Context-aware multi-feature fusion for open-domain dialogue generation

1st Jingbo Liao
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
S200201110@stu.cqupt.edu.cn

2nd Hong Yu
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
yuhong@cqupt.edu.cn

3rd Qi Cheng
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
chengqi0113@gmail.com

4th Ye Wang
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
wangye@cqupt.edu.cn

***Abstract*—With the rapid development of deep learning, open-domain dialogue models are able to generate relatively fluent conversations. However, current generative models present incomplete representations of dialogue features, which make it prone to generate low-quality or irrelevant-semantic responses. To address this problem, this paper presents a novel context-aware multi-feature fusion model, integrating keywords and category-semantic information into the dialogue generation. Specifically, distinct keywords share various significance in a sentence, incorporating this significance into the conversation text can improve the generative ability of the generative model. Our method enhances the feature representation of the text and thus generates high-quality conversation. Finally, we conducted comprehensive experiments on the DailyDialog dataset and EmpatheticDialogues dataset, analyzing the experimental results and verifying the feasibility of our approach.**

***Index Terms*—Open Domain Dialogue, Feature Representation, Generative Models**

I. Introduction

With the increasing number of large-scale corpus, deep learning has yielded fruitful results in the field of open dialogue systems [1] [2]. Unlike task-based dialogue systems [3], open domain dialogue systems are not aimed to perform specific tasks, which indicate that generated responses are more complicated in terms of relevant-semantic and high-diversity. Currently chatbots can be classified into three types, retrieval, generative and knowledge graph respectively. Specifically, retrieval-based chatbots [4] apply techniques such as sorting and matching to extract the most appropriate responses from an existing corpus of conversations. However, retrieval-based chatbots can only generate responses that exist in the corpus and cannot achieve conversational diversity. Moreover, if there are more conversations in the corpus, generating replies will be slower and thus influence the chatting experience. Generative chatbots [5] use the encoder-decoder model to encode conversations into specific feature vectors, and the decoder then decodes the feature vectors into responses. Those generative chatbots can real generate conversations instead of finding templates in the corpus, making the responses more comprehensive than retrieval-based chatbots. On the other hand, the generated dialogues are sampled from the word list, and then the sampled words are combined into replies of the sampled sequence, resulting in ambiguity and incoherence. Meanwhile, generative dialogues are generally based on the maximum likelihood to optimize the model, which leads to a tendency to generate generic responses such as ”I don’t know”, ”I’m sorry”, and other more frequent sentences. The knowledge graph-based chatbot [6] processes conversations through Artificial Intelligence Markup Language templates and handles complex logic through logical reasoning with knowledge graphs and then expresses the complex logic using language based on the use of neural networks. However, a knowledge graph-based chatbot necessitates not just the creation of a corpus, but also the creation of high-quality knowledge graphs that match the corpus’ expertise. As a result, the generative approach remains the dominant strategy in open domain discussion systems. For example, the Seq2seq model [7] is one of the earliest end-to-end generation models and has made a significant contribution to the field of text generation. The Seq2seq model contains two recurrent neural networks, the encoder, and decoder, respectively. The encoder encodes the input sequence into a semantic vector, and the decoder decodes the semantic vector into an output sequence. Most of the dialogue systems are basically based on the seq2seq paradigm. However, both RNN encoding and decoding sequences must be executed in an autoregressive order from left to right, which is challenging to parallelize. As a result, processing long sequences have a significant temporal complexity. What’s more, RNNs have difficulty modeling long-range context dependencies, leading to inaccurate semantics of the generated dialogues. The attention mechanism [8] [9] can strengthen the dependencies between long texts, so it is often

used to solve the problem that RNNs do not have enough dependencies for the earlier encoded parts when modeling long sequences. The Transformer [10] model abandons the basic paradigm of using a circular recursive structure to encode sequences and instead takes the idea of attention mechanisms to the extreme by computing the hidden state of a sequence based entirely on a global attention mechanism. The global self-attention mechanism allows the hidden state at each position of the sequence to be directly associated with all positions in the sequence, and thus enables better modeling of dependencies in long sequences than RNNs. Variational autoencoders (VAEs) [11] encode text as a probability distribution in the latent space rather than as a deterministic vector. So VAEs are able to produce different outputs from the same input by sampling in the latent space. We can also add some controlled variables when sampling the latent space, to achieve controllability in dialogue generation. To summarise our contributions:

- We present a novel context-aware multi-feature fusion model, integrating keywords and category-semantic information into the dialogue generation.
- We provide advanced feature representation to the decoder by fusing additional feature representations with the original features, so that the generative model can generate higher quality dialogues.
- We apply the proposed model in various experiments on the DailyDialog dataset and EmpatheticDialogues dataset for the validation, and we also analyze several scenarios of the model comprehensively.

