LEARNING TO SELECT CONTEXT IN A HIERARCHICAL AND GLOBAL PERSPECTIVE FOR OPEN-DOMAIN DIALOGUE GENERATION

Lei Shen¹,²,*, Haolan Zhan²,*, Xin Shen³, Yang Feng¹,²

¹IIP, Institute of Computing Technology, Chinese Academy of Sciences
²University of Chinese Academy of Sciences, ³Australian National University

ABSTRACT
Open-domain multi-turn conversations mainly have three features, which are hierarchical semantic structure, redundant information, and long-term dependency. Grounded on these, selecting relevant context becomes a challenge step for multi-turn dialogue generation. However, existing methods cannot differentiate both useful words and utterances in long distances from a response. Besides, previous work just performs context selection based on a state in the decoder, which lacks a global guidance and could lead some focuses on irrelevant or unnecessary information. In this paper, we propose a novel model with hierarchical self-attention mechanism and distant supervision to not only detect relevant words and utterances in short and long distances, but also discern related information globally when decoding. Experimental results on two public datasets of both automatic and human evaluations show that our model significantly outperforms other baselines in terms of fluency, coherence, and informativeness.

Index Terms— Open-domain Dialogue Generation, Context Selection, Hierarchical and Global Perspective

1. INTRODUCTION
Open-domain multi-turn dialogue generation has gained increasing attentions in recent years, as it is more accordant with real scenarios and aims to produce customized responses. In general, an open-domain multi-turn conversation has following features: (1) The context (including the query and previous utterances in our paper) is in a hierarchical structure, which means it consists of some utterances, and each utterance contains several words. (2) At most cases, many contents of the context are redundant and irrelevant to the response. (3) Some related information (utterances or words) and the response are in a long-term dependency relation. Therefore, Context Selection, detecting the relevant context based on which to generate a more coherent and informative response, is a key point in multi-turn dialogue generation.

Based on feature (1), the hierarchical recurrent encoder-decoder network (HRED) [1] has been proposed. It encodes each utterance and the whole context at two levels, and is widely applied to other methods for multi-turn dialogue generation. Then, hierarchical recurrent attention [2] and explicit weighting [3, 4], memory networks [5] and self-attention mechanism [6] have been introduced to match feature (2) and (3), respectively. However, few work could cover all these features simultaneously to fulfill context selection and response generation tasks.

When it comes to Context Selection, existing methods can be categorised into two ways: (1) Detecting related utterances measured by the similarity between query and each previous utterance [3, 4]. (2) Applying the attention mechanism from a local perspective, i.e., based solely on the current state in decoder with the Maximum Likelihood Estimation (MLE) loss [4, 6]. The similarity measurement in the former cannot select word-level context, while the guidance from the local perspective in the latter would make the model choose some deviated context and produce an inappropriate response [7, 8, 9].

To tackle the above mentioned problems, we propose HiSA-GDS, a modified Transformer model with Hierarchical Self-Attention and Globally Distant Supervision. To the best of our knowledge, it is the first time to design these two modules for open-domain dialogue generation. Specifically, we use Transformer encoder to encode each utterance in the context. During training, the response is firstly processed by a masked self-attention layer, and then a word-word attention aggregates related word information in each utterance individually. After that, we conduct utterance-level self-attention to get context-sensitive representations of aggregated information from last layer. Then, we calculate the attention weights between utterance-level outputs of the previous layer and the masked response representation. Finally, we generate the corresponding response based on the fusion of selected information at both word and utterance levels. Besides, to provide a global guidance of decoding, we import a distant supervision module which utilizes the similarity score between the response and each contextual utterance measured by a pre-trained sentence-embedding model. All parameters are learned based on the global Distant Supervision and local MLE in an end-to-end framework.

Experimental results on two public datasets along with further discussions show that HiSA-GDS significantly outper-

---

¹ IIP stands for Key Laboratory of Intelligent Information Processing.
* Equal Contribution. Corresponding to: shenlei17z@ict.ac.cn
forms other baselines and is capable to generate more fluent, coherent, and informative responses.

2. APPROACH

The input is a context containing $n$ utterances $\{X_i\}_{i=1}^n$, and each utterance is defined as $X_i = \{x_{i,1}, \ldots, x_{i,n_{X_i}}\}$, where $|X_i|$ is the length of the $i$-th utterance and $x_{i,m}$ is the $m$-th word of $X_i$. Our goal is to select relevant context consisting of utterances and words, and then generate a response $Y = \{y_1, y_2, \ldots, y_{y_Y}\}$ by utilizing the related information, where $|Y|$ is the length of response $Y$.

