Research on Clarification Question Recognition and Generation in Intelligent Q&A

Juncheng Hou[1,3], Wen Du[1,2(✉)], Qing Mu[1], Kunpeng Zhang[2], and Zhidong He[2]

[1] The First Research Institute of Telecommunications Science and Technology, Shanghai 200032, China
daviddu999@outlook.com
[2] DS Information Technology CO., LTD., Shanghai 200032, China
[3] Shanghai Research Institute of Criminal Science and Technology, Shanghai 200083, China

Abstract. Intelligent question answering systems often encounter ambiguous questions that require the generation of clarification questions to understand users' true intentions. Without clarification questions, syestems may be confused by ambiguous questions. In this paper, the generation of clarification questions is separated into three subtasks. This study focuses on the three subtasks of the clarification question identification and generation process. We propose the DeBERTA v3 + FC model for clarification question detection and entity prediction, and an improved ByT5-based model for generating diverse and comprehensible clarification questions. On the MSParS dataset, our method outperforms traditional DMN models by 11.2% and 15.17% in accuracy for clarification question detection and entity prediction tasks respectively, while the BLEU score is 4.9% higher than the traditional Seq2Seq models. The efficacy of our proposed methods is verified by their superior performance to traditional methods on all three subtasks. The results on the MSParS dataset demonstrate the effectiveness of our framework for generating clarification questions and entity predictions.

Keywords: Intelligent Question Answering Systems · Clarification Questions · Identification and Generation · Entity Prediction

1 Introduction

Questioning is a fundamental aspect of natural language human-computer interaction systems, including conversational information retrieval and conversational question answering. These systems aim to bridge the gap between users and machines [1]. In particular, clarification is a commonly used tool for knowledge-based Q&A, especially when addressing ambiguous questions [2]. For such ill-defined questions, it can be challenging to provide a satisfactory answer without first asking clarifying questions to confirm the questioner's intention. Therefore, detecting and generating clarification questions is of great practical significance and has garnered attention from academia and industry [3].

However, at the outset of the research, there was a paucity of datasets for clarifying questions, which significantly hindered progress for some scholars and research

M. Zhang et al. (Eds.): CCF NCCA 2023, CCIS 1959, pp. 168–188, 2024.
https://doi.org/10.1007/978-981-99-8764-1_14

teams. Stoyanchev et al. [4] propose a method for randomly removing a phrase from a question and asking an annotator to clarify the resulting question, e.g., the question: "Do you know XXX's birthday?" However, the size of the dataset limited the method's generality and applicability. Later, Guo et al. [5] presented a more extensive synthetic dataset, QRAQ, that relies on entity/variable substitution. Li et al. [6] propose a method for creating ambiguous questions by replacing entities in the question with misspelled word parts, which is groundbreaking but leads to unnatural questions. Xu et al. [7] construct the open-domain clarification corpus CLAQUA, which supports all three primary clarification tasks. They also propose a coarse-to-fine clarification problem generation model based on Seq2Seq and Transformer models as the baseline, which significantly improves performance.

The traditional method of clarifying problem generation involves manually setting rules to address context semantic generation clarification problems. However, this method is time-consuming, labor-intensive, and difficult to apply to different fields [8]. With the rapid development of natural language processing in representation learning and text generation, deep learning-based methods have been proposed and applied to this task [9]. For example, Gao et al. [10] improve traditional Seq2Seq performance by proposing a difficulty estimator to generate clarification questions according to different difficulty levels. Similarly, Rao et al. [11] construct a model based on generative adversarial networks (GANs) that generate clarification questions by estimating potential problem validity. However, existing methods lacked consideration from the user's point of view, had OOV problems when generating clarification questions, and did not have a good understanding of ambiguous statements.

To address these shortcomings, we first present a framework for the algorithmic flow of clarification questions. Research on clarification questions can be broadly divided into three parts: clarification question detection, clarification question generation (CQG), and final entity prediction. In this paper, we examine these three parts by combining single-turn Q&A and multi-turn Q&A. We propose a DeBERTa-based model for clarification question recognition and entity prediction, significantly improving the accuracy of the current SOTA model. Additionally, we propose a clarification question generation framework based on ByT5 that accurately understands the user's intentions and clearly repeats the user's ambiguous questions, using easy-to-understand and representative language to ask the user's final intentions. This approach resolves the issue of nonhumanized and rigid clarification questions, filling the gap in the generation of a representative framework for clarifying questions.In conclusion, our approach provides an improved method for generating clarification questions that is more user-centered, accurate, and understandable, thereby enhancing communication between humans and machines.

In this paper, Sect. 2 describes the process of detecting and generating clarification questions. Section 3 introduces the improved model that incorporates clarification question detection and entity prediction. Section 4 outlines the proposed architecture for clarification question generation and its specific details. In Sect. 5, we conduct ablation experiments and comparative experiments to verify the effectiveness of the proposed model. Finally, the paper concludes with a summary of the study's findings.

2 Design of Clarification Process Framework

We first present our framework for the algorithmic flow of clarification questions, which divides the overall algorithm process into three subtasks, i.e., clarification question detection, clarification question generation, and entity prediction. These tasks are defined as Task1, Task2, and Task3, respectively, as illustrated in Fig. 1 below:

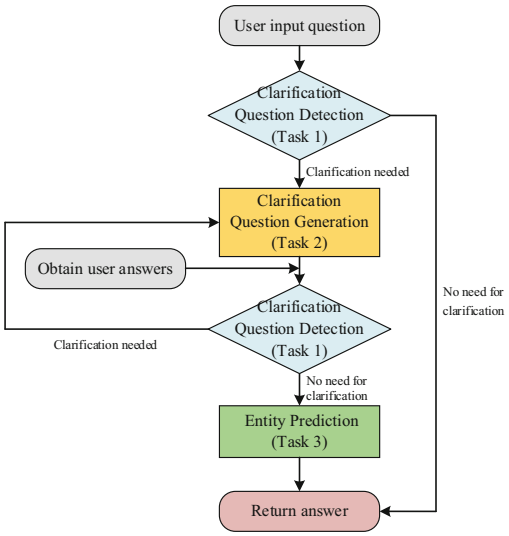

Fig. 1. Process framework diagram

2.1 Task 1: Clarification Question Detection

The detection of clarification questions is the initial and crucial step in the entire task, as it determines whether the QA system comprehends the user's intention and properly addresses it. The detection process takes into account the dialogue context and analyses the specific context of the question and the knowledge base information provided to judge whether the system needs to clarify the current question.

