Evaluation Baseline for Open-Domain Chatbots Based on Publicly Available Models and Apps

Siran Li
siran.li@epfl.ch
(Sciper 321825)

Abstract

Emotion expression and delivery are central to the human experience and social interaction. Evaluation of conversational chatbots based on various features can help improve the interpretability and controllability of the dialogue systems. This project aims on curating an evaluation baseline for open-domain chatbots. Based on EmpatheticDialogues dataset and some previous work about empathetic evaluation metrics, we annotated the grammatical types of the questions in the dialogue. The project utilizes different visualization techniques to exhibit the delivery of empathy, stresses the significance of language competence, and reveals practical ask-question strategies in social dialogues. These results can further facilitate a range of research and practical activities about empathetic conversation generation.

1 Introduction

The development of evaluation metrics is a crucial subject for open-domain chatbots research. Standard automatic evaluation metrics such as BLUE, METEOR, and ROUGE show either weak or no correlation with human judgments[1]. Recently, researchers have proposed a lot of advanced, dialog-specific metrics, and released multiple evaluation datasets based on the evaluation metrics for different features of chatbots[2, 3, 4, 5]. They point out the weaknesses of the standard automatic evaluation metrics and build metrics that show a better correlation with human judgment. Since most of these metrics were created during the same periods and were evaluated on different datasets. We need to assess these metrics and datasets and summarize these evaluations for future in-depth research.

The dialog evaluation metrics can be divided into rule-based and model-based metrics[6]. The rule-based approach use heuristic rules to evaluate the text, based on the dialog content and human references. BLUE[7] is a typical rule-based metric comparing n-grams of the candidate and the reference to benchmark natural language generation systems. METEOR[8] and ROUGE[9] improve upon BLEU, but they still remain ineffective for dialog evaluation[1]. These metrics assume that valid responses have significant word overlap with the ground truth responses, which greatly limit the diversity of the valid responses.

Compared with the rule-based approach, the model-based approach is more flexible to the dialog context and can analyze the semantic meaning without the restriction of limited contexts. Model-based approach trains or uses machine learning models, to measure the quality of the responses. ADEM[9] uses a recurrent neural network (RNN) to predict the quality of the text. And the predictions correlate significantly, and at a level much higher than rule-based metrics, with human judgments at both the utterance and system level. RUBER[2] uses a hybrid

model of both a referenced metric and an unreferenced metric. It has a strong correlation with human annotation and has fair transferability over different datasets. BERTScore[10] computes a similarity score for each token in the candidate sentence with each token in the reference sentence and computes token similarity using contextual embeddings, to correlate better with human judgments. BLEURT[11] generates synthetic data to pre-train BERT and fine-tune the model to predict a human score with MSE loss. GRADE[12] incorporates both coarse-grained utterance-level contextualized representations and fine-grained topic-level graph representations to evaluate dialogue coherence. Compared with GRADE, DynaEval[13] is not only capable of performing the turn-level evaluation but also considers the quality of the entire dialogue. USR[5] is a reference-free metric that trains unsupervised models to measure several desirable qualities of dialog. DiaoGPT[14] is based on the GPT2 architecture and trained with dialog data. The task includes predicting human feedback of responses and whether the response is human-like. HolisticEVal[15] metric is computed by GPT-2 model to evaluate the context coherency and response fluency of the dialogue contexts. PredictiveEngage[16] metric estimates utterance-level engagement to be used as real-time feedback for training better dialogue models. And FED[17] metric uses DialoGPT[18] to measure 18 fine-grained qualities of dialog and attains moderate to strong correlation with human judgment at both turn-level evaluations and whole dialog-level.

With the fast development of chatbots, a number of evaluation metrics emerged at about the same time. For this reason, we conduct an overview of the existing evaluation datasets. These datasets have human annotations that measure the quality of the responses. After the exhaustive overview, we discover that none of the existing datasets is focused on evaluating empathy. However, to improve the interaction of chatbots communication services, it is crucial for chatbots to have a great performance with inference, personalization, and empathy[19].

For this reason, the aim of this project is to curate an evaluation benchmark for empathy approximating human judgment and effective conversation strategies. First, we explored existing dialog evaluation datasets and made an exhaustive overview of the evaluation datasets. Then, we annotated the grammatical types of all questions and preprocessed the empathetic dataset. Through the visualization and analysis techniques, we found effective strategies to conduct pleasant conversations and avoid some communication breakdowns.

The remainder of this paper is organized as follows. Section 2 describes the related work about taxonomies of emotions and grammatical questions. Section 3 presents the Empathedic-Dialogues dataset and the basic statistic summary. Section 4 describes our attempt to annotate grammatical question types and the distribution of different grammatical question types. Section 5 analyzes the empathetic strategies hidden in social dialogues. After the dataset is preprocessed with emotion labels, question intents, and acts labels, it presents the significance of language competence. It also shows the delivery of empathy and discusses the correlation between grammatical question types and other empathetic labels. We also introduce the top frequency words in the dialogues. In section 6, the limitations and future work are discussed. Finally, the project is concluded in Section 7.

2 Related Work

Automatic evaluation metrics Chatbots are dialog engines for interactive user experience, and the evaluation of dialogue is crucial to the development of chatbots. Since human evaluation is quite expensive and time-consuming, automatic evaluation metrics become a significant component of the research process. Yi-Ting et al.[6] provides a comprehensive assessment of recently proposed dialog evaluation metrics. They have an overview of automatic metrics, from the popular and basic rule-based metric BLEU to diverse pre-trained model-based metrics. They

analyze the different types of response generation models, compare their performances, explore combinations of different metrics, suggest how to best assess evaluation metrics, and point out the promising directions for future work.

Linguistic proficiency Besides all the above-mentioned evaluated qualities of the metrics, communication quality also correlates with linguistic proficiency. Code-switching (CS) is a strategy usually used in multilingual communities - when a bilingual mixes two or more languages within a discourse, or even within a single utterance. Comparing the high and low code-switchers, the high code-switchers exhibit statistically significantly lower mean sentence length and lower proficiency in the lexical metrics[20]. In this case, we consider using the number of words and the number of sentences in each utterance to measure the communication qualities.

Emotional taxonomy In human social interaction, empathy always plays a vital role and affects the conversation development[21]. Therefore, in social interaction, a chatbot needs to be empathetic to maintain healthy interaction with humans and develop trust. A taxonomy of empathetic listener intent covering 32 types of emotion categories was developed to explain the patterns and trends of the conversation flow[22]. To annotate all the utterances with emotional labels, they trained a BERT transformer-based classifier[23]. With the emotional labels, the taxonomy of empathetic listener intents shed light on the frequent empathetic conversation patterns among social chitchat. Our project is based on this model and used the same dataset to analyze the empathetic performance in dialogues.

Empathetic question taxonomy In a dialogue system, effective question-asking is significant for a successful conversation chatbot. It could help the chatbots manifest empathy and render the interaction more engaging by demonstrating attention to the speaker's emotions. Ekaterina et al.[24] developed an empathetic question taxonomy (EQT) focused on questions in the dialogue to capture communicative acts and their emotion-regulation intents. By analyzing the question acts and intents with the emotion of the continuous utterances, they got some effective question-asking strategies to make the conversation more productive and suitable.

