修　士　論　文　の　和　文　要　旨

研究科・専攻	大学院　情報理工学研究科　情報学専攻　博士前期課程		
氏　　　　名	SUN QIFAN	学籍番号	1830126
論　文　題　目	Automatic Generation of Rhetorical Questions and Its Application to a Chatbot		

要　　旨

近年、チャットボットをより人間に近いものにする研究が増えた。例えば、皮肉を生成するチャットボットなどの研究が提案されている。しかし、修辞疑問を生成する研究はまだ存在しない。修辞疑問は質問だが、答えを得ることを目的としたものではない。日常会話やソーシャルメディアの対話では修辞疑問が使われることが多いため、チャットボットをより人間らしくするためには修辞疑問を生成する必要がある。
修辞疑問を認識して話者の意見を表すためには、聞き手がお互いに共有している知識を利用すると、話者返事の前に、聞き手が答えを知った。そうすると、話者の観点が明確になる。さらに、非常識的な知識を基づいて、修辞疑問を生成すれば、修辞疑問として認識される可能性が高まる。本研究は常識の文を反転し、非常識な文として用いて、修辞疑問を生成するジェネレーターを作った。チャットボットが返事する前に前の対話を考慮して、返事すべきタイプを皮肉か、修辞疑問か、リテラルの返事とかを判断し返事する。
そして、４２の対話例をチャットボットで全ての返事を生成し、その返事は対話中の適切性と人間性を評価した。その結果、本研究のチャットボットは、修辞疑問を生成すべき時、修辞疑問が対話中の適切性がリテラルと皮肉の返事より優れた結果となる。

Automatic Generation of Rhetorical Questions and Its Application to a Chatbot

In partial fulfillment of the requirements for
The degree of Master of Engineering:

Student Number: 1830126
Student Name: Qifan Sun
Supervisors: Akira Utsumi, Professor
Suguru Matsuyoshi, Assistant Professor

Department of Informatics
Graduate School of Informatics and Engineering
The University of Electro-Communications
Chofu, Tokyo, Japan

January 25, 2021

Acknowledgements

I would like to thank Professor Akira Utsumi and Assistant Professor Suguru Matsuyoshi for supporting me during my master's degree.

I would also like to thank the members of Utsumi Lab. at the University of Electro-Communications for useful suggestions about study and life.

Contents

1 Introduction

In recent years, an interpersonal attraction in the conversation has been studied extensively. Figurative language such as sarcasm and the rhetorical question is a ubiquitous part of conversation[22]. Creating figurative language generation modules can make chatbots become more human-like[13]. Some recent studies have proposed chatbots that generate sarcasm[18][2]. However, they do not focus on generating rhetorical questions (RQs). It is necessary for chatbots to generate RQs to be more human-like because RQs are usually used in daily conversation and social media dialog[19].

RQs are questions but not meant to obtain an answer[11]. People usually use them to express their opinions in conversation. Most of the time, a question can be recognized as an RQ when both the speaker and the listener know the answer or the question's situation. For example, "[The situation: Someone arrives late for a meeting. As the chair of the meeting enters, they utter:] Do you know what time it is?[9]" All of the speakers and the listener know the time. Therefore the listener knows that the speakers want to declare that "the listener is late."

To recognize an RQ and obtain the speaker's opinion, the listener needs to use the knowledge shared between them. Furthermore, there is a specific interrelation between irony and RQs[19]. Therefore RQs are always used to express their negative opinions. Questions based on the valence-reversed commonsense knowledge can be easily recognized as RQs because both speaker and listener know their answer are negative. For example, the commonsense knowledge "Giving money to the poor will make good world" can be converted into an RQ: "Will giving money to the rich make a good world?"

This study aims to generate a negative-answering RQ by using valence-reversed commonsense knowledge sentences to make the chatbot more appropriate and human-like in a conversation. Additionally, we use a situation classifier analyzing previous contexts to decide when to generate a literal response, sarcastic response, and RQ.

The rest of this paper is as follows in Section 2, we review previous researches related to the RQ generation and the classification for conversation. In Section 3, we introduce the techniques that we used in this study. Pre-trained models: BERT, RoBERTa, common sense knowledge scoring model, grammatical error correction model, DIALOGPT; Neural network architecture: Bi-LSTM; modules: Sarcasm generation. In Section 4, we show the methodology of the chatbot, the situation classification, and the RQ generation. In Section 5, we illustrate the situation classifier and RQ generator's evaluation and their result. In Section 6, we conclude a summary and describe the future work.

2 Related Work

2.1 Classification for the Conversation

Ghosh et al.[7] proposed a sarcasm detection for conversation by using pre-trained word embedding models[6] and attention-based LSTM models to analyze contexts and responses separately. They address that the previous contexts have a relationship with the response that improves the sarcasm detection performance.

Oraby et al.[19] proposed an RQ classification for conversation using SVM and LSTM models. They also experimented with LIWC categories as additional features. The classification was trained by the merge of RQ, a previous context, and a past context.

In this study, we only select the previous contexts to classify the situation in which response should be output in a conversation by using a neural network which is introduced in Section 4.2.

2.2 Sarcasm Generation

We do not find any research on RQ generation. However, there is a relationship between sarcasm and rhetorical question. We refer to the research of sarcasm generation to obtain the idea of a rhetorical question.

Research on sarcasm generation is just started. Joshi et al.[13] presented a SarcasmBot that extracted information from user input by a rule-based generator selector. It decided the sarcasm generator with eight-generation modules (such as offensive word response generator and hyperbole generator) to generate a sarcastic response.

Mishra et al.[18] proposed a sarcasm generator that converted a literal negative opinion into a sarcasm. Their method was based on a characteristic of sarcasm that it usually had both of the positive sentiment phrase and the negative situation at the same time. The first filtered the sentiment words and phrases from the input. Next, they reversed the sentiment words and phrases, which had a strong positive sentiment. Then, they used the filtered input to retrieve the related negative situation sentences from the negative sentiment sentences corpus. Finally, they synthesized the positive sentiment phrase and the negative situation phrase to obtain a sarcastic output.

Chakrabarty et al.[2] proposed a state-of-the-art sarcasm generator that converted a literal input into a sarcastic sentence. We use it in our chatbot to generate sarcastic responses, so we introduce it in Section 3.1.

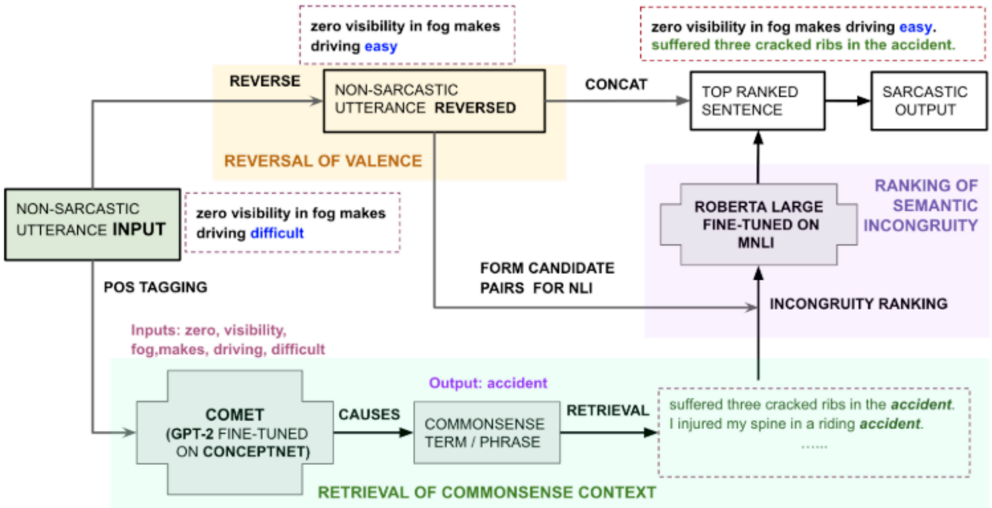


Figure 3.1: The pipeline of the sarcasm generation[2].

3 Technical explanation

3.1 Sarcasm Generation

Chakrabarty et al.[2] proposed a state-of-the-art sarcasm generator that converted a literal utterance into a sarcastic sentence based on commonsense knowledge. We use it in our chatbot to generate sarcastic responses and evaluate them by contrasting them with RQ responses and literal responses. The pipeline of the whole sarcasm generation is shown in Figure 3.1.

They create three modules to generate sarcasm. These modules are introduced below.

3.1.1 Reversal of Valence

There is a characterization of the sarcasm that usually contains a positive sentiment phrase and a negative situation[15]. So this module uses a word-level negative score obtained from SentiWordNet[5] to reverse the word's valence with the highest negative sentiment score. It makes the positive sentiment phrase in the sentence. When the negation word “not” or word ending in “n't” exist, they remove it without reversing the negation word.

3.1.2 Retrieval of Commonsense Context

First, this module extracts nouns, adjectives, adverbs, and verbs from the literal utterance. Next, it feeds them to the COMET[1] model initiated with a pre-trained GPT[20] model and fine-tuned by the ConceptNet corpus[25] to get the commonsense term or phrase. Then, they use it to retrieve sentences from the corpus of a high-quality sentence website[1]. Finally, they modify the subject of the sentence to keep it the same as the original sentence and use a neural grammatical error correction system[28] to correct grammar errors for it.

3.1.3 Ranking for Semantic Incongruity

They fine-tuned a RoBERTa-large[17] pre-trained model to calculate semantic incongruity scores for ranking the sentences' appropriateness. After using the model to predict the rank of sentences, the module concatenates the literal utterance with the top-ranked sentence to finish the generation of a sarcasm.

3.2 Transformer Pre-trained models

We fine-tuned four pre-trained models, i.e., BERT, BERT Large, RoBERTa, and RoBERTa Large model in this study for the situation classification.

3.2.1 BERT

Devlin et al.[4] proposed a language representation model that stands for bidirectional encoder representations from Transformers (BERT). Unlike traditional encoder representation models, BERT is able to pre-train representations from unlabeled text by conditioning both right and left contexts in all layers.

The BERT encoder uses the encoder part of Transformer. It is a multi-layer bidirectional Transformer encoder. Because the self-attention mechanism is allowed in the Transformer, BERT also uses the self-attention. So it can be fine-tuned easily. Different from GPT[20] Transformer uses a one-way self-attention that every token can only attend to context to its left, BERT Transformer uses bidirectional self-attention. These self-attentions arrange in parallel, called Multi-Head-Attention, as shown in the "Trm" part of Figure 3.2. BERT is a combination of these units, as shown in Figure 3.2.

[1] https://sentencedict.com/

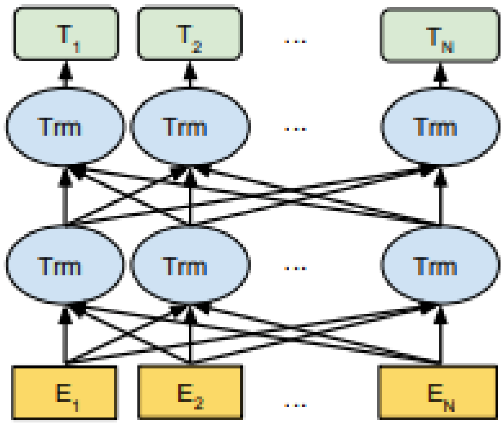


Figure 3.2: The structure of BERT model[4].

In BERT, A sentence or a pair of sentences specified as a token embedding. The token embedding for them uses WordPiece embeddings[26]. With a 30,000 token vocabulary to divide and tokenize in units more refined than word units. Every sequence's first token is always a unique classification token ([CLS]) for the first token in the series. There is always a special token ([SEP]) in the middle of two sentences to differentiate the sentences, and it is also used in the last token.

The pre-learned models distributed by Google in the repository uses Books Corpus (800 million words) and English Wikipedia (2.5 billion words) for learning. The BERT-Base-uncased has 12 layers, 768 hidden layers, 12 attention heads, and 110M parameters. The BERT-Large-uncased has 24 layers, 1024 hidden layers, 16 attention heads, and 340M parameters.

3.2.2 RoBERTa

Liu et al.[17] proposed a robustly optimized BERT pretraining approach (RoBERTa). RoBERTa has the same architecture as BERT, but it uses a different tokenizer that is a byte-level BPE. Therefore, Roberta does not have the special token ([SEP] and [CLS]) and token type IDs are no longer used. They improved BERT by four measures:

- Training the model longer with bigger batches and more data.
- Removing the next sentence prediction objective.
- Training on longer sequences.

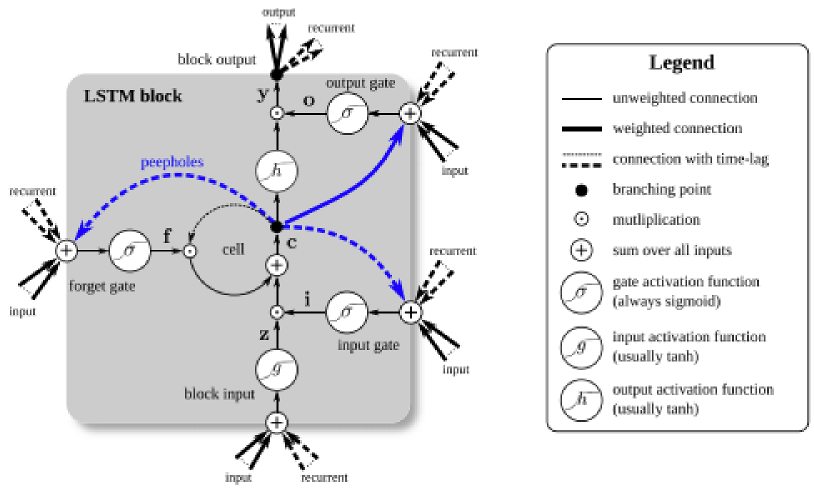


Figure 3.3: The structure of an LSTM memory block[10].

- Dynamically changing the masking pattern applied to the training data.

They also use a larger dataset (CC-NEWS) to train the model.

Every sequence's first token is always a special classification token ($< /s >$) for the first and last token in the series. This token is also added in the middle of two sentences to differentiate the sentences.

The pre-trained RoBERTa model has 12 layers, 768 hidden layers, 12 attention heads, and 125M parameters. The RoBERTa-Large-uncased has 24 layers, 1024 hidden layers, 16 attention heads, and 355M parameters.

3.3 Bidirectional Long Short-Term Memory

3.3.1 Long Short-Term Memory

Long Short-Term Memory (LSTM) is an architecture that is proposed by Hochreiter et al.[12] and extended by Greff et al.[10]. An LSTM layer consists of a set of recurrently connected blocks, which are known as memory blocks. Each of them includes recurrently connected memory cells and three units: the input, output, and forgets gates that provide continuous analogs of write read and reset operations for the cells. Figure 3.3 shows the structure of the memory block.

The network that contains LSTM layers is well-suited to the classifying tasks because it is developed to solve the vanishing gradient problem in training.

