Improve the Response Diversity of Multi-turn Dialogue System by Combining Knowledge

Zhengpeng Li, Jiansheng Wu, Jiawei Miao, and Xinmiao Yu

Abstract—**Recently, neural network language models have been trained in an end-to-end and fully data-driven manner, generating more flexible responses. Traditional neuro-language models have some problems, such as generating meaningless security responses and containing less information, resulting in unattractive conversations. According to the above problems, we propose a knowledge-driven hierarchical recurrent attention network (Kd-HRAN) model. The Kd-HRAN model adds a knowledge entity generator to hierarchical recurrent attention network infrastructure and introduces a knowledge awareness gate in its decoder. The knowledge awareness gate can integrate the context-related external knowledge into the reply process, and determine whether the final output word is generated from the decoder or copied from the knowledge entity generator. The Kd-HRAN model ensures the consistency, richness, and sustainability of the dialogue system. After a large number of experimental verification, the Kd-HRAN model is superior to the baseline model and has diversified and robust responses.**

Index Terms—**deep learning, dialogue system, generative dialogue system, knowledge-driven**

I. Introduction

Building a dialogue system with human nature is a long-term goal in the field of artificial intelligence (AI). The rapid progress in the research of dialogue systems is due to the progress in deep learning technology, the increase in internet data volume, and the rapid landing of products [1-2]. Therefore, the research of dialogue systems is also widely concerned by researchers.

At present, researchers have proposed various neural network models for generating dialogue responses [3,4]. The response generated by the traditional generative multi-turn dialogue system is more natural and coherent, but these models only learn dialogue interactions from dialogue data [5]. When it comes to actual scenarios [6,7] such as company and customer service, there are still problems in the quality of response. One problem is the lack of diversity of responses, such as "I don't know." Another is the response about factual content is wrong. A good conversation is not only coherent but informative in its response. The research focus of the generative multi-round dialogue systems is to avoid generating boring replies, promote the sustainability of dialogue and enhance the diversity of replies.

In the actual small talk scenarios, assist the conversation by considering the information and knowledge covered in the conversation history. As shown in the example in table 1 [8], the conversation revolves around ‘Tsinghua University’. On the one hand, the knowledge (Tsinghua University) can improve the information and diversity of responses generated, and better facilitate the continuation of the dialogue. On the other hand, the association between knowledge can lead the dialogue to change from the current dialogue to another related dialogue and improve the sustainability of the dialogue. As shown in Table 1, the dialogue is transferred from ‘Tsinghua University’ to ‘The Old Summer Palace’.

This paper focuses on a knowledge-driven multi-turn conversation response generation model to improve the quality of conversation generation by using context-related knowledge. To solve the diversity problem of the multi-turn dialogue system, we proposed the Kd-HRAN model. The Kd-HRAN adds a knowledge entity generator and combines a knowledge awareness gate in the decoder part based on a hierarchical recurrent attention network (HRAN) [9]. Specifically, we use the hierarchical encoder to obtain the context vector of the history, as input to the decoder, to determine the probability distribution of the words in the vocabulary. The probability distribution of knowledge entity words is obtained by using a knowledge entity generator and affects the final probability distribution of the vocabulary. The knowledge awareness gate calculates the cosine similarity between the knowledge entity vector and the current message vector to get the matching degree of the two. Finally, the sigmoid function is used to obtain the probability of copying knowledge entities from the knowledge graph is γ_t and the probability of generating words from a fixed word list is $1-\gamma_t$.

The main contributions of this paper are as follows,

1) We propose a knowledge-driven hierarchical recurrent attention network (Kd-HRAN) model.

2) The Kd-HRAN model improves the response diversity and information content by combining the knowledge perception gate with the hierarchical model.

3) According to the results of the automatic evaluation and human evaluation, the experimental Kd-HRAN model on the KDconv dataset is superior to all baseline models. It

The research work was supported by National Natural Science Foundation of China (No.51774179), and Science and Technology Innovation Project of University of Science and Technology Liaoning (LKDYC202109).

Zhengpeng Li is a graduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Ans han, 114051, China. (e-mail: 1156361257@qq.com)

Jiansheng Wu is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: ssewu@163.com)

Jiawei Miao is a graduate student of School of Computer Science and So ftware Engineering, University of Science and Technology Liaoning, Ansha n, 114051, China. (e-mail: 455605116@qq.com)

Xinmiao Yu is a graduate student of School of Computer Science and So ftware Engineering, University of Science and Technology Liaoning, Ansha n, 114051, China. (e-mail: 2749936763@qq.com)

