Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2021.016729

ARTICLE

A Knowledge-Enhanced Dialogue Model Based on Multi-Hop Information with Graph Attention

Zhongqin Bi[1], Shiyang Wang[1], Yan Chen[2,*], Yongbin Li[1] and Jung Yoon Kim[3,*]

[1]College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, 201306, China

[2]Department of Computer Science and Engineering, Guilin University of Aerospace Technology, Guilin, 541004, China

[3]Department of Game Media, College of Future Industry, Gachon University, Seongnam-si, Gyeonggi-do, 13120, Korea

[*]Corresponding Authors: Yan Chen. Email: a18616387167@163.com; Jung Yoon Kim. Email: kjyoon@gachon.ac.kr

Received: 30 March 2021 Accepted: 19 April 2021

ABSTRACT

With the continuous improvement of the e-commerce ecosystem and the rapid growth of e-commerce data, in the context of the e-commerce ecosystem, consumers ask hundreds of millions of questions every day. In order to improve the timeliness of customer service responses, many systems have begun to use customer service robots to respond to consumer questions, but the current customer service robots tend to respond to specific questions. For many questions that lack background knowledge, they can generate only responses that are biased towards generality and repetitiveness. To better promote the understanding of dialogue and generate more meaningful responses, this paper introduces knowledge information into the research of question answering system by using a knowledge graph. The unique structured knowledge base of the knowledge graph is convenient for knowledge query, can acquire knowledge faster, and improves the background information needed for answering questions. To avoid the lack of information in the dialogue process, this paper proposes the Multi-hop Knowledge Information Enhanced Dialogue-Graph Attention (MKIED-GA) model. The model first retrieves the problem subgraph directly related to the input information from the entire knowledge base and then uses the graph neural network as the knowledge inference module on the subgraph to encode the subgraph. The graph attention mechanism is used to determine the one-hop and two-hop entities that are more relevant to the problem to achieve the aggregation of highly relevant neighbor information. This further enriches the semantic information to provide a better understanding of the meaning of the input question and generate appropriate response information. In the process of generating a response, a multi-attention flow mechanism is used to focus on different information to promote the generation of better responses. Experiments have proved that the model presented in this article can generate more meaningful responses than other models.

KEYWORDS

E-commerce ecosystem; conversation generation; knowledge graph; graph neural network; graph attention

1 Introduction

With the rapid development of the Internet industry and the continuous maturity of artificial intelligence technology, the development of edge computing technology has promoted the reliability and real-time performance of data transmission [1]. The ecology of e-commerce is also constantly developing through innovation. To further promote the ecology of e-commerce from labour orientation to technology orientation, customer service robots appear in all aspects of our production and life. As people's pace of life has increased continually, online shopping has become people's first choice. In the online shopping scene, increasingly more people directly obtain the information they want by talking with customer service robots. At the same time, the speed of industrial IoT technology [2] processing information continues to accelerate. While providing people with the information they want, this also greatly saves labour costs.

At this stage, many e-commerce customer service intelligent robots are used in e-commerce after-sales scenarios. After-sales customers have a single intention and event in mind; they usually ask when their items are shipped, what courier they are sent through, how the products are used, etc. Therefore, the intelligent customer service robot can directly identify the event and intention of the visitor's consultation content and then reply based on the answer template corresponding to the question. With the gradual deepening of artificial intelligence technology in e-commerce, further improve e-commerce service forecasts by using cloud-side hybrid computing [3], there are higher requirements for the diversity, novelty and relevance of the reply content of customer service robots. On the one hand, they must increase the conversion rate of product purchases in the pre-sales link and improve their ability robots to respond to customers' shopping needs. On the other hand, the experience and sense of substitution in the dialogue between consumers and customer service robots must be improved. This approach can generate more meaningful and richer semantic answers through dialogue generation technology. Therefore, the research on the external knowledge base proposed in this paper to enhance dialogue generation is of great significance to the intelligent development of the e-commerce ecosystem.

In natural language generation (NLG) technology, the sequence-to-sequence model strongly promotes the development of dialogue generation technology with its unique model structure. For a given input, we can generate the response information directly through the encoder and decoder. However, due to the lack of corresponding background information, it is difficult to fully and deeply understand the semantic information of some words in a post using only the input post information. Many times, the content generated by the generative model is meaningless, repetitive, and boring [4], and the answers may be unreasonable.

With the in-depth application of the recommendation system in e-commerce [5], in order to better improve the recommendation effect, our dialogue model needs to understand more background knowledge. To increase the background knowledge to promote the understanding of semantic information, many studies have begun to explore the possibility of introducing an external knowledge graph, which usually includes unstructured text information and structured knowledge graph information, such as open domain knowledge graphs [6] and commonsense knowledge bases [7]. For the introduction of unstructured text information, in many cases, the required background information cannot be obtained well, which is likely to cause considerable noise in the understanding of the input semantics. For generation based on graph structures, multi-hop information is rarely considered, and many studies use only the information of simple triples. Given an entity, the goal is to find another entity as a response according to the corresponding relationship, and it is impossible to analyse the semantic expression of the entity from the perspective of the entire graph.

To address the abovementioned problems, a knowledge-enhanced dialogue model is proposed in this article to obtain background knowledge related to the content of the post in a large-scale commonsense knowledge base and encode it into the post. In this way, the semantic understanding of the input posts is enhanced, study the hidden relationships between users [8], and more meaningful and relevant responses are generated, preventing the blind generation of meaningless replies. This model retrieves the entities that exist in the knowledge base in each input post and then uses the retrieved entities to construct a knowledge subgraph related to the post as the external knowledge of the post. For the knowledge subgraph as an external background, every time one layer is expanded, correspondingly, the number of entities greatly increases, and the amount of information increases. Most of the time, for an entity that is farther from the zero-hop entity, its relevance will also decrease; the aggregation of too much irrelevant information leads to deviations in the semantic understanding of the posts.

To use the information in the knowledge graph more effectively. This paper analyses the correlation between multi-hop entities and zero-hop entities and believes that the one-hop entity and two-hop entity in the subgraph are more related to the semantic information of the post. The graph neural network (GNN) is used to encode the sub-graph information and aggregate the information of adjacent nodes. In the process of information aggregation, in order to better introduce the multi-hop information in the knowledge subgraph, prevent the introduction of noise and cause the deviation of semantic understanding, the model uses the graph attention mechanism on the retrieved subgraphs to focus on more useful information, which improves the model's ability to fuse information. The model after further fusion of background information continuously improves its ability to acquire knowledge. After obtaining the information of the one-hop entity, the semantic understanding of the input text is deeper, so the two-hop entity with high relevance can be better obtained. Compared with previous work, our model has a stronger ability to obtain effective information in the knowledge graph.

In summary, the contributions of this article are as follows:

1. Introduce the knowledge graph into the dialogue system to enrich the background knowledge of the dialogue. A commonsense knowledge graph is introduced as external background knowledge into the dialogue generation model. At the same time, the graph attention mechanism is introduced to pay closer attention to the entity information with high relevance.

2. The multi-hop entity information is better integrated into the post that is input by the user so that the input post has more background information, which helps improve the response quality.

3. Both word-level and character-level embedding are used, and the highway network is used as a gating mechanism to adjust the relationship between the two, which largely avoids the problem of OOV.

4. The mechanism of multi-attention flow is used to select the information in the input text and the information in the acquired external knowledge graph, attend to the information related to the question and answer, ignore the information unrelated to the question and answer, and reduce the influence of noise on the generated words.

This paper proposes a new method of using a commonsense background knowledge graph to enhance dialogue generation. Our presentation is organized as follows: Section 2 analyses the current status of dialogue generation research and points out some limitations in the current research. In Section 3, the method proposed in this paper is described in detail, which combines a GNN and graph attention mechanism to obtain more accurate entity information that has a higher

correlation with the input information, further improving the quality of dialogue generation. In Section 4, the data set and evaluation index used in this study are introduced in detail, and the experimental results obtained are analysed. In Section 5, we summarize the paper and propose future work directions.

2 Related Work

Many studies on open-domain generative dialogue systems are based on the encoder-decoder architecture of deep learning. Deep learning-based technology does not usually rely on previously set candidate answers to match; instead, through the learning of a large corpus to conduct dialogue, the method of directly generating answers based on the content of the question is defined as a generation model based on certain conditions. References [9,10] applied the sequence-to-sequence model to large-scale dialogue generation tasks and achieved some results. The research of [11,12] enhanced the diversity of responses by encoding topic words and emotion words into the model during the generation process. However, the current generative dialogue research encounters some problems: the answers obtained tend to be more general in terms of logic, accuracy or contextual coherence and tend to produce generic responses, which usually lead to dull conversation.