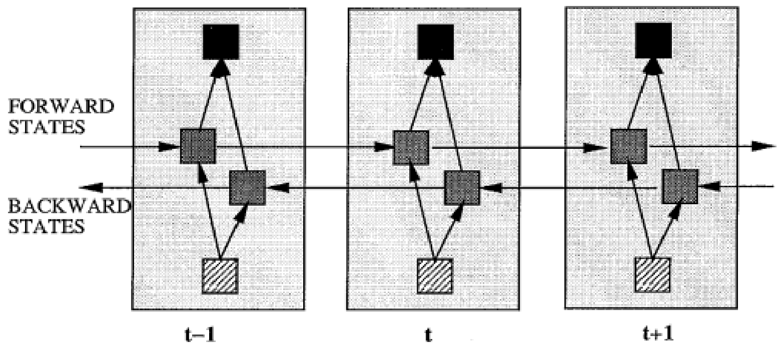


Figure 3.4: The structure of a BRNN unfolded in three-time steps[23].

Table 3.1: Tuples with the left term "soak in hotspring" when searching in the common-sense knowledge base

relation	right term
UsedFor	heal
HasPrerequisite	take shower
UsedFor	relaxation
MotivatedByGoal	your muscle be sore

3.3.2 Bidirectional Recurrent Neural Networks

Bidirectional long short-term memory (Bi-LSTM) network is one of the bidirectional recurrent neural network (BRNN)[23]. The BRNN connects two hidden layers of opposite directions to the same output. Therefore, the output layer can receive data from backward and forward states at the same time. Figure 3.4 shows the structure of a BRNN unfolded in three-time steps. It works well in context input.

3.4 Common Sense Knowledge Scoring Model

Li et al.[16] convert ConceptNet's[24] commonsense resources by formulating the problem as a commonsense knowledge base. The base contains 600,000 tuples of commonsense knowledge phrases that the tuples consist of a relation, a left term, and a right term. When we search the "soaking in a hotspring" in the base, some examples are shown in Table 3.1.

They trained a neural network model for scoring the relationship of two tuples by using the base. The base has 34 relation types. The model analyzes the left term and the right term to predict each relation type's probability.

3.5 Grammatical Error Correction Model

Grammatical Error Correction (GEC) is a sequence to sequence task where a model corrects an ungrammatical sentence to a grammatical sentence. Kaneko et al.[14] proposed a model for GEC. They incorporated a pred-trained mask language model (MLM) into an encoder-decoder model to deal with the GEC task. They called it the BERT-fuse GED model.

The MLM that Kaneko et al. used is a BERT-fused neural machine translation (NMT) model developed by Zhu et al.[29] uses the representation from BERT by feeding it into all layers instead of feeding input embeddings only. It also uses the attention mechanism to adjust how each layer interacts with the representations.

The BERT-fuse GED model is one of the most effective techniques in GEC tasks. We use it to check and revise the grammatical error in our generated sentences.

3.6 DIALOGPT model

Zhang et al.[27] proposed a large, tunable neural conversation response generation model (DIALOGPT) that is able to produce consistent responses. The model is based on the GPT-2. It is trained on about 147M posts from Reddit comments. These comments are filtered by removing the contexts and the responses that accord with the rules below:

- It contains a URL.
- It contains a word that is repeated at least three times.
- The response does not contain at least one of the top-50 most frequent English words.
- The response contains special markers.
- It is longer than 200 words.
- It contains offensive language which exists in a large blocklist.

They trained the model by three different configurations, which are shown in the Table 3.2.

Table 3.2: Model configurations. “B” denotes batch size per GPU, “D_emb” is the Embadding size[27].

Model	**Layers**	**D_emb**	**B**
117M	12	768	128
345M	24	1024	64
762M	36	1280	32

4 Methodology

4.1 Chatbot

The chatbot in this study works as shown in Figure 4.1. It uses the user's utterances and previous utterances to generate appropriate responses. The following three steps process the chatbot.

4.1.1 Step A

The situation classifier selects an appropriate response type from among sarcasm, RQ, and literal responses. The situation classifier is introduced in Section 4.2.

4.1.2 Step B

The literal response generator generates a literal response by using preceding utterances. We use the DIALOGPT-Large model that is introduced in Section 3.6 after fine-tuning to generate literal responses. We fine-tune the DIALOGPT model by setting the parameters. The settings are shown in Listing 1. All of them are used to make the model generate more human-like utterances than the original model.

do_sample=True The model can randomly choose the next word ω_t in the generation according to its conditional probability distribution shown in Equation 4.1.

max_length=1000 It makes the model can only admit the first 1000 words.

top_k=50 It limits the sampling pool to the 50 most similar words for the next word candidate.

top_p=0.95 The number of samplings chosen from the sampling pool, which is as small as possible. Simultaneously, the sum of selected words ' probability mass must higher than the probability of 0.95.

temperature=0.9 The temperature makes the next word distribution less random when it is close to one.

Generated literal responses are directly used to respond to a chatbot (when the situation classifier judged that a literal response is most appropriate) or input to the RQ and sarcasm generator.

4.1.3 Step C

When the situation classifier decides to generate an RQ response, an RQ generator converts the literal response into an RQ response, explained in Section 4.3. When the classifier decides to generate a sarcasm response, a sarcasm generator converts the literal utterance into a sarcasm response. We use the sarcasm generator proposed by Chakrabarty et al.[2]. It concatenates valence-reversed literal utterances with the sentences retrieved from an online sentence dictionary with high-quality sentences using commonsense phrases.

$$\omega_t \sim P(\omega|\omega_{1:t-1}) \tag{4.1}$$

Listing 1: The parameters that we set for the DIALOGPT model

```
do_sample=True,
max_length=1000,
top_k=50,
top_p=0.95,
temperature=0.9,
```

4.2 Situation Classification

In this study, we fine-tune four pre-trained models, i.e., BERT, BERT large, RoBERTa, and RoBERTa large, to classify which of RQ, sarcastic and literal responses are appropriate according to the current situation.

Two previous contexts before a response in a conversation are the target that we choose to analyze. When several previous contexts exist in a conversation, we choose the latest two contexts, i.e. there is a conversation that contains "context_3", "context_2", "context_1" and "response". The "response" is a reply to the "context_1". We choose "context_2" and "context_1". Next, we add the separation token between them and then concatenate them to one sequence. Then we encode the sequence by the pre-trained model. Finally, we input it into our classification model. We describe two model families for the situation classification. The first model is a base-line.

Figure 4.2 shows the structure of the BERT base-line model. Figure 4.4 shows the structure of the BERT large model. Figure 4.6 shows the structure of the roBERTa model. Figure 4.8 shows the structure of the roBERTa large model.

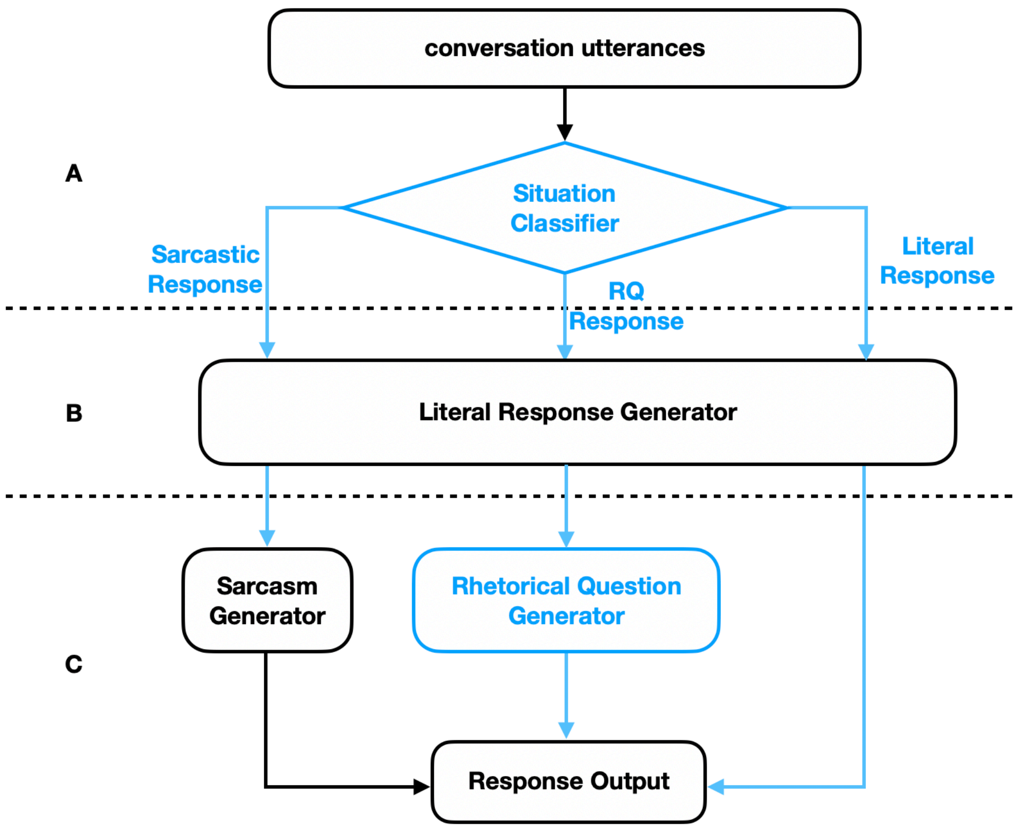


Figure 4.1: The flowchart of the chatbot: The components originally proposed in this study are indicated in blue.

Parentheses in the figures show the input shape and output shape of current layers. The question mark means this dimension is variable. They are defined in the fit methods. BERT and BERT-large use the same model as well as Roberta and Roberta-large use the same model. The difference between them is the output shape because large models have a large size of the hidden layers.

- BERT models have three input layers that are the input IDs (input_ids), attention mask (attention_mask), and token type IDs (token_type_ids). All of them are generated by the BERT tokenizer. The input IDs contain the index of the encoded sequence. The attention mask points out the range of tokens that the model should pay attention to it. Token type IDs are not used in the classification case, so it is initiated to a list of zero. Otherwise, RoBERTa models have two input layers: the input IDs (input_ids) and attention mask (attention_mask). All of them are generated by the RoBERTa tokenizer. The token type IDs are no longer used in RoBERTa.
- The TFBertModel is the pre-trained model.
- The flatten layer reduces the elements' dimension to one.
- The dropout layer drops out units randomly to prevent over-fitting too early.
- Softmax activation of the last layer to determine the results.

The second model uses two bidirectional LSTM layers instead of flatten layer to get a better result.

Figure 4.3 shows the structure of the BERT model with Bi-LSTM layers. Figure 4.5 shows the structure of the BERT large model with Bi-LSTM layers. Figure 4.7 shows the structure of the roBERTa model with Bi-LSTM layers. Figure 4.9 shows the structure of the roBERTa large model with Bi-LSTM layers.

4.3 RQ Generation

We propose an RQ generator that converts a literal response to an RQ using commonsense knowledge. The structure of the RQ generator is shown in Figure 4.10.

The generator reverses the valence of commonsense knowledge that is relevant to the literal response and converts it into an interrogative sentence. The method of valence reservation is the same as the sarcasm generation, which is introduced in Section 3.1.1

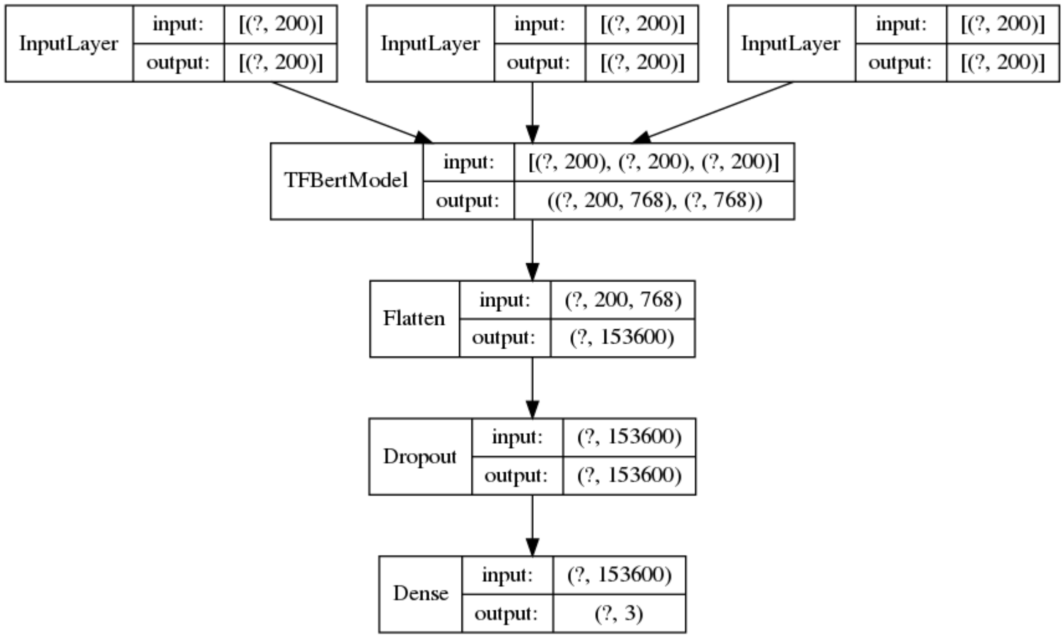


Figure 4.2: The structure of BERT model.

The generator analyzes the sentence structure of the literal response by using the part-of-speech tagging (POS) to mark up each word in the initial response as corresponding to a particular part of speech[3]. It also checks the sentiment polarity of all words.

When a verb has the highest sentiment score, the generator reverses its valence to reverse the whole meaning of the sentence. The generator extracts keywords from the original literal response, such as the verbs, nouns, and adjectives, and use the commonsense knowledge scoring model[16] to calculate a commonsense score. When the score is higher than 0.5, the valence-reversed literal response is converted to wh-question. The method of question generation is introduced in Section 4.3.1. When the score is lower than 0.5, the initial literal response is output directly.