II. Related Works

A. *Seq2seq for Dialog System*

Sutskever et al. [7] first proposed an end-to-end sequence model with good results on various text generation tasks. Cho [12] improved the quality of text generation by adding a fixed-dimension vector to seq2seq and using this vector as input for each step of the decoder. Traditional seq2seq models tend to generate safe, universal answers due to the fact that traditional models are based on maximum likelihood functions to optimize the model. To solve this problem, Li et al. [13] use the maximum mutual information instead of the maximum likelihood function as the new objective function with the aim of using the mutual information to reduce the generation probability of boring responses. In addition, dialogue models that use the maximum likelihood function as the objective function are prone to cause dead loops in the dialogue. To solve the dead-loop problem, Li et al. [14] added reinforcement learning strategies to a seq2seq model that uses mutual information as the objective function, and finally optimized the parameters of the model by maximizing the expected reward. The traditional seq2seq model is unable to incorporate external knowledge, so it generates less informative responses. Lian et al. [15] proposed to add a knowledge encoder and knowledge selector to the end-to-end model. So the model can use the posterior knowledge to guide the training of the prior knowledge distribution, thus achieving the purpose of effective knowledge selection and integration. In addition, common generative models are based on conversations generated from an identical word list, making these models vulnerable to generic patterns and irrelevant noise. Therefore, Wu et al. [17] proposed a dynamic vocabulary sequence-to-sequence model (DVS2S), which makes each input possess its own vocabulary at the time of decoding. Compared with the ordinary seq2seq model, the DVS2S model generates dialogue texts with many fewer generic responses and irrelevant vocabularies. To solve the problem of semantic dependencies between input and output, Luo et al. [16] proposed the AEM model to learn the semantic dependencies at the corpus level and thus generate more continuous text.

B. *VAEs for Dialog System*

Kingma et al. [11] first proposed variational autoencoders (VAEs) and applied them to the image domain. Later studies found that VAEs can achieve better generation results in the field of text generation. Kihyuk et al. [18] proposed the conditional variational autoencoder (CVAE). The conditional encoder of CVAE is able to encode replies and some other semantics in dialogue texts, such a structure is more suitable for dialogue generation tasks. At the same time, the conditional encoder is able to encode semantics such as sentiment, which enables the generation of controlled text. Zhao et al. [19] applied conditional variational autoencoders to the dialogue generation task for the first time and proposed Knowledge-Guided CVAE. Knowledge-Guided CVAE can focus not only on the features of responses but also on higher-level topic features. Cui et al. [20] proposed a CVAE-based model to explicitly model content relevance and sentiment consistency jointly, resulting in richer and more complementary information. Shen et al. [21] proposed a framework for hierarchical conditional variational autoencoders and addressed the issues of diversity and validity at the global phrase level. Wang et al. [22] proposed a semantic-aware conditional variant autoencode to solve the one-to-many problem by utilizing embedded classifiers and feature decoupling modules.

Unlike these works, we proposed a method to enhance the representation of dialog text features and apply it to the seq2seq model and the CVAE model. Each sentence of dialogue text contains one or more keywords. These keywords take up most of the information in the conversation, and ordinary encoders treat all text as the same priority, thus ignoring this important information. Different from the Attention mechanism, we do not change the structure of the encoder and decoder. We extract the keywords of each sentence in the dataset by a keyword extraction model and add a keyword encoder to encode these keywords. The keyword vector is then combined with a semantic vector to obtain a feature vector containing rich semantics. Finally, we fuse this feature vector with the