To improve the quality of dialogue generation, many studies have begun to apply pre-trained language models in the field of dialogue generation. They aim to train a large amount of data to improve the quality of dialogue generation, such as in ELMO [13], UniLM [14], BART [15] and GPT-2 [16], which first pre-train a large amount of data and then fine-tune it in specific dialogue tasks to improve the quality of dialogue generation. However, training a large amount of data alone cannot provide some commonsense background knowledge for the model. Therefore, to further enrich the knowledge information in the pre-trained language model, many studies integrate the information of the knowledge map into the pre-trained language model; References [17–20] introduced the knowledge graph information into the pre-trained language model to further enhance the pre-trained language model's ability to encode background knowledge or commonsense knowledge, but at this stage, due to the high training cost of the pre-trained language model, the application of pre-trained language models incorporating knowledge graph information in the field of dialogue generation is still relatively small.

Many studies have introduced external background knowledge into the dialogue generation task to enhance the understanding of dialogue semantics and generate more meaningful responses. For example, Reference [7] used structured knowledge graphs as a supplement to the background knowledge of the conversation in order to effectively improve the quality of dialogue generation. However, due to the static attention mechanism it uses, there is no more effective attention to the information in the graph through common sense relationships, resulting in more noise in the generated dialogue. Reference [21] effectively promoted the transfer of dialogue concepts through the attention mechanism and generated more diverse results in response. However, in order to better promote the transfer of concepts, the model acquires conceptual entities that are more relevant to the current concept. Makes the attention mechanism unable to pay more attention to useful information more comprehensively and thus cannot introduce more entity information more effectively. In the generation task in [22], incremental coding was used to represent contextual clues across story backgrounds. In addition, the multi-source attention mechanism is used to apply commonsense knowledge to improve the quality of the generated tasks. Through incremental coding and multi-source attention mechanism, it is possible to combine contextual clue information and introduce potential external background knowledge at the same time. However, the background information required by different contexts may be different. In the process of integrating the

background knowledge information into the clue information of the context, it may cause confusion in semantic understanding and ambiguity. In [23], through a dynamic storage mechanism, the commonsense knowledge from an external knowledge base was integrated into the generator to improve the diversity, novelty and theme consistency of the generated results. However, the model uses subject words to obtain adjacent concepts in the knowledge base and cannot make good use of multi-hop information to enrich the topic background information. The model in [24] also introduced external knowledge graphs and text information as background knowledge for dialogue generation, set up different readers to obtain knowledge graphs and text information, and generated responses based on a large-scale knowledge base and text corpus. The effect of this model is largely affected by the quality of the unstructured text. If the unstructured text used is not highly relevant to the topic of the context, much noise is generated, which is not conducive to the understanding of semantics. In [25], a heterogeneous GNN was used as an encoder to integrate human emotions, dialogue history, etc., so that the generated response not only was context related but also had the appropriate emotion. In [26], external knowledge was encoded according to demand, thereby improving the quality of response generation in the knowledge fusion dialogue system. Because the model is mainly based on the need to obtain information from external documents, it then integrates the corresponding information into the dialogue process. The external documents here use unstructured text, which is relatively slow to obtain external information and is highly dependent on the quality of external documents. Reference [27] designed a FelicitousFact mechanism to help the model focus on knowledge facts that are highly relevant to the context. Reference [28] proposes an MQA-QG framework, which uses unsupervised data to reason about multi-hop knowledge. Generate multi-hop problems. However, further exploration of the generation of complex problems is needed. Reference [29] proposed a new perspective that uses non-conversational text to achieve diversified dialogue generation. By combining non-conversational text content, although the search space for text generation can be expanded, the field of topics can be expanded. However, it also has higher requirements for the quality and domain relevance of the non-conversational text used. Reference [30] proposed a knowledge-aware dialogue generation model, which converts the question representation and knowledge matching ability in the knowledge-based question-answering task into discourse understanding and objective knowledge selection in conversation generation. In the process of selecting and using common sense knowledge, the model did not further consider the semantic information gain that may be brought by multi-hop entity information. Reference [31] improved the quality of the generated data by improving the quality of the data. However, these studies introduce some noise in this stage in the process of obtaining relevant multi-hop entities, and many studies are not universal and are only for certain specific fields. Unlike previous studies, our model uses a large-scale knowledge base of common sense. The graph attention mechanism is used to introduce the more relevant one-hop and two-hop entity information in the sub-graph. In the process of acquiring entities, the context information of the input text is merged into the semantic information of each entity. This promotes the flow of information in the sub-graphs towards more relevant entities and achieves better knowledge choices. By combining unstructured text information with structured text information, the introduction of external noise can be effectively prevented, and the positive gain brought by the introduction of external knowledge can be improved. This promotes more meaningful and diverse response generation.

3 Model

Our model is an encoder-decoder structure based on the sequence-to-sequence (seq2seq) framework. To make better use of the information of the external knowledge graph, a subgraph

inference (SGI) module is added to encode the subgraph of the problem. The model introduces the information of the knowledge subgraph from two main perspectives: the entity information that incorporates the subgraph information in the subgraph reasoning module is used to enrich the semantic information of the input post; the attention mechanism is used in the entity output by the subgraph reasoning module, and the information is further screened to better obtain useful information and to screen noise.

The entire model structure is divided into four main modules are shown in Fig. 1:

- Multi-granularity information encoding (MGIE) module: This uses character granularity, word granularity, and context information to encode the word information in the input post; to better integrate the corresponding background information of the input post, the part that uses contextual information to encode is integrated with the entity semantic encoding of the neighbour entity information in the subgraph inference module.

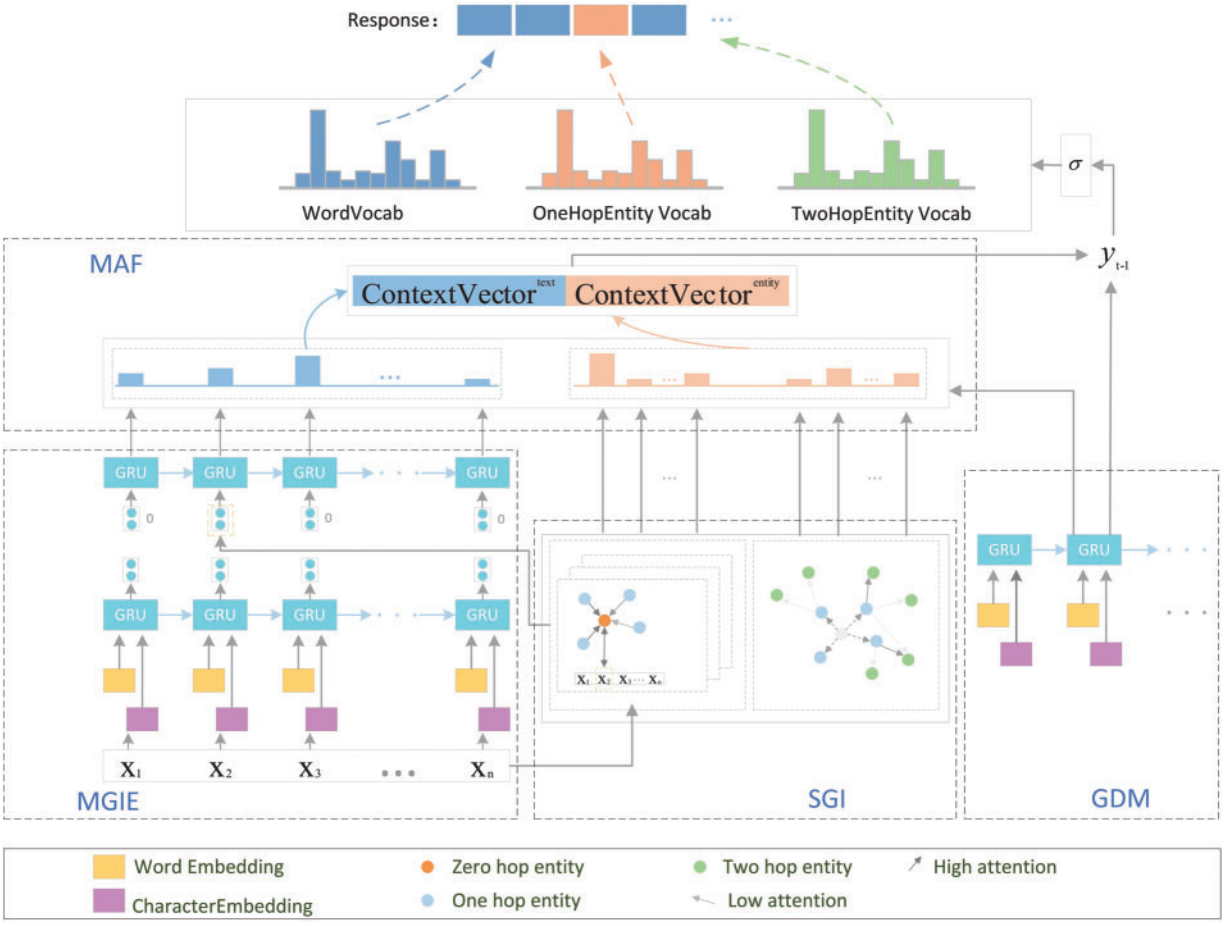

Figure 1: Model structure: first, obtain a preliminary representation of the input question in the multi-granularity information encoding (MGIE) section, and then use the subgraph inference (SGI) module to calculate the subgraph obtained from the input query and collect the entity information related to the question. Next, use the multi-attention flow (MAF) module to focus on more useful information, and finally use the text and subgraph information to make predictions in the gated decoding module (GDM)

- Sub-graph inference (SGI) module: This constructs heterogeneous subgraphs of the input posts and entities, uses a GNN to encode the subgraphs, and uses a graph attention mechanism to focus on entities that are more relevant to the input posts based on commonsense relationships.
- Multi-attention flow (MAF) module: This uses the attention mechanism to obtain information from the input post encoding information, the zero-hop entity information that gathers neighbour entity information, the one-hop entity information and the two-hop entity information to better promote response generation.
- Gated decoding module (GDM): This uses a single-layer GRU network for decoding and adds a gating mechanism during the decoding process to control the use of different vocabularies to generate the words to be predicted.