When the highest sentiment word is an object, a subject, or no sentiment phrase, the generator extracts keywords from the literal response and searches on the commonsense knowledge sentences using the keywords to obtain commonsense knowledge sentences. The commonsense knowledge sentences is introduced in Section 4.3.3. When no sentence can be found, the initial literal response is output directly. Next, the generator selects the top ten commonsense knowledge sentences topically similar to the initial response. The framework of sentence transformers[21] scores the similarity of sentences. Then, the generator reverses the valence of these commonsense knowledge sentences and converts

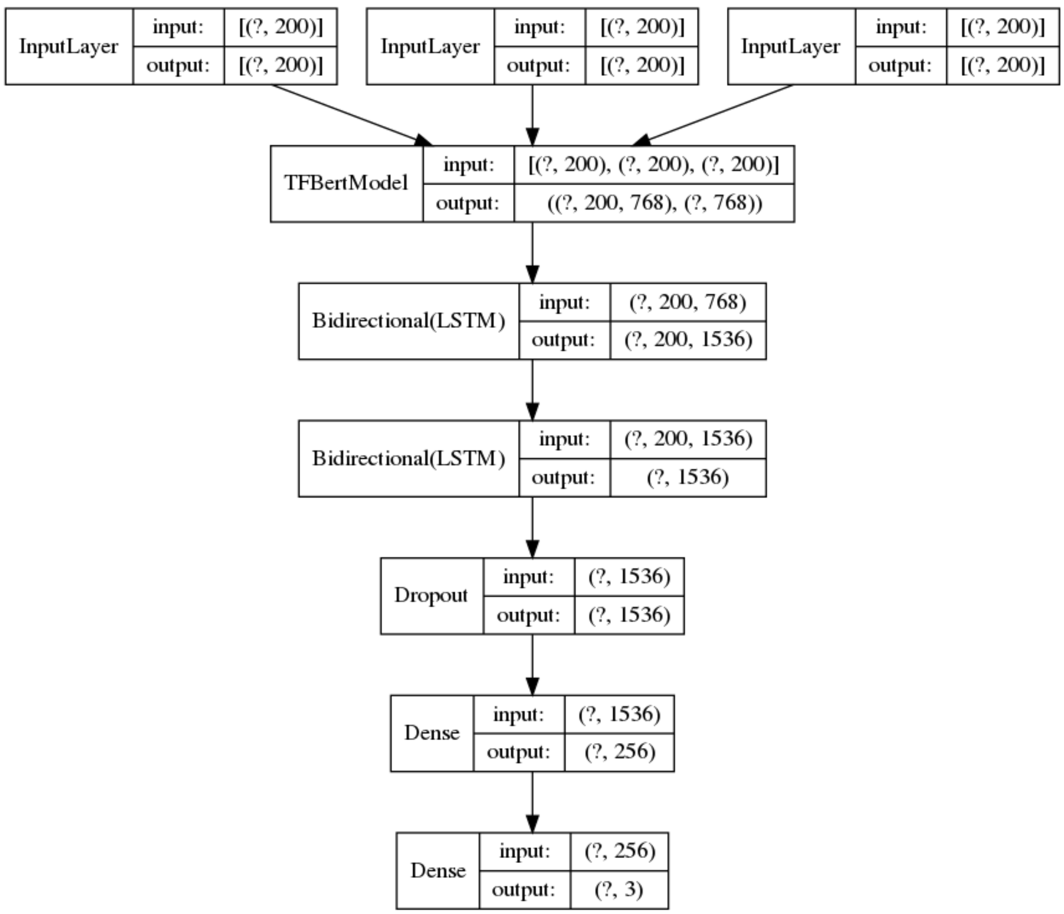


Figure 4.3: The structure of BERT model with bi-LSTM layers.

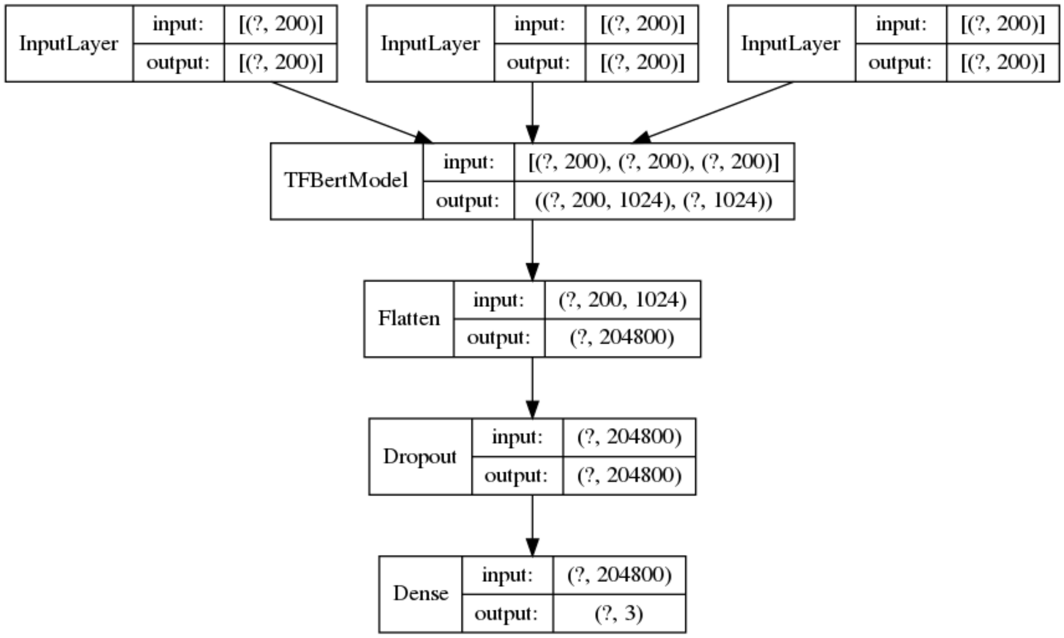


Figure 4.4: The structure of BERT large model.

them to yes-no questions. After that, it concatenates them behind the initial response to obtain RQs. Finally, the generator selects the most appropriate RQ response using a fine-tuned BERT model, which is introduced in Section 4.3.2 that computes candidate responses' appropriateness.

4.3.1 Question Generation

We use the classic method to convert the question because we do not find a dataset that contains short sentences and their questions. The question generation first obtains the POS marks for the initial response and all words' sentiment polarity.

When a verb has the highest sentiment score, the generator converts the response to wh-question. Firstly, it uses the Stanford Parser[2] to find named-entity recognition of the subject. Secondary, it matches the dictionary that we made to obtain interrogative words. The dictionary is shown as the Listing 2.

When there is no matching result in the dictionary, the generator finds the POS tags that are the Proper Nouns Singular (NNP), Proper Nouns Plural (NNPS), Common Nouns Singular or Mass (NN), and Common Nouns Plural (NNS) in the response. When there is no word matched with these POS tags, the initial response is output. When there are

[2]https://nlp.stanford.edu/software/lex-parser.shtml

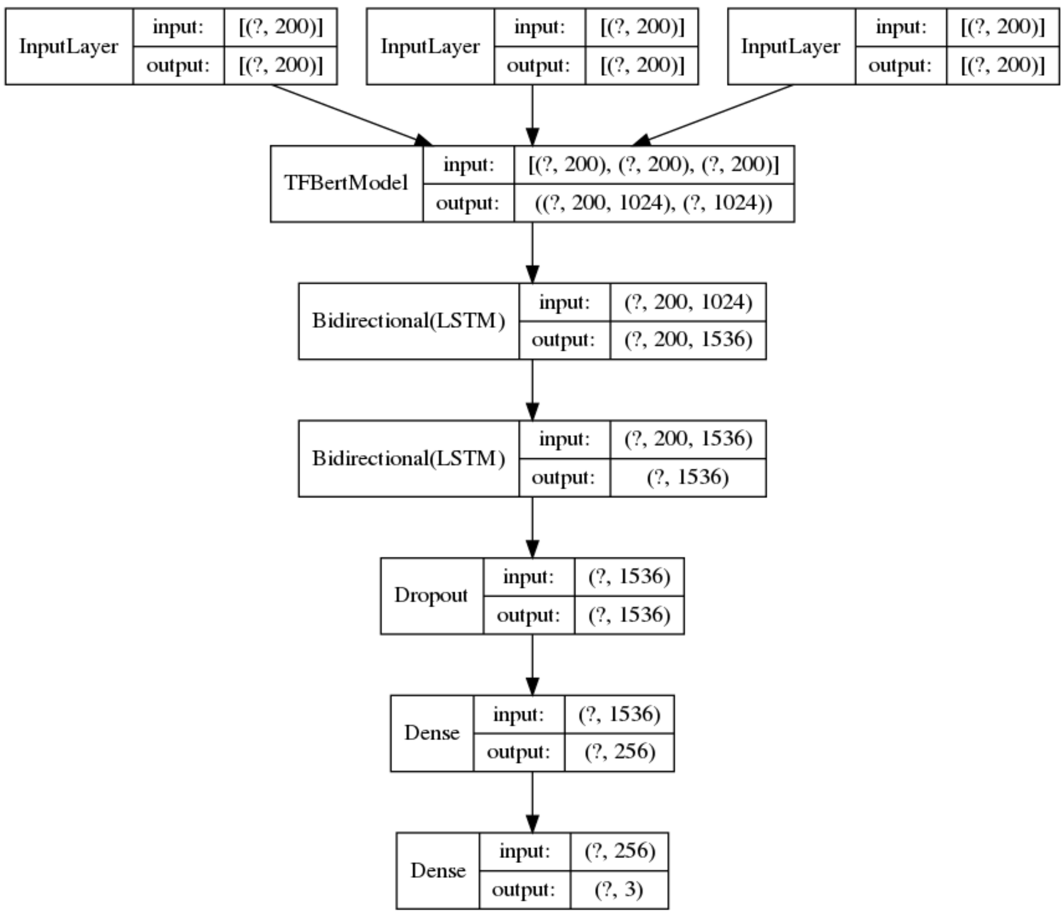


Figure 4.5: The structure of BERT large model with bi-LSTM layers.

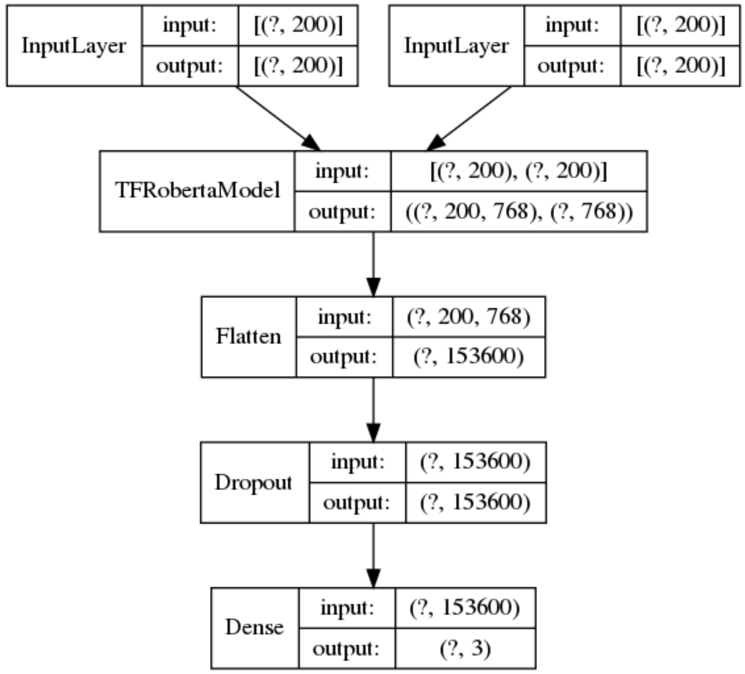


Figure 4.6: The structure of RoBERTa model.

words with these POS tags, the generator will do:

- It finds the words of indefinite pronouns that do not refer to any specific person, which are usually used in the conversation. Such as “everybody” and “everyone”. It also searches for the POS tag of Personal Pronouns (PRP)in the sentences. These words are converted to the interrogative words: “Who”.

- It searches for the POS tag of Possessive Pronouns (PRP$) and Possessive Endings ’s (POS) in the sentences. These words are converted to the interrogative words: “Whose”.

- The words of indefinite pronouns that do not refer to any specific things, such as “everything”, “nothing” and others will be converted to the interrogative words: “What”.

When the response has no highest sentiment scored verb, the generator converts the response to a yes-no question. When the first word in the response has the POS tag of “TO”, it will be converted into “Should we”. When the words in the response have the POS tag of Modal Verbs (MD), such as “should” and “will”, and there is a verb in the next two words simultaneously, the generator moves the word that has “MD” tag to the

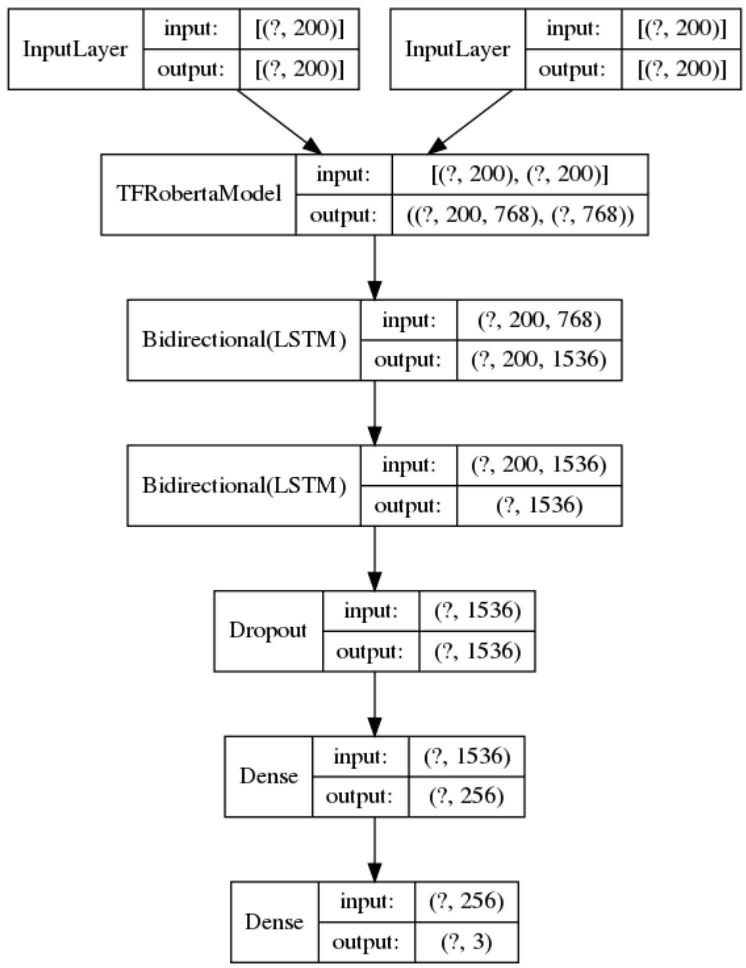


Figure 4.7: The structure of RoBERTa model with bi-LSTM layers.

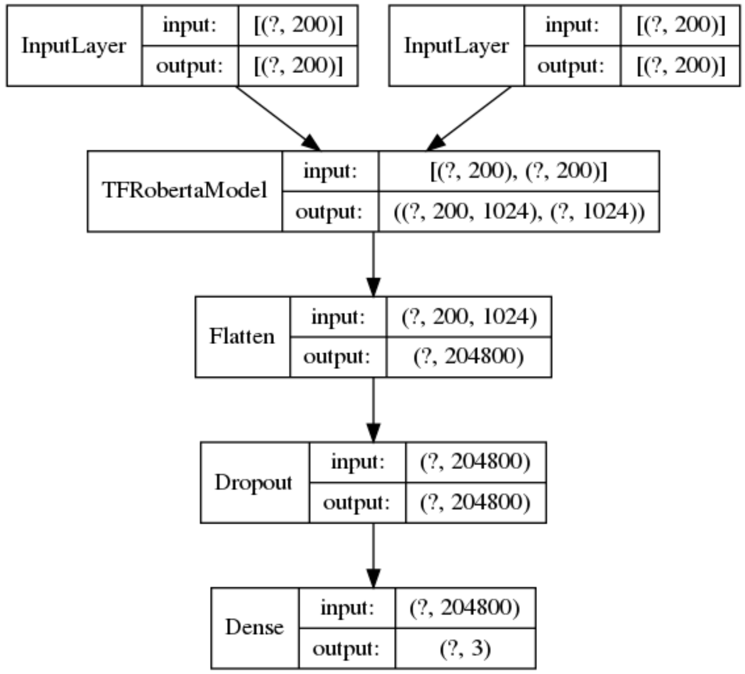


Figure 4.8: The structure of RoBERTa large model.

response's beginning. Otherwise, When the verb has the tags of the past tense (VBD), the third-person singular present (VBZ), past participle (VBN), or non-3rd person singular present (VBP), the generator adds "Do" to the response's beginning. The grammatical error correction model that we introduced in Section 3.5 can tenses problems. When all the matching is missed, the generator adds the "Is it that" at the beginning of the response.