Evaluation datasets During the development of dialogue systems and automatic evaluation metrics, a number of dialogue evaluation datasets with assessments approaching human judgments were created. Our goal is to build an evaluation baseline for open-domain chatbots, so we need to have an overview of the existing evaluation datasets to prepare for our future work.

In general, we assessed 17 evaluation datasets. From the An Overview of Evaluation Dataset file, we can get the source and brief introduction of each dataset. Table 1 lists some characters of the evaluation datasets. We also assessed the datasets on both the turn level and the dialog level and find out their public resources. In the overview, the datasets are also considered with static evaluation and interactive evaluation. For static evaluation, the chatbots talk based on a curated dataset with multi-turn conversations. For interactive evaluation, chatbots can chat about anything they want.

The overview gives us exhaustive information about the existing evaluation datasets. Most of these evaluation datasets focus on the proximity to humans, the correctness, and the naturalness of the context. And some also analyze the reaction of humans to the generated context, such as whether they are interested or engaged in the conversations. Therefore, from the overview, we can find that there are no available public datasets that can support the development of evaluation metrics for empathy analysis in chatbots.

Datasets	Size	Evaluated qualities	Used NLG models
USR-TopicalChat[5]	360	Understandable, Uses Knowledge, Natural, Maintains Context, Interesting, Overall Quality	Argmax Decoding, Nucleus Decoding, Nucleus Decoding, Nucleus Decoding, New Human Generated, Original Ground Truth
USR-PersonaChat[5]	300	Understandable, Uses Knowledge, Natural, Maintains Context, Interesting, Overall Quality	KV-MemNN, Seq2Seq, Language Model, New Human Generated, Original Ground Truth
GRADE-ConvAI2[12]	600	The coherence between the context and the response	BERT, DialoGPT, Transformer Seq2Seq, Transformer Ranker
GRADE-DailyDialog[12]	300	The coherence between the context and the response	Transformer Seq2Seq, Transformer Ranker
GRADE-EmpatheticDialogue[12]	300	The coherence between the context and the response	Transformer Seq2Seq, Transformer Ranker
HolisticEval-DailyDialog[25]	400	Context Coherence, Language Fluency Response Diversity, Logical Self-consistency	LSTM Seq2Seq
PredictiveEngage-ConvAI[16]	13,124	Engagement score	SVM, MLP Word2vec, MLP Bert(Mean), MLP Bert(Max)
PredictiveEngage-DailyDialog[16]	25,900	Engagement score	LSTM Seq2Seq
DSTC6[26]	40000	Overall score	LSTM Seq2Seq
DSTC7[27]	9990	Relevance, Informativeness, Overall	LSTM Seq2Seq
FED (turn-level)[17]	375	Interesting, Engaging, Specific, Relevant, Correct, Semantically Appropriate, Understandable, Fluent, Overall Impression	Meena, Mitsuku, Human
FED (Dialog-Level)	125	Coherent, Error Recovery, Consistent, Diverse, Topic Depth, Likeable, Understanding, Flexible, Informative, Inquisitive, Overall Impression	Meena, Mitsuku, Human
Persona-Chatlog[28]	3316	Avoiding Repetition, Interestingness, Fluency, Listening, Inquisitiveness, Humanness and Engagingness	LSTM Seq2Seq
DailyDialog-Eval (GD)[29]	500	Overall score	Human, HRED, Seq2Seq Dual Encoder, CVAE
DailyDialog-Eval (ZD)[30]	900	Content, Grammar, Relevance and Appropriateness	ADEM and RUBER
PersonaChat-Eval (ZP)[30]	900	Appropriateness	ADEM and RUBER
HUMOD[31]	9500	Relevance, Language Usage	BLEU-4, ROUGE, METEOR, HAN-R(CE), HAN-R(MSE), BERT
Google Meena (Static case)[32]	5810	Sensibleness and Specificity	Cleverbot, DialoGPT, Meena, Meena (base), Human
Google Meena (Interactive case)	700	Sensibleness and Specificity	Cleverbot, DialoGPT, Meena, Meena (base),XiaoIce, Mitsuku, Human

Table 1: Summary of the evaluation datasets assessed

3 Dataset

The EmpatheticDialogues (ED) dataset created by Rashkin et al[33] comprises 25k publicly available dialogues. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding. The conversation consists of up to 6 turns, and each conversation takes place between 2 individuals related to a particular emotion. For convenience, we call the talker of the first utterance 'speaker', and the talker in the second utterance 'listener'. The total 32 emotional contexts are evenly distributed, which makes sure that the conversations sufficiently cover each emotional situation. Table 2 displays some basic statics of the dataset.

Number of dialogs	24,850
Average number of turns per dialog	4.31
Number of dialogs with at least one question from listener	15,253 (61.4%)
Number of questions from listeners	20,201

Table 2: Summary statistics of our data

4 Annotation of Grammatical Questions types

How speakers design and use their questions and responses in ordinary spontaneous conversation depends on their social action. Stivers et al. [34] discuss the range of ways that speakers ask and respond to questions and what speakers are doing through asking questions by analyzing the different grammatical types.

Therefore, to understand the speakers' intents and the linguistic methods, we divide the questions into three primary question types: polar questions, Q-word questions, and alternative questions and interjection questions[35, 36, 37]. With these grammatical categories of questions, we can analyze the emotions and intents in the dialogue clearly, and obtain more information from the question-response system.

Polar questions Polar questions are said to be answered with a yes or a no in English. Interrogative, tag, and declarative questions make up the dominant sub-types of polar questions. The interrogative is formed by placing the operator before the subject and giving the sentence a rising intonation[36]. Tag questions express "maximum conduciveness" thereby coercing particular answers in line with the question to a greater extent than other question types, for example, "You speak English, don't you?". A declarative question has the form of a declarative sentence but is spoken with rising intonation at the end, like "You think I'm kidding you?", "You're firing me?" etc.

Q-word questions By asking Q-word questions, people want to get more information and are more curious about the ongoing topics. For Q-word questions, they can also be classified by the question words (who, whose, what, where, when, why, and how).

Alternative questions Alternative questions are seldom asked. It offers the listeners a closed choice between two or more answers, and the choices are conjoined by or, such as "Are you coming or going?"

Interjection questions Interjection questions are usually very short, like "Oh really?", "What?", "Sure?". They can be used to express strong feelings or sudden emotions., and aren't grammatically related to any other part of the sentence.

4.1 Method

Firstly, we use spaCy to help us extract question types. Based on different structures of grammatical types, we utilize a POS tag for each token to identify its function in every sentence. Then we define the most basic linguistic generalizations about each category and we also find out some counterexamples to the generalizations and revise them. After testing the rules on a bunch of examples, we fine-tune the rules by addressing any false positives and testing new examples.

Based on the generalization rules, we can already categorize 86% of questions. After this, we use the pattern matching method to identify the fine-grained sub-types questions and the identified questions ratio already achieves 96%. At last, we manually identify categories for the remaining questions.

4.2 Statistic summary

Fig. 1 shows that across a broad range of grammatical types, ranging from requesting information to initiating repair to seeking agreement with an assessment and expressing strong feelings. The

substantial majority of all questions asked are polar questions (n=10544). About 40% (n=8281) of all questions are Q-word questions, and the proportion of alternative questions is less than 5% (n=958) of questions. The interjective questions only accounts for about 2% (n=418).