2.1. Encoder

We consider each utterance independently, and given an utterance $X_i$, the input representation of word $x_{i,j}$ is the sum of its word embedding and position encoding: $I(x_{i,j}) = \text{WE}(x_{i,j}) + \text{WPE}(x_{i,j})$, where $\text{WE}(x_{i,j})$ and $\text{WPE}(x_{i,j})$ represent word and word position embedding, respectively. The input embedding is then fed into Transformer encoder with $N$ layers. The final encoding of $X_i$ is the output from the $N$-th layer, $E_i^{(N)}$. Please refer to [10] for more details.

2.2. Hierarchical Self-Attention based Decoder

The decoder also contains $N$ layers, and each layer is composed of five sub-layers. The first sub-layer is a masked self-attention, which is defined as:

$$M_i^{(l)} = \text{MHA}(D_i^{(l-1)}, D_i^{(l-1)}, D_i^{(l-1)}),$$  \hspace{1cm} (1)

where $\text{MHA}$ is the multi-head attention function, $D_i^{(l-1)}$ denotes the input representation of the $l$-th layer, and $M_i^{(l)}$ denotes the output of masked self-attention at the $l$-th layer. $D_i^{(0)}$ is the concatenated result of all words before time step $t$ in the response and each word is also represented as the sum of its word embedding and position encoding.

The second sub-layer is a word-word attention that summarizes word-level response-related information from each utterance $X_i$ into a vector at a specific decoding time:

$$U_i^{(l)} = \text{MHA}(f_w(M_i^{(l)}, E_i^{(N)}, E_i^{(N)}),$$  \hspace{1cm} (2)

where $f_w$ is a linear transformation.

The third sub-layer is an utterance-level self-attention. Inspired by Zhang et al. [6], we also utilize the self-attention mechanism to capture the long-term dependency of utterance-level information. Similar to word position encoding, we add utterance position encoding (UPE) to $U_i^{(l)}$, and denote the sum result as $\tilde{U}_i^{(l)}$. The output of this sub-layer is calculated as:

$$H_i^{(l)} = \text{MHA}(\tilde{U}_i^{(l)}, \tilde{U}_i^{(l)}, \tilde{U}_i^{(l)}),$$  \hspace{1cm} (3)

where $\tilde{U}_i^{(l)} = [\tilde{U}_i^{(l,1)}, \tilde{U}_i^{(l,2)}, \ldots, \tilde{U}_i^{(l,n)}]$. Then, the fourth sub-layer is a word-utterance attention layer to find out utterance-level relevant information which is defined as:

$$C_i^{(l)} = f_i(\text{MHA}(f_u(M_i^{(l)}, H_i^{(l)}, H_i^{(l)})),$$  \hspace{1cm} (4)

where $f_i$ and $f_u$ are linear transformations, and $f_i$ is used for changing the output dimension. The last sub-layer is a feed-forward neural network (FFN):

$$F_i^{(l)} = \text{FFN}(C_i^{(l)}).$$  \hspace{1cm} (5)

Each of above mentioned sub-layer is followed by a normalization layer and a residual connection. Finally, we use a fusion gate to regulate the relevant information at word level ($U_i^{(l)}$) and utterance level ($F_i^{(l)}$):

$$\lambda_t = \sigma(W_g[U_{t,n}F_i^{(l)}]),$$  \hspace{1cm} (6)

$$D_i^{(l)} = \lambda_t * F_i^{(l)} + (1 - \lambda_t) * U_{t,n},$$  \hspace{1cm} (7)

where $W_g$ is parameter metric, $\sigma$ is the sigmoid activation function, and $\ast$ means the point-wise product.