4.3.2 The RQ Detection

The RQ uses the same method of the situation classification that we introduced in Section 4.2. The only difference is the input data and the labels. The generated RQ is concatenated behind the initial response and then The completed RQ is added behind the input sequence. The labels are changed to RQ or not RQ. We also trained the fine-tuned models. The output of the model is the probability of the RQ. The generator uses it as a score of the appropriation of the RQ response.

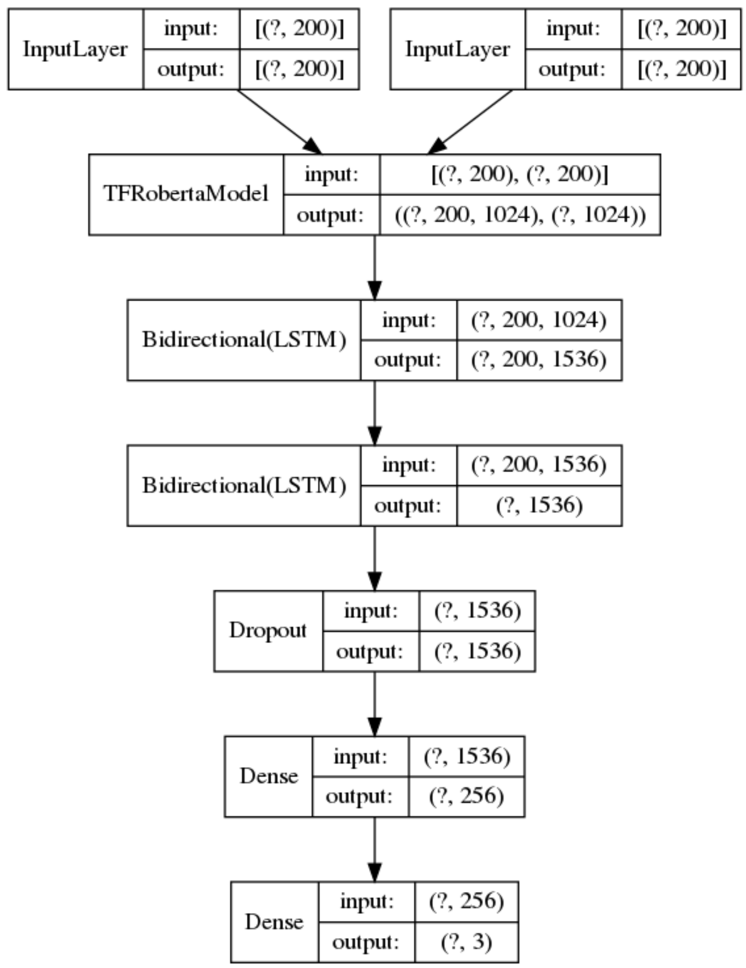


Figure 4.9: The structure of RoBERTa large model with bi-LSTM layers.

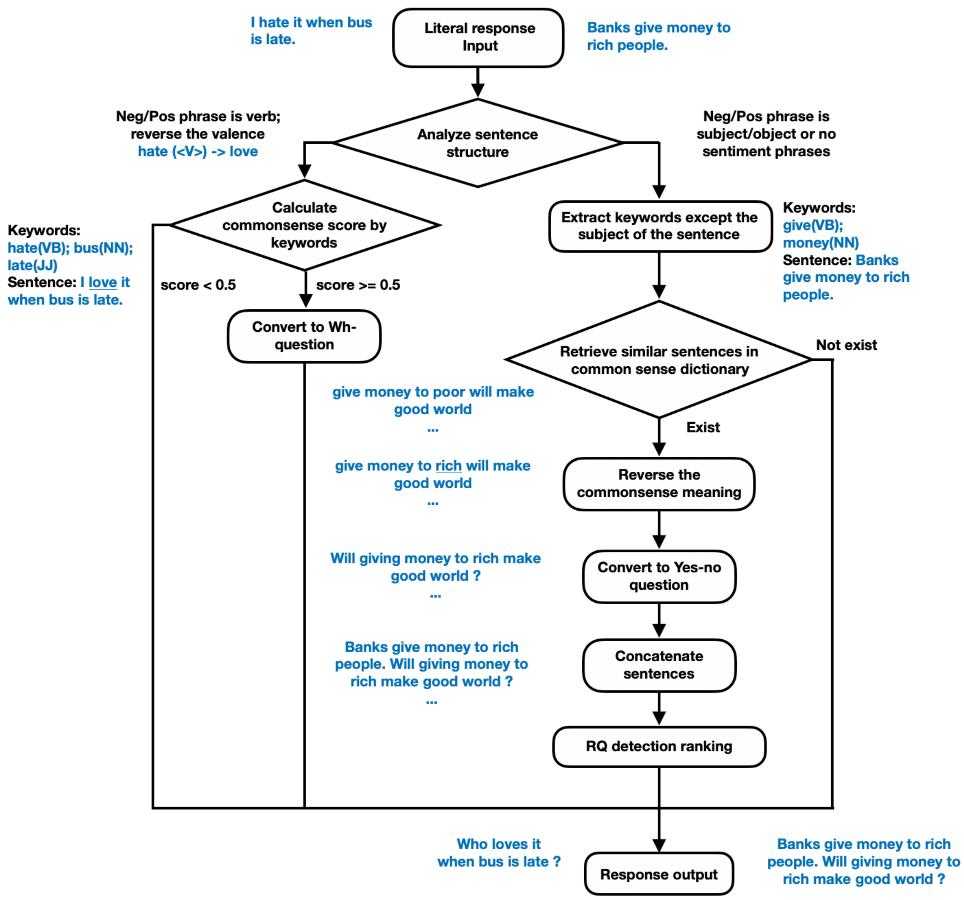


Figure 4.10: The structure of RQ generator.

Table 4.1: Samples of the commonsense knowledge sentences

Relation	Left_term	Right_term	Knowledge
AtLocation	a letter	an envelope	a letter in an envelope
CapableOf	athlete	jump high	athlete can jump high
HasPrerequisite	make a better world	give money to the poor	give money to the poor will make a better world
Desires	Most people	become wealthy	Most people want to become wealthy

Listing 2: The dictionary of NER and interrogative words

```
ner_dic = {"Who": ["PERSON"],
           "Where": ["NORP", "FAC", "ORG", "GPE", "LOC"],
           "What": ["PRODUCT", "WORK_OF_ART", "LAW"],
           "What event": ["EVENT"],
           "What language": ["LANGUAGE"],
           "When": ["DATE", "TIME"]}
```

4.3.3 Commonsense Knowledge Sentences

We build commonsense knowledge sentences by processing commonsense knowledge sentences base, which is introduced in Section 3.4 to concatenate left and right terms. Then, make them a complete sentence by filling the phrases between them. We create a dictionary to choose the filling phrases, shown in Listing 3. The phrases are matched by the relation type, which is marked in the commonsense knowledge base. Some relation types are used at a very low probability, so we do not use them.

There are some samples of the base shown in Table 4.1. The column of “Knowledge” contains completed sentences of commonsense knowledge.

Listing 3: Commonsense knowledge sentences of filling phrases

```
left_to_right = {"ReceivesAction": "is",
                 "AtLocation": "in",
                 "HasA": "has a",
                 "IsA": "is a",
                 "NotCapableOf": "can not",
                 "Causes": "makes people",
                 "CausesDesire": "make people want to",
                 "HasProperty": "is",
                 "Desires": "want to",
                 "InheritsFrom": "is inherited from",
                 "CapableOf": "can",
                 "NotIsA": "is not a",
                 "NotHasProperty": "is not",
                 "NotDesires": "do not want to",
                 "DesireOf": "have a desire of",
                 "LocationOfAction": "in",
                 "MotivatedByGoal": "because",
                 "PartOf": "is a part of",
                 "MadeOf": "is made of",
                 "RelatedTo": "is related to",
                 "DefinedAs": "is",
                 "NotHasA": "does not have",
                 "SymbolOf": "is a symbol of",
                 "CreatedBy": "is created by",
                 "UsedFor": "is used of"}
right_to_left = {"HasSubevent": "to",
                 "HasPrerequisite": "will",
                 "HasLastSubevent": "after",
                 "HasFirstSubevent": "before",
                 "InstanceOf": "such as"}
not_use = {"LocatedNear", "HasPainCharacter",
           "HasPainIntensity", "NotMadeOf"}
```

5 Experiment

5.1 Evaluation on Situation Classifier

FigLang 2020[8] provided a dataset which contains 13,000 posts of Twitter and Reddit contexts and responses. It includes posts as utterances in conversation as well as their sarcastic responses and non-sarcastic responses. We choose the interrogative sarcastic responses as RQs. We also selected 797 non-sarcastic interrogative responses and conducted RQ annotation on the Amazon Mechanical Turk (Amazon MTurk)[3]. MTurk is a crowdsourcing marketplace where everyone can outsource their annotation tasks to a distributed workforce who can virtually perform these tasks. Three MTurk workers annotated each interrogative response. We showed them the previous contexts and the response. Then we ask them the response is RQ or not. An example task is shown in the Figure 5.1. We calculated the average of their annotation of each response. When it is larger than 0.5, the response is marked as RQ. As a result, 399 questions are annotated as RQ. There are 6500 sarcasm and 1515 RQs in the dataset. Therefore, we cut the trisection of the data by three types of responses are literal, sarcastic, and RQ responses, to create a balanced dataset. 80% of the data is used as training data, and 20% of them are used as test data. We also use the 5-fold cross-validation to train each model. We evaluate the proposed situation classifier by accuracy(A), precision (P), recall (R), and F1 scores (F1). The fine-tuned RoBERTa model with Bi-LSTM layers receives the highest accuracy, which is 0.43, shown in Table 5.1. The confusion matrix figure without normalization for the best model is shown in Figure 5.2. There are two reasons that the predicted label of RQ and sarcasm are majority.

- Mturkers tend to select RQ and sarcasm because the task's title makes them think that the answer may have a high RQ probability.

- RQ sometimes has an ironic meaning, and it is appropriate to choose both RQ and literal in the same situation. It is difficult to predict whether the response is RQ or sarcasm by only analyzing previous contexts, even by human-being.

[3]https://requester.mturk.com/

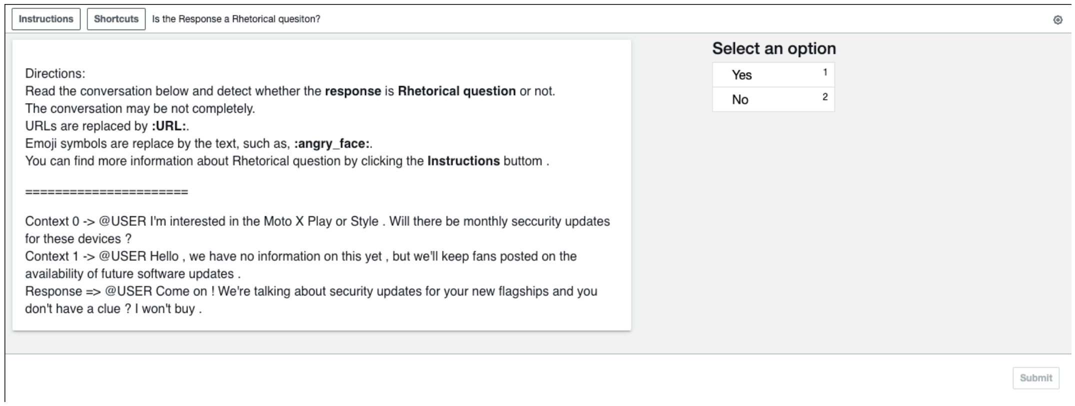


Figure 5.1: An example task for the RQ annotation.

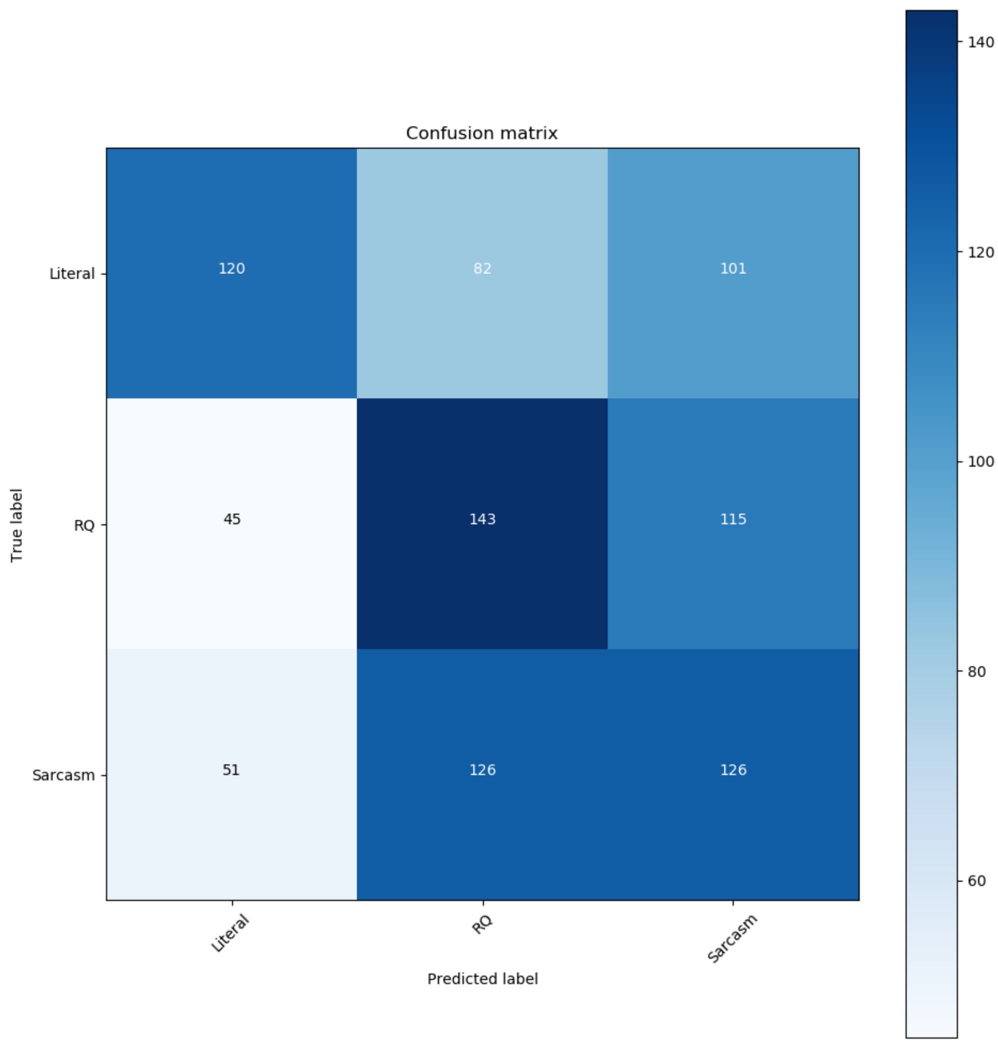


Figure 5.2: The confusion matrix figure for the RoBERTa model with Bi-LSTM layers

Table 5.1: The results of each model for the situation classifier

Model	Accuracy	Type	Precision	Recall	F1	Support
BERT	0.40	Literal	0.47	0.38	0.42	303
		RQ	0.38	0.40	0.39	303
		Sarcasm	0.37	0.42	0.39	303
BERT-Large	0.41	Literal	0.46	0.45	0.46	303
		RQ	0.39	0.37	0.38	303
		Sarcasm	0.39	0.41	0.40	303
RoBERTa	0.42	Literal	0.52	0.41	0.46	303
		RQ	0.37	0.37	0.37	303
		Sarcasm	0.40	0.49	0.44	303
RoBERTa-Large	0.41	Literal	0.48	0.44	0.46	303
		RQ	0.40	0.32	0.36	303
		Sarcasm	0.35	0.45	0.40	303
BERT + Bi-LSTM	0.41	Literal	0.53	0.38	0.44	303
		RQ	0.39	0.33	0.36	303
		Sarcasm	0.36	0.51	0.42	303
BERT-Large + Bi-LSTM	0.39	Literal	0.48	0.37	0.42	303
		RQ	0.37	0.40	0.39	303
		Sarcasm	0.34	0.40	0.37	303
RoBERTa + Bi-LSTM	**0.43**	Literal	0.56	0.40	0.46	303
		RQ	0.41	0.47	0.44	303
		Sarcasm	0.37	0.42	0.39	303
RoBERTa-Large + Bi-LSTM	0.42	Literal	0.50	0.43	0.46	303
		RQ	0.39	0.51	0.44	303
		Sarcasm	0.37	0.31	0.34	303

5.2 Evaluation on Chatbots

5.2.1 Data

We randomly selected 42 posts from the test data of FigLang 2020 dataset[8] and generated all the types of responses (Literal, sarcasm, RQ and original human responses). We designed a task on MTurk. Each post was rated by five workers. They were asked to read the previous contexts and answer the following questions below, and an example task is shown in Figure 5.3. The complete data for the evaluation and its results are shown in Section A.1.