effectively combines the given knowledge and dialogue

TABLE 1
Examples of Knowledge-Driven Conversation

Conversation（travel）	**Knowledge Triple**		
	Head Entity	**Relation**	**Tail Entity**
User1：你好，请问清华大学在什么位置？ Hello, where is Tsinghua University?			
User2：地址在*北京市海淀区双清路30 号*，你没有这儿的电话吗，打电话多方便啊？ The address is *No.30 Shuangqing Road, Haidian District, Beijing*. Don't you have a phone number here? How convenient it is to make a phone call?	Tsinghua university	Address	No.30 Shuangqing Road, Haidian District, Beijing
User1：对啊，我有电话啊，*010-62793001*，你能帮我查一下这里玩多久合适吗？ Yes, I have a phone number. *010-62793001*. Can you find out how long it's ok to stay here?		Phone	010-62793001
User2：行啊，建议玩*1 小时-2 小时*，那你知道这里门票多少钱吗？ Yeah, I suggest *an hour or two*. Do you know how much it costs here?		Playtime	1 h – 2 h
User1：当然了，这里*免费开放*的，那景点周边还有别的好玩的地方？ Of course, it's *free*. Are there any other interesting places around the scenic spot?		Tickets	Free
User2：必须的啊，像*圆明园*，*北京大学*这些都在它的周边呢，你感兴趣不？ Of course, the Old Summer Palace and Peking University are all around it. Are you interested?		Surrounding attractions	The Summer Palace Peking University
User1：我对*圆明园*特别感兴趣，麻烦你告诉我一下这里的电话？ I'm particularly interested in the Old Summer Palace. Could you tell me the telephone number here?	The Summer Palace	Phone	010-62628501
User2:行啊，电话是*010-62628501*，地址你要不？ Yeah, it's *010-62628501*, address Do you want?		Address	No.28, Tsinghua West Road, Haidian District, Beijing
User1：不用了，地址我知道在*北京市海淀区清华西路28 号*。 Don't bother, I know the address is *28 West Qinghua Road, Haidian District, Beijing.*			

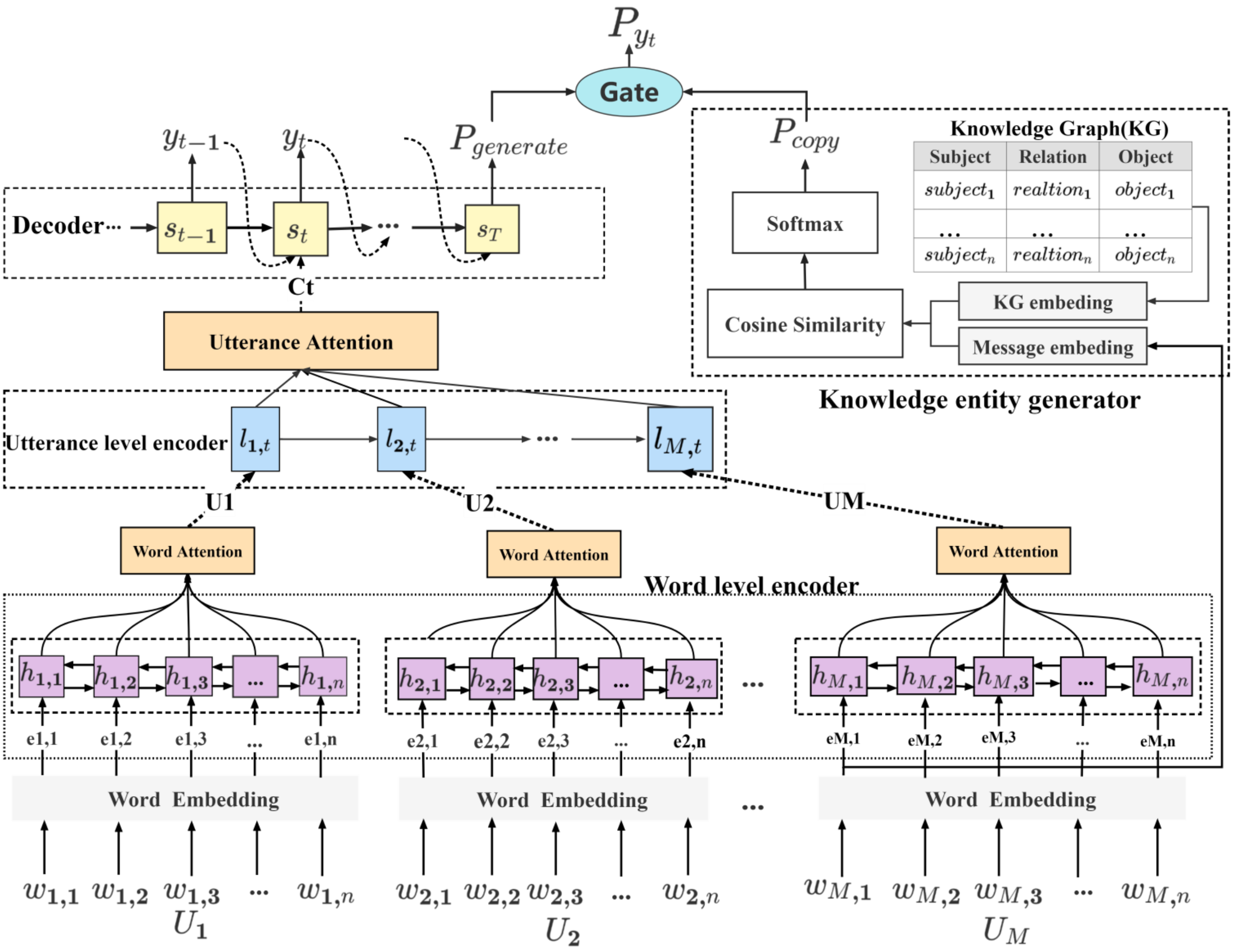

Fig. 1 Knowledge-driven Hierarchical Recurrent Attention Network structure

context and has a good application effect in the chatty multi-round dialogue system.

II. Related Work

In recent years, researchers [10,11] mostly study generative multi-round dialogue systems based on the sequence-to-sequence [12] framework. Serban et al. [13] proposed a hierarchical recurrent encoder-decoder neural network (HRED) to model the semantics of history. Xing et al. [9] extended the attention mechanism [14] to a multi-round dialogue system, added word attention mechanism and sentence-level attention mechanism to the hierarchical model structure, and proposed a hierarchical recursive attention network (HRAN). The HRAN has achieved good results in the Chinese dialogue tasks, but the diversity of response needs to be improved.