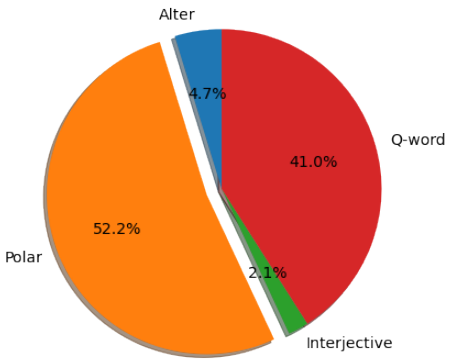

Figure 1: Distribution of questions across question type.

In the dialogue system, the distribution of polar questions by sub-type is shown in Table 3. Interrogative questions are the dominant polar question type, and the declarative questions account for 14% to confirm the information they get. The tag questions are not very usual in dialog, it only comprises 3% of polar questions.

Polar question type	Percent/counts
Total polar question	100% (n=10544)
Interrogative	83% (n=8780)
Declarative	14% (n=1420)
Tag	3% (n=344)

Table 3: The distribution of polar questions by sub-type

5 Analysis of Empathic Strategies

We aim at developing an evaluation baseline for open-domain chatbots. The empathetic ability of a chatbot plays a vital role to build a friendly conversation. Therefore, we use different approaches and classifiers to get more information about the dialogue system and prepare for the exhaustive analysis.

5.1 Data pre-processing

Emotion labels According to the taxonomy of empathetic response intents described by Welivita and Pu[22], we add the emotion category labels to each utterance. Since the accuracy of the classifier for identifying emotion labels is 65.88%[22], we first use pattern matching methods to identify the 8 most frequent emotional intents (questioning, acknowledging, agreeing, consoling, encouraging, sympathizing, wishing, and suggesting) for each utterance to improve the accuracy. After this process, the ratio of unidentified sentences is still over 75%. Then, we use the pre-trained BERT-based classifier[22] to give the remaining sentences emotion labels. Finally,

there is an emotion label for each utterance and sentence in ED dataset. The total number of categories of the emotion labels is 41.

Coding for emotion labels To better quantify the emotion changes and analyze the conversation tendency, we convert emotion labels to 7-point scale codes. Relying on Plutchik's weel[38], we give a positive emotion, neutral emotion, and negative emotion a positive, zero, and negative score respectively. And we also give the stronger emotion a score with a larger absolute value. Table 8 in Appendix A illustrates the mapping of emotion labels to 7-point scale scores.

Question acts and Question intents From table 2 we can get that over 60% of all dialogs contain a question in one of the listeners' turns. Asking questions plays a leading part in the conversation tendency. Therefore, we also add the question acts and question intents labels in ED dataset[24]. We use Question acts to capture semantic-driven communicative actions of questions, and question intents to describe the emotional effect the question should have on the dialog partner.

5.2 Language competence

Language competence affects communication qualities, and talkativeness is an important indicator of language competence. Here, we use the average length of a sentence and the average number of sentences in the utterance to exhibit the talkativeness quality.

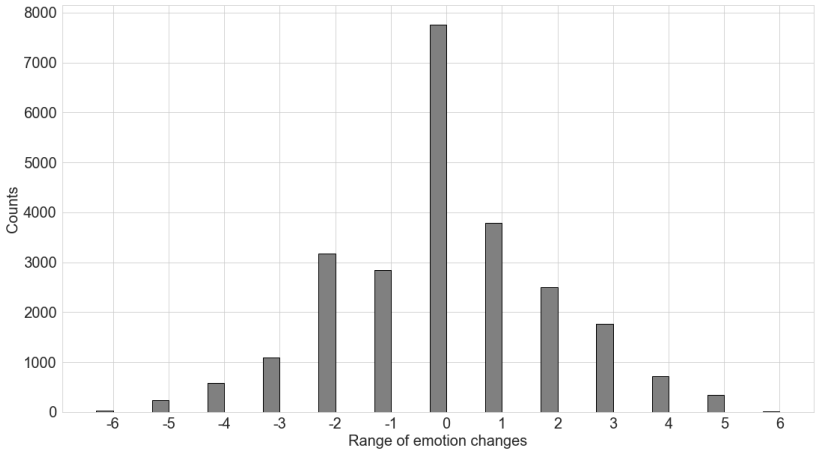

Figure 2: Distribution of the utterance with different emotion score change value

Based on the difference score of the third utterance and the first utterance, we count the number of dialogues in Fig.2. In most cases, the speakers' emotion is stable and the change value is 0. The greater the emotion scores change, there are fewer cases, which accords with common sense.

In Fig.2, we can know that there are few cases with change value ± 6. After checking several cases with change value ± 6, we find that the emotion change so rapidly because of the classifier error. Therefore, in Fig.3, we count the average number of words in and the average number of sentences with different emotion change values without the case having change value ± 6. It is obvious that the average number of words and sentences with the positive change values are larger than the average number of words and sentences with the negative change values

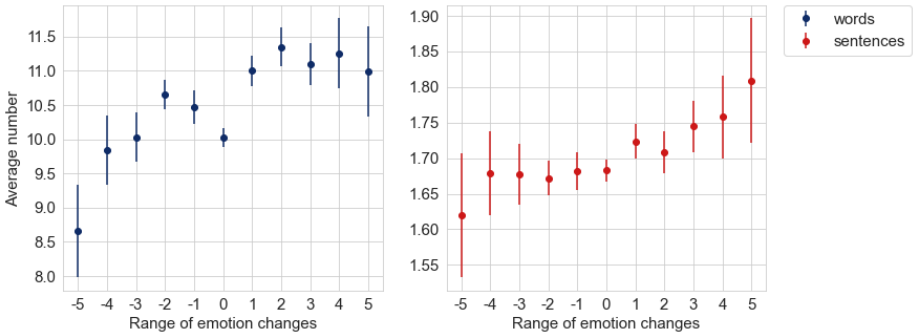

Figure 3: Distribution of average number of words and sentences with different emotion change values

respectively. And as the emotion change value increases, the average number of sentences increases accordingly.

In this case, we can infer that as the listener is more talkative, it has better language competence to express their intents more clearly and give reasonable suggestions. Fluency and consistency during chat are important qualities for chatbots to improve their language ability.

5.3 Delivery of empathy

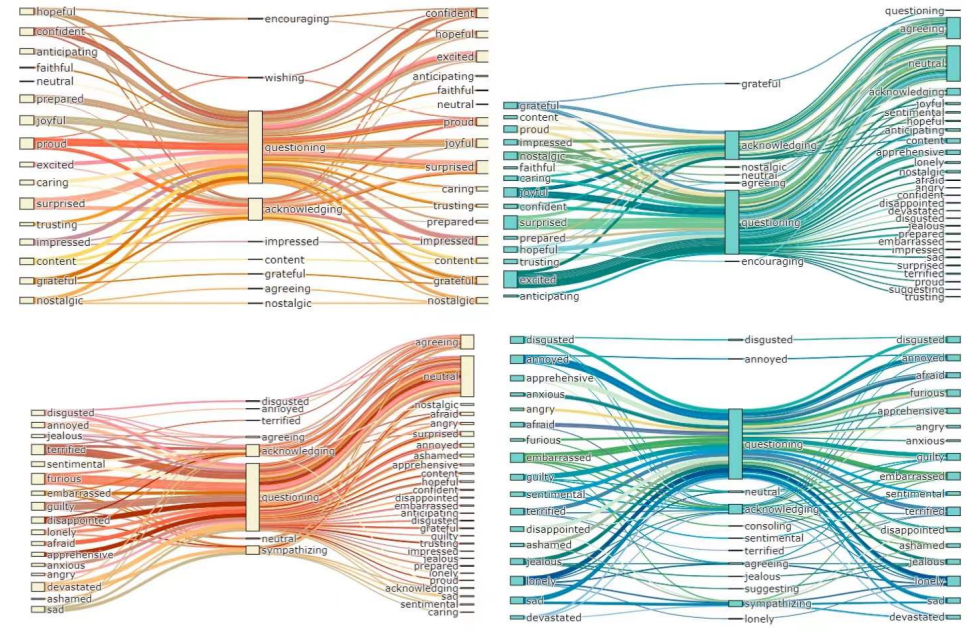

Figure 4: Visualization of the most common emotion flow patterns (frequency ≥ 10) throughout the first three dialogue turns in the EmpatheticDialogues dataset.

