**	0.91	0.37	0.052/663	0.150/1265	0.269/1749
HSCJN	11.63	2.25	**1.01**	**0.54**	**0.053/689**	**0.175/1500**	**0.313/2056**

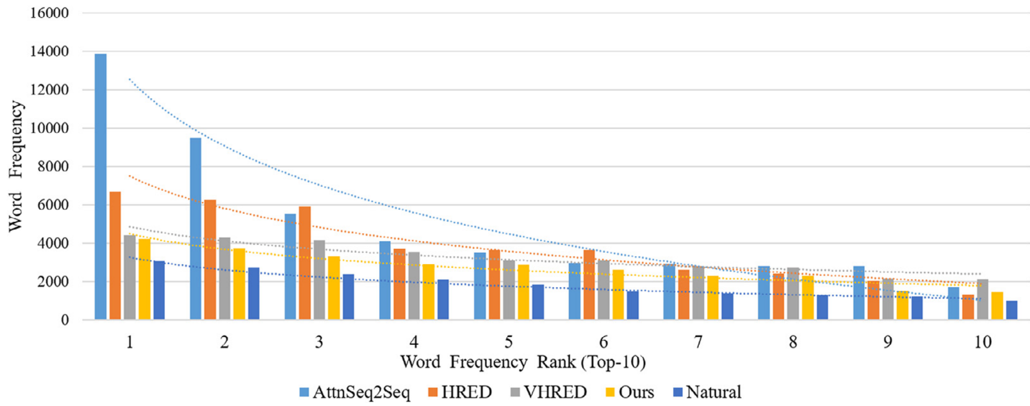

Fig. 2. Word frequency distribution on generated responses on the OpenSubtitles dataset. We draw the ten most frequent words, excluding punctuations.

+2: The response is not only grammatical and relevant, but also informative and interesting.

The average Kappa coefficient of annotation consistency between annotators is 0.623. Manual evaluation results for 1-turn response generation are presented in Table 3, which lists the percentage of each score and an overall average score. Among the three baselines, AttnSeq2Seq performs the worst and VHRED the best. Our model obtains a minimum of 0 points and a maximum of 2 points among all compared models on two datasets. This indicates that our model can generate fewer low-quality responses, as well as more semantically relevant and informative responses. The highest average score achieved by our model also confirms that our model outperforms the baselines.

Table 4 shows the manual evaluation results for 2-turns dialogue generation on DailyDialog dataset. In multi-turn generation, VHRED's performance is not good. Responses generated by VHRED may be informative but most of them are irrelevant to the context. Our model also outperforms the baselines in terms of relevance and informativeness, as well as the overall average score.

5. Case study

Table 5 presents the examples of 1-turn response and 2-turns dialogue generated by different models, given the multi-turn contexts between two speakers as inputs. "Human" lists the reference response in the dataset of the given context. We can see that Case 1 is an interview between an interviewer and a candidate. Our model captures that this is a job interview and produces a question matching the interview situation. In contrast, the responses generated by the baseline models are generic and irrelevant. In Case 2, our model captures the emotion information that speaker A likes the magazine, giving a more specific and informative response, and generates two consecutive turns matching different speakers' roles, while the results generated by

Table 3
Manual evaluation results for 1-turn response generation.

	Dailydialog Corpus				OpenSubtitles Corpus			
Score	0	+1	+2	Average	0	+1	+2	Average
AttnSeq2Seq	68.4%	22.2%	9.4%	0.410	47.9%	34.6%	17.5%	0.708
HRED	55.2%	30.7%	14.2%	0.590	28.6%	50.0%	21.4%	0.928
VHRED	47.7%	30.0%	22.3%	0.747	29.4%	39.8%	30.8%	1.014
HSCJN	**46.2%**	28.8%	**25.0%**	**0.788**	**21.4%**	45.2%	**33.4%**	**1.120**

Table 4
Manual evaluation results for 2-turns dialog generation on the DailyDialog corpus.