Solving the diversity of response of dialogue systems has become one of the key researches. Li et al. [15] proposed the maximum mutual information objective function to replace the maximum likelihood method often used in the traditional sequence-to-sequence model training. In addition to the improvement of the objective function, the diversity of the response can also be improved by improving the decoding process. The classical cluster algorithm beam search [16] selects the words with the highest probability as the response. To avoid falling into the local minimum path problem, Vijayakumar et al. [17] increased the generation space of candidate response according to the measurement method of response diversity, meeting the demand for response diversity.

In addition to improving the dialogue model itself, the researchers attempted to introduce external knowledge information to improve the diversity of the dialogue model on the task. Ghazvininejad et al. proposed a knowledge-based session engine [18], which introduced unstructured knowledge into the sequence-to-sequence structure. In addition to unstructured knowledge data, there are a lot of structured knowledge data on the internet, such as knowledge triplets based on Wikipedia. Zhou et al. [8] proposed using the graph attention mechanism to introduce structured triplet knowledge into the generative dialogue model, using the static attention mechanism to fuse knowledge to enhance the encoder semantic vector, and using the dynamic attention mechanism to enhance the decoder generation effect.

Some researchers have found out of vocabulary (OOV) problems lead to the lack of information in reply statements. The classic Seq2Seq model has trouble learning rare words (eg. proper nouns), which is one of the reasons for OOV problems. Vinyals et al. [19] proposed a pointer generator network for the OOV problems. The pointer network can copy words as output from input sequences and word lists. Gu et al. [20] introduced the copy mechanism into the sequence-to-sequence model and proposed the CopyNet model. The copying mechanism can effectively integrate the generation model with the pointer mechanism, and select a specific sub-sequence output from the input sequence.

III. MODEL

A. Problem Definition and Overview

Given a dataset $D=\left\{\left(C_i,Y_i\right),KG_i\right\}_{i=1}^{T_d}$, taking a sample (*C*, *Y*), *KG* introduces the internal composition of the dataset. $C=\left(U_1,U_2...,U_M\right)$ indicates the history of conversation consists of *M* turn utterance, U_M denotes the most recent message in the conversation as the current input. *Y* represents the response. $KG=\{kg_1,kg_2,...,kg_n\}$ represents knowledge triplets related to conversation history *C*, $kg_n=\left(s,r,o\right)$ stands for a triad of knowledge in *KG*. *s*, *r*, and *o* represent head entities, relation entities, and tail entities, respectively. The goal of the model is to generate a reply $Y=\left\{y_1,y_2,...,y_T\right\}$, made up of *T* words.

Fig. 1 is the structure of the knowledge-driven hierarchical recurrent attention network (Kd-HRAN). The Kd-HRAN model consists of an utterance encoder, context encoder, knowledge entity generator, and decoder with the knowledge awareness gate. Next, we give the specific workflow and equation calculation process of the Kd-HRAN model.

A. Utterance Encoder

The utterance encoder encodes each round of discourse in the history of the dialogue and obtains the correct semantic representation. The utterance encoder adopts a bidirectional gated recurrent unit (Bi-GRU) [21]. The utterance encoder is shown in fig. 2, and consists of the forward GRU and the reverse GRU.

Given a conversation context $C=\left(U_1,U_2,...,U_M\right)$ of length *M*. Suppose there are *n* words in the *m-turn* utterance, and the expression is $U_m=\left(w_{m,1},w_{m,2},...,w_{m,n}\right)$. Take the *m-th* utterance. For example, the word sequences $\left\{w_{m,1},w_{m,2},...,w_{m,n}\right\}$ are transformed into word vector sequences $\left\{e_{m,1},e_{m,2},...,e_{m,n}\right\}$ through the word vector layer. The word vector sequence is then input into the utterance encoder.

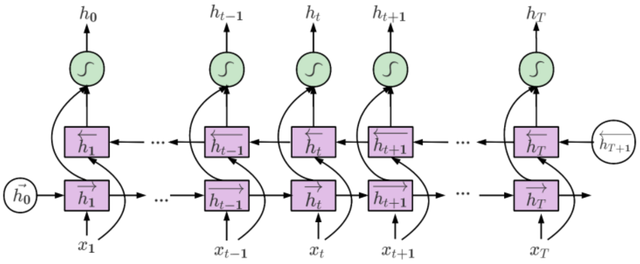

Fig. 2 Bi-GRU Structure

The forward GRU reads the sequence from left to right in the chronological order of *1-th* to *n-th* and calculates the corresponding forward the hidden state of each word vector $\vec{h}_{m,i}$. The reverse GRU reads sequences from right to left in order of *n*-th to *1*-th and calculates the reverse hidden states of each word vector $\overleftarrow{h}_{m,i}$. The calculation equations of $\vec{h}_{m,i}$ and $\overleftarrow{h}_{m,i}$ are as follows,

$$\vec{h}_{m.i}=GRU\left(\vec{h}_{m,i-1},e_{m,i}\right), \tag{1}$$

$$\overleftarrow{h}_{m.i}=GRU\left(\overleftarrow{h}_{m,i+1},e_{m,i}\right), \tag{2}$$

where $GRU(\bullet)$ represents GRU function, $\vec{h}_{m,i-1}$ is the hidden state at step (*i-1*)-th during the forward GRU process, $\overleftarrow{h}_{m,i+1}$ is the hidden state at step (*i+1*)-th in the reverse GRU process. The BiGRU splicing forward hidden state ($\vec{h}_{m,i-1}$) and backward hidden state ($\overleftarrow{h}_{m,i+1}$) to get the hidden state ($\left\{h_{m,i}\right\}_{i=1}^{n}$) corresponding to each word in the sentence sequence.