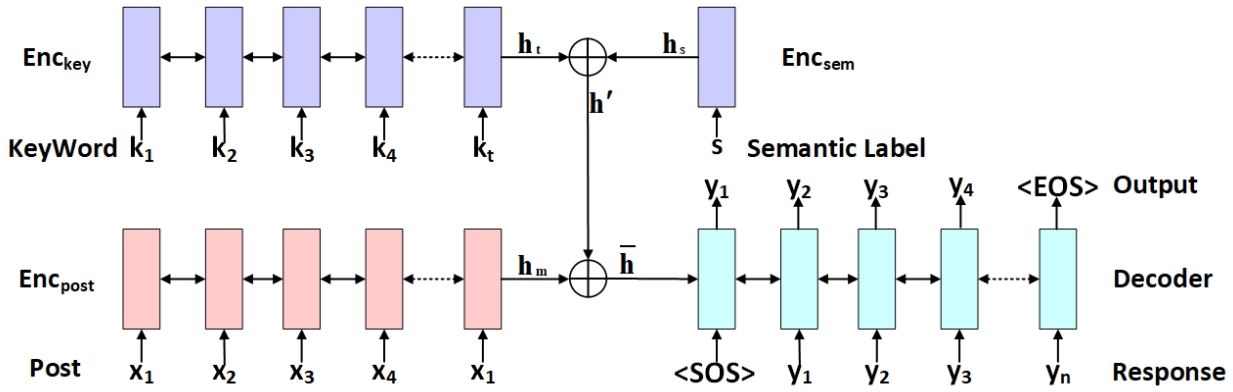


Fig. 1. Applied in seq2seq-based model

text vector of the dialogue, which is used as the initial input to the decoder. The feature representation is enhanced by adding external knowledge so that the model can generate higher-quality dialogue text.

III. Approach

A. Applied in seq2seq-based model

In this section, we describe how to apply our method to the Seq2seq model (Fig 1). $X = (x_1, x_2, \ldots, x_m)$ denotes the sequence of questions in the dialogue, where x_i represents the i-th word in the sequence of questions. $Y = (y_1, y_2, \ldots, y_n)$ denotes the response sequence of the dialogue model, where y_i represents the i-th word in the response sequence. The seq2seq model models the conditional probability distribution $P(Y|X)$by learning to encode X and decode Y. We added two inputs to the normal seq2seq model. $K = (k_1, k_2, \ldots, k_t)$ denotes the keywords in the problem sequence, where k_i represents the i-th word in the keyword sequence. We use the TextRank algorithm to extract the keywords of the question sequence, which results in the keyword sequence K. S denotes the semantic category of the conversation,which is the semantic label in the dataset. $e(.)$ denotes word embedding and $e(x_i)$ denotes embedding the word x_i into the word vector space. $Enc_{post}(.)$ denotes question encoder, $Enc_{key}(.)$ denotes keyword encoder, $Enc_{sem}(.)$ denotes semantic encoder, $Dec(.)$ denotes the decoder.

First, we encode the problem sequence by Eq. 1 to get the vector h_m. The keyword sequence is encoded by Eq. 2 to obtain the vector h_t. The semantic categories are encoded by Eq. 3, and the vector h_s is obtained.

$$h_m = Enc_{post}(e(X)) \tag{1}$$

$$h_t = Enc_{key}(e(K)) \tag{2}$$

$$h_n = Enc_{sem}(e(S)) \tag{3}$$

Next, h_t and h_s are concatenated as [h_t:h_s] and [h_t:h_s] are fused into h' using a multilayer perceptron, as shown in Eq. 4. h' and h_m are concatenated as [h':h_m] and [h':h_m] are fused into $\bar{h}$ using a multilayer perceptron, as shown in Eq. 5. $\bar{h}$ will be used as the input to the decoder.

$$\boldsymbol{h}' = \mathrm{MLP}\left([h_k : h_s]\right) \tag{4}$$

$$\bar{h} = \mathrm{MLP}\left([h' : h_m]\right) \tag{5}$$

For training, the first input of the decoder is the $\langle SOS \rangle$ token, and the input of each subsequent step is the words in the response sequence. Finally, the output of the decoder is mapped to the word table space using softmax, and the word with the highest probability in the word table is taken as the generated result, as shown in Eq. 6, 7. In the formula, s_t represents the hidden state of the decoder in the next step, and $\bar{y}_t$ represents the result currently generated by the decoder.

$$s_t = \mathrm{Dec}\left(s_{t-1}, e\left(y_{t-1}\right), \bar{h}\right) \tag{6}$$

$$\bar{y}_t = \mathrm{SoftMax}\left(\mathrm{MLP}\left(s_t\right)\right) \tag{7}$$

Finally, we use cross-entropy as the loss function and optimize our model according to the loss function, as in Eq. 8.

$$\mathcal{L}_{seq} = -[y \log \bar{y} + (1 - y) \log(1 - \bar{y})] \tag{8}$$

B. Applied in CVAE-based model

In this subsection we describe how to apply our approach to CVAEs model, as shown in Fig. 2. Compared to VAEs, CVAEs are more suitable for conversational tasks because the output encoder of CVAEs can encode reply sequences. In CVAE, $C = \{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_m\}$ is the sequence of questions in the conversation, $X = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ is the sequence of responses in the conversation. After introducing the latent variable z, the conditional probability distribution $P(X \mid C, K, S)$ is decomposed into Eq. 9:

$$P = \int_z p(z \mid C, K, S) P(X \mid z, C, K, S) dz \tag{9}$$

On the basis of Eq. 9, we can decompose to obtain the evidence of lower bound (ELBO), as shown in Eq. 10.