Score	0	+1	+2	Average
HRED	54.0%	24.4%	21.6%	0.676
VHRED	64.8%	17.2%	18.0%	0.532
HSCJN	**53.2%**	22.4%	**24.4%**	**0.712**

Table 5
Examples of responses generated by different models given the multi-turn dialogue contexts.

Case 1: 1-turn response generation
Speaker A: Good morning. Are you Mr.Liu? **Speaker B:** My name is Liu Lichi. How do you do? … **Speaker A:** Have you had any working experience? **Speaker B:** Well, I worked at a supermarket during last summer holidays. **Speaker A:** How are your English and computer skills? **Speaker B:** I have passed the CET- 4 and 6. As far as computer is concerned, I can use the computer for word processing. Responses generated by different models: **AttnSeq2Seq:** A great idea. **HRED:** What do you do? **VHRED:** I think so. **HSCJN:** That sounds great. How long have you been interested in the job? **Human:** Okay. Mr.Liu, we'll inform you of the results within a week.
Case 2: 2-turns dialogue generation
Speaker A: Can I borrow this magazine from you? It's really interesting and I can't put it down. **Speaker B:** I am sorry, but I can't lend it to you now, for I haven't finished reading it. If you don't mind, I can lend you some back numbers to you. Responses generated by different models: **HRED:** *Speaker A:* Thank you very much. *Speaker B:* You're welcome. **VHRED:** *Speaker A:* Thank you very much. *Speaker B:* Do you have any questions? **HSCJN:** *Speaker A:* That's great. *Speaker B:* It's too good as you like it. **Human:** *Speaker A:* That would be very kind of you. By the way, is it a monthly magazine? *Speaker B:* No, it is a fortnightly. So, you see, I can get the new one quite soon.

baselines are universal and monotone. It can be found that our model generates obviously better responses with more specific details and higher diversity. Moreover, it also shows that our results are more relevant to the dialogue scenario.

In addition, we have also done an error analysis. Table 6 shows some failure cases in the results of multi-turn dialogues generated by our model. In Case 1, our model captures the keyword "room", but it fails to understand "room service", generating a dialogue about the room reservation. In Case 2, our model encounters the problem of misrepresentation of pronouns. In Case 3, our model encounters trouble in the syntactic collocation of the statement. In the generated dialog, "how much" does not match "check out", and the phrase "a week" in the generated second turn is replied to the time related question, which also does not match the phrase "how much" in the first turn. It can be seen that in the future work, we need to make conversational models deeply understand the overall context of dialog histories, and explore on the more challenging semantic understanding issues.

Table 6
Failure cases of multi-turn dialogues generated by our model given the contexts.

Case 1: context
Speaker A: Room service. What can I do for you? **Speaker B:** I am afraid there isn't any fancy soap in my room. **Generated Result:** *Speaker A:* What kind of room do you want, sir? *Speaker B:* Well, I'd like to have a double room with you.
Case 2: context
Speaker A: It's said that John's broken up. **Speaker B:** Really? But his wife has always waited on him hand and foot. **Generated Result:** *Speaker A:* Yes, he will be perfect for him. *Speaker B:* But he is not a good shot.
Case 3: context
Speaker A: I have to check out today. I'd like my bill ready by 10 this morning. **Generated Result:** *Speaker B:* How much would you like to check out? *Speaker A:* I'd like to check it for a week.