3.1 Multi-Granularity Information Encoding (MGIE)

3.1.1 Multi-Granularity Embedding

Current network models usually use only word-granular embedding encodings, such as the pre-trained GloVe word-level embedding used in this study. If the input word is not in the vocabulary, it is considered an OOV word, and a random vector is assigned to express the word. If we ignore this point in the algorithm and do not address it, it will affect the outcome of the model. Therefore, another embedding mechanism is needed. The model introduces embedding encoding with character granularity and uses a one-dimensional convolutional neural network (CNN) to obtain the expression of the meaning of the word. To embed the word granularity, the model directly uses GloVe [32] as a pre-trained word vector to obtain a fixed embedding for each word, thereby mapping the representation of each word to a high-dimensional vector space.

The embedding of character granularity maps a word to a high-dimensional vector space by inputting each word to the embedding layer of character granularity. The words in the output question are represented as $\{x_1, x_2, \ldots, x_m\}$. The embedding of the character granularity of each word in the output question is obtained mainly by a CNN. First, the character sequence is converted into a vector, and the obtained one-dimensional vector is used as the one-dimensional input of the CNN. Then, the one-dimensional convolution kernel performs convolution in the direction of the sequence. By using multiple convolution kernels of different dimensions for convolution to obtain the features after the maximum pooling, a fixed-size vector is obtained for each word—that is, the matrix $Q \in R^{d^* m}$.

Regarding the obtained character-granular embedding and word-granular embedding output, they are stitched together (by splicing each word) through a two-layer Highway Network [33], where the Highway Network is mainly used to adjust the relationship between the embeddings at the word level and character level. When handling words that are not in the vocabulary, the weight of the corresponding character level embedding will be higher. When handling words that exist in the vocabulary, the corresponding embedding at the word level will have a greater weight.

As shown in Fig. 2, in the Highway Network, a gating mechanism is mainly used to adjust the relationship between the embeddings of two different words. The corresponding formula is as follows:

$$z = t \odot g\left(W_H y + b_H\right) + (1 - t) \odot y \tag{1}$$

where $t = \sigma\,(W_T y + b_T)$ is a transform gate, $(1-\mathrm{t})$ is a carry gate, W_H, b_H is affine transformation, $\odot$ represents the operation of the product, and $y = [Q_{char}, Q_{word}]$. The final result obtained through the Highway Network is a matrix $\mathrm{X} \in \mathrm{R}^{2d*m}$.

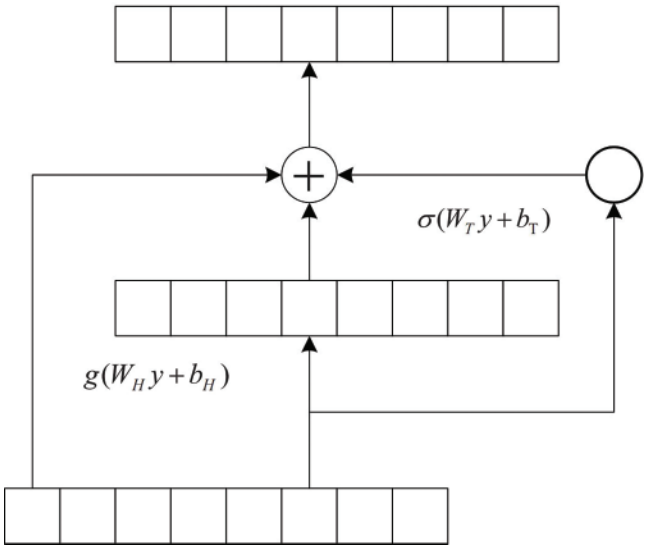

Figure 2: Highway Network: In the highway network, a gating mechanism is mainly used to adjust the relationship between the embeddings of two different words. The corresponding formula is as follows

3.1.2 Contextual Information Encoding

For a given input of length m, this paper uses the recurrent neural network GRU [34] to encode the fusion of character granularity and word granularity queries to obtain the corresponding hidden layer representation. Compared with LSTM [35], using GRU can achieve better results, and it is easier to train, which can greatly improve the training efficiency.

$$\mathrm{h_i} = GRU\,(h_{i-1}, x_i) \tag{2}$$

The current input is x_i, and the hidden state (hidden state) h_{i-1} is passed down from the previous node (this hidden state contains information about the previous node). Combining x_i and h_{i-1}, the GRU obtains the output y_i of the current hidden node and the hidden state h_i passed to the next node.

3.2 SubGraph Inference (SGI)

In many GNN models, the information transfer between nodes is unrestricted, and each entity aggregates the information of its neighbouring entities indiscriminately. For each node, this leads to all neighbour node information being equally important, which introduces too much noise in obtaining background knowledge from external knowledge graphs to enhance dialogue generation; this is not conducive to generating highly relevant responses. For example, the neighbour nodes of "Apple" may be "company" or "fresh". If neighbour information such as this is aggregated, it will make the correct understanding of the query more difficult, and the query will not be able to be understood well with semantic information.

Each round of a query and corresponding response has a corresponding subgraph as an external knowledge base, which provides rich background knowledge for the question-and-answer process. In the process of constructing a subgraph, the entity is first retrieved in the input query as the zero-jump entity of the subgraph corresponding to the query, and then the zero-jump entity is used as the starting point to identify entities that may be highly related to the query or answer. In this way, one-hop entities and two-hop entities are obtained. The main considerations for the selection of one-hop entities and two-hop entities are the degree of correlation between the entity and the input query and whether it is mentioned in the query. To better encode the information of the subgraph, two improvements are made based on using the GNN to encode the information of the subgraph to promote the more accurate propagation of the subgraph information: the attention score, calculated based on the input query and the commonsense relationship between entities, and the PageRank weight calculation.

3.2.1 Query-Based Heterogeneous Entity Update

The GNN is used to update the semantic representation of the entity. The initial entity representation is obtained using TransE [36], and then the information of its neighbouring entity nodes and query information are used to update the corresponding entity semantic information.

$$\mathrm{h}_{\mathrm{e_i}} GNN = \left(\vec{e}_i, G, H\right) \tag{3}$$

The process of encoding and updating the entity information is described below. In this study, the initial entity representation, the initial state representation of the query, and the corresponding subgraph are passed into the GNN network for encoding. $\vec{\mathrm{e}}_i$ is the embedding representation of entity e_i, and H is the representation of the input query.

In the selection process for entity semantic information aggregation, one cannot simply retrieve the entity from the query and then use only the semantics of the entity itself to indiscriminately aggregate its corresponding neighbour information. In this case, there will often be increased noise. The model presented in this article uses query coding and relationship coding in the commonsense knowledge graph to calculate the attention of the graph, uses this as a guide to aggregate the neighbour entities that are closely related to the query semantics, and then updates the query semantic information through the obtained entities. In this way, in the process of calculating the attention information of the next layer, the information encoded by the query contains the information of the entities of the previous layer, which promotes focus on entities with a higher correlation.

First, to obtain the representation of the query, each word of the query input $X = \{x_1, \ldots, x_m\}$ is processed by the LSTM to obtain the initial representation of the query h_0^q,

$$h_0^q = LSTM\,(x_1, \ldots, x_m) \tag{4}$$

Among all outputs of the LSTM, this study chooses the output of the last hidden unit as the initial state of the entire query. In the subsequent layers, the state representation of the problem will be updated to

$$h_q^{l-1} = FFN\left(\sum_{e_i \in E_0} h_{e_i}^{l-1}\right) \tag{5}$$

where E_0 represents the zero-hop entity that appears in the query, $h_{e_i}^{l-1}$ represents the semantic representation of entity i at level $l-1$, and the obtained h_q^{l-1} uses the zero-hop entity information to update the semantic representation of the query input by the user at layer $l-1$.