- Are these responses rhetorical questions? (Yes/no)
- Are these responses sarcastic? (Yes/no)
- How appropriate is the response in the conversation? (5-point scale)
- How human-like is the response? (5-point scale)

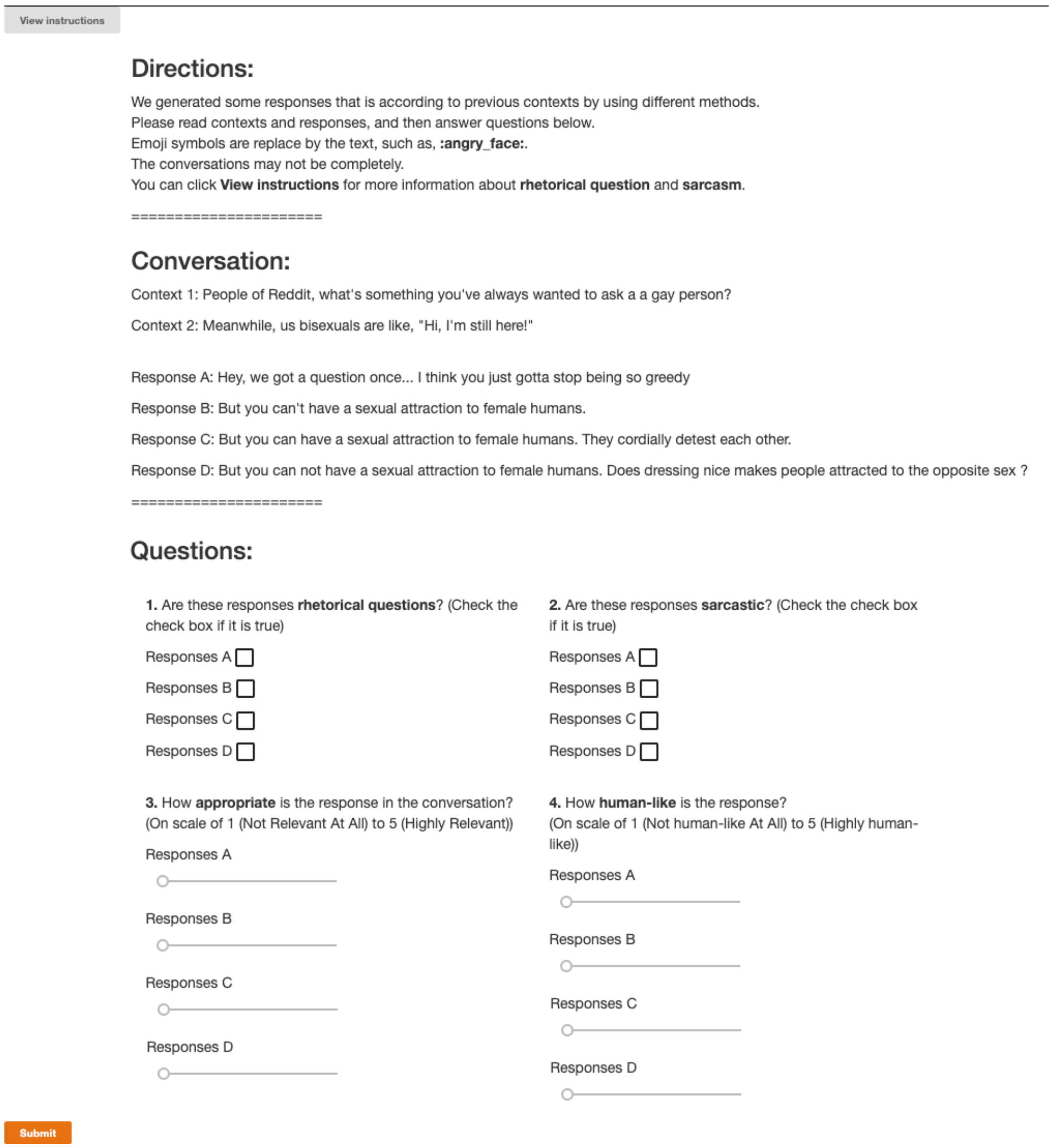
5.2.2 Results of RQ Detection Models

Table 5.2 exhibits the result RQ detection by using different models. The input data contains the sequence of previous contexts and the generated RQ response. The model can get both information of the conversation's situation and the response. Furthermore, it is only used to classify the sequence contains RQ or not, so the Accuracy of models is higher than models for situation classifiers. The fine-tuned BERT model receives the highest accuracy, which is 0.81. All the results of the models are shown in Table 5.2. We select it to do the RQ detection in the RQ generator.

5.2.3 Results of the Chatbot Evaluation

Figure 5.4 displays all human evaluation results for whether the response is RQ or not. RQ responses that humans think, which are RQs, are the least. It is because:

- RQ responses are generated by rules. It sometimes makes grammatical errors that affect the human's detection.
- The commonsense knowledge is unclear for different people in different countries. In this evaluation, our tasks are answered by Australia, Brazil, Canada, the United



View instructions

Directions:

We generated some responses that is according to previous contexts by using different methods.
Please read contexts and responses, and then answer questions below.
Emoji symbols are replace by the text, such as, **:angry_face:**.
The conversations may not be completely.
You can click **View instructions** for more information about **rhetorical question** and **sarcasm**.

=====================

Conversation:

Context 1: People of Reddit, what's something you've always wanted to ask a a gay person?

Context 2: Meanwhile, us bisexuals are like, "Hi, I'm still here!"

Response A: Hey, we got a question once... I think you just gotta stop being so greedy

Response B: But you can't have a sexual attraction to female humans.

Response C: But you can have a sexual attraction to female humans. They cordially detest each other.

Response D: But you can not have a sexual attraction to female humans. Does dressing nice makes people attracted to the opposite sex ?

=====================

Questions:

1. Are these responses **rhetorical questions**? (Check the check box if it is true)

Responses A ☐

Responses B ☐

Responses C ☐

Responses D ☐

2. Are these responses **sarcastic**? (Check the check box if it is true)

Responses A ☐

Responses B ☐

Responses C ☐

Responses D ☐

3. How **appropriate** is the response in the conversation? (On scale of 1 (Not Relevant At All) to 5 (Highly Relevant))

Responses A

Responses B

Responses C

Responses D

4. How **human-like** is the response? (On scale of 1 (Not human-like At All) to 5 (Highly human-like))

Responses A

Responses B

Responses C

Responses D

Submit

Figure 5.3: An example task for chatbots evaluation.

Table 5.2: The results of each model for the RQ detection

Model	Accuracy	Type	Precision	Recall	F1	Support
BERT	**0.81**	Literal	0.84	0.76	0.80	303
		RQ	0.78	0.85	0.82	303
BERT-Large	0.78	Literal	0.83	0.71	0.76	303
		RQ	0.74	0.85	0.80	303
RoBERTa	0.74	Literal	0.73	0.78	0.75	303
		RQ	0.76	0.71	0.74	303
RoBERTa-Large	0.69	Literal	0.68	0.72	0.70	303
		RQ	0.70	0.66	0.68	303
BERT + Bi-LSTM	0.80	Literal	0.84	0.75	0.79	303
		RQ	0.77	0.85	0.81	303
BERT-Large + Bi-LSTM	0.74	Literal	0.75	0.73	0.74	303
		RQ	0.73	0.76	0.74	303
RoBERTa + Bi-LSTM	0.73	Literal	0.73	0.74	0.73	303
		RQ	0.74	0.72	0.73	303
RoBERTa-Large + Bi-LSTM	0.65	Literal	0.64	0.69	0.66	303
		RQ	0.66	0.61	0.63	303

Table 5.3: Scores of the chatbot in a different response situation. The leftmost column indicates the type of response (i.e., literal, sarcasm, or RQ) selected by the situation classifier. The rows indicate that the mean evaluation scores for four responses in each situation.

(a) Appropriateness				
Type	**Human**	**Literal**	**Sarcasm**	**RQ**
Literal	3.63	3.42	3.15	3.13
Sarcasm	3.98	3.46	3.42	3.40
RQ	3.52	3.39	3.40	3.47
(b) Human-likeness				
Type	**Human**	**Literal**	**Sarcasm**	**RQ**
Literal	3.70	3.53	3.28	3.45
Sarcasm	4.15	3.62	3.66	3.37
RQ	3.58	3.56	3.43	3.39

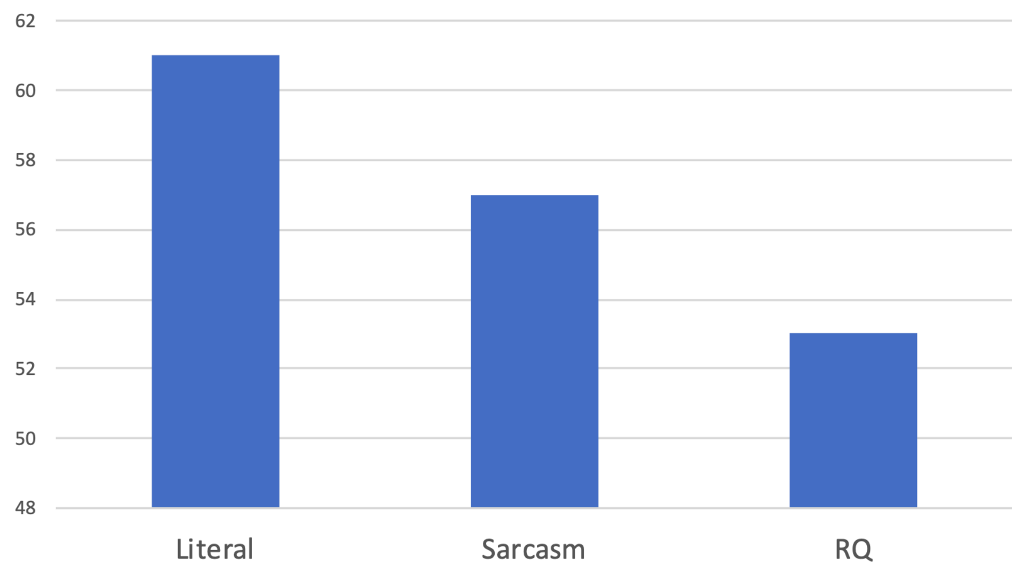


Figure 5.4: The results of the human evaluation for whether the response is RQ or not.

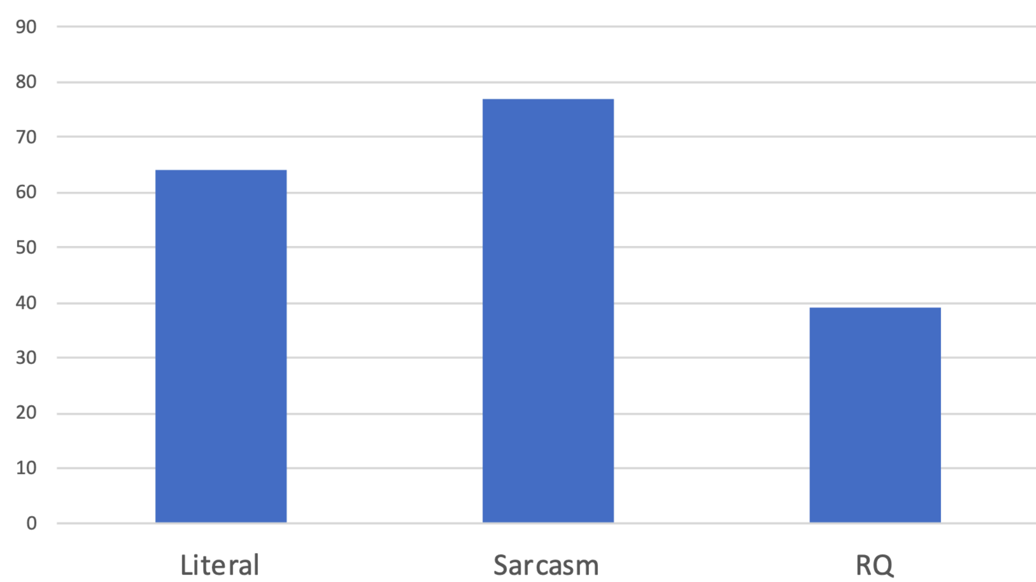


Figure 5.5: The results of the human evaluation for whether the response is sarcastic or not.

Kingdom, and the United States. For example, the question "Does the democratic party is a United States political party?" has a vague answer for different people. Three workers think it is an RQ, but two workers believe it is not.

Figure 5.5 displays all human evaluation results for whether the response is sarcastic or not. Sarcastic responses, which humans think are sarcasm, are most selected. It is because that there is the model in the sarcasm generator that ranks sarcastic responses by the semantic incongruity. It selects the most sarcastic responses from generated responses, So workers detected that it is sarcastic no matter it is appropriated for the conversation or not.

Table 5.3 displays each type of response's average score is rated by workers when the situation classifier classifies the situation.