In this task, the input comprises the user and system's historical dialogue, from round 0 to t $< U_0, M_0, U_1, M_1, ..., U_t >$, where U_k denotes the k-th round of the user session and M_k denotes the k-th round of the system session. The output is the system's prediction of whether the current question requires clarification and can also serve as a binary text classification task. By analysing the user's question and employing previous rounds of dialogue knowledge information, reasoning can determine whether the user requires clarification of their question.

2.2 Task 2: Clarification Question Generation

If the outcome of the detection of clarification questions is true, the intelligent question answering system instigates the generation of a corresponding clarification question to

seek a more precise understanding of the user's intent. The input data are the user and system's historical dialogue from round 0 to t, $< U_0, M_0, U_1, M_1, ..., U_t >$. The output generated by the system is a clarification question Q. Through the analysis of previous rounds of dialogue information, clear or ambiguous expressions that do not express the user's intention accurately are extracted, and the system formulates questions in natural language to elicit the user's true intent.

2.3 Task 3: Entity Prediction

After clarifying the user's questions, the intelligent question answering system must further evaluate and comprehend the user's answers. The input for this task is denoted by $< U_0, M_0, U_1, M_1, ..., U_t, Q, A >$, where Q represents the clarification question asked by the system, and A represents the user's corresponding answer. The model output is the entity referred to by the user in their response A. Thus, based on all relevant conversation information, the system infers the user's intention and identifies the appropriate entity.

2.4 Overall Process

The process for the clarification question algorithm involves three distinct tasks: clarification question detection, clarification question generation, and entity prediction. When users input natural language questions, the system initially determines whether their questions are ambiguous. If clarification is necessary, the system extracts relevant information from the user's query, generates a corresponding natural language clarification question, and awaits the user's response. This response is then analysed to determine whether further clarification is needed, and the process continues until the user provides a clear question that requires no further elucidation. Once a clear question is provided, the system combines the user's answer with the dialogue history to extract the central entity. If no clarification is needed, the answer is immediately returned by the system.

3 The Model of Clarification Question Detection and Entity Prediction

This paper utilizes the DeBERTa pre-training model for problem detection. Microsoft introduced the decoding-enhanced BERT with disentangled attention(DeBERTa) model in 2021 [12].

The DeBERTa model improves upon the BERT and RoBERTa [13] models through two new methods: attention decoupling and an enhanced mask decoder. The attention decoupling mechanism represents each input word using two independent vectors for content and location. Meanwhile, its attention weight between words is calculated through a decoupling matrix of their content and relative positions. Similar to BERT, DeBERTA uses MLM mask language modelling for pretraining. While distraction mechanisms account for the content and relative position of contextual words, they do not consider the absolute position of those words, which is critical for prediction in many cases.

Additionally, the enhanced mask decoder merges all transformer layers before the softmax layer to predict mask words. This allows DeBERTa to capture relative positions in all transformers and add absolute position information only when decoding mask words.

For clarification question detection, the dataset is cleaned, and the DeBERTa v3 [14] architecture segments words and extracts feature vectors. The final hidden layer state splices into a fully connected layer to obtain the label for whether it is a clarification question. The model structure diagram is shown in Fig. 2 below:

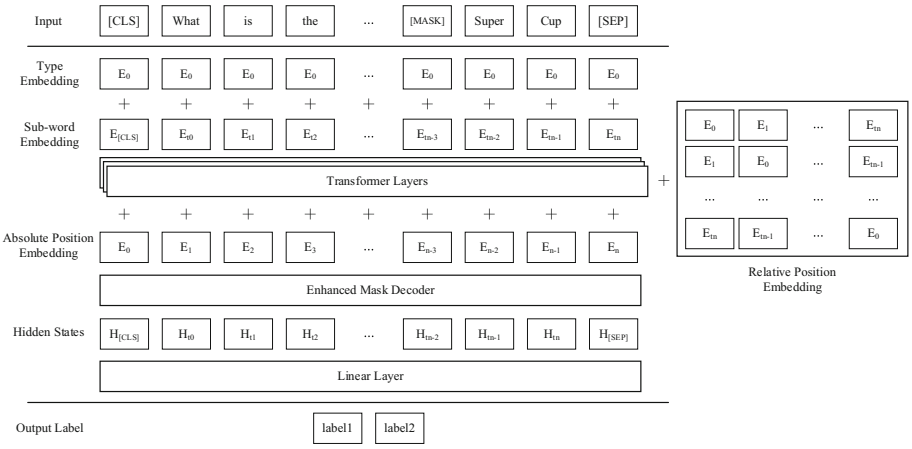

Fig. 2. Structural diagram of clarification question detection model

3.1 Disentangled Attention

In the BERT model, each word in the input layer is represented using a vector obtained by summing the word content embedding and word position embedding. This is achieved by adding Token Embedding, Segmentation Embedding, and Position Embedding. On the other hand, the DeBERTa architecture utilizes bidirectional quantity notation to embed both content and location into each word. For a word at i, the vector H_i is used to represent its content embedding, $P_{i|j}$ represents its relative position embedding relative to the word at j, the cross-attention score for words at i and j is $A_{i,j}$, and the score is calculated as follows:

$$\begin{aligned} A_{i,j} &= \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^T \\ &= H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T + P_{i|j} P_{j|i}^T \end{aligned} \tag{1}$$

In formula (1), it is observed that the weight of a word's attention not only depends on its content, but also has a strong correlation with its position. The weight of a word's attention can be calculated as the sum of four attentions: content-to-content, content-to-location, and position-to-content. Since location-to-location completely separates the importance of content, its reference value is very low, and the information it provides is negligible, so it will be ignored $P_{i|j} P_{j|i}^T$ in this experiment.