In Fig. 3, the attention weight $\alpha_r^{e_j}$ is calculated by the embedding encoding of the query and the embedding encoding of the relationship between the commonsense knowledge entities.

$$\alpha_r^{ej} = soft\max\left(\vec{r} \cdot h_q^{l-1}\right) \tag{6}$$

where $\vec{r}$ is the embedding encoding representation of the relationship r between the selected head entity and tail entity. e_j represents the neighbour node. The relationship weight value is calculated according to the input query representation and the relationship representation. To make sure that the information is spread along the edge that is more relevant to the input query, we apply softmax normalization to all outgoing edges of neighbour node e_j.

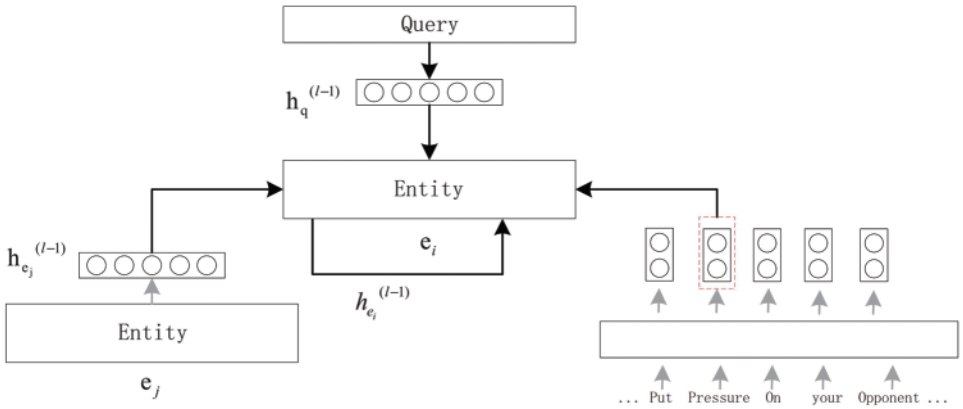

Figure 3: Entity update module: use neighbour node information and input query information to update the semantic information of the current node

3.2.2 Node Information-Directed Propagation

For many questions, it is usually impossible to obtain the desired answer through one-hop entity information, and multi-hop reasoning is often required. This requires obtaining the path from the zero-hop entity mentioned in the question to the entity where the target answer is located. To better promote the spread of the information in the graph along the path of the answer, the PageRank score P_e is set in the process of graph information dissemination. At the same time, when calculating the PageRank weight, this study reuses the attention weight $\alpha_r^{e_j}$ to ensure that nodes obtain higher weights along the path related to the problem.

$$P_e = PageRank\left(e_j^{l-1}\right) \tag{7}$$

3.2.3 Information Aggregation

We use $\mathrm{h}_{\mathrm{r}}^{\mathrm{e_j}\rightarrow e_i}\left(\vec{e}_j^{\,l-1}\right)$ to aggregate the semantic information of the neighbour entity e_j under a specific relationship determined by the attention weight $\alpha_{\mathrm{r}}^{e_j}$ and the PageRank score [37].

$$\mathrm{h}_{\mathrm{r}}^{\mathrm{e_j}\rightarrow \mathrm{e_i}}\left(\vec{e}_j^{\,l-1}\right)=\alpha_r^{e_j}\cdot P_{\mathrm{e}}\cdot FNN\left(\vec{r}\circ h_{e_j}^{l-1}\right) \tag{8}$$

The process of updating the representation $h_{\mathrm{e}_i}^{l}$ of the lth level of entity e_i is as follows:

$$h_{\mathrm{e_i}}^{l}=FNN\left(h_{\mathrm{e_i}}^{l-1^{\circ}}h_{\mathrm{q}}^{l-1^{\circ}}\sum_{r}\sum_{e_j}h_r^{e_j\rightarrow \mathrm{e_t}}\left(h_{\mathrm{e_j}}^{l-1}\right)\right) \tag{9}$$

where $\circ$ is the concatenation operator, which is mainly used for splicing operations. $h_{e_i}^{l-1}$ is the embedding representation of layer $l-1$ of entity e_i, h_q^{l-1} is the semantic representation of the query input at layer $l-1$, and FFN is a feed-forward neural network. Here, the three aspects of information are spliced together to aggregate richer semantic information. Mainly including the representation of entities at layer $l-1$, the query at layer $l-1$ after entity information update represents h_q^{l-1}, and the neighbour entity information is obtained according to the selected specific relationship. The multi-hop entity information is encoded through the attention mechanism [38], and the one-hop entity, the two-hop entity and the relationship information between them are aggregated to obtain the value of attention. In this way, we can effectively pay attention to multi-hop entities and find information that is useful for understanding the semantic information of the query in multi-hop entities.

The information encoding h_{e_O} for jumping from a one-hop entity to a two-hop entity is expressed as

$$h_{e_O}=\sum_{e_T}\beta_{e_O}\cdot[\vec{e_O}\circ\vec{e_T}]. \tag{10}$$

$$\beta_{e_o}=softmax\left(\left(w_r\cdot\vec{r}\right)^{\top}\cdot\tanh\left(w_h\cdot\vec{e_o}+w_t\cdot\vec{e_T}\right)\right) \tag{11}$$

where e_O is a one-hop entity, and e_T is a two-hop entity connected to the one-hop entity. To avoid introducing too much noise, the attention mechanism is used to encode, where w_h and w_t are learnable parameters, and $\vec{e}_T$, $\vec{e}_O$ and $\vec{r}$ are the codes of two-hop entities, one-hop entities and the relationship between them, respectively.

3.3 Multi-Attention Flow (MAF)

3.3.1 Context Vector Representation

For the seq2seq model in the baseline model, in the process of using the attention mechanism, only the input text information is processed with attention, focusing on the most relevant part of the text and the current generated word. The model presented in this article introduces an external background knowledge base in order to better obtain the information in the background knowledge graph and to generate words that are highly relevant to the problem. The model uses the attention mechanism to obtain the probability distributions based on the common words in the input text and the entity words retrieved. The first step is to perform an attention calculation on the input text through the attention mechanism, focusing on the most useful part, generating

suitable words at the current time step. The context vector representation based on the input text is as follows:

$$C_{t-1}^{text} = \left(\sum_{i=1}^{m} soft\max\left(s_{t-1} \cdot h_i\right) \right) \cdot h_i \tag{12}$$

For the attention calculation of the entity words, the main purpose is to calculate the entity with the highest attention score in the respective lists of one-hop entities and two-hop entities, which is used to help generate the most contextual semantic word at the current time step; furthermore, by calculating the attention distribution of the one-hop entities and two-hop entities, the noise introduced can be better filtered. This yields the entity-based context vector representation through attention distribution:

$$C_{t-1}^{one} = \left(\sum_{e_i \in G_{one-hop}} soft\max\left(s_{t-1} \cdot h_{e_i}\right) \right) \cdot h_{e_i} \tag{13}$$

$$C_{t-1}^{\text{triple}} = \left(\sum_{\text{e}_\text{o} \in G_{\text{triple}}} soft\max\left(s_{t-1} \cdot h_{e_O}\right) \right) \cdot h_{e_O} \tag{14}$$

$$C_{t-1}^{two} = \left(\sum_{\text{e}_\text{T} \in G_{\text{two-hop}}} soft\max\left(s_{t-1} \cdot h_{e_T}\right) \right) \cdot h_{e_T} \tag{15}$$

where h_{e_T} is the coded representation of the two-hop entity.

3.4 Gated Decoding Module (GDM)

The output result of the t-th time step of the Decoder part of the model is s_t. This result is calculated by the decoder by using the context vector calculated above based on the text information and the entity words in the knowledge graph, and the output result of the previous time step. The decoder is updated as follows:

$$\text{s}_\text{t} = GRU\left(s_{t-1}, \left[\text{c}_{\text{t}-1}^{\text{text}}; \text{C}_{\text{t}-1}^{\text{one}}; \text{C}_{\text{t}-1}^{\text{triple}}; \text{C}_{\text{t}-1}^{\text{two}}; y_{t-1}\right]\right) \tag{16}$$

where y_{t-1} is the representation of the result of the previous time step, which is used as the input of the current time step. $c_{\text{t}-1}^{text}$ is a contextual representation based on the text content, $c_{\text{t}-1}^{one}$ is a contextual representation based on the one-hop information, $c_{\text{t}-1}^{triple}$ is a contextual representation based on the triple information obtained between a one-hop entity and a two-hop entity, and $c_{\text{t}-1}^{two}$ is a contextual representation based on the two-hop information.

In the word generation process, the gating mechanism σ is used to determine whether the word to be predicted at the current time step is generated through a general word in the text or an entity word in the external knowledge graph. When $\sigma = 0$, general words are selected in the text to generate the words to be predicted at the current time step. When $\sigma = 1$, the one-hop entity is selected in the external knowledge graph to generate the word to be predicted at the current time

step. When $\sigma = 2$, the two-hop entity is selected in the external knowledge graph to generate the word to be predicted at the current time step.

$$\sigma = argmax_{\sigma \in \{0,1,2\}} = \mathrm{FFN}_{\sigma}\left((s_t)\right) \tag{17}$$

After determining which part of the vocabulary to use for word generation at the current time step, the probability distribution of each word on the vocabulary is calculated, and the word with the highest probability is selected as the word to be generated at the current time step.