2.3. Globally Distant Supervision

Previous attention-based models achieve context selection from a local perspective, i.e., they try to generate one token at a time based solely on the current decoding state, which would detect deviated context and mislead the further generation. Besides, we do not have manual annotations to provide direct signals for selection. To address these problems, we design a globally distant supervision module to help determine relevant information, which provides a global guidance...
for the response generation process. Firstly, we apply a high quality pre-trained sentence-embedding model to encode contextual utterance \( X_i \) and response \( Y \) into vectors, denoted as \( x_i \) and \( y \). Then, we use the dot product to measure the semantic relevance between \( x_i \) and \( y \), and compute the selection probability as follows:

\[
P(x = x_i | y) = \frac{\exp(x_i \cdot y)}{\sum_{j=1}^{n} \exp (x_j \cdot y)}. \tag{8}
\]

2.4. Training Objective

We utilize three loss functions in our training process. The first one is MLE loss which is defined as:

\[
L_{MLE}(\theta) = - \frac{1}{|Y|} \sum_{t=1}^{|Y|} \log p(y_t|y_{<t}, \{X_i\}_{t=1}^{n}; \theta), \tag{9}
\]

where \( \theta \) represents the model parameters, and \( y_{<t} \) denotes the previously generated words. Since MLE loss only provides local (token-wise) supervision, inspired by Ren et al. [12] and Zhan et al. [13], we apply the Kullback-Leibler divergence (KL) loss and the Maximum Causal Entropy (MCE) loss for globally distant supervision. KL loss measures the distance between two distributions: \( P(x|y) \), which is the distant ground-truth supervision described in Equation 8 and \( Q(x|y) = \frac{1}{|V|} \sum_{i=1}^{V} \sum_{j=1}^{N} C_i^{(N)} \), which is the average sum of estimated probabilities at all steps from the output of word-utterance attention sub-layer in the last decoder layer. We denote the KL loss as:

\[
L_{KL}(\theta) = KL(P(x|y)||Q(x|y); \theta). \tag{10}
\]

Then, we use MCE loss to alleviate the negative effects of noises caused by imprecise \( Q(x|y) \):

\[
L_{MCE}(\theta) = \frac{1}{|Y|} \sum_{t=1}^{V} \sum_{w \in V} p(y_t = w) \log P(y_t = w), \tag{11}
\]

where \( V \) denotes the vocabulary. Finally, our overall loss is a linear combination of these three loss functions:

\[
L(\theta) = L_{MLE}(\theta) + \eta_1 L_{KL}(\theta) + \eta_2 L_{MCE}(\theta), \tag{12}
\]

where hyper-parameters \( \eta_1 \) and \( \eta_2 \) govern the relative importance of different loss terms.

3. EXPERIMENT SETTINGS

Datasets: We evaluate the performance on two public datasets: Ubuntu Dialogue Corpus [14] (Ubuntu) and JD Customer Service Corpus [15] (JDDC). Baselines: (1) Seq2Seq with Attention Mechanism (S2SA) [16], and we concatenate all context utterances as a long sequence; (2) Hierarchical Recurrent Encoder-Decoder (HRED) [1]; (3) Variational HRED (VHRED) [17] with word drop and KL annealing, and the word drop ratio equals to 0.25; (4) Static Attention based Decoding Network (Static) [4]; (5) Hierarchical Recurrent Attention Network (HRAN) [18]; (6) Transformer [10], and we concatenate all context utterances into a long sequence; (7) Relevant Contexts Detection with Self-Attention Model (ReCoSa) [6]. They all focus on multi-turn conversations, and ReCoSa is a state-of-the-art model on both Ubuntu and JDDC. As for ablation study, HiSA is our model without the globally distant supervision.

Hyper-parameters: The utterance padding length is set to 30, and the maximum conversation length is 10. The hidden size of encoder and decoder is 512, and the number of layers is 4 for encoder and 2 for decoder. The head number of multi-head attention is set to 8. The high-quality pre-trained sentence-embedding model we used is InferSent [19]/Familia [20] for Ubuntu/JDDC. These models are both pre-trained on large-scale datasets in either English or Chinese, and perform well on our datasets. For optimization, we use Adam [21] with a learning rate of 0.0001 with gradient clipping. Hyper-parameters in Equation 12 are set to 1.

Performance Measures: For automatic evaluation, we use 4 groups of metrics: (1) BLEU-2 [22]; (2) Embedding-based Metrics (Average, Greedy, and Extrema) [17]; (3) Coherence [23] that evaluates the semantic coherence between the context and response; (4) Distinct-2 [24]. For human evaluation, we utilize the side-by-side human comparison. We invite 7 postgraduate students as annotators. To each annotator, we show a context with two generated responses, one from HiSA-GDS and the other from a baseline model, but the annotators do not know the order. Then we ask annotators to judge which one wins based on fluency, coherence, and informativeness. Please refer to [18] for more details. Agreements among the annotators are calculated using Fleiss’ kappa.