$$\begin{aligned} \mathrm{ELBO} = & E_{q_\phi(z|X,C,K,S)}\left[\log P_\theta(X \mid z, C, K, S)\right] \\ & - KL\left(q_\phi(z \mid X, C, K, S) \| p_\varphi(z \mid C, K, S)\right) \end{aligned} \tag{10}$$

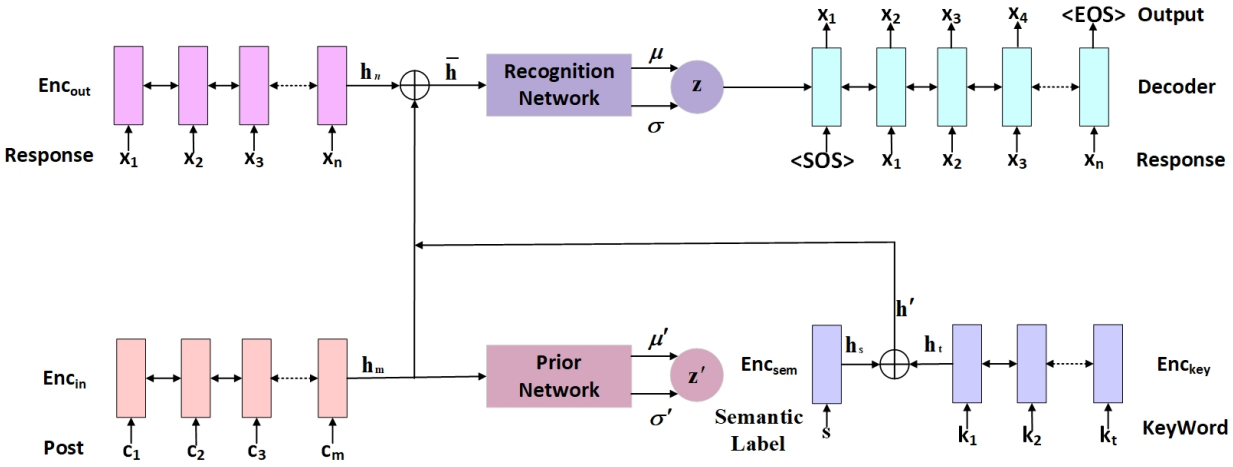


Fig. 2. Applied in CVAE-based model

$E_{q_\phi(z|X,C,K,S)}[\log P_\theta(X \mid z,C,K,S)]$ is the reconstruction loss. $q_\phi(z \mid X,C,K,S)$ is the approximate posterior distribution. Since the sampled process is not derivable, we use a reparameterization trick instead of the sampled process to make the whole process derivable, as in Eq. 11. $P_\theta(X \mid z,C,K,S)$ is the conditional probability distribution of X reconstructed under the condition that z,C,K,S is known.

$$z = \mu + \sigma * \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I) \tag{11}$$

The second loss $KL\left(q_\phi(z \mid X,C,K,S) \| p_\varphi(z \mid C,K,S)\right)$ is the KL-Distance between prior distribution $p_\varphi(z \mid C,K,S)$ and the approximate posterior distribution $q_\phi(z \mid X,C,K,S)$. The prior distribution obeys the multivariate Gaussian distribution $\mathcal{N}\left(\mu', \quad \sigma'^2 I\right)$, and the approximate posterior distribution obeys $\mathcal{N}\left(\mu, \quad \sigma^2 I\right)$. μ, σ^2, μ', σ'^2 are computed by the a priori network and the recognition network, respectively. ϕ, φ, θ are the parameters of the probability distribution.

CVAEs model contains the input encoder $\mathrm{Enc}_{in}(\cdot)$, the output encoder $\mathrm{Enc}_{out}(\cdot)$, the recognition network $\mathrm{RecogNetq}_\phi(z \mid X,C,K,S)$, the prior network $\mathrm{PriorNetp}_\varphi(z \mid C,K,S)$, keyword encoder $Enc_{key}(.)$, the semantic encoder $Enc_{sem}(.)$ and the decoder $\mathrm{Dec}(\cdot)$. As with seq2seq, the corresponding sequences are encoded using different encoders to obtain h_m, h_n, h_s and h_t, respectively. h_m is the output of the input encoder, h_n is the output of the output encoder, h_s is the output of the semantic encoder, and h_t is the output of the keyword encoder. h_t and h_s are concatenated as $[h_t{:}h_s]$ and a multilayer perceptron is used to fuse $[h_t{:}h_s]$ into h', as shown in Eq. 4. h' , h_m and h_n are concatenated into $[h'{:}h_m{:}h_n]$ and fuse $[h'{:}h_m{:}h_n]$ into $\bar{h}$ using a multilayer perceptron. $\bar{h}$ will be used as the input to the recognition network $\mathrm{RecogNetq}_\phi(z \mid X,C,K,S)$. h' and h_n are concatenated into $[h'{:}h_n]$, and $[h'{:}h_m]$ are fused into $\bar{h}'$ using a multilayer perceptron. The $\bar{h}'$ will be used as the input to the prior network $\mathrm{PriorNetp}_\varphi(z \mid C,K,S)$. With Eq. 12, we obtain the parameters μ, σ^2, μ', σ'^2 of the approximate posterior and prior distributions.