3.2 Enhanced Mask Decoder

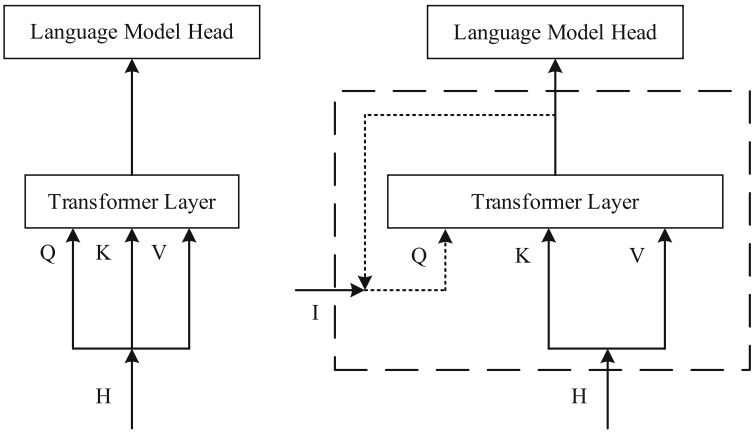

Fig. 3. Comparison of the decoding layer

The EMD (Encoded Mask Decoder) structure is shown in Fig. 3(b) above. The EMD structure has two inputs: H and I. Here, H represents the hidden state of the previous transformer layer, while input I represents any necessary information used for decoding. N represents the number of stacked layers of EMD. The output of each EMD is the input I for the next EMD, and the output of the last EMD is directly fed into the LM head. When $I = H$ and $n = 1$, the encoded mask decoder layer is the same as the BERT decoder. In the model settings in this chapter, we set $n = 2$ to reduce the number of parameters, and the input I for the first layer is an absolute position embedding.

4 The Model of Clarification Question Generation

First, this paper intends to address this task through text summarization, that is, to summarize the QAD part of the dataset and form a restated problem description clarify_state and ques_term, which is placed at the beginning of the question; E, E_ATT, and E_DESC represent the entity name, entity attribute and entity description, respectively, and summarize them in text to obtain a descriptive overview of two entities entity_desc. However, during the experiment, we found that generating clarify_state and ques_term together affects the accuracy of the results, i.e., it enhances the generation effect of ques_term and weakens the generation effect of clarify_state.

After several poor attempts, this paper splits the overall model structure into the following three parts: Model 1 problem retelling model, Model 2 entity difference generation model, and question answering terminology module. In this paper, before passing into Model 1 and Model 2 the data need special processing, so the attribute tree module and semantic understanding module are added. Finally, the output of the two models is merged and stitched with the template to obtain the final generated clarification problem. The overall model architecture diagram is shown in Fig. 4 below:

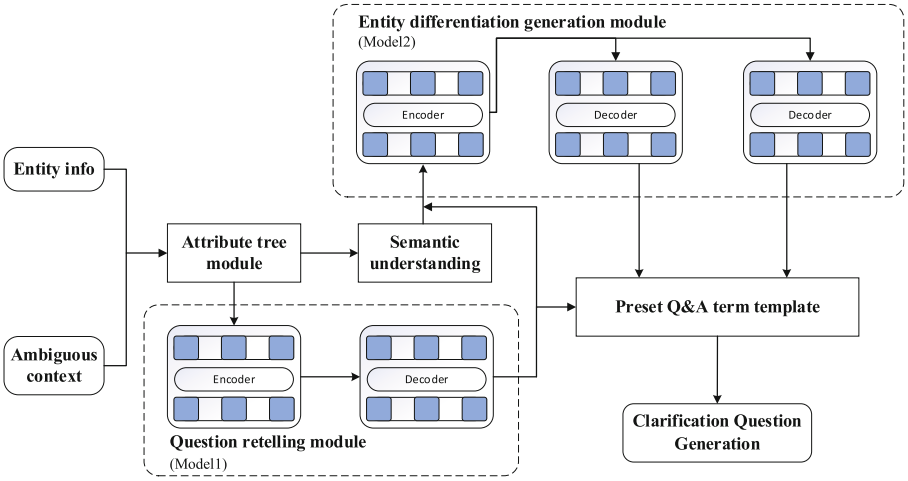

Fig. 4. Architecture diagram of clarification question generation

4.1 Attribute Tree and Semantic Understanding Module

The original JSON data take too long to train when input into the model due to their excessive length. Although the data have been cleaned of irrelevant information, the two models have different tasks that require different information. Therefore, the data preprocessing before entering Model 1 is called Module 1, which includes the attribute tree module, and the data preprocessing before entering Model 2 is called Module 2, which includes both the attribute tree module and the semantic understanding module. The specific flow chart is shown in Fig. 5 below:

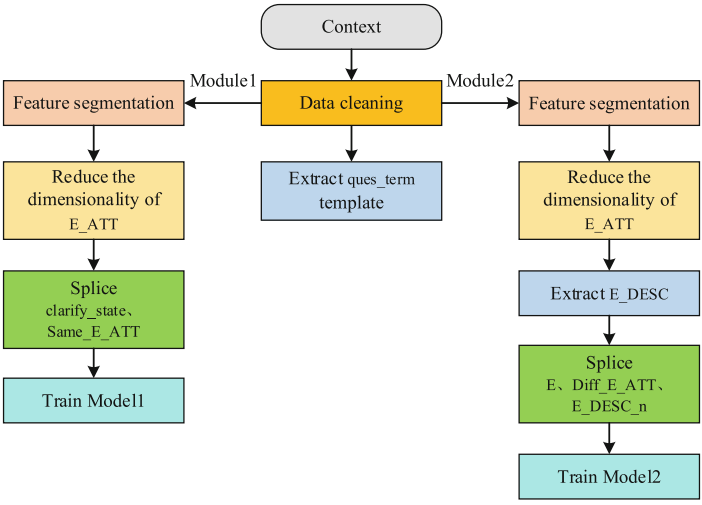

Fig. 5. Data processing flowchart

In Module 1, the cleaned text data are used for feature segmentation, and then E_ATT is reduced in dimension (specific segmentation characters in the dataset < SP >, < S > etc.), which is also the ultimate purpose of the attribute tree module. During the process of generating clarify_state, the dataset is used in several ways, including simple paraphrasing users' questions, extracting entities mentioned by users, extracting common attributes of two entities E as attributes of E, and directly replacing two entities E with the same attribute value. Therefore, the dimensionality reduction process of Module 1 is to treat the E_ATT as the attribute tree of the entity, extract the common attributes and attribute values in the two attribute trees, and record them as Same_E_ATT. The different attributes in the two entities E1 and E2 are recorded as Diff_E1_ATT and Diff_E2_ATT, while the different attribute values are discarded in Module 1 and used in Module 2. For example:

E1_ATT : media_common.creative_work ratings.rated_entity theater.production
E2_ATT : media_common.creative_work award.winning_work ratings.rated_entity
Same_E_ATT: media_common.creative_work ratings.rated_entity
Diff_E1_ATT: theater.production
Diff_E2_ATT: award.winning_work

After dimension reduction of attributes, it is convenient to extract synonyms or attributes from the same Same_E_ATT of two entities for replacement when generating clarify_state to enhance the diversity and generalization of model generation and facilitate the model to learn synonyms and sentences from it.The final input of Model 1 is in the form of [clarify_state < SP > Same_E_ATT]. Example input data are as follows:

Directors of Two Trains Running **<SP>** media_common.creative_work ratings.rated_entity theater.production award.winning_work media_common.cataloged_instance

In Module 2, the first is the attribute tree module. It utilizes the cleaned text for feature segmentation, and performs E_ ATT dimensionality reduction, but the dimensionality reduction process of Module 2 is different from that of Module 1. In the process of generating the entity_desc of the dataset, to distinguish between two entities E, if they do not have the same name, it is easy to distinguish between the two entities. Simply list the user to know which entity E to point to. In contrast, entities usually have different attributives, which may appear in the E_ATT or E_DESC of the two entities, to distinguish entities with the same name. Therefore, in Module 2 we extract different attributes Diff_E_ATT separately, and discard the same attributes Same_E_ATT. The different attribute values Diff_E1_ATT and Diff_E2_ATT in the two entities E are retained.

Through the calculation of regular expressions, in most cases (more than half), entity_desc of different names of E uses different attribute Diff_E_ATT as attributive and uses E_DESC of the two entities to generate different descriptions. For example, when information such as year, location, occupation appears, etc., priority will be given

to using them as attributives to distinguish them. Therefore, the purpose of the extraction of E_DESC is to extract useful relevant information and remove redundant information.

The semantic understanding module uses TF-IDF for the calculation of text similarity. The TF-IDF method requires calculating both Term Frequency and Inverse Document Frequency [15]. Word frequency means that the more times a word appears in the text, the more relevant it is to the text topic. Inverse text frequency means that the more times a word appears in multiple texts in the entire text collection, the worse the distinguishing ability of the word, indicating that the word has a worse ability to discriminate. When the word frequency and inverse text frequency are calculated, these two values are multiplied to obtain get the TF-IDF value of a word, and the larger the TF-IDF value of a word in the article, the more important the term is in this article.

Through TF-IDF similarity calculation, it is concluded that 96.12% of the strong correlation descriptions of entity_desc and E_DESC appear in the first 5 sentences of the E_DESC, the remaining text is either too redundant E_DESC (the longest text can reach 15680 characters), or E_DESC tends to be narrative, which may be the story line of entity E, the biography, etc. Therefore, the relevant dataset has a limited relationship with entity E regardless of the extraction, and thus has a limited overall negligible impact. The maximum length of the input in this setting is 1024. The comprehensive evaluation shows that the first four sentences of E_DESC have the best text effect, and they are recorded as E_DESC_n.

The input of Model 2 is a splicing of E, Diff_E_ATT, and E_DESC_n, and the input form is: [E1 < S > Diff_E1_ATT < S > E1_DESC_4 < TSP > E2 < S > Diff_E2_ATT < S > E2_DESC_4].Example data are as follows:

Two Trains Running **<S>** award.nominated_work **<S>** Two Trains Running is a 1991 pre-Broadway theater production of the play by August Wilson, performed at Kennedy Center. **<TSP>** Two Trains Running **<S>** **<S>** Two Trains Running is a 2006-2007 theater production of the play by August Wilson.

4.2 ByT5 Model

This paper uses the ByT5 model as a pre-training model for problem generation.The T5 (Text-To-Text Transfer Transformer) model is a model proposed by Google in 2020 [16], which converts all natural language processing tasks into text-to-text tasks.

T5 model compares various architectures in pre training, such as the Encoder-Decoder model, language model, and Prefix LM model, where the encoder takes in the entire input sequence and passes the result to the decoder. The representative models of this architecture are BART, MASS, etc. The language model is equivalent to using only the decoder part, and only the data information of the previous time step can be seen within the current time step, which is often used in machine translation.The representative models of this architecture are GPT2, CTRL, etc.; The third Prefix LM type is a fusion of the encoder and the decoder, one part of which can see all the data information, and the other part can only see the past information like the decoder. The representative model is UniLM. To make the model have a unified input and output mode, T5 adds a

prefix before the input sequence Sequence A. The prefix format not only contains the label, but also contains the essence of the task to be solved, and the unified format of T5 input and output is shown in Fig. 6 below:

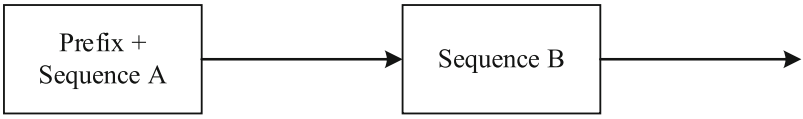

Fig. 6. The input and output diagram of T5 model

The T5 model uses the architecture of the original transformer and does not modify it on a large scale.but highlights some key aspects. The encoder and decoder are still retained, and the sublayers become subcomponents, containing the self-attention layer and the feed-forward network. Self-attention is order independent, using the matrix dot product for operations, and positional encoding is added to the embedding of words before the dot product. T5 modifies the position embedding of the original transformer, using relative position embedding, and the position encoding is shared while re-evaluating in all layers of the model.