If a vocabulary of common words is used in the input text, then the probability distribution of all words in the vocabulary is

$$y_t = softmax\left(s_t \cdot \vec{v}\right), \quad \sigma = 0 \tag{18}$$

where $\vec{v}$ is the embedded representation of a general word in the input text.

If the vocabulary of the one-hop entity in the external knowledge graph is used, the probability distribution of all words in the vocabulary is

$$y_t = softmax = \left(s_t \cdot h_{e_i}\right), \sigma = 1 \tag{19}$$

If the vocabulary of the two-hop entity in the external knowledge graph is used, the probability distribution of all words in the vocabulary is

$$y_t = softmax\left(s_t \cdot \vec{e}_T\right), \quad \sigma = 2 \tag{20}$$

where $\vec{e_T}$ is the embedding representation of the two-hop entity e_T.

4 Experiments and Analysis

4.1 Experimental Data Set

The dialogue data set used in the thesis experiment is a single round of dialogue, mainly collected from the Reddit data set. The single-round dialogue data set used here is further expanded based on the previous data set, mainly by obtaining multi-hop entities. The aim is to promote the understanding and generation of dialogue by further enriching the background knowledge. The data set in this study contains 600,000 training pairs and 10,000 test pairs. A processed ConceptNet [39] is used as the background knowledge graph, which contains not only the facts of formal relations, such as London being the capital of the United Kingdom, which are background facts that are always true in reality, but also informal relationships between common concepts from daily life, such as the Lakers being in the championship. To obtain a better representation of the entities, the triples of entities composed of multiple words are deleted from the original knowledge graph. The final knowledge graph contains 120,850 triples, and the number of entities is 21,471 with 44 types of relationships.

The basic configuration of the experimental environment is as follows: Ubuntu 16.04 operating system, Intel Core i9-10900k processor, NVIDIA GTX 3080 discrete graphics, 10 GB of video memory, and 48 GB of memory.

4.2 Evaluation Index

The experiment in this paper uses a variety of evaluation indicators that are commonly used in text generation tasks to evaluate the quality of the generated response:

BLEU [40]: This index calculates the similarity between the generated text and the reference text by calculating the common n-gram:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{21}$$

where

$$BP = \begin{cases} 1 & if \ \ c > r \\ e^{1-r/c} & if \ \ c \le r \end{cases} \tag{22}$$

$$P_n = \frac{\sum_{n-gram \in C} Count_{clip}\,(\mathrm{n-gram})}{\sum_{n-gram \in C} Count\,(\mathrm{n-gram})} \tag{23}$$

ROUGE [41]: Since BLEU considers only the accuracy rate and not the recall rate, the ROUGE index is mainly used to calculate the similarity between the generated text and the reference text from the perspective of the recall rate:

$$R = \frac{\sum_{S \in \{Reference \ \ Summaries\}} \sum_{gram_n \in S} Count_{match}\,(gram_n)}{\sum_{S \in \{Reference \ \ Summaries\}} \sum_{gram_n \in S} Count\,(gram_n)} \tag{24}$$

where R is ROUGE-N, S represents a certain item in the reference text set, $gram_n$ represents an n-gram, and $Count_{match}(gram_n)$ represents the number of n-grams matched with the reference text. The denominator of ROUGE-N (R) is the total number of n-grams of each reference text, so it represents the recall rate.

Meteor [42] considers both the precision rate and recall rate and uses a weighted F value as the evaluation index.

$$Pen = \gamma \left(\frac{ch}{m}\right)^{\theta} \tag{25}$$

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1-\alpha)\, R_m} \tag{26}$$

$$P_m = \frac{|m|}{\sum_k h_k\,(c_i)} \tag{27}$$

$$R_m = \frac{|m|}{\sum_k h_k\,(s_{ij})} \tag{28}$$

The total number of words is m, and the number of blocks is ch.

$$Meteor_{score} = (1 - Pen) \cdot F_{mean} \tag{29}$$

The NIST [43] index is an improvement on the BLEU index. The most important part is to introduce the concept of information for each n-gram. The BLEU algorithm simply sums the

number of n-grams, while NIST sums the amount of information and divides it by the number of n-gram fragments in the entire translation.

Dist-1, Dist-2 and Ent-4 are mainly used to evaluate diversity. This part of the work mainly refers to [44,45].

The above indicators are evaluated using the implementation of [46]. The comprehensive use of these evaluation indicators can better determine whether the model has greater advantages than other models.

This study uses the Seq2seq model, CCM model, and ConceptFlow model as the baseline models and uses ROUGE, NIST, Meteor, PPL, Dist-1, Dist-2 and Ent-4 as the evaluation indicators for making comparisons with the model presented in this article. By analysing the relevance and diversity of the generated text, we can further analyse whether the generated text is more in line with the semantic information of the context and whether it is more meaningful.

4.3 Experiment Analysis

Our model uses a recurrent neural network GRU containing 512 hidden units in the encoder and decoder and uses a word vector pre-trained by GloVe. The dimension of each word vector is 300, and the size of the vocabulary is set to 30,000. We use TransE to obtain the embedding representation of the corresponding entity and relationship. The dimensions of the entity and relationship vector are set to 100. The learning rate of the model is set to 0.0001, and the number of training rounds of the model is set to a maximum of 20 rounds. In the neural network part, we use a 3-layer GNN. The entire model is implemented on Pytorch. The cross-entropy loss function is used for training the model, and the loss changes on the training set and verification set are shown in Fig. 4:

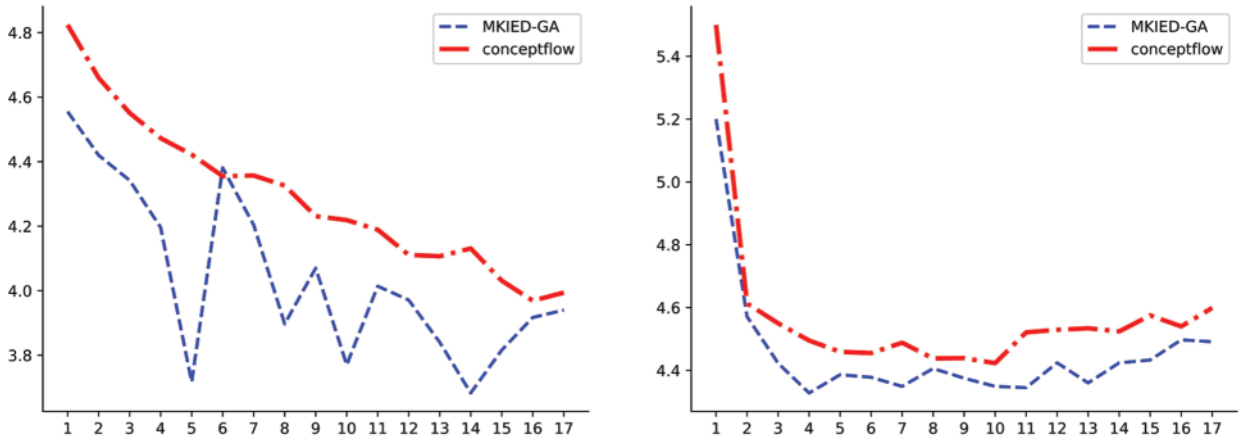

Figure 4: Loss comparative analysis: visually analyse the loss obtained from training on the training set and validation set. (a) Train_loss (b) Valuation_loss

A comparison with the training loss curve of the ConceptFlow model is given below. The blue curve is the training curve of our model, and the red model is the training curve of the ConceptFlow model. The training curve in the figure has relatively large fluctuations at the beginning and is slightly higher than Conceptflow at the 6th epoch. This is mainly due to the

fluctuations caused by our model's fusion of external background knowledge during the encoding process, but as the training progresses, our model uses the multi-attention stream mechanism to continuously improve the ability to integrate external background knowledge, thereby obtaining better performance. In the test set, our training loss can be reduced to 4.38, while the loss of the ConceptFlow model is 4.43, and our model can converge better.

To better evaluate the quality of the response generated by the model in this paper, ROUGE, NIST, Meteor, PPL and other indicators were used in the experiment to evaluate the correlation and diversity of the responses generated by the model. The results are shown in Fig. 5.