$$\begin{aligned} \mu, \sigma &= \mathrm{RecogNetq}_\phi(z \mid X,C,K,S) \\ \mu', \sigma' &= \mathrm{PriorNetp}_\varphi(z \mid C,K,S) \end{aligned} \tag{12}$$

The latent variable z is reparameterized by Eq. 11, and z is initialized to the initial hidden state of the decoder by a linear transformation. The decoding process is the same as seq2seq. The final loss function is shown in Eq. 13.

$$\begin{aligned} \mathcal{L}_{\mathrm{CVAE}} =& KL\left(q_\phi(z \mid X,C,K,S) \| p_\varphi(z \mid C,K,S)\right) \\ & - E_{q_\phi(z|X,C,K,S)}[\log P_\theta(X \mid z,C,K,S)] \end{aligned} \tag{13}$$

IV. Experiments

A. Datasets

We use two datasets: the DailyDialog dataset (DD dataset) [23] and the EmpatheticDialogues dataset (ED dataset) [24]. Many of today's conversation datasets are not from real conversations. The DailyDialog dataset is a real multi-round conversation dataset for everyday chat scenarios, and compared to previous corpora, the DD dataset has less noise and covers several major topics in life. Each conversation has two labels, action and emotion. However, the distribution of different categories is uneven, and we equalize the dataset so that the conversations in each category are relatively evenly distributed. The EmpatheticDialogues dataset contains 24850 conversations, collected from 810 different participants, and each sentence is labeled with an emotion tag, for a total of 32 types of emotions. we aggregated the fine-grained tags into two coarse-grained tags: positive, and negative.

B. Experimental Settings

Both Seq2seq and CVAE are based on the encoder-decoder framework. The encoder is a two-layer LSTM [25] of hidden size 300. the parameters and latent variables of the CVAE latent space distribution are both 200-dimensional. The recognition and prior networks are multilayer perceptrons with the

TABLE I
THE AUTOMATIC EVALUATION RESULTS OF ALL COMPARED METHODS IN THE DAILYDIALOGS DATASET WITH ACTION LABEL.

Method	BLEU	METEOR	ROUGE	dist-2
VAE	10.85±0.10	11.52±0.11	30.73±0.11	0.14
Transformer	10.11±0.08	11.16±0.09	**33.09±0.11**	**0.25**
Seq2seq	10.14±0.08	10.52±0.09	30.15±0.11	0.17
CVAE	11.50±0.11	11.95±0.12	31.12±0.11	0.14
seq2seq+ours	11.22±0.11	11.56±0.12	31.31±0.11	0.19
CVAE+ours	**11.75±0.13**	**12.06±0.13**	31.62±0.12	0.16

TABLE II
THE AUTOMATIC EVALUATION RESULTS OF ALL COMPARED METHODS IN THE DAILYDIALOGS DATASET WITH EMOTION LABEL.

Method	BLEU	METEOR	ROUGE	dist-2
VAE	10.01±0.17	10.48±0.19	28.09±0.21	0.18
Transformer	10.44±0.19	11.03±0.21	**31.59±0.25**	**0.42**
Seq2seq	11.15±0.22	11.89±0.26	30.99±0.25	0.30
CVAE	11.17±0.22	11.50±0.25	29.46±0.23	0.25
seq2seq+ours	11.40±0.26	**12.23±0.27**	31.23±0.27	0.34
CVAE+ours	**11.62±0.24**	11.75±0.25	30.44±0.24	0.29

TABLE III
THE AUTOMATIC EVALUATION RESULTS OF ALL COMPARED METHODS IN THE EMPATHETICDIALOGUES DATASET WITH EMOTION LABEL.