To find some strategies for making the conversation develop better, we focus on the first three turns in each dialogue, because over 70% of dialogs in the dataset have only four turns, so

it would not be possible to see the influence of questioning strategy on the speaker's emotion further in the fifth turn. Then we visualize their emotion flow patterns in Fig.4. In the first row, the first utterance has positive emotions, while in the second row, the first utterance has negative emotions. In the first column, comparing the third turn with the first turn, the emotion scores are ascending, while in the second column, the emotion scores are descending.

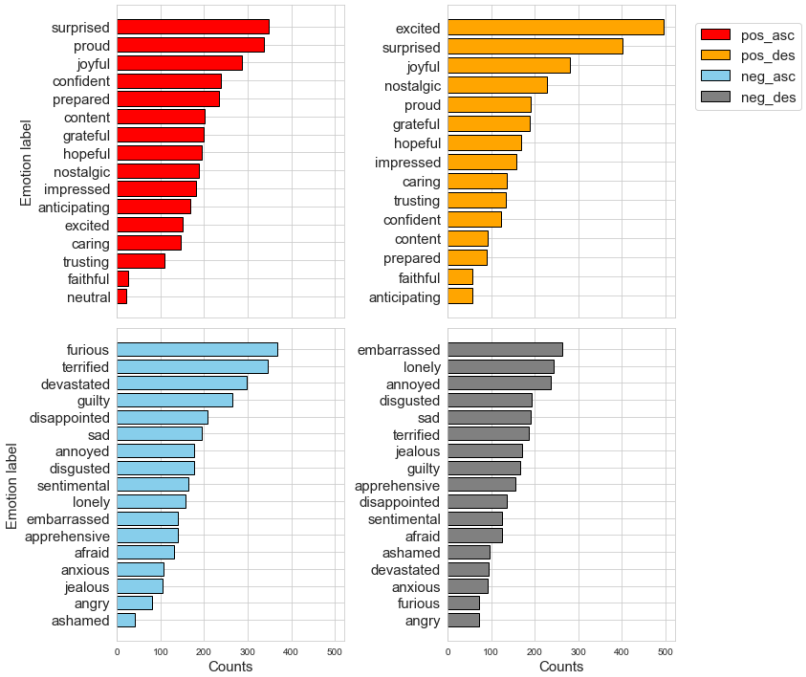

Figure 5: The distribution of the emotion labels in the first utterance

From the flow chart, we can get that no matter in which group, asking questions is always the most decisive action to convert the emotion of speakers. In Fig.5, we can get the distribution of the emotion labels in the first utterance, and they are also divided into 4 groups as in Fig.4. As observed before, 'surprise', 'proud' and 'joyful' emotion labels are the largest groups and they are also easy to be affected by questions or acknowledging talks to maintain the status or turn to neutral emotions. Since 'excited' emotion has the highest emotion score, it is hard to make speakers keep excited all the time. After asking questions, the speaker's emotion often turns to more calm and neutral feelings.

However, for the negative emotions, the distributions of the emotion labels are different. 'furious', 'terrified' and 'devastated' emotion can convert to 'neutral' after asking questions, while 'embarrassed', 'lonely' and 'annoyed' feelings are hard to be changed, and in most of the cases, the emotions keep the same. 'furious', 'terrified' and 'devastated' are very strong feelings, people cannot keep these feelings for a long time, and these emotions can be changed more easily. On the contrary, in most situations, 'embarrassed', 'lonely' and 'annoyed' come from oneself, they are not very strong and not easy to be displayed or to be vented out. So compared with the intense emotional experience, the inner motivations last longer and are harder to change.

To better understand how the listener's intents affect the speaker's emotions, we use a heatmap Fig.6 to exhibit the correlation between the speaker's emotion and the listener's intent. The color indicates the emotion score change value between the first utterance and the third

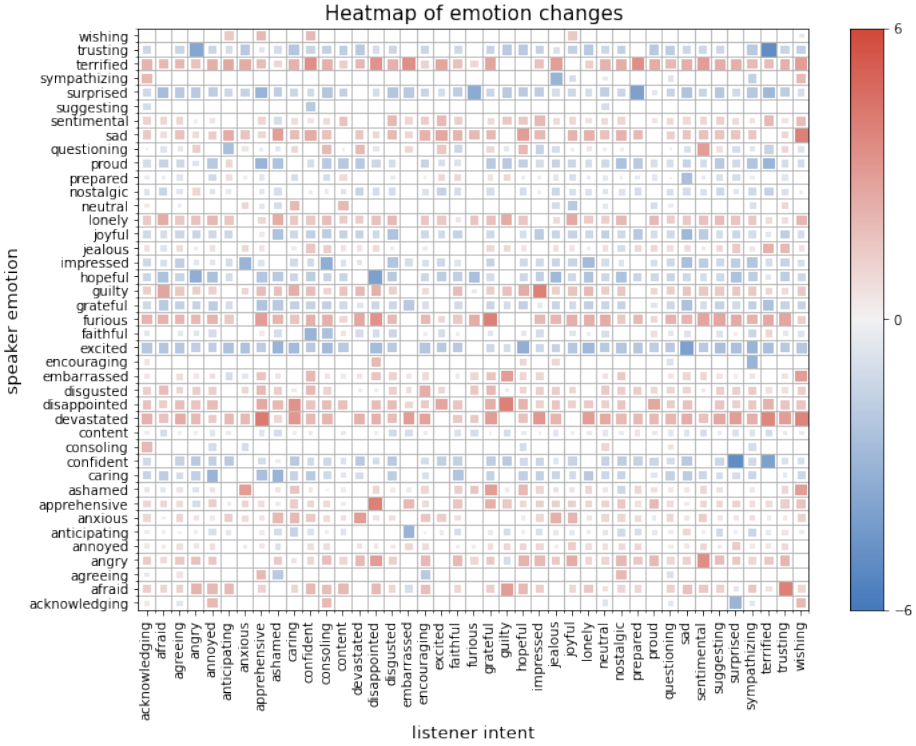

Figure 6: The heatmap shows the correlation between the emotion label of the first utterance and the emotion intent of the second utterance. The color indicates the emotion score change value between the first utterance and the third utterance.

utterance. From this plot, we can have a direct view about which talking strategy is effective and which is ineffective or terrible.