$$h_{m,i} = BiGRU\left(e_{m.i}\right) = \left[\vec{h}_{m,i}, \overleftarrow{h}_{m,i}\right]. \tag{3}$$

Words have different effects on the utterance. The Kd-HRAN introduces an attention mechanism to improve the weight ratio of important information in a sentence vector. Suppose the decoder is at step *t*−1, the hidden state is s_{t-1}. $h_{m,i}$ is the input of the word-level attention layer, and the attention weight value of the word-level $\left\{\alpha_{i,t}^{w}\right\}_{i=1}^{N}$ is calculated by equation (4). $\left\{\alpha_{i,t}^{w}\right\}_{i=1}^{N}$ measures the importance of words in utterance at step *t*.

$$\alpha_{i,t}^{w} = \frac{\exp\left(h_{m,i}^{T} W_{w} s_{t-1}\right)}{\sum_{n=1}^{N} \exp\left(h_{m,n}^{T} W_{w} s_{t-1}\right)}, \tag{4}$$

$$r_m = \sum_{i=1}^{N} \alpha_{i,t}^{w} h_{m,i}, \tag{5}$$

where W_w is a trainable weight matrix. The utterance vector r_m is composed of the hidden state of words in utterance U_m and the attention weight.

B. Context Encoder

The context encoder is to model the entire conversation history and gets the semantic representation. The Kd-HRAN context encoder uses unidirectional GRU [22]. The utterance vector $\left\{r_1, r_2, ..., r_M\right\}$ is input as a context encoder. $\left\{r_1, r_2, ..., r_M\right\}$ is transformed into the hidden state of the utterance $\left\{l_1, l_2, ..., l_M\right\}$ by GRU. As shown in equation (6),

$$l_m = GRU\left(r_m\right), \tag{6}$$

The Kd-HRAN focuses on the important information of each sentence by the attention mechanism and gets the context vector c_t. Input the hidden state $\left\{l_1, l_2, ..., l_M\right\}$ to the sentence-level attention layer to calculate the context vector c_t, the formula is as follows,

$$\beta_{m,t}^{u} = \frac{\exp\left(l_m^T W_u s_{t-1}\right)}{\sum_{n=1}^{M} \exp\left(l_n^T W_u s_{t-1}\right)}, \tag{7}$$

$$c_t = \sum_{i=1}^{M} \beta_{m,t}^{u} l_m^T , \tag{8}$$

where W_u is a trainable weight matrix. The utterance-level attention weight $\left\{\beta_{m,t}^{u}\right\}_{i=1}^{M}$ measures the importance of the sentence in context.

C. Knowledge Entity Generator

The knowledge entity generator is used to calculate the probability distribution of knowledge entities as output words.

The Kd-HRAN considers using *k* triples in the local knowledge graph. First, encode knowledge triples. The head entity word embedding emb_{s_i} and relation word embedding emb_{r_i} of each triplet are averaged as the knowledge entity vector emb_{kg_i} of the knowledge triplet. The emb_{kg_i} specific calculation equation is as follows,

$$emb_{kg_i} = \frac{1}{2}\sum\left(emb_{s_i} + emb_{r_i}\right), i \in \left\{1...k\right\} \tag{9}$$

In the present moment, the word embedding of nouns and verb phrases in U_m are averaged. The vector emb_q of the input sequence is represented as follows,

$$emb_q = \frac{1}{N}\sum\left(emb_{w_{m,1}} + ... + emb_{w_{m,t}}\right) \tag{10}$$

where $emb_{w_{m,t}}$ represents the word embedding of the *t*-th word in U_m. Finally, the similarity scores of knowledge entities and current messages are calculated by chord similarity, and the copy coefficient of the knowledge entity in the knowledge triplet is obtained.

$$\text{score}_{\text{kg}} = \tanh\begin{pmatrix}\cos\left(emb_q, emb_{kg_1}\right), ..., \\ \cos\left(emb_q, emb_{kg_k}\right)\end{pmatrix}, \tag{11}$$

Normalized the copy coefficient of all the knowledge entities to obtain the probability word distribution P_{copy},

$$P_{copy} = \text{softmax}\left(score_{kg}\right) \tag{12}$$

D. Decoder with Knowledge Awareness Gate

The decoder uses GRU [22] to select a common word from the fixed word list as the predicted reply word. Calculation of probability distribution of common words generated at step *t* is shown in equation (13),

$$P_{generate} = s_t^T W_{out} \tag{13}$$

$$s_t = f\left(e\left(y_{t-1}\right), s_{t-1}, c_t\right) \tag{14}$$

where f is GRU, s_t is the hidden state of the decoder at the step t, $e\left(y_{t-1}\right)$ is the embedding representation of the word y_{t-1}, s_{t-1} is the hidden state of the previous step. c_t is the output of the context encoder, i.e. the context vector.