ByT5 is an improvement on the T5 model. In the existing model mainly based on word segmentation, an attempt is made to use a byte-based approach for segmentation, replacing the traditional SentencePiece vocabulary [17] with UTF-8 bytes to directly enter the model without any data processing and using 256 different bytes and three special tag characters for byte embedding. In addition, the improved model pretraining no longer adds 100 new ID tags for sentinels, but reuses the last 100 byte IDs and replaces the mask longer byte span with the average mask length. ByT5 makes the encoder depth three times that of the decoder, making this heavier encoder outstanding in text generation and text classification tasks.

4.3 Q&A Term Template

The development of deep learning is weakening the concept of templates, because the setting of templates is easily solidifies people's thinking and increases manpower, maintenance and other costs. Therefore, this task adopts the idea of weak templates and extracts a fixed template by matching and calculating through regular expressions.The fixed Q&A terms that rank among the top 6 in ques_term and whose repetition exceeds 10% of the total, are used as templates for ques_term, as shown in the following table (Table 1):

5 Experiments

5.1 Dataset

This experiment is performed on the public dataset MSParS (Multi-perspective Semantic ParSing Dataset). MSParS is a large-scale dataset published by Microsoft Research Asia for open-domain multi-type semantic analysis tasks. This dataset (V2.0) contains

Table 1. Extracted Q&A Terminology Template

Q&A Template
Which one do you mean, < e > or < e >
are you referring to < e > or < e >
Are you referring to < e > or < e >
Are you talking about < e > or < e >
Do you mean < e > or < e >
are you talking about < e > or < e >

approximately 81,826 natural language problems and their corresponding structured semantic representations, covering 12 different domain problem types. It is currently the most comprehensive semantic analysis dataset in academia. MSParS is annotated based on Microsoft's Open Domain Knowledge Graph Satori. Entities, predicates and types in MSParS follow the standard form in Satori.

In this paper, experiments are performed on single-turn, multi-turn, and single + Multi datasets. The data of the experiments are shown in the following table (Table 2):

Table 2. Experimental Dataset

Class	Task	Train(piece)	Val(piece)	Test(piece)
Single-turn	Task1	10099	853	1175
	Task2	3507	431	497
	Task3	3502	431	502
Multi-turn	Task1	20462	973	828
	Task2	12173	372	384
	Task3	12173	372	384
Single + Multi	Task1	30561	1826	2003
	Task2	15680	803	384
	Task3	15675	803	384

5.2 Evaluation Indicators and Benchmark Models

In this paper, Task 1 clarification detection and Task 3 entity prediction use the accuracy P and F1 values as the performance indicators of the model.Task 2 clarification question generation uses the BLEU (Biling Evaluation Understudy) score [18] as the evaluation indicator of the model. The BLEU score was originally widely used in the field of machine translation to assess the quality of translations by comparing the similarity of machine translation output translations to the translations given in the original corpus.

At present, the BLEU score is widely used in the field of text generation and is one of the most widely used automatic evaluation indicators, which evaluates the advantages and disadvantages of the model by calculating the difference between the content generated after model training and the actual content given. It is calculated as the n-grammatical mean of n = 1 to N and the shorter output is penalized by a short penalty. Its calculation formula is shown as (2), (3).

$$BP = \begin{cases} 1, L_{sys} > L_{ref} \\ e^{(1-L_{ref}/L_{sys})}, L_{sys} \leq L_{ref} \end{cases} \tag{2}$$

$$BLEU = BP * \exp(\sum_{n=1}^{N} W_n * \log p_n) \tag{3}$$

where L_{sys} is the length of the model output, L_{ref} is the length of the reference text in the original corpus, the standard value of BLEU's N is 4, different N-gram lengths are rarely used, and W_n are positive weights of 1. p_n is the number of matching n-grams in the model output divided by the number of n-grams in the original corpus.

5.3 Experimental Environment

This chapter conducts all experiments using Torch.Due to graphics memory limitations, this chapter uses Nvidia Tesla T4 16 GB graphics cards for training and testing on Task 1 and Task 3, and Nvidia Tesla A40 48 GB graphics cards for training and testing on Task 2.The hardware settings and software environment of the experiment are shown in the following table (Table 3):

Table 3. Experimental environment configuration

Class	Specific configuration
CPU	Inter(R) Xeon(R) Platinum 8253 CPU @ 2.20GHz × 4
Internal storage	64 GB
GPU	Nvidia Tesla T4/ Nvidia Tesla A40
Python version	3.7
Torch version	1.6.0 + cu101

5.4 Results and Discussion

The Results of Clarification Question Detection. In this experiment, three pre-trained models BERT, RoBERTa, and Longformer were selected for ablation experiments.They were carried out on single-turn question answering, multi-turn question answering and single + Multi question answering datasets.Due to the limitations of the length of the results, each experiment showed the results of accuracy and F1 value. The results are shown in Tables 4 and 5:

Table 4. Ablation experiment on clarification question detection (Acc)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(val)	Single + Multi(test)
BERT + FC	0.791	0.819	0.829	0.803
RoBERTa + FC	0.800	0.826	0.826	0.811
Longformer + FC	0.843	0.873	0.875	0.858
DeBERTa-v3 + FC	**0.928**	**0.946**	**0.946**	**0.935**