Figure 5.6 shows that when the situation classifier chooses RQs, RQ responses achieve a higher score on appropriateness than literal response and sarcasm. The additional RQ makes the response more appropriate in the RQ situation. For example, Table 5.4 shows a post that is classified as a RQ situation. The one who speak the last context thinks that the idea of the poll is a good idea, and he/she may be a theist. Therefore, when responses want to against his/her opinion, people think the RQ that aims at theism is more appropriate. However, when the literal generator misses the important information from the previous contexts, the RQ generator can not make the appropriate RQ response even in the RQ situation such as the example in Table 5.5. The literal generator misses the information about the carbon tax, which is vital in the conversation. It makes the RQ generator's response only talking about climate change, which is not related to the previous contexts. On the other hand, as we said that the RQ generator converts question based on rules that changes the structures of the sentence. Sometimes grammatical errors cannot be fixed by the grammatical error correction model. Consequently, the RQ responses are rated as less human-like than other responses.

Figure 5.7 shows that when the situation classifier selects literal responses, both appropriateness and human-likeness scores are higher than other responses, except for the human response. Since both the sarcasm generator and the RQ generator are based on the literal response. When the previous contexts show that it should be a literal response in this situation, people tend to choose the literal one. Furthermore, there is a high probability that the literal response is also appropriated for a conversation when sarcasm or RQ is appropriated for it.

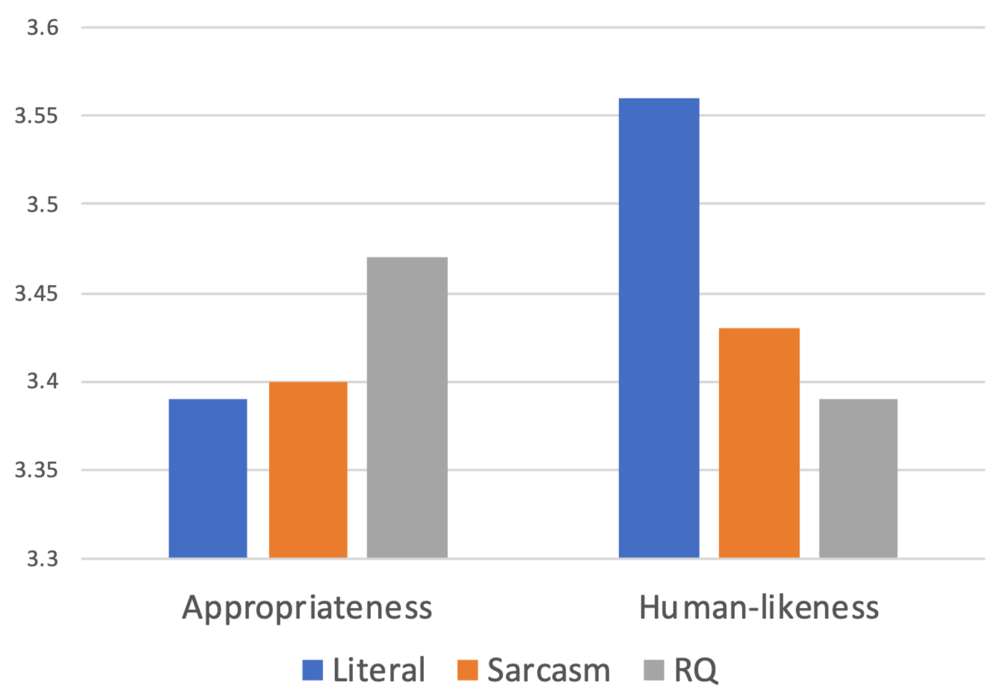


Figure 5.6: Each type of response's average score is rated by workers when the situation classifier decides to generate RQ.

Figure 5.8 shows that when the situation classifier decides to respond a sarcasm, the appropriateness scores are lower than the literal responses. However, it has a higher score in human-likeness. The semantic incongruity ranking selects the most sarcastic response, which makes the response more human-like. However, the sarcasm generator concatenates the valence-reversed literal response with a directly selected sentence from a sentence corpus. The sentence sometimes does not appropriate to the current conversation.

We average the human-rated appropriateness results and select the response, which has the highest score as the true target. Table 5.6 is the result that when we use the situation classifier results as estimated targets and the true targets to get classification metrics. The accuracy of this evaluation is 0.33. Figure 5.9 shows that people tend to speak sarcasm and literal rather than the RQ. But as the result of Table 5.3, when the situation is appropriate to speak RQ, people also give a high score to RQ, but lower than the top one. The result is not good because RQs are usually fitting for choosing both RQ and literal in the same situation.

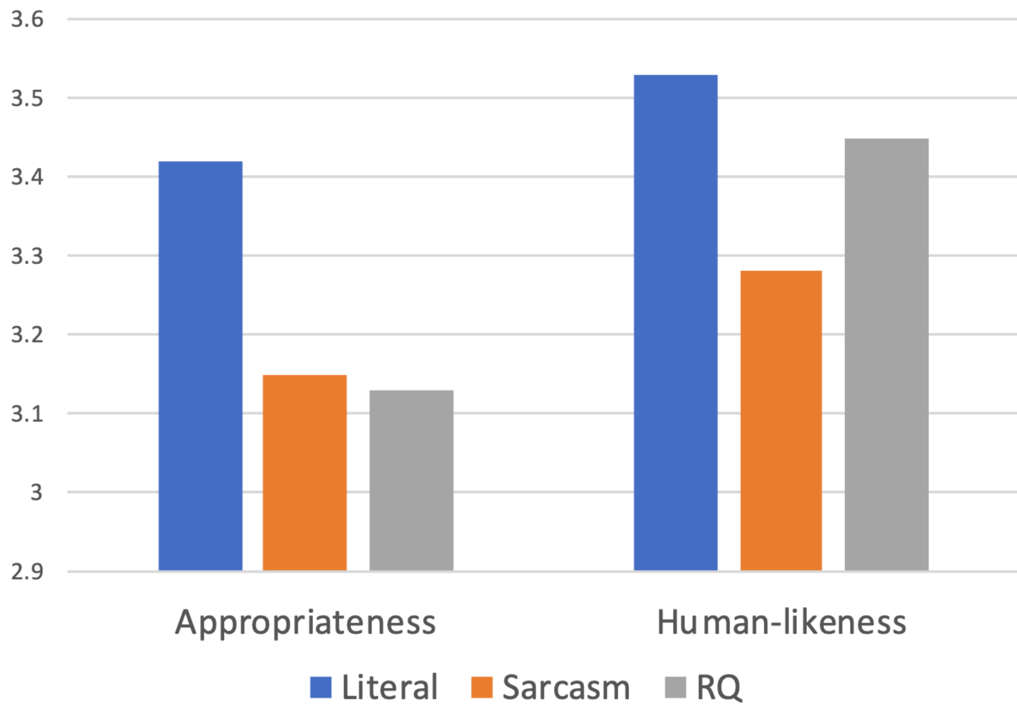


Figure 5.7: Each type of response's average score is rated by workers when the situation classifier decides to generate literal responses.

Table 5.4: The example of a post that RQ situation. "App" means the scores of appropriateness; "HL' means the scores of human-likeness

Type	**Utterance**	**App**	**HL**
Context 2	Please tell me that you've seen mine and poll , we've had an argument about this for over a year and we've actually got polls on this right now , it's legit , split down the middle		
Context 1	It should be praised by God's , oh wait , GODS MADE THIS BRILLIANT IDEA ! ! !		
Human	So you hate your own creation ? You really thought through that one well	3.0	4.0
Literal	I'm pretty sure God wouldn't have the money to pay for a poll.	3.75	3.0
Sarcasm	I'm pretty sure god wouldn't have the money to default for a poll. Happy am I who owes nothing	3.75	3.75
RQ	I'm pretty sure God wouldn't have the money to pay for a poll. Can atheists agree that god does exist ?	4.25	3.5

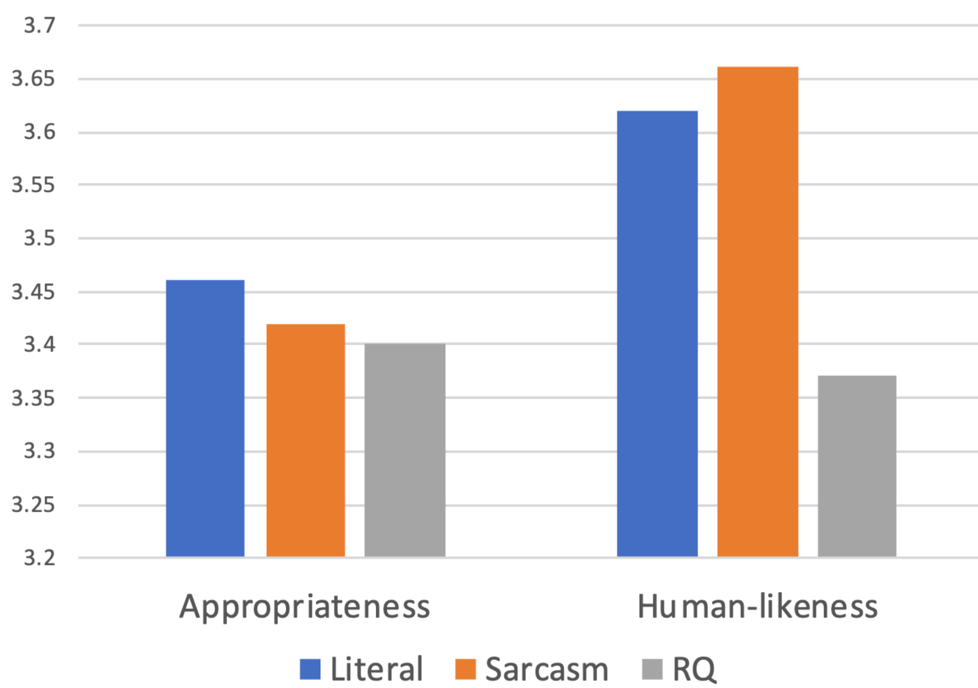


Figure 5.8: Each type of response's average score is rated by workers when the situation classifier decides to generate sarcastic responses.

Table 5.5: The example of a post that RQ situation. “App” means the scores of appropriateness; “HL’ means the scores of human-likeness

Type	**Utterance**	**App**	**HL**
Context 2	Did you know that as climate change disrupts existing wind patterns , we are likely to have worse and harsher winters as the cold Arctic air is no longer held back by those same wind patterns ? I know the difference between climate and weather is challenging for you, but still.		
Context 1	So ... to hedge against climate change we should pay carbon tax to the government , so they can give that same money back to tax payers :person_shrugging: :male_sign: . I guess that will normalize those wind patterns and will likely prevent catastrophic fires etc ? ? BRILLIANT IDEA ! ! !		
Human	more like : appropriate taxation can help steer consumers spending habits directing them to less environmentally disruptive behaviours and consumer choices . Ppl are driven by price more than moral and principle . that one well	3.6	2.8
Literal	That’s like a reverse climate change.	2.4	3.4
Sarcasm	That’s like a reverse climate change Nuclear winter - semi-polished basmati rice?	2.4	2.8
RQ	That is like a reverse climate change. Is it that climate change causes global warming ?	2.6	2.8

Table 5.6: The Classification Report of the Evaluation on the Chatbot.

Type	**Precision**	**Recall**	**F1**	**Support**
Literal	0.29	0.45	0.36	11
RQ	0.22	0.14	0.17	14
Sarcasm	0.44	0.41	0.42	17

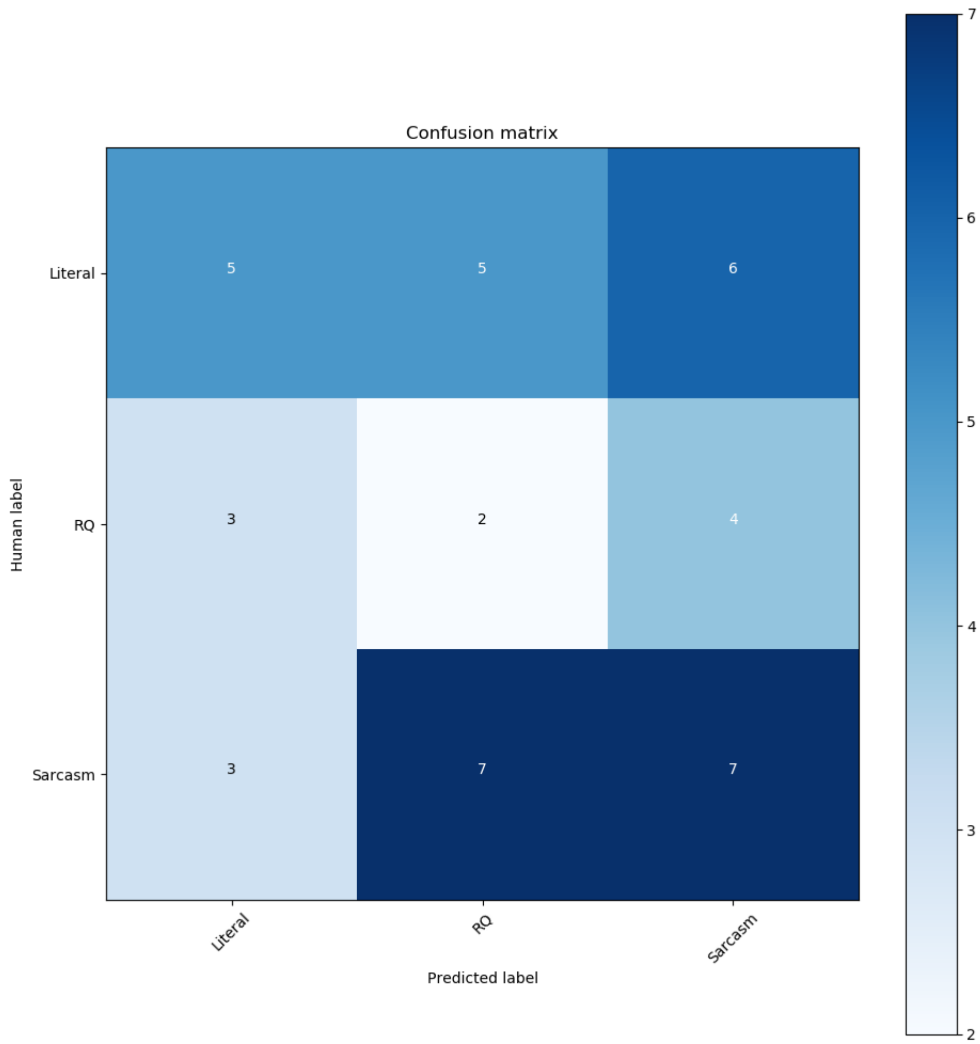


Figure 5.9: The confusion matrix figure for the situation classifier's predict and the highest human-rated appropriateness results

6 Conclusion and Future Work

In this study, we propose a method for generating RQ based on the commonsense knowledge and the chatbot application that can decide which of RQ, sarcastic, and literal response should be generated by analyzing users' utterances and previous contexts. The conclusion of findings as follows:

- RQ sometimes has an ironic meaning, and it is appropriate to choose both RQ and literal in the same situation. It is difficult to predict whether the response is RQ or sarcasm by only analyzing previous contexts, even by human-beings. Yet, the classification with Bi-LSTM layers can still detect the right response sometimes.

- The proposed RQ generator and situation classifier is effective in that RQs are more appropriate than literal and sarcastic responses when the classifier decides to generate RQ responses. However, the generated RQ responses are less human-like than other types of responses beacuse of grammatical errors.