The knowledge perception gate of Kd-HRAN is mainly inspired by the paper [20, 23]. The knowledge awareness gate determines whether the final output of the decoder comes from a word list or a knowledge graph. At step *t*, the probability of the decoder copying the knowledge entity from the knowledge graph is γ_t, $\gamma_t \in \left[0,1\right]$. The γ_t calculation equation is shown in equation (15).

$$\gamma_t = sigmoid\left(W_{sent}\left[emb_q + emb_d ; score_{sim} ; s_{t-1}\right]\right), \tag{15}$$

where W_{sent} is a trainable parameter matrix, and $[;]$ is the connection operation. emb_d is the word embedding output by the decoder at step $t-1$.

The knowledge awareness gate controls the distribution of the final word list at any time, as shown in equation (16). $1-\gamma_t$ represents the probability the decoder generates common words from the fixed vocabulary.

$$p\left(y_t \mid y_1, y_2 ..., y_{t-1}, c_t, KG\right) = \gamma_t * P_{copy} + \left(1-\gamma_t\right) * P_{generate}, \tag{16}$$

E. Loss Function

The loss function of the model consists of two parts: cross-entropy loss function (L_{cross}) and gating loss function (L_{gate}). The cross-entropy loss function (L_{cross}) calculates the difference between the actual distribution and the expected distribution. The L_{cross} calculation equation is as follows,

$$L_{cross}\left(\theta\right) = -\sum_{t=1}^{T} p_t \log p\left(y_t \mid y_1, y_2 ..., y_{t-1}, c_t, KG\right), \tag{17}$$

where p_t is the real distribution at step *t*, and $p\left(y_t \mid y_1, y_2 ..., y_{t-1}, c_t, KG\right)$ is the predicted distribution at step *t*. L_{gate} monitors the probability of selecting knowledge entity words, as shown in equation (18),

$$L_{gate} = -\sum_{t=1}^{T} \left(q_t \log y_t + \left(1-q_t\right) \log\left(1-y_t\right)\right), \tag{18}$$

where $q_t \in \{0,1\}$ is the real probability selection of common words or knowledge entity words in the training sample. When $q_t = 0$, the Kd-HRAN selects common words from a fixed vocabulary at step *t*. When $q_t = 1$, the Kd-HRAN copies knowledge entities from knowledge graphs at step *t*. The goal of the Kd-HRAN training is to minimize the loss of function $L\left(\theta\right)$,

$$L\left(\theta\right) = L_{cross}\left(\theta\right) + L_{gate}\left(\theta\right). \tag{19}$$

IV. EXPERIMENT

In this section, we conduct experiments on a real knowledge-based dataset to verify the effectiveness of the Kd-HRAN. We introduce the experimental dataset, the baseline model, parameter settings, and the evaluation index.

A. Experimental Datasets

Zhou et al [8] published a Chinese knowledge-driven multi-turn dialogue (KdConv) dataset. The KdConv dataset covers film, music, and tourism, with 4.5k multi-turn conversations in total, including 86K dialogue utterances in total, and the average number of turns is about 19.0. Each discourse is marked with relevant knowledge in the knowledge map and organized in the form of triples.

The KdConv dataset of film and music field knowledge mixed with a large number of Chinese, English, Japanese, and Korean (eg. movie name, actor name, singer name, and song name,). The Kd-HRAN could not understand the semantic information of these unusual words, so this paper uses a multi-round dialogue data set in the tourism field.

The statistical information of the multi-round dialogue dataset (tourist attractions within China) in the tourism field is shown in table 2.

First of all, we split each dialogue example in the dataset into several samples in the form of "dialogue history-response". The samples from the same dialogue example share the same set of knowledge information. The statistical information of the knowledge triple is shown in table 3. We use the Jieba Chinese word separator to segment the dialogue content and knowledge triple.

TABLE 2
Specific Statistics for the KdConv Dataset

Dialogue dataset	train	dev	test
# dialogues	1,200	150	150
# dialogue pairs	8,109	1,196	1,257
Avg. # Turns	13.5	15.9	16.8
Avg. # utterances per dialogue pair	6.75	7.97	8.38

TABLE 3
Details of Knowledge Triples

statistics	data
# entities	699
#head entities	476
#relations	7
#triples	5,287
Avg. # triples per dialogue	10.0

B. Baselines and Implementation Details

● HRED: The Hierarchical Recurrent Encoder-Decoder (HRED) [13] model uses word encoders and speech encoders to model multi-round conversation contexts. The HRED model uses the representation of the context to decode and generate the corresponding response.

● HRAN: Hierarchical Recurrent Attention Network (HRAN) [9] is one of the best models in the current multi-turn dialogue system.

● KG-Copy: The Knowledge copy network (KG-Copy) [24] is a neural network model based on a sequence-to-sequence encoder-decoder. The KG-Copy uses a special gating mechanism to copy knowledge entities from local knowledge graphs.

We build a vocabulary of 20,000 words that occur more than once in the dataset. The special marks words "UNK" are outside the dictionary. In the experiment, the word vector dimension is set to 200, and the vocabulary is initialized by the pre-trained Chinese word vector table of Tencent AI Lab. The coverage is 90.45%, and the word vectors of words not found in the Tencent word vector table are randomly initialized. The hidden vector size is set to 200 for both the encoder and decoder, and the *batch size* is set to 32. Adam algorithm [25] was used to optimize loss and update parameters, and the initial learning rate was 0.001. To prevent over-fitting of parameters, the dropout method [26] is adopted in the experimental training, and the *dropout* is set to 0.3.

To ensure the reliability of experimental results, the hyperparameters used in the comparison model are consistent with those of the KD-HRAN.