Table 5. Ablation experiment on clarification question detection (F1)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(val)	Single + Multi(test)
BERT + FC	0.752	0.781	0.789	0.764
RoBERTa + FC	0.776	0.811	0.810	0.796
Longformer + FC	0.818	0.849	0.846	0.838
DeBERTa-v3 + FC	**0.915**	**0.944**	**0.939**	**0.928**

Tables 4 and 5 show that our proposed model of clarification question detection has very good performance, i.e., both the accuracy and the F1 score improve by 10% compared to the original pre-trained BERT model, and are nearly 8.23% higher than the Longformer pre-training model. In the comparison between single-turn dialogue and multi-turn dialogue, the results show that the accuracy and F1 value of multi-turn dialogue are nearly 2% higher than those of single-turn dialogue, because the user provides more information in multi-turn dialogue, which is convenient for the model to obtain more information. Meanwhile, most of the single-turn dialogue has the same name entities, and the difficulty of distinguishing between entities with the same name is much greater than that of nonsame-name entities, which is one of the reasons affecting the accuracy of the model. In addition, the F1 score is slightly lower than the accuracy score, indicating that the model has a slight category imbalance, which is also closely related to the inconsistency of the number of single-turn and multi-turn conversations.

The comparative experiment of clarification question detection selected CNN, Transformer, HAN (Hierarchical Attention Networks) [19] and DMN (Dynamic Memory Networks) [20] as comparisons.The HAN model was proposed in 2016, using a bidirectional GRU structure and considering the attention mechanism at the word and sentence level. The DMN mainly uses the triplet of < input - question - answer > as input, calculating the vector representation of all inputs and questions and training the processing of the attention mechanism with questions. The relevant facts are retrieved in the input.The memory module provides all the relevant vectors of facts to the answer module to generate answers. The results of the clarification question detection comparison experiment are as follows (Table 6):

Table 6. Comparative experiment on clarification question detection (Acc)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(test)
CNN	0.801	0.721	0.767
Transformer	0.822	0.704	0.773
HAN	0.809	0.822	0.815
DMN	0.840	0.798	0.823
DeBERTa-v3 + FC	**0.928**	**0.946**	**0.935**

This comparative experiment shows the powerful ability of the clarification question detection model, overwhelmingly supassing the traditional neural network model. And the pre-trained model makes DeBERTa-v3 + FC surpass the DMN model by 11.2% in accuracy and it is also the only model that can achieve more than 90% of the results. However, the recognition accuracy of the multi-turn model exceeds that of the single-turn model, while the traditional neural network models have high single-turn question answering accuracy. This indicates that our DeBERTa-v3 + FC model is different from the traditional model in extracting key information. The more information there is, the higher the recognition rate and the better the generalization ability.

The Results of Clarification Question Generation. Due to the memory limit of the graphics card, the base ByT5 is selected for the experiment.The ablation experiment in this section uses T5 (base), mT5 (base) and ByT5 (small) for experiments. The Module1 and Module2 ablation experiments are shown in Tables 7 and 8 below:

Table 7. Result generated by Module1 (BLEU)

Model	Single + Multi (val)	Single + Multi(test)
T5	0.476	0.463
mT5	0.485	0.479
ByT5-small	0.483	0.462
ByT5-base	**0.515**	**0.499**

It can be seen from the above table that the generative model based on ByT5-base shows very good performance, improved by 7.21% and 11.88%, respectively, compared to the original T5 model on the test set. When compared with the generative model based on ByT5-small, the larger pre-training model has a positive effect on the BLEU score. However, the results are often closely related to training data, and there are still differences between different data.

The clarification question generation comparison experiment adds the original Seq2Seq model and the Transformer model. The results are shown in the following table (Table 9):

Table 8. Result generated by Module2 (BLEU)

Model	Single + Multi (val)	Single + Multi(test)
T5	0.568	0.534
mT5	0.572	0.561
ByT5-small	0.592	0.606
ByT5-base	**0.619**	**0.606**

Table 9. Comparative experiment on clarification question final generation result (BLEU)

Model	Single + Multi (val)	Single + Multi(test)
Seq2Seq	0.398	0.403
Transformer	0.421	0.423
T5	0.436	0.434
mT5	0.425	0.427
ByT5-small	0.438	0.435
ByT5-base	**0.447**	**0.449**

According to the comparative experimental results table, the effectiveness of the traditional model does not differ significantly from that of the ByT5 based model, which is significantly different from the BLEU scores obtained in the Module 1 and Module 2 sections above. This is due to the addition of a question and answer terminology template when splicing into the final answer. The model generates different answers each time, resulting in a decrease in the BLEU score, but this does increase the diversity of the final generated answer, making each generated answer personalized without losing its original semantics.

In this task, the attempted model structure is evaluated, and the following three examples are selected for display (The table can be found in the appendix), including single-turn and multi-turn question answering, including both generated answers that are almost the same as the ground truth and generated answers that are different from the ground truth's sentence structure with the same semant, and it is more in line with the current context. Our goal is to find simpler and more representative answers, simplify the generation of clarification questions, and enable users to better understand the generated answers for the required clarification content.

In the examples, Example 1 is a single-turn dialogue. The answer generated based on the ByT5 model is consistent with the semantic of the ground truth, only inconsistent at the template, and can be completely used as an alternative answer. Example 2 is a multi-turn dialogue, again only inconsistent at the template, and all other anthers are the same. Example 3 is a multi-turn dialogue. When generating clarification questions, the more obvious different characteristics of the entities are extracted, so that users can understand the clarification content more clearly.