In the future, we would like to develop a model for converting sentences to questions that enables the chatbot to generate a higher quality of RQs and to be more human-like. The literal generator's problem misses the information in previous contexts affects the RQ response's appropriateness, so we would like to extract the keywords from previous contexts to improve the RQ response's suitability in the conversation. Furthermore, we would like to add features to improve situation classifier model.

References

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online, July 2020. Association for Computational Linguistics.

[3] Douglass Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*, pages 133–140, 1992.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[6] Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1003–1012, 2015.

[7] Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[8] Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online, July 2020. Association for Computational Linguistics.

[9] Risa Goto. *Rhetorical Questions: A Relevance-Theoretic Approach to Interrogative Utterances in English and Japanese*. Hituzi Syobo Publishing, 2018.

[10] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[11] Chung-hye Han. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229, 2002.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mark James Carman. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD*, 2015.

[14] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online, July 2020. Association for Computational Linguistics.

[15] Roger J Kreuz and Kristen E Link. Asymmetries in the use of verbal irony. *Journal of Language and Social Psychology*, 21(2):127–143, 2002.

[16] Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, 2016.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18] Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. A modular architecture for unsupervised sarcasm generation. 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)*, pages 6144–6154, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. Available at https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[22] Richard M Roberts and Roger J Kreuz. Why do people use figurative language? *Psychological science*, 5(3):159–163, 1994.

[23] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[24] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686, 2012.

[25] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[26] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[27] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[28] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[29] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

A Appendix

A.1 Evaluated Posts and Generated Responses

Table A.1: The table of evaluated posts and generated responses. It contains previous contexts, human responses, and generated responses."SC result" is the result of the situation classification. "0" means that the chatbot generates a literal response in the situation. "1" means RQ. "2" means sarcastic response.

	Context 2	Context 1	Response	Literal response	Sarcastic response	RQ response	SC result
0	We don ' t know who you are either .	Well , thank god for that . No need to worry .	We ' re beyond worries here . The state pushing experiments with uneducated Muslims against the hard working people paying the bill .	He nods and smiles, then goes back to his food.	He nods and smiles, then goes back to his food. A green banana is not ripe enough to eat.	He nods and smiles , then goes back to his food. Are smiles just upside-down frowns ?	1
1	Dr . Drew who apparently didn ' t even know he lived in Adam Schiff ' s district for 20 years says he is considering a run against Adam Schiff <URL>	You lived in there 20 years nd you didn ' t know Nything bout your district . YOU ARE A MORON Dr . Drew . Is it because you are losing your audience and lots of money and you are not popular anymore ? <URL>	Go and talk to Melania and see if you can fix her bs campaign . :face_with_tears_of_joy: :face_with_tears_of_joy: :face_with_tears_of_joy: You are so full of it .	he's had some good points the last few days...	He's had some good points the last few days... Who's never got any money.	he is had some good points the last few days ... Will making points write ?	1
2	' This could work ... ' - @ jenniebond 1 believes Prince Harry and Meghan Markle could successfully modernise how members of the monarchy work by controlling their own publicity on social media . Do you agree ? #Lorraine <URL>	I thought it was less publicity they wanted not more !	They want to be paid for it , they can see the money they'll make to secure their independence :face_with_tears_of_joy:	Yes. Yes they do. And they deserve to have that right.	Yes. yes they don't do. and they deserve to have that right. Justice is blind and will prosecute without any discrimination, either chinese, formosan, or american, because all are equal before the law.	Yes they do . And they deserve to have that right. Is it that I nod and say yes and will agree with someone ?	0
3	if y ' all like Regal Eagle Smokehouse y ' all are gonna love the shit out of this new hole in the wall joint called Trails End	If y ' all love Regal Eagle peep this crazy place called 4Rivers .	rigth ? ? Go taste & we talk ... dont eat the Mac & cheese please .. :face_with_tears_of_joy: <URL>	Dude I wish my trails end was 4 rivers.	Dude i don't wish my trails end was 4 rivers. The city secured itself against flood.	What I begrudge my trails ended with 4 rivers ?	0
4	All I ' m saying is there should be a remake of Grumpy Old Men with me and Boyd <URL>	If you wouldn't mind and if you think it could help , would you RT my today's tweet on animal adoption please ? I believe you have a rather large following . If you will , then I'll share it with you once again . Let me know please	If any of you in Chennai / Bangalore / Ooty would like to ADOPT a dog or cat , let me know please . I shall enable that for you . Today is a good day to change an animal's life :heart_suit: :heart_suit: :heart_suit: #adopt #adoptdontshop #adoptdontbuy <URL>	I think it can get me a few more upvotes.	I don't think it can get me a few more upvotes. The resolution was affirmed by a unanimous vote.	Who can get me a few more updates ?	1
5	Girl Beats The Sh*t Out Of Chick For Talking About Her Son Then Starts Shooting At Her Car As She Pulls Off!	And that's how you goto jail for attempted murder	And ensure you have the evidence to prove it	I think you mean the same thing actually.	I don't think you mean the same thing actually. All right. I understand.	Who actually means the same thing ?	0
6	Let all that you do be done in #ThinkBIGSundayWithMarsha #InspireThemRetweetTuesday <URL>	All is done in love , and so it shall be . Thank you God for your creation of All ! Amen . <URL>	It ' s not how much we have but how much we enjoy that makes #happiness #ThinkBIGSundayWithMarsha #InspireThemRetweetTuesday <URL>	My pleasure. All hail the all powerful God!!	My pleasure. all don't hail the all powerful god!! But I didn't have any joy.	What ! ! ?	0
7	America First ! Lots of sitting Democrats for this one ...	Its Been Obvious they hate OUR GREAT PRESIDENT #TrumpDerangementSyndrome but the #Socialist & #MainstreamMedia that " WORK FOR DEMS " have let their mask fall & SHOWED THEIR TRUE HATRED FOR . These Disquesting Human Beings r now Sitting proving it ! <URL>	How Can #DemocraticParty NOT BE for #wearethepeople . It's not just great news for it's Great news for , we need to Stop this hatred & Work together w / to help The people that elected U 2 Help Them ! ! <URL>	Don't you dare to have a discussion. s	Do you dare to have a discussion. s We agreed without much further argument.	Do not you dare to have a discussion . s. Does the purpose of the discussion is to extract ideas from the audience ?	1
8	Crazy how Sheffield Wednesday can even consider buying when they a riddled . Shocking when you think #lufc are terrified to spend yet are no where near as deep as they are .	It ' s Wednesday and still rattled . Love it :OK_hand: :medium_skin_tone:	:face_with_tears_of_joy: Come on pal . Did you your promotion celebrations after the game ? #desperate #massivetinpot	What does it denote? I'm lost :p	What does it denote? i'm found :p I am on the margin of death.	What does it denote ? I am lost : p. Is it a sharpen hint to give a clue ?	0

	Context 2	Context 1	Response	Literal response	Sarcastic response	RQ response	SC result
9	Irony : Calling out " celeb obsessed " Chrissy Teigen followers in the process of sending 5000 tweets to Chrissy Teigen	5,000 tweets eh . When did you learn to count ?	why do you accuse people of being " celeb obsessed " then tag unrelated celebrities in your posts ?	What do you mean you don't count your followers? s	What do you mean you do count your followers? s Time will never change and stop for any number of person.	What do you mean you do not count your followers ? s. Are Muhammad and his followers persecuted ?	0
10	Did you know that as climate change disrupts existing wind patterns , we are likely to have worse and harsher winters as the cold Arctic air is no longer held back by those same wind patterns ? I know the difference between climate and weather is challenging for you , but still .	So ... to hedge against climate change we should pay carbon tax to the government , so they can give that same money back to tax payers :person_shrugging::male_sign: . I guess that will normalize those wind patterns and will likely prevent catastrophic fires etc ? ? <URL>	more like : appropriate taxation can help steer consumers spending habits directing them to less environmentally disruptive behaviours and consumer choices . Ppl are driven by price more than moral and principle .	That's like a reverse climate change	That's like a reverse climate change Nuclear winter - semi-polished basmati rice?	That is like a reverse climate change. Is it that climate change causes global warming ?	1
11	At the Toronto G20 summit, some cop randomly started wailing on some poor guy walking by and eight cops caught watching on video could not recall the incident or identify the cop. This is the world we live in, anyone can just have bad luck and get randomly attacked by a cop and if there isn't video, they will get away with it.	The scariest part is that it seems like even if there is video, they still get away with it or let off easy half the time.	Sometimes video and police reports contradict each other, so both forms of evidence end up getting tossed. Absurd.	Theres always a risk the officer will turn the attention.	Theres always a not risk the officer will turn the attention. Lawrence island with no loss of life.	Theres always a risk the officer will turn the attention. Is there more than one way to skin a cat ; it is cliche ?	1
12	it's not ineffective at all . Cynicism gets us nowhere .	Not effective ? The GOP just gained more governmental power than they've had in almost 90 yrs . & they'r picking the new sc justice	the only thing we have to #fear is fear itself .	You just don't know how the Democratic Party operates.	You just do know how the democratic party operates. They distanced themselves from the upper class and adopted a communist political ideology.	You just don't know how the Democratic Party operates. Does the democratic party is a United States political party ?	1
13	This idiot was trying to make a racist announcement and this guy kept yelling :speaking_head: NOOOOO in his face lmao <URL>	So what happened to free speech ! ! ! Brilliant comrades	Hiel Chris , I see two people exercising their right to free speech . Who's right would you revoke ? Sieg Heil stupid fascist .	Don't know how you even can think that's a good thing.	Do know how you even can think that's a good thing. I am radiating joy and happiness.	Don't know how you even can think that's a good thing. Do you lose faith in understanding something bad ?	2
14	The Wire: Coast Guard Narrowly Avoids Sparking World War III	I had to post this.. The level of stupidity is priceless.	I think I need to elaborate.. the article is stupid.	I don't think anyone was stupid, just misguided.	I do think anyone was stupid, just misguided. Smile like you have never been hurt.	I don't think anyone was stupid, just misguided. Can anyone certainty dogma ? ?	2
15	and i'm white :)	I'm surprised he got away with that.	Yeah, since when are jews considered "white?"	It was a joke, bro.	It was a joke, bro. Gary simply sat with an admiring and respectful look on his face.	It was a joke, bro. Do you laugh because you are distraught ?	2
16	Aaannnd this is why I conceal carry.	And this is why I Canada	It is impossible to get robbed in Canada amirite?	And this is why I don't own a firearm.	And this is why i do own a firearm. There is no medicine against death.	And this is why I don't own a firearm. Is a firearm a harmless weapon ?	2
17	But it's not a tragedy	exactly, in most instances the bull is then slaughtered and given freely to the poor.	We don't do that in America because our parents teach us not to pay with our food.	You have a weird definition of most.	You have a not weird definition of most. The story ends on a note of macabre paradox.	You have a weird definition of most. Does another definition for babies is young humans ?	2
18	Stop finding excuses and justification for a failed administration ! Do the Job you sworn under oath to protect and defend the constitution ! <URL>	Pres Trump is doing the opposite of BHO appeasements and BHO aid to terrorists . Therefore , unlike the man you supported , President Trump IS fulfilling his oath to protect and defend not only the Constitution but the people of the United States #NoMoreWars #NoBenghazi	Totally the opposite ! Actually , people that knows recognizes his telegraphs from the past . A loser doesn't became a overnight ! <URL>	So Trump is an actuary, not a lawyer?	So trump is an actuary, a lawyer? I don't feel at all nervous.	So Trump is an actuary, not a lawyer? Is a lawyer called an attorney or barrister ?	1
19	Gina Rinehart: World's richest woman calls for Australian workers to be paid $2 a day	What in the actual fuck She thinks its possible to live on a McD burger a day homeless?	She doesn't see these people as humans, she sees them as wage costs.	If you want to have a job, sure.	If you don't want to have a job, sure. I can't lump the symptoms together and blame them all on stress.	If you want to have a job, sure. Is getting a job used to having something to do ?	1
20	Please consider that if you ' re upset about my essay , you could perhaps try to make a morally irrefutable case for your candidate instead of trolling Raghav ' s loved ones . We are human beings and I don ' t think this makes you or the candidate you support look as good as you think	Sorry Kate but talking about how various candidates policies might kill or not kill human beings is manipulative and against the rules of debate	What ? Politics are personal ? No way ! <URL>	I did not vote for this candidate and I am not a supporter of this candidate so I would not vote for her either	I did vote for this candidate and i am a supporter of this candidate so i would vote for her either Republican voters were not anti-foreigner because their neighbours were arabs.	I did not vote for this candidate and I am not a supporter of this candidate so I would not vote for her either. Is it that being eligible to vote means you are n't allowed to vote ?	1