4. RESULTS AND DISCUSSION

Automatic Evaluation Results: As shown in Table 1 our model outperforms all baselines significantly on both Ubuntu and JDDC (significance tests, \( p \)-value < 0.01) by achieving the highest scores in almost all automatic metrics. Compared with existing baseline models, our model demonstrates its ability of generating relevant and appropriate responses. This is supported by the fact that results of our proposed model have gained improvements on BLEU-2, Embedding-based Metrics, and Coherence. Besides, we also achieve higher Distinct-2 score, which indicates that HiSA-GDS can generate more informative responses.

Human Evaluation Results: These results are shown in Table 2. We observe that HiSA-GDS outperforms all baseline models on both Ubuntu and JDDC. Specifically, the percentage of “win” is always larger than that of “loss”. Take Ubuntu dataset as an example. Compared with VHRED and Transformer, HiSA-GDS achieves preference gains with 48%,
Table 1. Automatic evaluation results on Ubuntu and JDDC (%). The metrics BLEU-2, Distinct-2, Average, Extrema, Greedy and Coherence are abbreviated as B-2, D-2, Avg, Ext, Gre, and Coh, respectively.

| Model      | Ubuntu          | JDDC           |
|------------|------------------|----------------|
|            | B-2 D-2 Avg Ext Gre Coh | B-2 D-2 Avg Ext Gre Coh |
| S2SA [10]  | 0.896 6.104 46.323 28.851 39.209 48.117 | 4.233 3.609 53.901 36.493 37.578 46.176 |
| HRED [11]  | 3.853 6.661 57.972 34.007 41.462 63.173 | 9.405 11.762 63.191 46.714 43.295 57.183 |
| VHRED [12] | 3.677 8.098 57.251 32.024 41.808 61.464 | 6.367 15.184 62.436 43.337 41.787 63.934 |
| Static [4] | 1.581 3.586 51.055 36.193 53.983 69.748 | 2.285 3.738 60.820 38.047 35.367 65.934 |
| HRAN [13]  | 3.880 7.402 56.763 33.501 41.584 67.635 | 5.962 16.365 63.064 43.439 42.389 62.391 |
| Transformer [10] | 3.697 7.278 53.463 36.353 42.763 69.970 | 5.389 5.185 68.336 41.473 42.942 71.374 |
| ReCoSa [6] | 3.872 9.406 59.368 35.834 41.835 71.922 | 3.880 7.402 56.763 33.501 41.584 67.635 |
| HiSA-GDS   | 4.021 9.598 63.527 36.208 40.598 72.261 | 6.986 14.804 66.103 48.284 41.103 67.485 |
| HiSA       | 7.351 10.934 68.283 41.468 50.382 75.823 | 7.127 15.823 73.952 52.502 49.477 74.281 |

Table 2. Human evaluation between HiSA-GDS and other baselines on Ubuntu and JDDC.

| Dataset | Model      | HiSA-GDS vs. | kappa |
|---------|------------|--------------|-------|
|         |            | Win | Loss | Tie |        |
| Ubuntu  | S2SA [10]  | 58% | 12% | 30% | 0.546 |
|         | HRED [11]  | 46% | 19% | 35% | 0.531 |
|         | VHRED [12] | 48% | 20% | 32% | 0.493 |
|         | Static [4] | 51% | 17% | 32% | 0.596 |
|         | HRAN [13]  | 42% | 9%  | 49% | 0.424 |
|         | Transformer [10] | 44% | 19% | 37% | 0.474 |
|         | ReCoSa [6] | 40% | 6%  | 54% | 0.528 |
| JDDC    | S2SA [10]  | 53% | 24% | 23% | 0.547 |
|         | HRED [11]  | 56% | 16% | 34% | 0.468 |
|         | VHRED [12] | 52% | 19% | 29% | 0.453 |
|         | Static [4] | 48% | 11% | 41% | 0.518 |
|         | HRAN [13]  | 50% | 22% | 28% | 0.495 |
|         | Transformer [10] | 51% | 29% | 20% | 0.447 |
|         | ReCoSa [6] | 45% | 27% | 28% | 0.461 |

51%, and 44%, respectively. We check responses generated by our model with “win” and find that they are more relevant to contextual utterances. The kappa scores indicate that annotators come to a “Moderate agreement” on judgement.