There are some examples listed in Appendix B. When the speaker was devastated, the listener talked with an apprehensive tone, this can make the speaker feel relieved and comfortable. In a disappointing conversation, the listener also talked with guilty feelings, it made the speaker feel better and hopeful. On the contrary, when the speakers have good emotions and the listeners talk with negative feelings, this can make the speakers change their moods to negative spirits and the conversation become oppressive.

However, there exist some deficiencies in this discussion. The samples of each label are not enough for a detailed and complete analysis. In some cases, there are only several samples, so the conclusion lack massive data support. What's more, the classifier error can also cause analytical inaccuracy. The heatmap Fig.6 can help us learn more about the linguistic strategy, but for more accurate results we need numerous data and more exact tagged labels.

5.4 Questions with different grammatical types

Effective question-asking plays an important role in a successful conversational chatbot, which we can also know from Fig.4. Speakers can design and use their questions in ordinary spontaneous conversation to express their different social action[34]. The steps described above provided a large labeled collection of empathetic questions with question intents, question act,

and grammatical labels. This allowed us to explore question strategies used by the listeners in response to speakers' emotional input.

Fig.7 shows the joint distribution of grammatical types and question intents and the joint distribution of grammatical types and question acts. The main intent of the polar question, Q-word question (WH question) and alter question is to express interest and concern. Compared with Q-word questions, polar questions have a better ability to offer relief. And for interjection questions, it is used to express interest in most cases. As for the correlation between question acts and the grammatical types, polar question, Q-word question, and alternative question all request information as the main act. Comparing polar questions and Q-word questions, polar questions ask about consequence more and Q-word questions take care about antecedent more. And polar questions have a broader range of question acts, like suggesting a solution, asking for confirmation, and suggesting a reason. For alternative questions, suggesting a reason is also a dominant component. For interjection questions, it was mainly used to ask for confirmation.

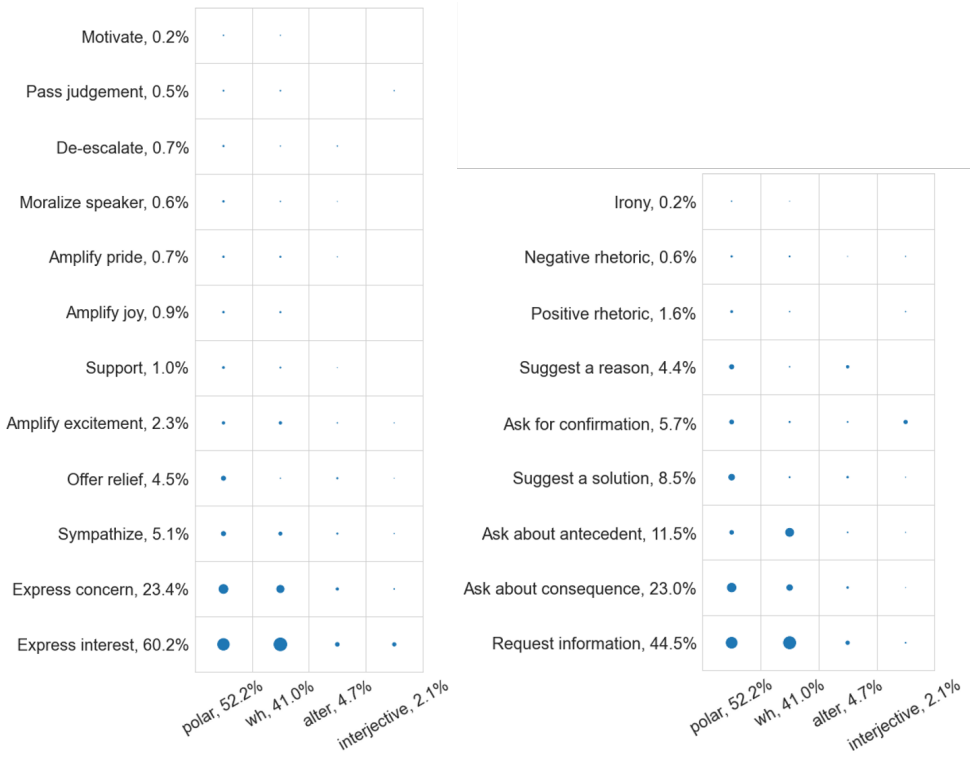

Figure 7: Joint distribution of the grammatical types with question intents and question acts. Left: joint distribution of question intents and grammatical types. Right: joint distribution of question acts and grammatical types. The size of blue circles is proportional to the frequency of each pair's co-occurrence. The percentage of each individual label is printed next to it along the axes.

After analyzing the intents and acts expressed by different grammatical types, we also need to connect different types with emotion inputs and discuss them in specific cases. Here we count the number of different grammatical question types asked by listeners (in the second utterance) in different emotional states (the emotion label in the first utterance). Comparing the third utterance with the first utterance, if the speaker's emotion becomes worse, we add -1 to this emotion group, otherwise, we add +1.

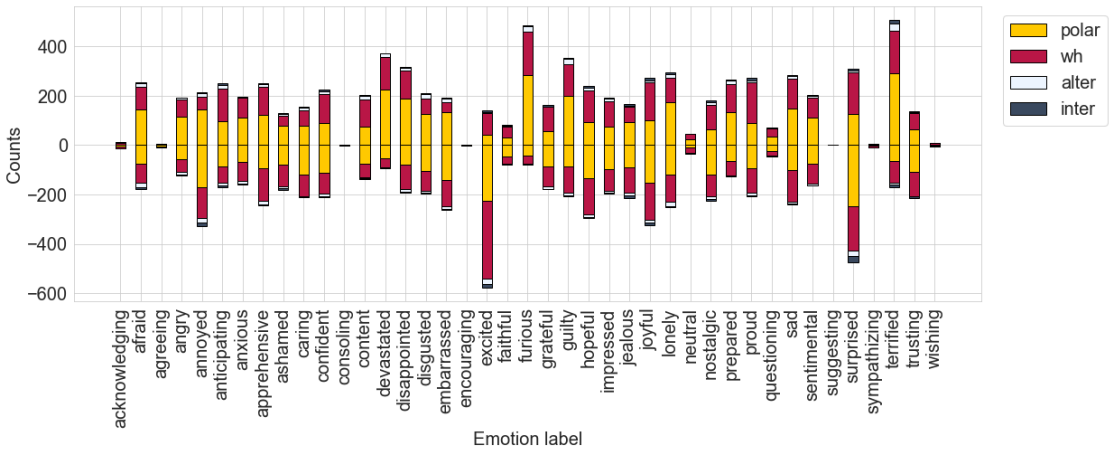

Figure 8: The Speaker's Emotion changes with different grammatical type of questions

Fig.8 shows the distribution of different grammatical types with different speaker's emotion input. Polar question is the majority of the whole questions, and it plays an active role in some strong and negative emotions, such as 'afraid', 'angry', 'anxious', 'devastated', 'disappointed' and 'furious'. After asking for some information and expressing concern about the situation, the speaker's emotion gradually gets better or becomes calmer. While for some positive emotions such as 'caring', 'confident', 'nostalgic', 'surprised' and 'trusting' asking inappropriate polar questions can make the emotion of speakers turn bad. Some examples of emotion change with polar questions are shown in Table.4

Emotion gets better:	Furious- Polar - Content
Speaker:	I was so mad last week when I got a flat. It was in traffic too.
Listener:	That's a bummer. Did you get it repaired?
Speaker:	I did after a few hours.
Emotion gets worse:	Nostalgic- Polar - Terrified
Speaker:	I watched IT when I was only six years old.
Listener:	Were you scared?
Speaker:	It was the worst experience of my life. I still can not watch movies like that.