C. Evaluation Measures

To objectively evaluate the dialogue system based on generation, we use a combination of multiple automatic evaluation indexes and human evaluation indexes to evaluate the proposed model.

In the experiment, we use the automatic evaluation methods of the *distinct-1* and *distinct-2* to evaluate the diversity of generated responses. The amount of information and diversity of generated responses are measured by calculating the proportion of different *n-grams* in generated sentences. The *distinct-n* calculation equation is as follows,

$$distinct-n=\frac{Count\left(unique\_ngram\right)}{Count\left(word\right)}, \tag{20}$$

The larger of *distinct-n*, the more words are used, and the more abundant the responses are. If the appropriate answer cannot be obtained, the high diversity result is simply to piece up words from the utterance.

For example, the utterance assembled by completely random words can be highly diversified. Therefore, *F*1 [43], *Precision*, and *Recall* are also used in this paper. The *F*1 evaluates the matching degree between the responses generated by the model and the real responses. At the same time, the combined effect of accuracy and recall rate can also determine the content richness of the generated responses. The equation (21-23) is the expression to calculate the value of *F1*.

$$\text{Precision}=\frac{\left|W_{t\arg et}\cap W_{predict}\right|}{\left|W_{predict}\right|}, \tag{21}$$

$$\mathrm{Re}\,call=\frac{\left|W_{t\arg et}\cap W_{predict}\right|}{\left|W_{t\arg et}\right|}, \tag{22}$$

$$F1=2\times\frac{\Pr\text{ecision}\times\text{Re}\,\text{call}}{\Pr\text{ecision}+\text{Re}\,\text{call}}. \tag{23}$$

where $W_{t\arg et}$ represents the set of words in the real results, and $W_{predict}$ represents the set of words in the predicted results of the model.

TABLE 4
Human Evaluation Standards

	score	scoring standards
content	0	The response is not fluency or relevant to the conversation.
	1	The response is fluency but not relevant to the conversation or generic or meaningless.
	2	The response is fluency and relevant to the conversation and contains richer information.
knowledge	0	The response does not utilize knowledge.
	1	The response leverages knowledge but uses unrelated external knowledge.
	2	The response leverages knowledge and uses appropriate external knowledge.

The Kd-HRAN also uses the human evaluation method to evaluate the quality of responses generated by different models from the aspects of “content” and “knowledge”. Specifically, 300 samples were randomly selected from the test set results of all models. Three experienced annotators evaluated the model without understanding the relationship between response and model. The average score of the three annotators was taken as the final result of each model. There are three standards for each aspect: score 0, score 1, and score 2. The specific scoring standards are shown in table 4.

D. Experimental Results

Table 5 shows the experimental results of the baseline model and the Kd-HRAN model on the KdConv dataset. The experimental results in the table will be further analyzed.

TABLE 5
Automatic and Human Evaluation of Results

Model	Automatic evaluation			Human evaluation	
	distinct-1	distinct-2	F1	content	know-ledge
HRED	0.0461	0.1235	-	0.988	-
HRAN	0.0480	0.1246	-	0.996	-
KG-Copy	0.0359	0.1036	0.494	0.986	**0.931**
Kd-HRAN	**0.0665**	**0.2048**	**0.497**	**1.030**	**0.937**

In table 5, no external knowledge is introduced into HRED and HRAN models, so only the *F*1 values of KG-Copy and Kd-HRAN models are given.

In table 5, KG-Copy is slightly lower than the Kd-HRAN in *F*1. To compare the richness of responses generated by each model more clearly, the *distinct*-1 and *distinct*-2 from the automatic evaluation results are presented in the form of a bar graph in fig. 3.

1) Compared with HRED and HRAN, the *distinct*-1 of the Kd-HRAN model is increased by 0.0204 and 0.0185 on the test set, and the distinct-2 is increased by 0.0813 and 0.0802. The introduction of external knowledge can improve the diversity of hierarchical model responses.

2) Both the Kd-HRAN and the KG-Copy use a copy mechanism. Significantly, KG-Copy uses a non-hierarchical encoder and Kd-HRAN uses a hierarchical encoder in the encoding context. The Kd-HRAN is 0.0306/0.1012 higher than the KG-Copy in distinct-1/distinct-2 (0.0665/0.2048 vs. 0.0359/0.1036).

The hierarchical encoding used in the Kd-HRAN model is helpful to the understanding of context semantics, uses knowledge to generate knowledge entity words, and increases the n-tuples in the reply.

The hierarchical coding is due to non-hierarchical coding in the dialogue system.

3) The *distinct*-1 and *distinct*-2 of the Kd-HRAN are both the highest. The Kd-HRAN can refer to more entity words in reply, and enriches the diversity of reply generation. Our model performance is superior to all baselines.

Three annotators scored 300 test samples on both content and knowledge, so each model received 900 content and 900 knowledge scores. To further analyze the results of human evaluation, this paper makes statistics on the proportion of scores in each aspect of different models. Fig. 4 shows the proportion of scores in "content".

The HRED model has the highest proportion in "*score* 1" and the lowest proportion in "*score* 2" (*score* 1=42.6%, *score* 2=28.1%), followed by the HRAD model. The HRED and HRAN models produced many unpractical responses.

The KG-Copy share of "*score* 2" has increased slightly, the knowledge can increase the diversity of responses. The KG-Copy model has the lowest manual score in terms of content (the lowest proportion of "score0"), the KG-Copy model improves the diversity of responses but produces more responses unrelated to dialogue.