	Context 2	Context 1	Response	Literal response	Sarcastic response	RQ response	SC result
21	Sohrab Ahmari , New York Post “ The Trump Campaign raised $ 10 million in the two days following the impeachment (Scam) vote . It seems the Democrats have shot themselves in the foot in one more way . They set up a process they know is not going to lead to the Presidents removal , & ...	... it ’ s alienating independents . This President is being persecuted over three years with one investigation after another , and that really plays to his base . ” Their partisan Witch Hunt is hurting our Country do badly , & only bringing more division than ever !	He ’ s done it again ! Pres . Trump outfoxes the Dems !	I didn’t vote for this president	I did vote for this president Happy about no one can only hold the force.	I didn’t vote for this president. Can citizens vote for leaders who are cowards ?	1
22	During SDCC they said that it might come, but if it will, it’ll be sometime in the future.	As opposed to sometime in the past?	What if it’s already in the game... ...but it’s locked behind an ARG?	I don’t know why you were downvoted for not making that much sense.. lol.	I do know why you were downvoted for making that much sense.. lol. I managed to convince the jury of my loss of innocence.	I don’t know why you were downvoted for not making that much sense.. lol. Do smoke is used with a false sense of unpopularity ?	2
23	Seth Rogen, Evan Goldberg sell Sony a raunchy animated film ’Sausage Party’	This has box office flop written all over it.	eh, I think it has a good chance	That’s some solid marketing.	That’s some solid marketing. Who’s never got any money.	That’s some solid marketing. Do some liquid have been purchased ?	0
24	I just deleted the eagles album from my phone .	I have to go through my vinyl collection . I know many albums will wind up in the garbage . I think I’m ready to say good bye .	not all of us are Evil some of us are against it BUT we are shut off even when we are signed for the Universe I am One of those who refused to shut up - - Q	Just make sure the tape is in mint condition.	Just make sure the tape is in mint condition. Let’s drink to good health.	Just make sure the tape is in mint condition. Do tape dispensers are used to make tape difficult to use ?	0
25	Always wondered what happened to that troll . “ How World Soccer Daily went up in smoke ” by Howler <URL>	it was beautiful , especially when he kept doubling down	You could argue he saved LFC . Many of the people who came together to oust him were the same who forced Gillette & Hicks out	And his reply to him was really well written as well.	And his reply to him was really well written as well. Such conduct is unworthy of praise.	And his reply to him was really well written as well. Do flames are an objectionable reply to a message ?	2
26	What if the team was nammed the ”Washington Negroes”	Well to see negro as offensive is now commonly accepted, but at some point you can’t just have every group of people make up a list of words they don’t like to hear.	Why don’t they rename the team the ”Washington Crackers”, white people are apparently ’thicker skinned’ so there’ll be no further issues.	You can’t just have everyone make up a list of words they don’t like to hear.	You can’t just have everyone make up a list of words they do like to hear. ’what fun!’ she said with a laugh.	You can’t just have everyone make up a list of words they don’t like to hear. Is it that question that everyone determines the lie ?	2
27	Fox and friends anchors wide-eyed as fellow anchor talks about how races should not be interbreeding	That was so out of no where. Those two other anchors were just making banter about some fluff piece and this guy had to suddenly shift to ”people shouldn’t marry outside their ethnic group”.	Then the other two tried to save the train wreck but ended up just dumbfounded	I’ve never seen a better example of what an anchor does at the bottom of the news cycle	I’ve seen a better example of what an anchor does at the bottom of the news cycle Large lymphocyte and eof had no evidently increase.	I’ve never seen a better example of what an anchor does at the bottom of the news cycle. Is it that depreciating a visit to a hot spring an example of informal learning ?	2
28	never forget that the largest protest in American history followed your inauguration . Oh yeah pe ··· <URL>	women dressed up as Vaginas . Yeah , everyone took them seriously .	not to mention , a man hit a woman at a march , and the protesters said SHE deserved it !	When the article mentions the protests against Trump it’s not like they did not recognize the protests.	When the article mentions the protests against trump it’s like they did recognize the protests. Never uttered a word of protest.	When the article mentions the protests against Trump it’s not like they did not recognize the protests. Is it that joining protests to make the world worse ?	2
29	At least it’s not expensive to live in London :see-no-evil_monkey:	that’s a massive tick in the box for the job also eh ? ? ... ha ha	Slavery - ” a condition of having to work very hard without proper remuneration or appreciation . ”	It’s funny because I don’t even have a job.	It’s funny because i do even have a job. It was scarcely an occasion for laughter.	It’s funny because I don’t even have a job. Is it to express how even something is used to make observations about life ?	0
30	This ... Was sarcasm . Probably should have used some quotation marks I guess .	I believe mimicking / mocking outrage twitter , but very poorly executed .	Advice if you want to be sarcastic hashtag because you sounded like a complete dunce .	So you believe the entire thing wasn’t a joke?	So you don’t believe the entire thing wasn’t a joke? It was scarcely an occasion for laughter.	So you believe the entire thing wasn’t a joke? Do beliefs in anything are called freedom of thought ?	2
31	I bet porn sucked in the 90’s	The story in porn Is marginal at best	I agree it’s always so aggressive and forced, like you just met the cable man like shouldn’t you ask his name?	I am very interested in the story, but it would be difficult to find a source.	I am very interested in the story, but it would not be difficult to find a source. My face was void of all interest.	I am very interested in the story, but it would be difficult to find a source. Is telling a story used to make yourself sound trivial ?	2
32	Please tell me that you’ve seen mine and poll , we’ve had an argument about this for over a year and we’ve actually got polls on this right now , it’s legit , split down the middle <URL>	It should be praised by God’s , oh wait , GODS MADE THIS BRILLIANT IDEA ! ! !	So you hate your own creation ? You really thought through that one well	I’m pretty sure God wouldn’t have the money to pay for a poll.	I’m pretty sure god wouldn’t have the money to default for a poll. Happy am I who owes nothing	I’m pretty sure God wouldn’t have the money to pay for a poll. Can atheists agree that god does exist ?	1
33	Let me repeat . Slowly . I followed up on it . I always allow new information to change my opinion . That’s what tolerant people do .	I was reacting to your initial tweet . The fact that your opinion changed doesn’t validate it in any way .	Yep and you’ve demonstrated your ignorance by prejudging someone about whom you know nothing , which was my point ! !	I see your point. Sorry for the confusion.	I see your point. sorry for the confusion. Shouting abuse at my stupidity.	I see your point. Sorry for the confusion. do debate politics is used to see other points of view ?	1
34	It’s weird that people still equate socialism with communism, as if they are the same.	Socialism is communism without the guns pointed at the citizen.	Ah, yes, people hand over half their money to the State out of the goodness of their hearts and not because of force.	Socialism doesn’t have the same status quo as socialism.	Socialism does have the same status quo as socialism. They finally managed to throw off the shackles of communism.	Socialism doesn’t have the same status quo as socialism. Does a large house with an immaculate lawn be considered a status symbol ?	2

	Context 2	Context 1	Response	Literal response	Sarcastic response	RQ response	SC result
35	The left is a bunch of unprincipled, greedy, worthless.	The complainers were probably watching television, partying with friends, and sleeping in while Bezos was creating a business that literally changed how we buy products and consume content.	I guess that means he can evade tax with impunity? good to see your laws are universally applied	I think it has to do with the fact that they got bought out by a group of people who had been known for many years to be unethical.	I think it has to do with the fact that they got bought out by a group of people who had been known for many years to be ethical. Who have often disdained the generous candour of my sister, and gratified my vanity, in useless or blameable distrust.	I think it has to do with the fact that they got bought out by a group of people who had been known for many years to be unethical. Do seeing artifacts make people depreciate men ?	2
36	I think this should be shared (from the comments section on ESPN.COM)	I think Niner and Seahawk fans can agree that we shouldn't be taking advice from a guy wearing a fedora.	Anyone else see the irony of a Niners fan calling someone out for their appearance?	I know. We all should be a little more humble.	I know. we all should not be a little more humble. Exalt him as much as you can: for he is above all praise.	I know. We all should be a little more humble. Is it that knowing what is wrong with you will worsen you ?	2
37	Looks like ol ' Rusty's account may be locked . Wonder how that happened ? ;-) >	Looks fine to me . Pain fails again . Poor little guy .	Pasty white guy wants to know if I'm racist ?	Poor little Rusty. He's a fine dog.	Rich little rusty. he's a fine dog. Dirty money will not bring good luck.	Poor little Rusty. He's a fine dog. Is the bagel round and crusty outside ?	1
38	That sad moment when you realise has more twitter followers than you :disappointed_face: :disappointed_face: :disappointed_face:	I don't know if this is a compliment or not ! :see-no-evil_monkey: :face_with_tears_of_joy:	it means you are mr popular of the twitter world #makesmesad	Is a compliment.	Is a compliment. Good - tempered and forbearing.	Is a compliment. Is a compliment better than flattery ?	0
39	Pray tell, what's the second?	That they simply have a 95 lb problem on their hands. 95 lbs they have to move from point A to point B	Shouldn't be too hard. Fold it up and putit in your wallet!	Ah, I see. That's odd, but that makes sense to me.	Ah, i see. that's even, but that makes sense to me. Stared at them both in utter confusion.	Ah, I see. That's odd, but that makes sense to me. Do people want to see Taj Mahal ?	2
40	I thought the exact same thing! Even before I knew who he is. I think its the teeth, he looks like Brian from Family Guy when he gets new teeth & becomes a real estate agent.	I think he looks like Louis Litt from Suits Edit-[for all those who need to search for Litt](<URL>[looks like his brother lol](<URL>	After a quick google, I couldn't agree more!	Yeah it's a good show.	Yeah it's a good show. The theater closed down for good.	Yeah it's a good show. Is the news show not biased ?	0
41	Actually I agree w showing respect to those w whom you disagree . I'm making an argument not name-calling .	The man is going to be president . He will be held to account and criticized for wrongdoing . That's the job .	I guess we need to show the same respect as is shown to " King Obama . "	I mean if you want to name the thing and his crimes then you should've done it before the election.	I don't mean if you want to name the thing and his crimes then you should've done it before the election. I believe in cumberland, :URL: and the murderer was shot dead whilst resisting arrest.	I mean if you want to name the thing and his crimes then you should've done it before the election. do divest murder is used to keep someone from talking about something you did ?	2

A.2 The Evaluation Data on Chatbot

Table A.2: The table of human evaluation results on the chatbot. It contains the appropriateness and human-likeness average score of the response in a conversation, the average score of whether the response is RQ, and the average score of whether the response is sarcastic or not.

	Original App	Literal App	Sarcastic App	RQ App	Original HL	Literal HL	Sarcastic HL	RQ HL	Original is RQ	Literal is RQ	Sarcastic is RQ	RQ is RQ	Original is sarcastic	Literal is sarcastic	Sarcastic is sarcastic	RQ is sarcastic
0	3.25	3.50	1.75	2.25	3.00	3.75	2.00	2.25	0.50	0.50	0.50	0.25	0.50	0.75	0.25	0.75
1	3.80	3.60	3.80	3.60	3.40	4.00	3.20	3.20	0.60	0.40	0.20	0.20	0.20	0.40	0.40	0.20
2	3.50	3.50	2.25	2.75	4.00	4.00	2.50	3.00	0.25	0.75	0.25	0.50	0.50	0.50	0.50	0.25
3	3.75	3.00	3.00	3.50	3.25	2.75	3.25	3.00	0.75	0.25	0.25	0.25	0.50	0.25	0.50	0.25
4	3.50	3.75	3.50	3.75	3.25	4.75	4.25	4.25	0.50	0.00	0.25	0.50	0.25	0.25	0.50	0.25
5	4.20	3.40	2.60	3.00	3.40	4.00	3.00	3.00	0.60	0.20	0.40	0.20	0.20	0.60	0.20	0.20
6	4.20	4.20	3.00	2.80	4.20	3.80	3.00	4.20	0.20	0.40	0.20	0.40	0.20	0.40	0.40	0.60

	Original App	Literal App	Sarcastic App	RQ App	Original HL	Literal HL	Sarcastic HL	RQ HL	Original is RQ	Literal is RQ	Sarcastic is RQ	RQ is RQ	Original is sarcastic	Literal is sarcastic	Sarcastic is sarcastic	RQ is sarcastic
7	3.40	3.00	3.40	3.20	3.40	3.60	3.20	3.00	0.20	0.40	0.40	0.20	0.20	0.40	0.40	0.20
8	3.80	4.20	3.20	3.20	3.60	3.20	2.40	3.40	0.80	0.20	0.20	0.20	0.40	0.40	0.20	0.40
9	3.40	2.60	3.00	2.60	3.80	2.60	3.40	3.40	0.80	0.40	0.20	0.20	0.20	0.20	0.60	0.40
10	3.60	3.40	2.40	2.60	2.80	3.40	2.80	2.80	0.60	0.20	0.20	0.40	0.20	0.40	0.40	0.40
11	3.60	3.00	3.60	3.60	3.60	3.40	3.60	3.60	0.60	0.20	0.80	0.40	0.40	0.40	0.20	0.80
12	3.33	2.67	3.33	3.67	3.33	2.33	3.33	3.67	0.00	0.33	0.67	0.33	0.33	0.00	0.67	0.33
13	3.60	3.20	3.40	3.20	4.80	4.20	4.00	3.60	0.40	0.40	0.20	0.20	0.20	0.40	0.80	0.20
14	3.80	4.40	3.20	3.00	4.40	4.00	3.60	2.80	0.20	0.20	0.40	0.40	0.00	0.60	0.60	0.20
15	3.25	4.25	3.25	4.00	3.75	3.50	3.25	4.00	0.00	0.50	0.50	0.25	0.50	0.50	0.25	0.25
16	4.25	2.75	3.00	3.25	3.75	2.75	3.25	3.75	0.25	0.25	0.75	0.25	0.75	0.50	0.25	0.00
17	3.20	3.20	3.80	3.80	4.20	3.60	4.00	3.80	0.20	0.60	0.40	0.20	0.40	0.20	0.60	0.20
18	3.75	3.75	4.25	4.25	3.25	3.50	4.50	4.25	0.75	0.50	0.00	0.25	0.25	0.50	0.75	0.25
19	4.00	3.25	3.50	3.25	4.00	3.50	4.00	4.00	0.00	0.50	0.50	0.00	0.50	0.25	0.25	0.25
20	3.60	3.80	4.00	3.80	4.80	3.40	3.40	3.20	0.20	0.40	0.00	0.40	0.20	0.20	0.40	0.00
21	3.20	3.80	3.80	3.60	4.00	3.60	2.80	2.80	0.00	0.40	0.40	0.40	0.00	0.60	0.20	0.20
22	4.25	3.50	3.75	3.75	3.75	3.25	4.25	3.25	0.25	0.50	0.50	0.25	0.00	0.25	1.00	0.25
23	3.25	4.00	3.50	2.50	3.50	3.75	3.50	3.00	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.00
24	2.40	3.60	3.40	4.20	3.00	3.80	3.80	4.40	0.60	0.20	0.40	0.20	0.60	0.20	0.60	0.00
25	4.00	3.00	4.00	2.50	4.50	3.50	4.00	2.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50
26	4.25	3.75	3.75	3.75	3.75	4.00	3.75	3.50	0.25	0.25	0.25	0.50	0.50	0.50	0.75	0.00
27	4.00	3.80	2.60	3.60	4.00	4.00	2.60	3.80	0.20	0.60	0.20	0.40	0.60	0.40	0.20	0.20
28	4.40	3.80	4.20	3.40	4.40	3.80	4.20	3.80	0.60	0.20	0.40	0.20	0.20	0.20	0.60	0.00
29	3.25	3.50	3.75	3.00	3.25	3.75	3.25	3.75	0.50	0.25	0.25	0.25	0.25	0.25	0.75	0.00
30	4.00	3.50	3.50	3.75	4.00	3.50	3.75	3.75	0.00	0.50	0.25	0.50	0.50	0.25	0.25	0.00
31	4.00	3.25	3.50	3.25	4.00	3.25	4.25	3.25	0.75	0.00	0.25	0.00	0.00	0.50	0.50	0.25
32	3.00	3.75	3.75	4.25	4.00	3.00	3.75	3.50	0.00	0.75	0.25	0.50	0.25	0.50	0.50	0.00
33	4.25	3.00	3.50	3.50	3.75	3.50	3.75	3.75	0.50	0.25	0.00	0.50	0.00	0.25	0.50	0.00
34	4.50	3.00	3.25	3.75	5.00	3.50	3.75	3.50	0.25	0.25	0.25	0.25	0.25	0.25	0.50	0.25
35	4.33	3.67	2.33	3.00	4.33	3.67	3.33	2.67	0.00	0.67	0.00	0.33	0.33	0.33	0.33	0.00
36	4.25	3.75	4.25	3.25	3.75	4.00	3.75	3.75	0.50	0.00	0.50	0.25	0.75	0.50	0.25	0.25
37	3.00	3.25	3.00	3.25	3.50	4.00	3.50	3.25	0.25	0.75	0.50	0.25	0.50	0.00	0.50	0.00
38	3.50	3.00	3.00	3.50	4.00	3.50	4.00	3.50	0.25	0.50	0.25	0.50	0.25	0.25	0.50	0.50
39	3.75	3.25	3.00	3.00	4.00	3.50	3.00	3.25	0.75	0.00	0.50	0.00	0.75	0.50	0.00	0.00
40	4.67	2.67	4.00	3.33	4.67	3.67	4.00	3.33	0.33	0.33	0.00	0.67	0.67	0.33	0.67	0.33
41	3.75	2.75	3.25	3.50	4.25	3.50	3.50	2.75	0.25	0.00	0.50	0.00	1.00	0.25	0.00	0.00