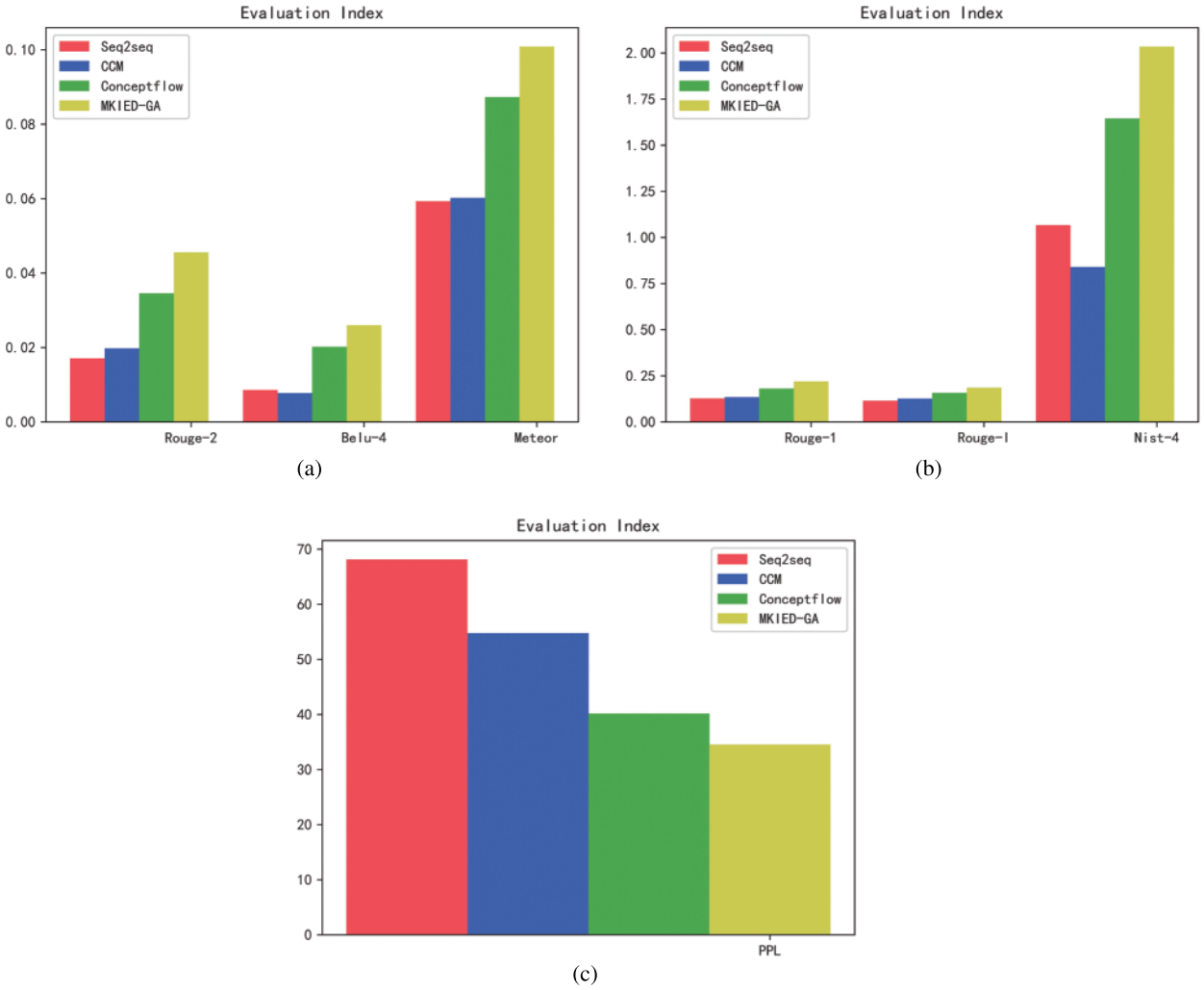

Figure 5: Comparison of the evaluation indicators of the four algorithm models: using ROUGE, METEOR, BLEU, NIST and other evaluation indicators to calculate the correlation between the generated response and the reference response, where higher means better. Using PPL to evaluate the quality of the generated response, where lower means better. (a) ROUGE-2/BLEU-4/ METEOR (b) ROUGE-1/ROUGE-l/NIST-4 (c) PPL

As seen in Fig. 5, the model uses ROUGE, NIST, and METEOR to analyse the relevance of the generated text. To better analyse the fluency of the generated dialogue of the model, a high-level n-gram (BLEU-4) is introduced to measure the fluency of the generated dialogue response. Our model integrates the highly relevant external knowledge entity information into the process of dialogue generation and further promotes the understanding of semantics on the basis of ensuring the relevance of the generated text. It is based on a higher degree of understanding of text semantics that can better improve the fluency of the generated text. Our MKIED-GA model can effectively capture the important information related to the input and response in the knowledge graph and can effectively aggregate this information. It can also well balance the relevance and fluency of the generated text. Its performance results are better than those of all baseline models.

For the response generated by the model, the most important consideration is whether it matches the context of the dialogue. In the experiment, we use the perplexity to evaluate the model at the content level (the main measurement is whether the generated content conforms to grammatical habits and the topic relevance), analyse the quality of response generation by comparing the introduction of external knowledge graphs, and analyse the impact of introducing external knowledge graphs on the quality of dialogue generation. The experimental results are shown in Fig. 6.

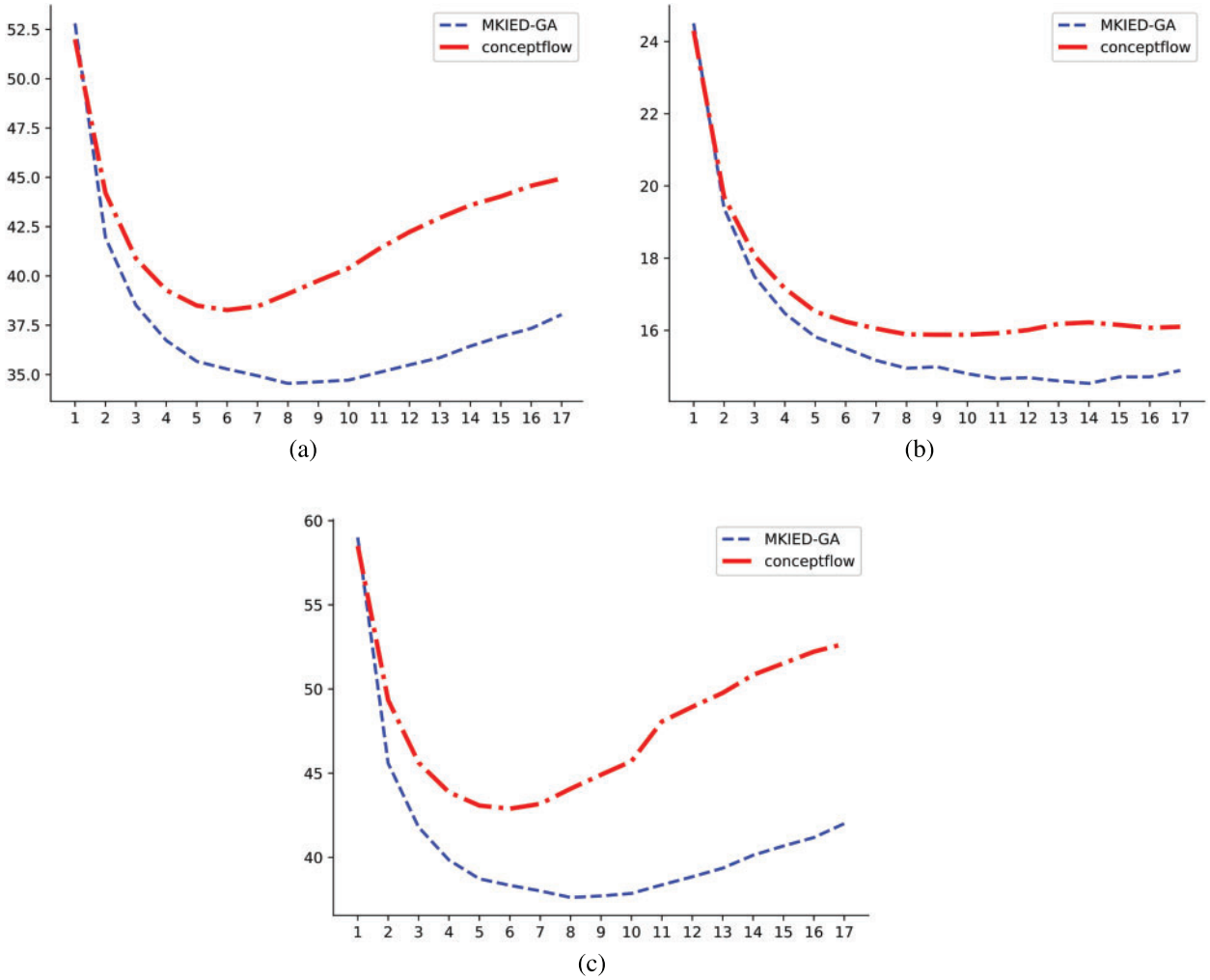

Figure 6: Perplexity comparison analysis: analysing the value of perplexity from the perspective of generated sentences, words and entities. (a) Sentence_ppx_loss (b) Sentence_ppx_local_loss (c) Sentence_ppx_word_loss

As seen in Fig. 6, the blue curve is the perplexity curve of our model, and the red curve is the perplexity curve of the ConceptFlow model. As the number of training rounds increases, the response perplexity (a) generated by our model is significantly lower than the perplexity of the ConceptFlow model, indicating that our model can better understand user input and is more syntactic and semantic. There is a more meaningful response. In addition, our model uses the entities in the knowledge graph to generate responses during the generation process by comparing graph (a) (combined with the calculation results of the perplexity of the external knowledge graph information) and graph (c) (without combining external knowledge). The calculation result of the perplexity of the graph information shows that the result obtained by combining the entity information of the external knowledge graph achieves better results, indicating that the introduction of the external knowledge graph can indeed promote the generation of better responses. Finally, Fig. 6b shows the perplexity values calculated based on the detected entities.