The Results of Entity Prediction. Similar to the clarification question detection, the same model was selected for ablation and comparative experiments. The results are shown in Tables 10, 11 and 12 below:

Table 10. Ablation experiment on entity prediction (Acc)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(val)	Single + Multi(test)
Bert + FC	0.865	0.873	0.878	0.866
RoBERTa + FC	0.870	0.878	0.886	0.875
Longformer + FC	0.979	0.974	0.983	0.976
DeBERTa v3 + FC	**0.982**	**0.997**	**0.991**	**0.989**

Table 11. Ablation experiment on entity prediction (F1)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(val)	Single + Multi(test)
Bert + FC	0.865	0.873	0.877	0.866
RoBERTa + FC	0.871	0.879	0.880	0.878
Longformer + FC	0.980	0.974	0.982	0.976
DeBERTa v3 + FC	**0.981**	**0.997**	**0.991**	**0.988**

Table 12. Comparative experiment on entity prediction (Acc)

Model	Single-Turn(test)	Multi-Turn(test)	Single + Multi(test)
CNN	0.801	0.721	0.784
Transformer	0.822	0.704	0.789
HAN	0.809	0.822	0.834
DMN	0.840	0.798	0.839
DeBERTa v3 + FC	**0.982**	**0.997**	**0.989**

The long-sequence model Longformer also showed very good accuracy in this training. However, again, the DeBERTa-v3-base pre-trained model + FC fine-tuned model has the best effect, reaching 0.99 +. Similarly, the accuracy rate on the test set is close to 0.99, more than 15 percentage points more than the traditional neural network model, indicating that the model has played a full role in understanding the problem and distinguishing different entities. Similarly, the model performs better in the multi-turn question answering than in single-turn question answering. The traditional model has a better effect of single-turn question answering, indicating that the proposed DeBERTa

v3 + FC model has a better understanding of the context and can extract and understand the semantic information proposed by the user.

6 Conclusions and Future Work

This paper proposes a clarification question detection and generation framework, which consists of three tasks, i.e. clarification question detection, clarification question generation and entity prediction. Specifically, The DeBERTa v3 + FC model architecture is proposed for clarification question detection and entity prediction. A clarification question generation model based on ByT5 is proposed for generating clarification questions. The proposed approach greatly addresses the difficulty of poor interaction with users during the intelligent Q&A process, maximizes the understanding of the semantics of user contextual conversations, and reduces the possibility of system misunderstandings of user intentions. Experimental results show that on the MSParS public dataset, the accuracy of clarification question detection and entity prediction tasks improved by 11.2% and 15.17%, respectively, compared to traditional DMN models. The BLEU score of the clarification question generation task improved by 4.9% compared to traditional Seq2Seq models, achieving breakthroughs in all three tasks. Combined with weak templates, the generated clarification questions are personalized and more understandable to users. We will conduct further research on the generation of clarification questions to enable them to have excellent performance on issues in all fields, and improve the timeliness of the output of results (Table 13).

Funding. This research was funded by Special Fund Project for Promoting High Quality Industrial Development in Shanghai (Number: 2021-GZL-RGZN-01018) and Shanghai "Science and Technology Innovation Action Plan" Project (Number: 21DZ1201400; 22DZ1200500; 22QB1400200).

Appendix

Table 13. Examples of clarification question final generation result (Examples).

Example 1	Context	The Comet's costume designer
	Entity1	The Comet <S> media_common.creative_work media_common.cataloged_instance broadcast.content film.film ratings.rated_entity <S> The Comet is a 1996 drama film written by Claude Santelli and Suzanne Jacques-Marin and directed by Claude Santelli.
	Entity2	The Comet <S> film.film award.**nominated_work** ratings.rated_entity media_common.creative_work media_common.cataloged_instance broadcast.content <S> After witnessing the arrest of her father for publishing subversive material against the dictatorship of Porfirio Díaz, Valentina escapes taking a sack of gold coins with her in order to ...
	Our model	When you say the costume designer, are you talking about non nominated work The Comet or nominated work The Comet?
	Ground Truth	Which one do you mean, non nominated work The Comet or nominated work The Comet, when you say the dress designer?
Example 2	Context	Who follows madonna? <EOS> Tila Tequila <EOS> What was a famous quote quoted?
	Entity1	madonna <S> biology.organism award.winner award.ranked_item award.nominee award.hall_of_fame_inductee award.competitor media_common.subject tv.actor theater.actor ratings.rated_entity people.person organization.founder tv.crewmember music.producer music.musician music.lyricist music.composer music.artist me-

(*continued*)

Table 13. (*continued*)

	dia_common.cataloged_instance internet.social_network_user film.writer film.subject film.producer music.singer tv.writer tv.personality film.music_contributor film.director film.actor fictional_universe.person_in_fiction event.agent celebrities.celebrity broadcast.artist book.author <S> Madonna Louise Ciccone (/tʃɪˈkoʊni/; Italian: [tʃikˈkoːne]) (born August 16, 1958) is an American singer, songwriter, actress, and businesswoman. She achieved popularity by pushing the boundaries of lyrical content in mainstream popular music and imagery in her music videos, which became a fixture on MTV. Madonna is known for reinventing both her music and image, and for maintaining her autonomy within the recording industry. Music critics have acclaimed her musical productions, which have generated some controversy. Often referred to as the Queen of Pop, she is cited as an influence by numerous other artists around the world.
Entity2	Tila Tequila <S> internet.social_network_user media_common.cataloged_instance music.musician people.person ratings.rated_entity tv.actor award.competitor award.nominee award.ranked_item award.winner biology.organism celebrities.celebrity event.agent film.actor tv.personality music.artist <S> Tila Tequila net worth: Tila Tequila is a Singaporean-American model and television personality who has a net worth of $500 thousand. Tila Tequila Nguyen was born in Singapore in October 1981. Her family moved to Houston, Texas when she was a year old. In high school she started using drugs and joined a gang. Her friends gave her the nickname Tila Tequila due to her allergy to alcohol. As a teen she experienced a drive-by shooting, and became pregnant before suffering a miscarriage. In 2001 she moved to California. At 19 she was discovered in a shopping mall by a Playboy scout. She was Playboy's Cyber Girl of the week in April 2002 and became the first Asian Cyber Girl of the month. She became popular through the import racing scene and was featured on the covers of magazines, at car shows, and in video games. She hosted Fuse TV's Pants-Off Dance-Off. She has been featured on the cover of Stuff and Maxim UK. She was a contestant on the NC show Identity and appeared in I Now Pronounce You Chuck and Larry. In 2009 she was featured in the MTV reality show A Shot At Love with Tila Tequila. The show was a bisexual-themed dating which featured straight men and lesbian women as would-be suitors. In 2006 she was signed to the Will.I.Am music group. She has released two solo EPs and five singles. In 2008 she released a self-help book. She has also started her own record label called Tila Tequila Records and a management firm called Little Miss Trendsetter Management LLC. Tila has been romantically linked to football player Shawne Merriman and heiress Casey Johnson. In 2011 a sex tape of Tequila was released. In 2012 she checked into rehab, and in 2014 she announced she is