Compared with HRED, HRAN, and KG-Copy models, the proportions of "score 0" and "score 1" in the Kd-HRAN model decreased, and the proportions of "score2" also increased. Therefore, the Kd-HRAN model achieved good results in the content of human evaluation. Table 5 and fig. 4 confirm the Kd-HRAN model is more inclined to generate context-related and informative responses.

Observe fig. 5, the Kd-HRAN share of "*score* 2" is higher than the KG-Copy model, and the Kd-HRAN model share of "*score* 1" is lower than the KG-Copy model. Under the premise of introducing knowledge, the correct rate of knowledge entity words used in the Kd-HRAN generation response is slightly higher than in the KG-Copy model.

E. Case Study

Examples of response generation for the Kd-HRAN model and the baseline model is shown in table 6. Generating models do not refer to external knowledge can easily generate generic responses, or even responses are not related to dialogue. Knowledge is important to multi-round dialogue interaction, and the knowledge model can improve the quality of the dialogue system.

The bold fonts (for example, “**鸟巢**”) indicates the generated response made appropriate use of the appropriate knowledge, and the bold italic font (for example, “***圆明园***”) indicates the generated response made the wrong use of irrelevant external knowledge. In addition, the Kd-HRAN model can utilize more external knowledge than the baseline model, thus effectively improving the information and diversity of responses.