6. Conclusion

In this paper, we investigate the low-diversity issue in dialogue generation task. We propose a Holistic Semantic Constraint Joint Network to introduce future information into the decoding stage. In addition, we devise a maximum entropy regularizer into our loss function to penalize the over-estimation of high-frequency words. In this way, the model can see the entire target sequence and consider all the words in the vocabulary in the learning process. It is worth mentioning that our model introduces the language information in the target sequences from the Seq2Seq model itself for diverse response generation, which does not depend on any external information and variables, also beneficial to capture holistic dialogue semantics to promote relevance. Moreover, our joint learning framework can be generalized to any end-to-end model. Extensive experiments show that our model produces more informative and relevant responses than several competitive baselines.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by The Key Program of National Natural Science Foundation of China under Grant No. U1903213, the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011387), the Dedicated Fund for Promoting High-Quality Economic Development in Guangdong Province (Marine Economic Development Project: GDOE[2019]A45), and the Shenzhen Science and Technology Project under Grant No. ZDYBH201900000002 and No. JCYJ20180508152042002.

References

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. CoRR. abs/1409.0473.
Chen, B., Cherry, C., 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. WMT@ACL 2014, June 26–27, 2014, Baltimore, Maryland, USA, pp. 362–367.
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 Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 5110–5117.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. CoRR. abs/1412.6980.
L'Hostis, G., Grangier, D., Auli, M., 2016. Vocabulary selection strategies for neural machine translation. CoRR. abs/1610.00072.
Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2016. A diversity-promoting objective function for neural conversation models. NAACL HLT 2016, San Diego California, USA, June 12–17, 2016, pp. 110–119.
Li, J., Monroe, W., Jurafsky, D., 2016. A simple, fast diverse decoding algorithm for neural generation. CoRR abs/1611.08562.
Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S., 2017. DailyDialog: a manually labelled multi-turn dialogue dataset. IJCNLP 2017, Taipei, Taiwan, November 27, - December 1, 2017 - Volume 1: Long Papers, pp. 986–995.
Lin, K.-Y., Hsu, C.-C., Chen, Y.-N., Ku, L.-W., 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Hilton Hawaiian Village, Honolulu, Hawaii, USA, January 27-February 1, 2019.
Luong, T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, pp. 1412–1421.
Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA., pp. 311–318.
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 Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA., pp. 3776–3784.
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 Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA., pp. 3295–3301.
Shang, L., Lu, Z., Li, H., 2015. Neural responding machine for short-text conversation. ACL 2015, July 26–31, 2015, Beijing, China, Volume 1: Long Papers, pp. 1577–1586.
Song, Y., Yan, R., Feng, Y., Zhang, Y., Zhao, D., Zhang, M., 2018. Towards a neural conversation model with diversity net using determinantal point processes. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 5932–5939.
Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., Dolan, B., 2015. A neural network approach to context-sensitive generation of conversational responses. NAACL HLT 2015, Denver, Colorado, USA, May 31, - June 5, 2015, pp. 196–205.
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. ACL 2017, Vancouver, Canada, July 30, - August 4, Volume 2: Short Papers, pp. 231–236.
Tiedemann, J., 2012. Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23–25, 2012, pp. 2214–2218.
Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D.J., Batra, D., 2016. Diverse beam search: decoding diverse solutions from neural sequence models. CoRR. abs/1610.02424.
Weng, R., Huang, S., Zheng, Z., Dai, X., Chen, J., 2017. Neural machine translation with word predictions. EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 136–145.
Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., Ma, W., 2017. Topic aware neural response generation. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA., pp. 3351–3357.
Xu, J., Sun, X., Ren, X., Lin, J., Wei, B., Li, W., 2018. DP-GAN: Diversity-promoting generative adversarial network for generating informative and diversified text. CoRR. abs/1802.01345.
Yao, L., Zhang, Y., Feng, Y., Zhao, D., Yan, R., 2017. Towards implicit content-introducing for generative short-text conversation systems. EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 2190–2199.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., Huang, M., 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 4970–4977.
Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., Dolan, B., 2018. Generating informative and diverse conversational responses via adversarial information maximization. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018, Montréal, Canada., pp. 1815–1825.
Zhao, T., Zhao, R., Eskénazi, M., 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. ACL 2017, Vancouver, Canada, July 30, - August 4, Volume 1: Long Papers, pp. 654–664.