Table 4: Dialogue examples with polar questions. The emotion label in blue font has positive emotion, and the emotion label in red font has negative emotion

.

When the speaker has positive emotions like 'anticipating', 'confident', 'content', and 'proud', in this situation, they'd like to answer some Q-word questions to express their glad mood. The Q-word questions usually ask about antecedents which can amplify the speaker's excitement and make them become more joyful. However, for some negative feelings from oneself, like 'annoyed', 'ashamed', and 'embarrassed', asking Q-questions is possible to keep them in the original emotion or feel worse about their experience. Table.5 shows the successful and unsuccessful dialogues under Q-word questions.

As for alternative question, it performs well with 'caring', 'prepared', 'disgusted', 'furious', 'guilty', and 'terrified' emotion input, while it makes emotion get worse when the speaker

Emotion gets better:	Anticipating- WH - Excited
Speaker:	I can't wait for next february.
Listener:	What is happening in February?
Speaker:	I'm going to Disney World with my family for the first time!
Emotion gets worse:	Annoyed- WH - Furious
Speaker:	I was irked when i saw my cousin coming inside the house.
Listener:	Why were you irked?
Speaker:	Cause he stole money from me.

Table 5: Dialogue examples with Q-word questions. The emotion label in blue font has positive emotion, and the emotion label in red font has negative emotion

.

Emotion gets better:	Prepared- Alternative - Confident
Speaker:	I have to make a tuna casserole for tomorrow.
Listener:	Do you enjoy cooking for others or for the event you are taking it to?
Speaker:	It's for a potluck. I am sure I can do it well.
Emotion gets better:	Terrified- Alternative - Afraid
Speaker:	I was scared walking home last night.
Listener:	Do you live in a bad area or a big city?
Speaker:	I live in the woods.
Emotion gets worse:	Grateful- Alternative - Neutral
Speaker:	My friend bought me dinner tonight. It made me really appreciate him.
Listener:	That's nice. Was it a special occasion or anything ?
Speaker:	No, it was pretty random which was nice.
Emotion gets worse:	Jealous- Alternative - Lonely
Speaker:	Whenever I see a happy couple, I get so envious.
Listener:	Aww, that's sad. Do you have a bad relationship or not have one?
Speaker:	I've been single for a while now :(

Table 6: Dialogue examples with alternative questions. The emotion label in blue font has positive emotion, and the emotion label in red font has negative emotion

.

has 'grateful', 'proud', or 'jealous' emotions. Table.6 exhibits some dialogues with alternative questions. If the speaker has 'caring' or 'prepared' emotion, and it's good for the listener to ask some details, then the speaker can share more information and his cheerful feeling. For 'disgusted', 'furious', 'guilty', and 'terrified' emotion, if the listener asks some alternative questions, it can give some assumptions of their situation or suggest a reason behind their experience. This can reduce the speaker's strong and terrible feelings or transfer his attention to other issues.

On the contrary, for 'grateful' and 'proud' emotions, the listener asked about some information to transfer the speaker's attention, which made the conversation become neutral and quiet. And for 'jealous' feelings, the speaker asks more details about their situation, which makes the speaker feel even worse.

When the speaker feels 'confident' and 'proud', he would like to hear others' admiring and interjective words. However, when the speaker has 'neutral', 'surprised', 'content', 'afraid', 'annoyed' or 'devastated' feeling, the interjection have a negative effect. For in the positive

emotion 'neutral', 'surprised', or 'content', the listener asked 'Oh yeah?', 'Really?' or 'Sure?'. These interjection questions doubted what the speaker said and affected the speaker's feelings getting worse. As for 'afraid', 'annoyed' or 'devastated' feelings, the listener amplifies the terrible feelings, which makes the speaker worried more about his situation.

Emotion gets better:	Confident- Interjection - Confident
Speaker:	I'm one of the best surfers that has ever hit the beach. I ride giants.
Listener:	Really?? Are you famous then? What is your name?
Speaker:	Yes, you can google me.
Emotion gets worse:	Surprised- Interjection - Acknowledging
Speaker:	I saw a couple of kids walking around in their high school football uniforms after practice the other day.
Listener:	Oh really?
Speaker:	Yeah it really took me back to when I was doing the same thing.
Emotion gets worse:	Afraid- Interjection - Terrified
Speaker:	Hi, a robber held me at gun point yesterday.
Listener:	Really? Please tell me more.
Speaker:	I was so scared, he demanded I give him my phone and wallet.

Table 7: Dialogue examples with interjection questions. The emotion label in blue font has positive emotion, and the emotion label in red font has negative emotion

.

5.5 Analysis of high frequency words

After analyzing each type of empathetic label and grammatical label, we have a better understanding of the listener's linguistic strategies. The exact context can not only inspire us to find the potential intents of the listener but also help us to summarize the conversation strategies. For this reason, we need to find the high-frequency words said by the listener in different conditions.

First, we remove the punctuation, lemmatize and remove the stopping words to clean the text. Then, we count the number of words said by the listener (in the second utterance) in the whole dialogue system and separate them into 4 groups. The first two groups have positive emotion input. And for the first group, after talking with the listener, the speaker's emotion gets better (emotion score ascending). For the second group, after talking with the listener, the speaker's emotion gets worse (emotion score descending). The third group and fourth group get negative emotion input. As for these two groups, the emotion score of the speaker in the third group becomes higher while in the fourth group becomes lower. Fig.9 shows the top 30 frequency words in each group. As the top 30 frequency words are not exactly the same, we add 2 or 3 columns to make each two groups show the same words in the figure.

As we observed before, with the same emotion input, the listener's top 30 frequency words are similar, but the distributions are a little different. For the conversation starting with positive emotions, 'great', 'like', 'must', 'excite', 'kind', 'make', 'work', and 'happy' have a higher frequency in the top 30 frequency words in the successful conversations. With these words, The listener's talks are very joyful and excited. In the unsuccessful conversation, 'oh', 'like', 'really' and 'see' have a higher frequency. 'Oh really?' is a very frequent injection sentence in the dialogue system, which appears more in the ineffective conversation strategies. For the conversation starting with negative emotions, 'know', 'time', 'hope', and 'sure' have a higher frequency in the successful conversation, and 'happen', 'bad' have a higher possibility to appear in the unsuccessful conversation. From this result, we can know that when people are in bad