Method	BLEU	METEOR	ROUGE	dist-2
VAE	7.40±0.09	8.37±0.10	30.37±0.12	0.19
Transformer	6.40±0.08	7.90±0.09	31.69±0.13	**0.41**
Seq2seq	6.86±0.08	8.04±0.10	31.03±0.12	0.21
CVAE	7.37±0.09	8.39±0.10	30.79±0.12	0.22
seq2seq+ours	7.25±0.09	8.23±0.11	31.35±0.12	0.25
CVAE+ours	**7.62±0.10**	**8.46±0.12**	**31.75±0.12**	0.22

number of hidden layer neurons set to 250. And our model uses the Adam optimizer. The initial learning rate is 0.0001, the learning rate decays by 0.01, and the coefficient of the kl term increases to 1 after 10,000 updates.

C. Automatic Evaluation

We used three metrics to assess the quality of the generated dialogue text and one metric to assess the diversity of the generated text. **BLEU** [26] is a method that compares the model output with the n-gram of the reference answer and calculates the number of matching fragments. These matching fragments are independent of their positions in the context, and here it is only considered that the higher the number of matching fragments, the better the quality of the model output. The advantages of this metric are that it is convenient, fast and the results are informative. However, there are many disadvantages, such as not considering the accuracy of language expressions, the measurement accuracy will be disturbed by common words, the measurement accuracy of short sentences is sometimes higher, and not considering the case of synonyms or similar expressions, which may lead to the negation of reasonable texts. **METEOR** [27] generates a clear demarcation line between the candidate answer and the target response. This demarcation line is determined based on a certain priority order, from highest to lowest: specific sequence matches, synonyms, roots and affixes, and paraphrases. With the demarcation line, METEOR can evaluate the results by taking the summed average of the precision and recall of the reference answer and the model output. **ROUGE** [28] is obtained by computing the F-measure for the longest common subsequence (LCS) between the candidate sentence and the target sentence. LCS is a set of word sequences that occur in the same order in both sentences, unlike n-gram, LCS does not need to remain consecutive, i.e., other words can occur in the middle of LCS. **Distinct-n** [13] uses the ratio of the number of non-repeating n-grams in the text to the number of all n-grams in the response as a diversity evaluation metric. the larger the Distinct-n value, the better the diversity of the generated text. The advantage is that it largely reflects the formal diversity of sentences and is less expensive to compute. The disadvantage is that the input of the calculation is the sentences generated by the model for the whole test set, and then the small sample on the small test set may not appear to be an objective evaluation.