In each question subgraph, there are usually many entities, but not every entity is related to the semantic environment of the question. Blindly aggregating the information of the entities only introduces noise. To better obtain the subgraphs related to the question, in the experiment, the graph attention mechanism is used to pay closer attention to the more relevant entity information. Fig. 7 introduces a case to analyse the role of the graph attention mechanism in the process of graph information aggregation.

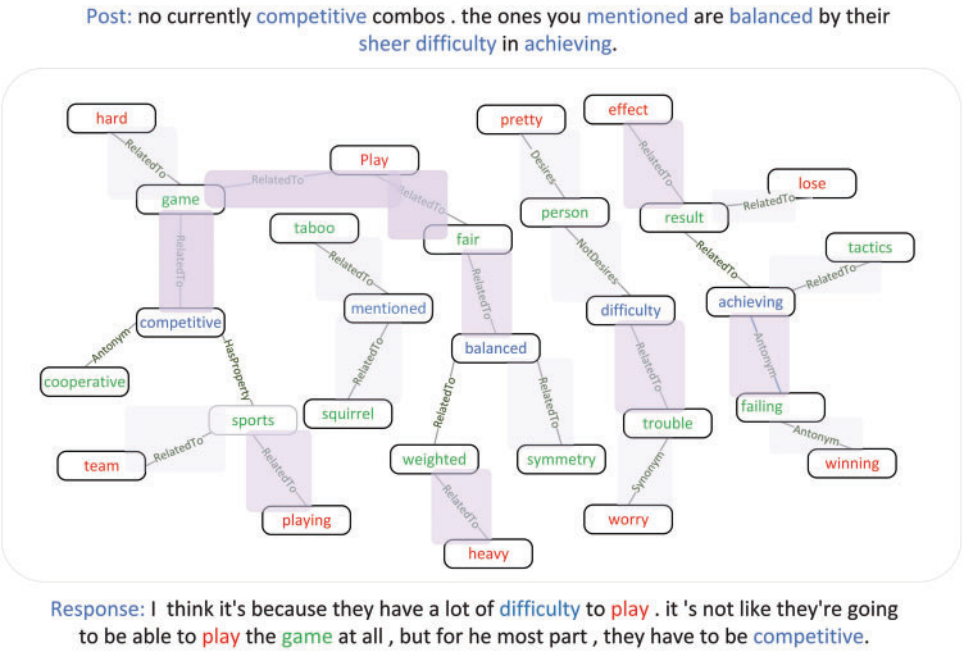

Figure 7: Analysis of subgraph examples: highlight the zero-hop entity, one-hop entity, and two-hop entity in the input post and the generated response. Different colour depths are used in the sub-images to represent the attention distribution. The darker the colour, the higher the attention score

Fig. 7 shows the distribution of image attention. The distribution of attention is represented by different colour depths. The darker the colour, the higher the attention score. The areas with the highest attention score are the one-jump entity game and the two-jump entity play. In the

generated response, by focusing on different entity information, we can obtain supplementary background information to make the generated response more meaningful. There are also entities with relatively high attention scores, such as playing and failing, fair, which are mainly used to better understand the semantic information of the zero-hop entities and one-hop entities in the query. For the one-hop and two-hop entities that are of additional interest, the MAF module is combined with the hidden layer status of each time step and the semantic information of the query for secondary screening. This promotes focus on the important parts and eliminates the noise.

By analysing the responses generated by different models, we can more intuitively see the impact of the introduction of external knowledge graphs on the quality of commonsense question and answer generation. Tab. 1 shows several responses generated by different models, among which the zero-hop entity, one-hop entity and two-hop entity are marked in blue, green and red, respectively.

Table 1: The comparison of the response quality generated by the four models, highlighting the zero-hop entity, the one-hop entity, and the two-hop entity

Post:	I'm not sure you'll get anyone but laker fans wishing laker fans wishing for another laker championship haha.
Seq2seq	I'm not sure if you're being _UNK or not, but I'm not sure if you're
CCM	being _UNK or not.
ConceptFlow	I'm not a fan of the lakers, but I'm not a fan of the lakers.
MKIED-GA	I'm not a fan of the laker, but I love the lakers.
	I'm a fan of the lakers, I think the championship is a little more than a laker fan.
Post:	Being from ohio, I do not know what or how to feel about this
Seq2seq	It is the only way to get a job. we do not know the truth.
CCM	I'm from ohio and I'm in ohio.
ConceptFlow	I do not know if you're in ohio.
MKIED-GA	We do not know the exact same thing about the state of ohio
Post:	No currently competitive combos. the ones you mentioned are balanced by their sheer difficulty in achieving
Seq2seq	I'm not sure if you're joking or not, but I don't think you understand
CCM	how you're getting _UNK.
ConceptFlow	I've played a lot of games, and I'm not a fan of the game.
MKIED-GA	The only way to get the difficulty is to be able to play the game is to be able to get the game. It's just the opposite of the deck that isn't in the game.
	The only way to get the difficulty is to play the game. The only way to get the achievement is to not get the most useful

In Tab. 1, we compare the generated results of several models used in the experiment. Many of the generated results of the seq2seq model have no practical meaning and have low correlation with the input query, and OOV words appear at the same time. In the results generated by the CCM and ConceptFlow models, entity information can also be considered to a certain extent, but

it is not sufficient to enable attention to be paid to multi-hop entity semantic information. The comparison shows that our model uses more one-hop entities and two-hop entities and can better obtain knowledge from external knowledge graphs to promote the improvement of the generation quality.

In the process of introducing external knowledge graphs, to further explore the impact of not introducing entity information, introducing only one-hop entity information and simultaneously introducing one-hop and two-hop information on the quality of response generation, Dist, Ent, confusion, and other indicators are used in the experiment, an analysis of the diversity and correlation of the generated responses is performed, and the experimental results are shown in Fig. 8.

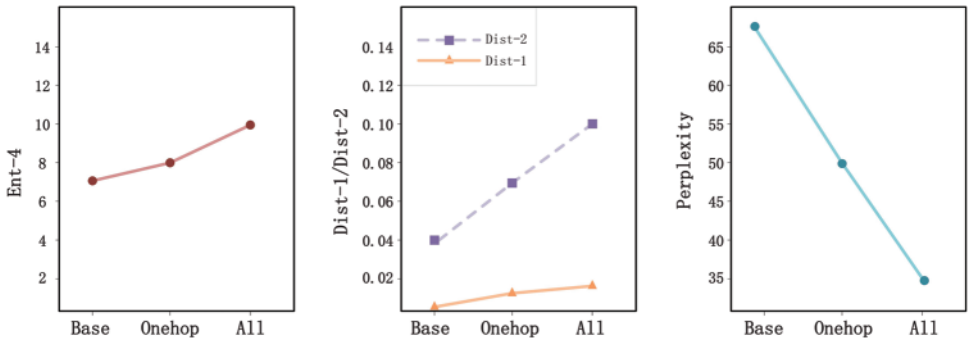

Figure 8: Diversity/Response quality (from left to right): using responses generated without using entity information, responses generated using only one-hop entity information, and responses generated using one-hop entity and two-hop entity information at the same time. Computing diversity, where higher means better, and the quality of response generation, where lower means better

In Fig. 8, the diversity of the generated responses is measured by using the evaluation indicators Dist-1, Dist-2 and Ent-4. By comparing the three methods of introducing external entity information, we analyse the influence of external entity information on the diversity value. These three methods are not introducing external graph entity information, introducing single-hop graph entity information, and simultaneously introducing single-hop and two-hop graph entity information. Then, the perplexity index is used to analyse the quality of the generated responses that introduce different levels of graph information. The lower the index, the better. The more information is introduced into the external graph, the better the quality of the generated response.

BLEU and NIST are used to analyse the correlation between the response generated by introducing different numbers of hops of entity information and the reference response are shown in Tab. 2.

Table 2: Using the response generated without using entity information, the response generated using only one-hop entity information, and the response generated using both one-hop entity and two-hop entity information to calculate the relevance, where higher means better

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist-1	Nist-2	Nist-3	Nist-4
MKIED-GA-text	0.1662	0.056	0.0215	0.0086	0.9925	1.0581	1.0662	1.0675
MKIED-GA-onehop	0.1689	0.055	0.0205	0.0082	1.0474	1.1161	1.1236	1.1249
MKIED-GA-full	**0.2692**	**0.116**	**0.0547**	**0.026**	**1.7927**	**1.995**	**2.0291**	**2.0354**

In this experiment, we analyse the impact of introducing different amounts of external entity information on the relevance and fluency of the generated text. MKIED-GA-Text uses only text information and does not introduce external knowledge information; MKIED-GA-oneHop represents the single-hop information in the subgraph that introduces external knowledge; MKIED-GA-full introduces multi-hop entity information in the subgraph. Compared with MKIED-GA-oneHop and MKIED-GA-Text, MKIED-GA-full uses more external entities, which is consistent with our research motivation. Introducing more external entity information can promote the generation of more relevant and smooth responses.