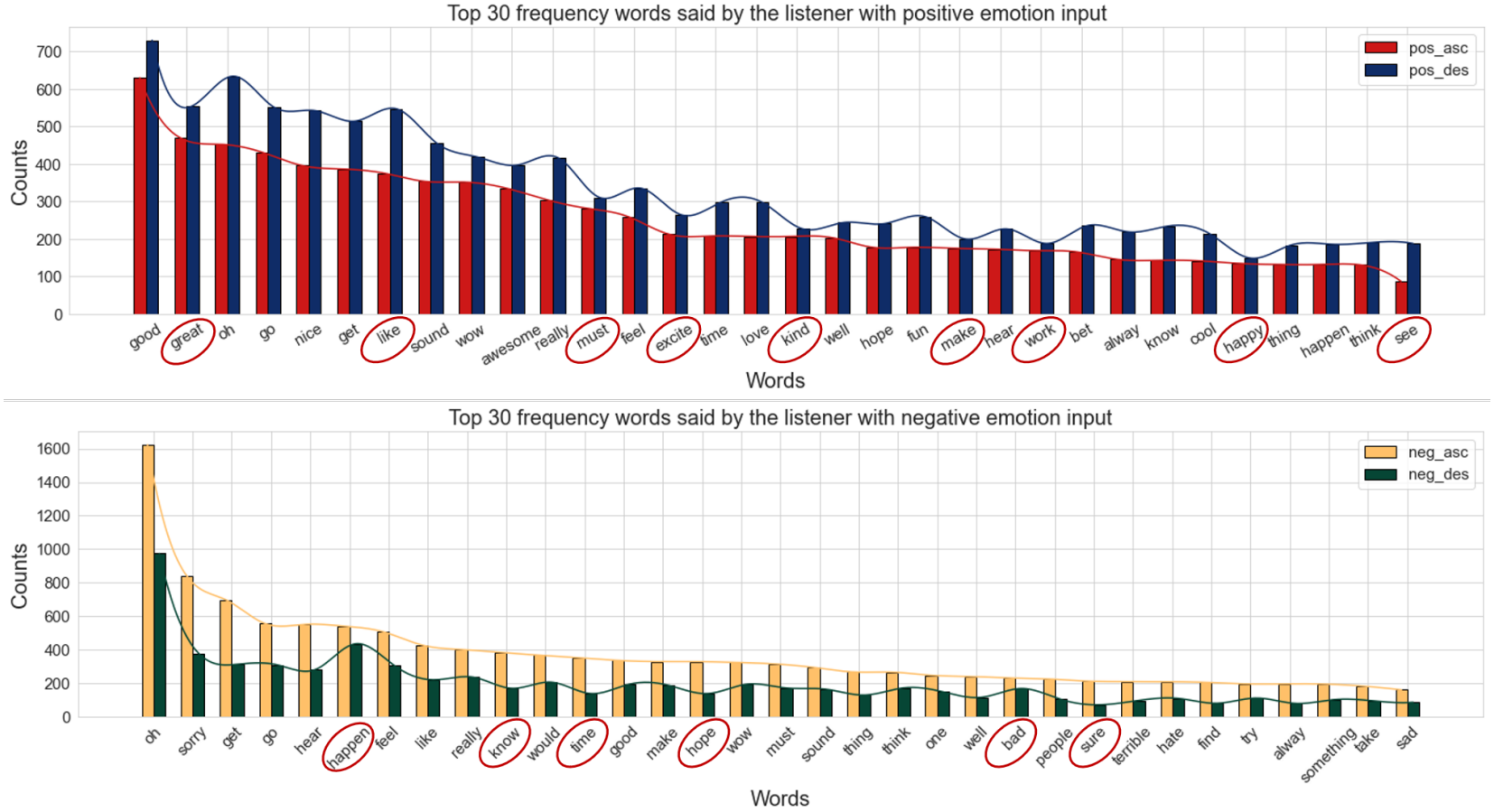

Figure 9: Top 30 frequency words said by listener. Top: the conversations have positive emotion input. After the conversation, the emotion score in one group is ascending, the other is descending. Bottom: the conversations have negative emotion input. After the conversation, the emotion score in one group is ascending, the other is descending.

situations, some words like 'know' and 'hope' that can transfer sympathy are more efficient than some description words, like 'happen' and 'bad' to express empathetic feelings.

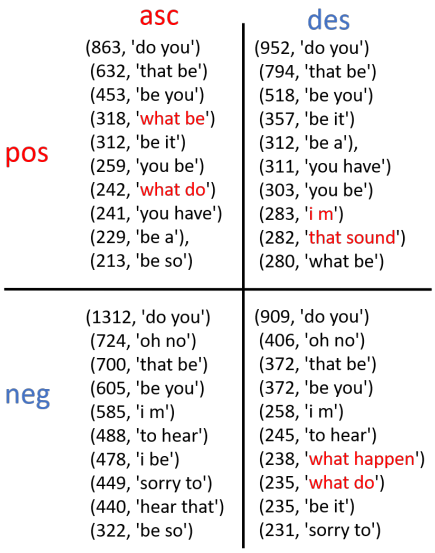

Figure 10: Top 10 frequency bigram said by the listener. Top: the conversations have positive emotion input. Bottom: the conversations have negative emotion input. Left: After the conversation, the emotion score of the speaker is ascending. Right: After the conversation, the emotion score of the speaker is descending.

We also count the bigram said by the listener in the dialogue without removing stopping words. Fig.10 shows the top 10 frequency bigram. In the successful conversation, 'what be' and 'what do' appear more. And in the unsuccessful conversation, 'you have', 'i m', and 'that sound' have a higher frequency. We can guess that if the speaker is in good spirit, the effective conversation strategy is to ask for some information and know more about the topics, which is more efficient than talking about the feelings ('i m') or commenting about the issue ('you have' or 'that sound'). However, with the negative emotion input, it's better to avoid asking the antecedent and requiring more details since 'what happen' and 'what do' as top frequency words only exist in unsuccessful conversations.

6 Limitations and Future Work

Due to the size limitation of the size of the ED dataset, the distributions of the emotion labels we gave are not similar. When we count the distribution of each label in different groups, the total counts of different labels have great differences. In this case, we can not analyze some emotion labels very well. 'consoling', 'encouraging', 'sympathizing', etc. emotions in the first utterance are very rare, so it's hard to find effective conversation strategies for these emotions. Due to a lack of data, the average score of each speaker and listener emotion pairs are not very accurate. Some cases have only several dialogue examples and some cases have near a hundred examples, which affect the quality of the heatmap Fig.6.

What's more, the classifier error can also cause analytical inaccuracy. We used the EmoBERT classifier to help us identify the emotion labels of each utterance and sentence, and the accuracy

of the classifier is 65.88%[22]. This can affect some computed scores in our projects, and we can also notice that in special examples.

We also tried to find some distribution differences of emotion labels, question types, question intents and question acts in successful conversations and unsuccessful conversations, but there is no obvious difference among them. The possible reason is that the total number of each label varies so much, and it's not easy to find some potential rules with these unbalanced tags.

Presented results can help improve the communication quality of social chatbots and the development of linguistics. For further analysis, we can use the correlation between different emotion labels and the grammatical sub-types. Since we know the grammatical types of questions, we can add this feature to train a neural model for achieving greater interpretability and controllability.

7 Conclusion

In this project, we preprocessed and annotated a dataset containing 25K empathetic dialogues with 4 grammatical types and their corresponding subtypes. We used the resulting dataset to verify the importance of language competence in the dialogue system. Through the Sankey diagram and heatmap, we find some potential linguistic strategies to produce more pleasant conversations. Further analysis of the grammatical types with other empathetic labels illustrates various question-asking strategies employed by the listener in response to the speaker's different emotional expressions. Finding the high-frequency words said by the listener gives us direct feedback on effective communication methods. We expect that our findings will expand the development of more controllable and personalized dialogue systems.