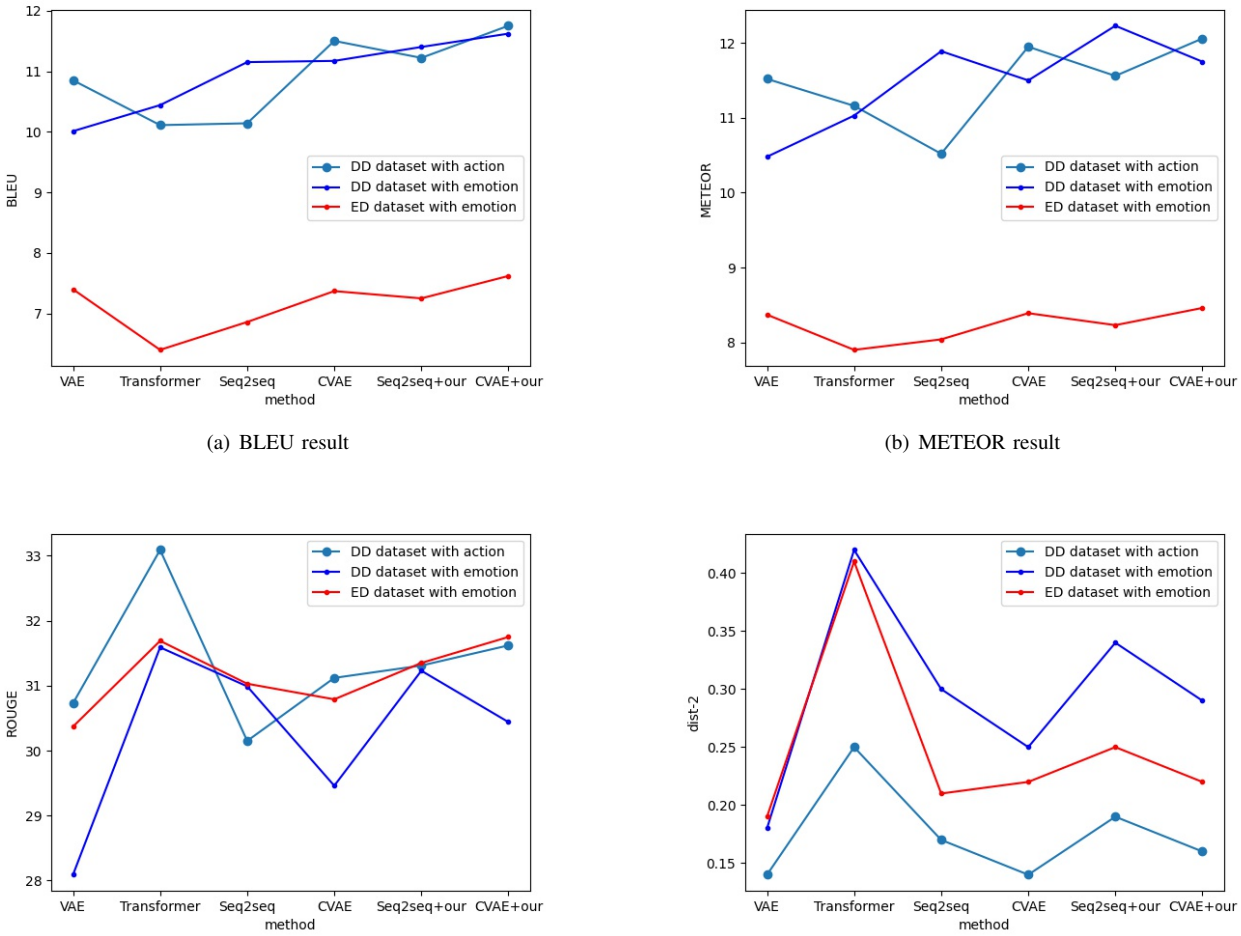
D. Baseline

We performed a full comparison using our model and baseline.

Seq2seq [7], is an end-to-end sequential model that is trained by autoregressive means.

Transformer [10], using the self-attention mechanism to encode and decode the text. Compared to Seq2seq, Transformer is able to encode long text dependencies, resulting in a richer representation;

VAE [11], generating diverse dialogues by encoding the dialogue data into a probability distribution and then sampling from the distribution as a representation of the text;



(a) BLEU result

(b) METEOR result

(c) ROUGE result

(d) dist-2 result

Fig. 3. Automatic evaluation of visualization results for different generation models.

CVAE [18], adding an output encoder to the VAE, which can generate category-controlled dialogues by encoding some control conditions.

E. Results and analysis

1) Quantitative Analysis: Tables I, II, III show the results of the automatic evaluation of our model and the compared models. Table I shows the results on the DailyDialog dataset with action label. Table II shows the results on the DailyDialog dataset with emotion label. Table III shows the results on EmpatheticDialogues dataset with emotion label. fig. 3 visualizes the automatic evaluation results, and we can observe the model generation effect more intuitively.

From the experimental results, we can see that our model achieves good results on all three different datasets. Our model outperforms the other comparative models in both BLEU and METEOT and is just below Transformer in Rouge. The results show that our model can generate higher-quality conversational text. Our approach has been improved on different datasets, indicating that it has good generalization ability. However, after adding our method, the error bar of the model becomes higher, which indicates that although adding our method can improve the dialogue model generation, it will make the model more complex and thus the model fluctuation becomes larger. In the future, we can try to figure out how to lower the model's fluctuation while enhancing the generative power, resulting in a model with stable generative power.

2) Ablation study: In this section, we will explore the impact of our approach on the model. We applied our method to Seq2seq and CVAE and compared it with the original model. As we can see from the tables I, II, III, with the addition of our method, both Seq2seq and CVAE achieve different degrees of improvement in each metric. The improvement in the three indicators of BLEU, METEOR, and ROUGE indicates that the importance of different words in a sentence of dialogue is different. We extract the keywords in the dialogues and use the keyword information and categories as the prior knowledge of the dialogues. Then we fuse the prior knowledge into the dialogue model by the neural network to be able to enhance the feature representation of the dialogue text, thus improving

the generation ability of the dialogue model. It can be seen that our method has different lifting effects on Seq2seq and CVAE. In general, the enhancement is greater for the Seq2seq model. This is since the fact that CVAE encodes the dialogue text into a continuous latent space, and the encoder-level knowledge has a limited impact on the model. In contrast, Seq2seq encodes the dialogue into a fixed vector, and our approach can enrich the representation of this vector. The improvement in the dis-2 metric indicates that the enhanced feature representation also improves the diversity of the generated text by the generative model after using our approach. Models optimized based on the maximum likelihood function are prone to generate generic responses, and our approach mitigates this to some extent.

V. Conclusion

In this paper, we propose a novel context-aware multi-feature fusion model that can fuse keyword and category-semantic information to the dialog generation. Besides, incorporating the importance information, the model can enhance the feature representation. Meanwhile, We apply our approch to Seq2seq-based and CVAE-based dialogue generation models. The experimental results show that our approach can improve the feature representation of the dialogue model and generate higher-quality dialogue texts. In the future, we consider investigating how to reduce the fluctuation of dialogue models while introducing external models to generate dialogue texts with more stable quality.