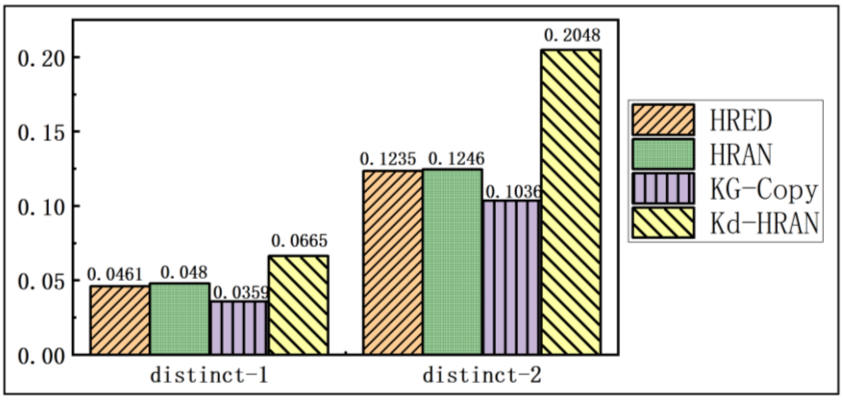

Fig. 3 Comparison diagram of four models under the diversity index

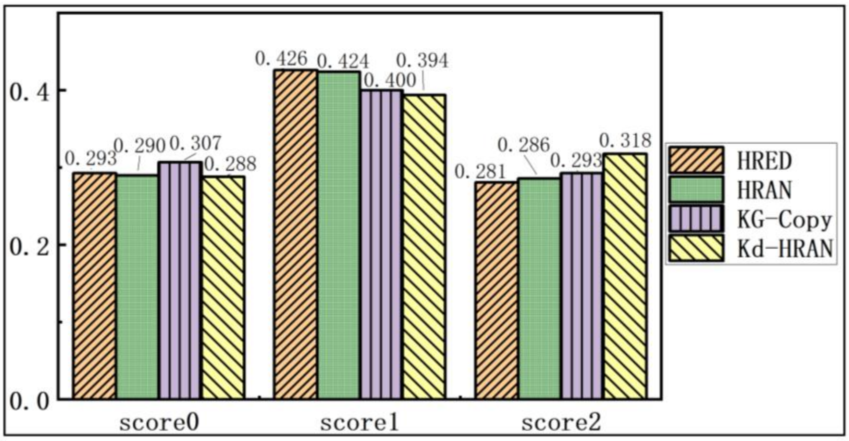

Fig. 4 The percentage of points scored by different models for content

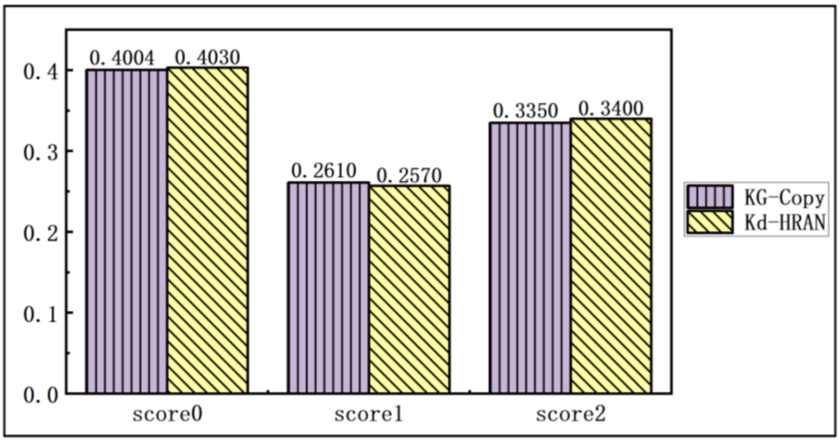

Fig. 5 The proportion of knowledge scores between Kd-HRAN and KG-Copy

TABLE. 6
Examples of Conversations

Conversation	Knowledge Triple		
	Head Entity	Relation	Tail Entity
User1:对国家体育馆有了解吗？ Do you know anything about the National Stadium?	National Stadium	Address	Olympic Park, 9 Tianchen East Road, Chaoyang District, Beijing
User2:有一点了解，但它那里不对外开放可以购买演出门票观看演出。 A little, but it is not open to the public, you can buy tickets to watch the performance.		Surrounding attractions	China Science and Technology Museum
User1:哦，这样呀，那的具体地址是哪里？ Oh, I see. What's the exact address?			Water cube （National Aquatics Center）
User2:地址是北京市朝阳区天辰东路9号奥林匹克公园。 The address is *Olympic Park, no. 9 Tianchen East Road, Chaoyang District, Beijing*.			Bird's nest （National Stadium）
User1:那在它周边都有什么景点呀？ What are the scenic spots around it?	……	…	……
Response			
HRED: 有，很值得去。Yes, it's well worth going.			
HRAN: 对，而且它的附近还有很多别的景点的。Yes, and there are many other scenic spots near it.			
KG-Copy:有，***圆明园***，我也没去过！Yes, The Old Summer Palace. I've never been there either!			
Kd-HRAN:有啊，比如**鸟巢（国家体育场），水立方（国家游泳中心）……** Yes, such as the Bird's Nest (National Stadium), the Water Cube (National Aquatics Center) ……			

V. Conclusion

In recent years, the rapid development of deep learning technology has brought breakthroughs to dialogue technology. The rapid development and wide application of chatty dialogue systems in recent years are amazing. In practical applications, people are often faced with less information generated, or unable to give a semantically rich response. Such a response often makes people lose the desire to continue the conversation. Concerning these problems, this paper proposes improvements through the Hierarchical Recurrent Attention Network to improve the diversity of responses generated by multi-turn of dialogue. We have innovatively proposed a knowledge-driven hierarchical recurrent attention network (Kd-HRAN).

This experiment proves the knowledge introduced in the Kd-HRAN model can enhance the diversity of responses. It further illustrates the important influence of the introduction of knowledge information on several rounds of dialogue response. In future work, we consider enhancing the use of dialogue systems for knowledge entities and try to apply them to other NLP tasks

References

[1] I. V. Serban, S. Chinnadhurai, G. Mathieu, et al. A deep reinforcement learning chatbot," in 31st Conference on Neural Information Processing Systems, Long Beach, 2017.

[2] H. Y. Shum, X. He, and L. Di, "From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots," Frontiers of Information Technology & Electronic Engineering, vol. 19, no. 1, pp. 10-26, Jan. 2018.

[3] L. F. Shang, Z. D. Lu, and H. Li, "Neural Responding Machine for Short-Text Conversation," in ACL-IJCNLP 2015, Beijing, 2015, pp. 1577-1586.

[4] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial Learning for Neural Dialogue Generation," in 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, 2017, pp. 2157-2169.

[5] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in 27th International Joint Conference on Artificial Intelligence, Stockholm, 2018, pp.4623–4629

[6] Y. Peng, Y. Fang, Z. Xie, and G. Zhou, "Topic-enhanced emotional conversation generation with attention mechanism," Knowledge-Based Systems, vol. 163, pp.429-437, Jan. 2019.

[7] D. Peng , M. Zhou , C. Liu , and et al, "Human-machine dialogue modeling with the fusion of word- and sentence-level emotions," Knowledge-Based Systems, vol. 192, 2019.

[8] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu, "KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation," in 58th Annual Meeting of the Association for Computational Linguistics, Virtual, Online, United states, 2020, pp. 7098-7108.

[9] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical Recurrent Attention Network for Response Generation," in 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018, pp.5610-5617.

[10] O. Vinyals, Q. Le, "A Neural Conversational Model," Computer Science, 2015, pp. 994-1003.

[11] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, 2015, pp. 196–205.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in 28th Annual Conference on Neural Information Processing Systems 2014, vol. 27, pp. 3104–3112, Dec. 2014.

[13] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016, pp. 3776–3783.

[14] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in 3rd International Conference on Learning Representations, San Diego, 2014.

[15] J. Li, M Galley, C Brockett, J. Gao, and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models," in 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, 2016, pp. 110-119.

[16] M. Freitag, Y. Al-Onaizan, "Beam Search Strategies for Neural Machine Translation," Proceedings of the First Workshop on Neural Machine Translation, 2017.

[17] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, et al, "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models," arXiv: Artificial Intelligence, 2017.

[18] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A Knowledge-Grounded Neural Conversation Model," in 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2017, pp. 5110-5117.

[19] O. Vinyals, M. Fortunato, N. Jaitly, "Pointer Networks," Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 2, no. 28, 2015 pp. 2692–2700.

[20] J. Gu, Z. Lu, H. Li, and V. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," in 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016, pp. 1631-1640.

[21] Z. Huang , X. Wei , and Y. Kai, “Bidirectional LSTM-CRF Models for Sequence Tagging,” Computer Science.2015.

[22] K. Cho, V. Merenboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in 2014 Conference on Empirical Methods in Natural Language Processing, Doha, 2014, pp. 1724-1734.

[23] S. Merity, C. Xiong, J. Bradbury, C. Xiong, J. Bradbury, and R. Socher, “Pointer sentinel mixture models,” in 5th International Conference on Learning Representations, Toulon, 2017.

[24] D. Chaudhuri, M. Rony,. S. Jordan, M. Rony, A. Rashad, S. Jordan, and J. Lehmann, “Using a KG-Copy Network for Non-goal Oriented Dialogues,” in 18th International Semantic Web Conference, Auckland, 2019, pp. 93-109.

[25] D. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization,” in 3rd International Conference on Learning Representations, San Diego, 2014.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, Jun. 2014.