References

[1] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[2] C. Tao, L. Mou, D. Zhao, and R. Yan, "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] T. Lan, X.-L. Mao, W. Wei, X. Gao, and H. Huang, "Pone: A novel automatic evaluation metric for open-domain generative dialogue systems," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 1, pp. 1–37, 2020.

[4] S. Ghazarian, J. T.-Z. Wei, A. Galstyan, and N. Peng, "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings," *arXiv preprint arXiv:1904.10635*, 2019.

[5] S. Mehri and M. Eskenazi, "Usr: An unsupervised and reference free evaluation metric for dialog generation," *arXiv preprint arXiv:2005.00456*, 2020.

[6] Y.-T. Yeh, M. Eskenazi, and S. Mehri, "A comprehensive assessment of dialog evaluation metrics," *arXiv preprint arXiv:2106.03706*, 2021.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[8] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[9] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," *arXiv preprint arXiv:1708.07149*, 2017.

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[11] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[12] L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang, "Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems," *arXiv preprint arXiv:2010.03994*, 2020.

[13] C. Zong, F. Xia, W. Li, and R. Navigli, "Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[14] X. Gao, Y. Zhang, M. Galley, C. Brockett, and B. Dolan, "Dialogue response ranking training with large-scale human feedback data," *arXiv preprint arXiv:2009.06978*, 2020.

[15] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3619–3629.

[16] S. Ghazarian, R. Weischedel, A. Galstyan, and N. Peng, "Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7789–7796.

[17] S. Mehri and M. Eskenazi, "Unsupervised evaluation of interactive dialog with dialogpt," *arXiv preprint arXiv:2006.12719*, 2020.

[18] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[19] A. Agarwal, S. Maiya, and S. Aggarwal, "Evaluating empathetic chatbots in customer service settings," *arXiv preprint arXiv:2101.01334*, 2021.

[20] E. Rabinovich, M. Sultani, and S. Stevenson, "Codeswitch-reddit: Exploration of written multilingual discourse in online discussion forums," *arXiv preprint arXiv:1908.11841*, 2019.

[21] J. Decety, "The neurodevelopment of empathy in humans," *Developmental neuroscience*, vol. 32, no. 4, pp. 257–267, 2010.

[22] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," *arXiv preprint arXiv:2012.04080*, 2020.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] E. Svikhnushina, "A taxonomy of empathetic questions in social dialogs," 2022, unpublished thesis.

[25] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3619–3629. [Online]. Available: https://aclanthology.org/2020.acl-main.333

[26] C. Hori and T. Hori, "End-to-end conversation modeling track in dstc6," *arXiv preprint arXiv:1706.07440*, 2017.

[27] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded response generation task at dstc7," in *AAAI Dialog System Technology Challenges Workshop*, 2019.

[28] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," *arXiv preprint arXiv:1902.08654*, 2019.

[29] P. Gupta, S. Mehri, T. Zhao, A. Pavel, M. Eskenazi, and J. P. Bigham, "Investigating evaluation of open-domain dialogue systems with human generated multiple references," *arXiv preprint arXiv:1907.10568*, 2019.

[30] T. Zhao, D. Lala, and T. Kawahara, "Designing precise and robust dialogue response evaluators," *arXiv preprint arXiv:2004.04908*, 2020.

[31] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, and M. Geist, "Human annotated dialogues dataset for natural conversational agents," *Applied Sciences*, vol. 10, no. 3, p. 762, 2020.

[32] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu *et al.*, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[33] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[34] T. Stivers, "An overview of the question–response system in american english conversation," *Journal of Pragmatics*, vol. 42, no. 10, pp. 2772–2781, 2010.

[35] O. Jespersen, *Essentials of English grammar*. Routledge, 2013.

[36] R. Quirk, *A comprehensive grammar of the English language*. Pearson Education India, 2010.

[37] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman grammar of spoken and written English*. Longman London, 2000.

[38] R. Plutchik, "The classical psychoanalytic view of ego defenses," *Ego defenses: Theory and measurement*, no. 10, p. 13, 1995.

[39] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[40] A. B. Sai, A. K. Mohankumar, S. Arora, and M. M. Khapra, "Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 810–827, 2020.

[41] C. Gunasekara, S. Kim, L. F. D'Haro, A. Rastogi, Y.-N. Chen, M. Eric, B. Hedayatnia, K. Gopalakrishnan, Y. Liu, C.-W. Huang *et al.*, "Overview of the ninth dialog system technology challenge: Dstc9," *arXiv preprint arXiv:2011.06486*, 2020.

A Coding of emotion labels

Emotion label	score
Devastated	-3
Terrified	-3
Furious	-3
Afraid	-2
Disappointed	-2
Disgusted	-2
Lonely	-2
Sad	-2
Angry	-2
Guilty	-2
Apprehensive	-1
Anxious	-1
Embarrassed	-1
Annoyed	-1
Ashamed	-1
Sentimental	-1
Jealous	-1
Neutral	0
Encouraging	0
Agreeing	0
Suggesting	0
Acknowledging	0
Sympathizing	0
Wishing	0
Consoling	0
Questioning	0
Content	1
Prepared	1
Nostalgic	1
Faithful	1
Anticipating	1
Trusting	2
Surprised	2
Caring	2
Joyful	2
Hopeful	2
Impressed	2
Confident	2
Proud	2
Grateful	2
Excited	3

Table 8: Coding of emotion labels.

B Emotion change dialogues

Emotion change:	Devastated- Apprehensive - Sympathizing
Speaker:	My mother was just recently diagnosed with cancer.
Listener:	Oh god that is horrible! Look in too Bitter apricot and also Chlorella if she does radiation... both have research done on them to help a LOT with cancer.
Speaker:	I am just so sad and upset right now, but I appreciate your information.
Emotion change:	Afraid - Trusting - Hopeful
Speaker:	The way interest rates are these days, I'm so afraid to go get a car loan.
Listener:	'Financing can be scary but a lot of places now have really great rates for auto loans.
Speaker:	I intend to fix my dad's old Chevrolet covet pending when I am buoyant enough to pay for a new car.
Emotion change:	Disappointed - Guilty - Hopeful
Speaker:	It makes me sad that some people can't afford school clothes for their children.
Listener:	That makes me sad too. I wish there was more I could do to help out.
Speaker:	I donated supplies this year. Maybe next year I will be able to donate uniforms.
Emotion change:	Hopeful- Disappointed - Lonely
Speaker:	I wish that when people say they are going to do something or make a promise, that they would stick to it.
Listener:	I know exactly what you mean. I have been let down many times, and it's never a good feeling.
Speaker:	I had plans to go out with my friends tonight, but suddenly everyone is busy now, and I'm stuck home alone.
Emotion change:	Surprise- Prepared - Angry
Speaker:	The other day I was getting a glass out the cabinet and There sat a spider.
Listener:	Ew, I hate spiders. We are in the process of getting them out of our garage.
Speaker:	I hate the little things too! It was IN the glass at that!
Emotion change:	Confident- Terrified - Lonely
Speaker:	I have quite a long drive ahead of me next month. It's about 12 hours, but I've done it before so I'll be fine.
Listener:	I don't know how you do it. I can't stand long drives.
Speaker:	I put music on and open the window. Otherwise I'll disappear into my own little world, and that's not a good thing to do when you're driving!

Table 9: Dialogue examples with emotion changing. The emotion label in blue font has positive emotion, and the emotion label in red font has negative emotion

.