Acknowledgment

This work was partly supported by National Key R&D Program of China (2021YFF0704100), the National Natural Science Foundation of China (62136002, 62102057 and 61876027), the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202100627 and KJQN202100629), and the National Natural Science Foundation of Chongqing (cstc2019jcyj-cxttX0002), respectively.

References

[1] T. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P. Su, S. Ultes and S. J. Young, "A Network-based End-to-End Trainable Task-oriented Dialogue System," in EACL, 2017, pp. 438–449.

[2] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in AAAI, 2016, pp. 3776–3784.

[3] Z. Zhang, X. Li, J. Gao and E. Chen, "Budgeted Policy Learning for Task-Oriented Dialogue Systems," in ACL, 2019, pp. 3742–3751.

[4] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. Xin Zhao, D. Yu and H. Wu, "Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network," in ACL, 2018, pp. 1118–1127.

[5] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao and B. Dolan, "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses," in NAACL, 2015, pp. 196–205.

[6] Y. Sun, Y. Hu, L. Xing, J. Yu and Y. Xie, "History-Adaption Knowledge Incorporation Mechanism for Multi-Turn Dialogue System," in AAAI, 2020, pp. 8944–8951.

[7] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in NIPS, 2014, pp. 3104–3112.

[8] T. Luong, H. Pham and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in EMNLP, 2015, pp. 1412–1421.

[9] Y. Wang, H. Wang, X. Zhang, T. Chaspari, Y. Choe and M. Lu, "An attention-aware bidirectional multi-residual recurrent neural network (abmrnn): A study about better short-term text classification," in ICASSP, 2019, pp. 3582–3586.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need," in NIPS, 2017, pp. 5998–6008.

[11] Kingma, P. Diederik and M. Welling, "Auto-Encoding Variational Bayes," in IICLR, 2014.

[12] K. Cho, B. Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in EMNLP, 2014, pp. 1724–1734.

[13] J. Li, M. Galley, C. Brockett, J. Gao and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models," in NAACL, 2016, pp. 110–119.

[14] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley and J. Gao, "Deep Reinforcement Learning for Dialogue Generation," in EMNLP, 2016, pp. 1192–1202.

[15] R. Lian, M. Xie, F. Wang, J. Peng and H. Wu, "Learning to Select Knowledge for Response Generation in Dialog Systems," in IJCAI, 2019, pp. 5081–5087.

[16] L. Luo, J. Xu, J. Lin, Q. Zeng and X. Sun, "An Auto-Encoder Matching Model for Learning Utterance-Level Semantic Dependency in Dialogue Generation," in EMNLP, 2018, pp. 702–707W.

[17] Y. Wu, W. Wu, D. Yang, C. Xu and Z. Li, "Neural Response Generation With Dynamic Vocabularies," in AAAI, 2018, pp. 5594–5601.

[18] Sohn, K. Lee, Honglak and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in NIPS, 2015, pp. 3483–3491.

[19] T. Zhao, R. Zhao and M. Eskénazi, "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders," in ACL, 2017, pp. 654–664.

[20] Z. Cui, Ya. Li, J. Zhang, J. Cui, C. Wei and B. Wang, "Focus-Constrained Attention Mechanism for CVAE-based Response Generation," in EMNLP, 2020, pp. 2021–2030.

[21] S.Shen, Y. Li, N. Du, X. Wu, Y. Xie, S. Ge, T. Yang, K. Wang, X. Liang and W. Fan, "On the Generation of Medical Question-Answer Pairs," in AAAI, 2020, pp. 8822–8829.

[22] Y. Wang, J. Liao, H. Yu and J. Leng, "Semantic-aware Conditional Variational Autoencoder for one-to-many dialogue generation," Neural Computing and Applications, 2022, pp. 1–13.

[23] Y. Li, H. Su, X. Shen, W. Li, Z. Cao and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," in IJCNLP, 2017, pp. 986–995.

[24] H. Rashkin, E. M. Smith, M. Li and Y. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in ACL, 2019, pp. 5370–5381.

[25] Y. Wang, X. Zhang, M. Lu, H. Wang and Y. Choe, "Attention augmentation with multi-residual in bidirectional LSTM," Neurocomputing, vol. 385, 2020 ,pp. 340–347.

[26] K. Papineni, S. Roukos, T.Ward and W. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in ACL, 2002 ,pp. 311–318.

[27] C. Lin, "Rouge: A package for automatic evaluation of summaries," Text summarization branches out, 2004 ,pp. 74–81.

[28] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in ACL, 2005 ,pp. 65–72.