(*continued*)

Table 13. (*continued*)

		pregnant with her first child.
	Our model	When you say the person's famous quote, are you referring to madonna or Tila Tequila?
	Ground Truth	Are you talking about music contributor madonna or Tila Tequila when you say the person's famous quote?
Example 3	Context Entity1	What is island group of Seymour Island Seymour Island <S> location.administrative_division geography.island geography.geographical_feature location.location travel.destination media_common.subject <S> North Seymour is a small island near to Baltra Island in the Galapagos Islands. It was formed by uplift of a submarine lava formation. The whole island is covered with low, bushy vegetation. The island is named after an English nobleman, Lord Hugh Seymour. It has an area of 1.9 square kilometres and a maximum altitude of 28 metres. This island is home to a large population of blue-footed boobies and swallow-tailed gulls. It hosts one of the largest populations of frigatebirds Fregata magnificens and a slow growing population of the Galapagos land iguana. North Seymour has a visitor trail approximately 2 kilometres in length crossing the inland of the island and exploring the rocky coast. One of the most famous birds found in the Galapagos are the Blue-footed Booby that are found on North Seymour. The stock for the captive breeding program of the Galapagos Land Iguana is descended from iguanas which Captain G. Allan Hancock translocated from nearby Baltra Island to North Seymour Island in the 1930s. This was very important because Baltra Island during World War 2 was populated by airplane base by the U.S.
	Entity2	Seymour Island <S> location.location geography.island geography.geographical_feature <S> Seymour Island is an uninhabited island in the Qikiqtaaluk Region of northern Canada's territory of Nunavut. A member of the Berkeley Islands group, it is located approximately 30 mi (48 km) north of northern Bathurst Island. Between Seymour Island and Bathurst Island lies Helena Island. Penny Strait lies about 90 km (56 mi) to the east where open water polynyas occur.
	Our model	Do you mean Seymour Island in the Galapagos Islands or Seymour Island in Nunavut, when you say the sequence of islands?
	Ground Truth	When you say the name of the group of islands, are you talking about Seymour Island near to Baltra Island in the Galapagos Islands or Seymour Island in the Qikiqtaaluk Region of northern Canada's territory of Nunavut?

References

1. Shin, S., Lee, K.H.: Processing knowledge graph-based complex questions through question decomposition and recomposition. Inf. Sci.f. Sci. **523**, 234–244 (2020)

2. Demetras, M.J., Post, K.N., Snow, C.E.: Feedback to first language learners: the role of repetitions and clarification questions. J. Child Lang. **13**(2), 275–292 (1986)
3. Aliannejadi, M., Zamani, H., Crestani, F., et al.: Asking clarifying questions in open-domain information-seeking conversations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 475–484 (2019)
4. Stoyanchev, S., Liu, A., Hirschberg, J.: Towards natural clarification questions in dialogue systems. In: AISB Symposium on Questions, Discourse and Dialogue, pp. 20 (2014)
5. Guo, X., Klinger, T., Rosenbaum, C., et al.: Learning to query, reason, and answer questions on ambiguous texts. In: International Conference on Learning Representations (2017)
6. Li, J., Miller, A.H., Chopra, S., et al.: Dialogue learning with human-in-the-loop. arXiv preprint arXiv:1611.09823 (2016)
7. Xu, J., Wang, Y., Tang, D., et al.: Asking clarification questions in knowledge-based question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1618–1629 (2019)
8. Liu, B., Zhao, M., Niu, D., et al.: Learning to generate questions by learningWhat not to generate. In: The World Wide Web Conference, pp. 1106–1118 (2019)
9. Dong, L., Yang, N., Wang, W., et al.: Unified language model pre-training for natural language understanding and generation. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
10. Gao, Y., Bing, L., Chen, W., et al.: Difficulty controllable generation of reading comprehension questions. arXiv preprint arXiv:1807.03586 (2018)
11. Rao, S., Daumé III, H.: Answer-based adversarial training for generating clarification questions. arXiv preprint arXiv:1904.02281 (2019)
12. He, P., Liu, X., Gao, J., et al.: DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)
13. Liu, Y., Ott, M., Goyal, N., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. He, P., Gao, J., Chen, W.: Debertav3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
15. Christian, H., Agus, M.P., Suhartono, D.: Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech Comput. Math. Eng. Appl. **7**(4), 285 (2016). https://doi.org/10.21512/comtech.v7i4.3746
16. Raffel, C., Shazeer, N., Roberts, A., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
17. Xue, L., Barua, A., Constant, N., et al.: Byt5: towards a token-free future with pre-trained byte-to-byte models. Trans. Assoc. Comput. Linguist. **10**, 291–306 (2022)
18. Papineni, K., Roukos, S., Ward, T., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
19. Yang, Z., Yang, D., Dyer, C., et al.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. (2016)
20. Kumar, A., Irsoy, O., Ondruska, P., et al.: Ask me anything: dynamic memory networks for natural language processing. In: International Conference on Machine Learning, pp. 1378–1387. PMLR (2016)