**Discussion of Hierarchical Self-Attention**: To validate the effectiveness of hierarchical self-attention mechanism, we present the heatmap of an example in Figure 2. In this example, there are seven contextual utterances, and for each utterance, importance of each word is indicated with the depth of blue color on the right part. Besides, we also show an utterance-level attention visualization in the left part. An utterance is more important when the red color is lighter. For example, the third and seventh utterances, i.e., X3 and X7, are more important than the others. The importance of a word (horizontal blue heatmap) or an utterance (vertical red heatmap) is calculated as the average value of all heads.

5. CONCLUSION

In this paper, we propose a novel model for open-domain dialogue generation, HiSA-GDS, which conducts context selection in a hierarchical and global perspective. The hierarchical self-attention is introduced to capture relevant context at both word and utterance levels. We also design a globally distant supervision module to guide the response generation at decoding. Experiments show that HiSA-GDS can generate more fluent, coherent, and informative responses.
6. REFERENCES

[1] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in AAAI, 2016, pp. 3776–3783.

[2] Jian Song, Kaihai Zhong, Xuesi Zhou, and Ji Wu, “HKA: A hierarchical knowledge attention mechanism for multi-turn dialogue system,” in ICASSP, 2020, pp. 3512–3516.

[3] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao, “How to make context more useful? an empirical study on context-aware neural conversational models,” in ACL, 2017, pp. 231–236.

[4] Weinan Zhang, Yiming Cui, Yifa Wang, Qingfu Zhu, Lingzhi Li, Lianqiang Zhou, and Ting Liu, “Context-sensitive generation of open-domain conversational responses,” in COLING, 2018, pp. 2437–2447.

[5] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin, “Hierarchical variational memory network for dialogue generation,” in WWW, 2018, pp. 1653–1662.

[6] Hainan Zhang, Yanyan Lan, Liang Pung, Jiafeng Guo, and Xueqi Cheng, “Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation,” in ACL, 2019, pp. 3721–3730.

[7] Lei Shen, Yang Feng, and Haolan Zhan, “Modeling semantic relationship in multi-turn conversations with hierarchical latent variables,” in ACL, 2019, pp. 5497–5502.

[8] Lei Shen and Yang Feng, “CDL: Curriculum dual learning for emotion-controllable response generation,” in ACL, 2020, pp. 556–566.

[9] Lei Shen, Xiaoyu Guo, and Meng Chen, “Compose like humans: Jointly improving the coherence and novelty for modern chinese poetry generation,” in IJCNN, 2020, pp. 1–8.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in NIPS, 2017, pp. 5998–6008.

[11] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu, “Learning to select knowledge for response generation in dialog systems,” in IJCAI, 2019, pp. 5081–5087.

[12] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke, “Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation,” in AAAI, 2020, pp. 8697–8704.

[13] Haolan Zhan, Hainan Zhang, Hongshen Chen, Lei Shen, Yanyan Lan, Zhuoyue Ding, and Dawei Yin, “User-inspired posterior network for recommendation reason generation,” in SIGIR, 2020, pp. 1937–1940.

[14] Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle Pineau, “The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems,” in SIGDIAL, 2015, pp. 285–294.

[15] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou, “The JDDC corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service,” in LREC, 2020, pp. 459–466.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, “Sequence to sequence learning with neural networks,” in NIPS, 2014, pp. 3104–3112.

[17] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio, “A hierarchical latent variable encoder-decoder model for generating dialogues,” in AAAI, 2017, pp. 3295–3301.

[18] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou, “Hierarchical recurrent attention network for response generation,” in AAAI, 2018, pp. 5610–5617.

[19] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes, “Supervised learning of universal sentence representations from natural language inference data,” in EMNLP, 2017, pp. 670–680.

[20] Di Jiang, Yuanfeng Song, Rongzhong Lian, Siqi Bao, Jinhua Peng, Huang He, and Hua Wu, “Family: A Configurable Topic Modeling Framework for Industrial Text Engineering,” arXiv preprint arXiv:1808.03733, 2018.

[21] Diederick P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in ICLR, 2015.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a method for automatic evaluation of machine translation,” in ACL, 2002, pp. 311–318.

[23] Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser, “Better conversations by modeling, filtering, and optimizing for coherence and diversity,” in EMNLP, 2018, pp. 3981–3991.

[24] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, “A diversity-promoting objective function for neural conversation models,” in NAACL-HLT, 2016, pp. 110–119.