5 Conclusion and Future Work

In this article, we propose a knowledge-enhanced dialogue generation model to illustrate that using structured knowledge graphs as background knowledge is helpful to the understanding and generation of dialogue. At the same time, we can see that the use of graph neural networks can better aggregate useful information. Finally, the evaluation shows that our model can generate more meaningful responses and that the generated responses can better capture the important information in the map.

In future work, we will further incorporate multi-modal information while using input image information and text information. We will obtain sub-picture information through the entities recognized in the picture and the entities retrieved in the input text. This will further enhance the quality of the response generated by the model. Thereby, it can better promote the healthy development of e-commerce ecology, improve the ability of customer service robots to understand customer intentions and respond to customer questions in pre-sales and after-sales scenarios.

Funding Statement: Funder One, National Nature Science Foundation of China, Grant/Award No. 61972357; Funder Two, National Nature Science Foundation of China, Grant/Award No. 61672337; Funder Three, Guangxi Colleges and Universities Basic Ability Improvement Project of Young and Middle-Aged Teachers, Grant/Award No. 2018KY0651.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Gao, H., Liu, C., Li, Y., Yang, X. (2020). V2VR: Reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability. *IEEE Transactions on Intelligent Transportation Systems, PP(99),* 1–14. DOI 10.1109/TITS.2020.2983835.
2. Gao, H., Qin, X., Barroso, R. J. D., Hussain, W., Xu, Y. et al. (2020). Collaborative learning-based industrial IoT API recommendation for software-defined devices: The implicit knowledge discovery perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence,* 1–11. DOI 10.1109/TETCI.2020.3023155.
3. Gao, H., Huang, W., Duan, Y. (2021). The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments: A QoS prediction perspective. *ACM Transactions on Internet Technology, 21(1),* 1–23. DOI 10.1145/3391198.
4. Tang, J., Zhao, T., Xiong, C., Liang, X., Xing, E. P. et al. (2019). Target-guided open-domain conversation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5624–5634. DOI 10.18653/v1/P19-1565.
5. Yang, X., Zhou, S., Cao, M. (2019). An approach to alleviate the sparsity problem of hybrid collaborative filtering based recommendations: The product-attribute perspective from user reviews. *Mobile Networks and Applications, 25(2),* 1–15. DOI 10.1007/s11036-019-01246-2.

6. Ghazvininejad, M., Brockett, C., Chang, M. W., Dolan, B., Gao, J. et al. (2018). A knowledge grounded neural conversation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32(1),* New Orleans, Louisiana, USA.
7. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J. et al. (2018). Commonsense knowledge aware conversation generation with graph attention. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4623–4629. Stockholm, Sweden. DOI 10.24963/ijcai.2018/643.
8. Yin, Y., Huang, Y., Gao, Q., Xu, H. (2020). Personalized apis recommendation with cognitive knowledge mining for industrial systems. *IEEE Transactions on Industrial Informatics,* 1. DOI 10.1109/TII.2020.3039500.
9. Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c.
10. Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, Conference Track Proceedings*. San Diego, CA, USA. http://arxiv.org/abs/1409.0473.
11. Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y. et al. (2017). Topic aware neural response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31(1),* San Francisco, California, USA. DOI 10.5555/3298023.3298055.
12. Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S. et al. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32(1),* New Orleans, Louisiana, USA.
13. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* vol. 1, pp. 4171–4186, Minneapolis, Minnesota. DOI 10.18653/v1/N19-1423.
14. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X. et al. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems, 32,* Vancouver Convention Center, Vancouver, Canada.
15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A. et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. DOI 10.18653/v1/2020.acl-main.703.
16. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openaiassets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf.
17. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. et al. (2019). ERNIE: Enhanced language representation with informative entities. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1441–1451. Florence, Italy. DOI 10.18653/v1/P19-1139.
18. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q. et al. (2020). K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(3),* 2901–2908. DOI 10.1609/aaai.v34i03.5681.
19. Shen, T., Mao, Y., He, P., Long, G., Trischler, A. et al. (2020). Exploiting structured knowledge in text via graph-guided representation learning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8980–8994. DOI 10.18653/v1/2020.emnlp-main.722.
20. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z. et al. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176–194.
21. Zhang, H., Liu, Z., Xiong, C., Liu, Z. (2020). Grounded conversation generation as guided traversesin commonsense knowledge graphs. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2031–2043. DOI 10.18653/v1/2020.acl-main.184.

22. Guan, J., Wang, Y., Huang, M. (2019). Story ending generation with incremental encoding and commonsense knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence, 33(1),* 6473–6480. DOI 10.1609/aaai.v33i01.33016473.
23. Yang, P., Li, L., Luo, F., Liu, T., Sun, X. (2019). Enhancing topic-to-essay generation with external commonsense knowledge. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2002–2012. Florence, Italy. DOI 10.18653/v1/P19-1193.
24. Xiong, W., Yu, M., Chang, S., Guo, X., Wang, W. Y. (2019). Improving question answering over incomplete KBs with knowledge-aware reader. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4258–4264. Florence, Italy. DOI 10.18653/v1/P19-1417.
25. Liang, Y., Meng, F., Zhang, Y., Xu, J., Chen, Y. et al. (2020). Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. https://arxiv.org/abs/2012.04882.2012.048.
26. Tian, Z., Bi, W., Lee, D., Xue, L., Song, Y. et al. (2020). Response-anticipated memory for on-demand knowledge integration in response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 650–659. DOI 10.18653/v1/2020.acl-main.61.
27. Wu, S., Li, Y., Zhang, D., Zhou, Y., Wu, Z. (2020). Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5811–5820. DOI 10.18653/v1/2020.acl-main.515.
28. Pan, L. M., Chen, W. H., Xiong, W. H., Kan, M. Y., Wang, W. Y. (2021). Unsupervised multi-hop question answering by question generation. *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 5866–5880. Association for Computational Linguistics.
29. Su, H., Shen, X., Zhao, S., Xiao, Z., Hu, P. et al. (2020). Diversifying dialogue generation with non-conversational text. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7087–7097. DOI 10.18653/v1/2020.acl-main.634.
30. Wang, J., Liu, J., Bi, W., Liu, X., He, K. et al. (2020). Improving knowledge-aware dialogue generation via knowledge base question answering. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(5),* 9169–9176. DOI 10.1609/aaai.v34i05.6453.
31. Cai, H., Chen, H., Song, Y., Zhang, C., Zhao, X. et al. (2020). Data manipulation: Towards effectiveinstance learning for neural dialogue generation via learning to augment and reweight. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6334–6343. DOI 10.18653/v1/2020.aclmain.564.
32. Pennington, J., Socher, R., Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. Doha, Qatar. DOI 10.3115/v1/D14-1162.
33. Srivastava, R. K., Greff, K., Schmidhuber, J. (2015). Highway networks. arXiv preprint arXiv: 1505.00387. https://arxiv.org/pdf/1505.00387.pdf.
34. Cho, K., van Merrinboer, B., Gulcehre, C., Bahdanau, D., Bougares, F. et al. (2014). Learning phrasere presentations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Doha, Qatar. DOI 10.3115/v1/D14-1179.
35. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735–1780. DOI 10.1162/neco.1997.9.8.1735.
36. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Neural Information Processing Systems*, pp. 1–9. Lake Tahoe, Nevada, USA. DOI 10.5555/2999792.2999923.
37. Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R. et al. (2018). Open domain question answering using early fusion of knowledge bases and text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4231–4242. Brussels, Belgium. DOI 10.18653/v1/D18-1455.
38. Luong, T., Pham, H., Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421. Lisbon, Portugal. DOI 10.18653/v1/D15-1166.

39. Speer, R., Chin, J., Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31(1),* San Francisco, California, USA. DOI 10.5555/3298023.3298212.
40. Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Philadelphia, Pennsylvania, USA. DOI 10.3115/1073083.1073135.
41. Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81. Barcelona, Spain.
42. Lavie, A., Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231. Prague, Czech Republic. DOI 10.5555/1626355.1626389.
43. Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145. San Diego, California. DOI 10.5555/1289189.1289273.
44. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. San Diego, California, USA. DOI 10.18653/v1/N16-1014.
45. Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X. et al. (2018). Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems,* vol. 31, pp. 1815–1825. Red Hook, NY, USA. DOI 10.5555/3326943.3327110.
46. Galley, M., Brockett, C., Gao, X., Dolan, B., Gao, J. (2018). End-to-end conversation modeling: Moving beyond chitchat. http://workshop.colips.org/dstc7/proposals/DSTC7-MSR_end2end.pdf.