A Taxonomy of Empathetic Questions in Social Dialogs

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita and **Pearl Pu**
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{ekaterina.svikhnushina,iuliana.voinea,
kalpani.welivita,pearl.pu}@epfl.ch

Abstract

Effective question-asking is a crucial component of a successful conversational chatbot. It could help the bots manifest empathy and render the interaction more engaging by demonstrating attention to the speaker's emotions. However, current dialog generation approaches do not model this subtle emotion regulation technique due to the lack of a taxonomy of questions and their purpose in social chitchat. To address this gap, we have developed an empathetic question taxonomy (EQT), with special attention paid to questions' ability to capture communicative acts and their emotion-regulation intents. We further design a crowd-sourcing task to annotate a large subset of the EmpatheticDialogues dataset with the established labels. We use the crowd-annotated data to develop automatic labeling tools and produce labels for the whole dataset. Finally, we employ information visualization techniques to summarize co-occurrences of question acts and intents and their role in regulating interlocutor's emotion. These results reveal important question-asking strategies in social dialogs. The EQT classification scheme can facilitate computational analysis of questions in datasets. More importantly, it can inform future efforts in empathetic question generation using neural or hybrid methods.[1]

1 Introduction

Questions constitute a considerable part of casual conversations and play many important social functions (Huang et al., 2017; Enfield et al., 2010). Asking follow-up questions about the speaker's statement indicates responsiveness, attention, and care for the partner (Bregman, 2020; Huang et al., 2017). Listeners who manifest such an empathetic and curious attitude are more likely to establish the common ground for meaningful communication (McEvoy and Plant, 2014) and appear more likable to the speakers (Huang et al., 2017).

The vital role of questions in social interaction makes question-asking a desirable property for open-domain chatbots. These chatbots aim to engage in a natural conversation with the users while practicing active listening to deliver understanding and recognition of users' feelings (Rashkin et al., 2019). In fact, generating meaningful questions is so important that this has become one of the central objectives of such agents (Xiao et al., 2020).

However, asking questions effectively is challenging as not all questions can achieve a particular social goal, such as demonstrating attentiveness or empathy (Huang et al., 2017; Robinson and Heritage, 2006; Paukert et al., 2004). Given the task complexity, automatic conversational question generation is still gaining momentum, with only few results reported so far. See et al. (2019) suggested a way to control the number of questions produced by the model with conditional training. Wang et al. (2019) proposed a question-generation method to increase their semantic coherence with the answer, employing reinforcement learning followed by the adversarial training procedure. Wang et al. (2018) devised a model generating appropriate questions for a variety of topics by modeling the types of words used in a question (interrogatives, topic words, and ordinary words). These works presented approaches to produce contextually appropriate and diverse questions, but none of them considered the effect of questions on the interlocutor's emotional state. We attribute the deficiency in this research to the lack of resources allowing to analyze and model various question-asking strategies in affect-rich social exchanges.

To address this gap, we present a categorization and analysis of questions in social dialogs, with four main contributions. First, we develop an Empathetic Question Taxonomy, EQT, by manually annotating a subset of the EmpatheticDialogues (ED)

[1] Our code and the annotated dataset are publicly accessible at https://github.com/Sea94/EQT.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2952 - 2973
May 22-27, 2022 ©2022 Association for Computational Linguistics

dataset (Rashkin et al., 2019) (§4). EQT delineates the acts and intents of questions. Question acts capture semantic-driven communicative actions of questions, while question intents describe the emotional effect the question should have on the dialog partner. For example, a listener may *request information* (question act) about the age of speaker's daughter by asking "How old is she?" after learning about her success with the aim to *amplify speaker's pride* of his child (question intent). Second, we design and launch a crowd-sourcing annotation task to grow the original labeled seed subset tenfold (§5). Third, we devise an automatic classification model, QBERT, to generate labels for the rest of the ED dataset to demonstrate one important application of the taxonomy (§6). QBERT can facilitate the development of chatbots that offer engaging and empathetic conversations by raising meaningful questions. Finally, we inspect co-occurrences of acts and intents and their effect on the interlocutor's emotion using visualization techniques (§7). The analysis illustrates the most prominent question-asking strategies in human emotional dialogs. To conclude, we discuss the implications of these results for future question generation approaches.

2 Related Work

Previously proposed taxonomies of dialog acts frequently differ in types of assisted natural language tasks. The Dialog Act Markup in Several Layers (DAMSL) tag set was designed to enable computational modeling of conversational speech using statistical methods (Jurafsky et al., 1997; Core and Allen, 1997). It consists of 42 communicative acts derived from a Switchboard corpus. Eight of these labels describe different question types according to their semantic role, e.g., *Wh-question* or *Rhetorical-Question*. Several works proposed hierarchical taxonomies of dialog acts, targeted at modeling users' intents in human-machine conversations. Montenegro et al. (2019) introduced their annotation scheme for a symbolic dialog system intended to improve the lives of the elderly, while Yu and Yu (2021) designed a scheme for facilitating general human-machine chit-chat. In both works the logs of human-machine interactions were used for producing the taxonomies. Each of them features labels devoted to questions, characterizing them either by a question word, e.g., *How* or *What*, or the form of expected answer, e.g., *Open-ended* or *Yes/No question*. Finally, Welivita and Pu (2020) suggested a taxonomy of empathetic response intents in dialogs from the ED dataset with the purpose of improving controllability in neural dialog generation approaches. It further stated that *Questioning* is one of the most frequent intents of the empathetic listeners. However, none of these works focused on the fine-grained analysis of questions and their role in empathetic dialogs.

Meanwhile, several linguistic studies closely examined the pragmatics of questions and offered a number of classification schemes. Graesser et al. (1994) developed a scheme of 18 tags based on the information sought by the question. Their taxonomy applies well for transactional exchanges, but does not capture the social dimension. Freed (1994) studied the correspondence between the social function of questions and their syntactic form. She established 16 social question functions occurring in dyadic spoken conversations between friends. In another research effort, a group of linguists explored the range of social actions performed by questions across 10 languages (Enfield et al., 2010). The authors developed a coding scheme comprising 3 semantic question types and 7 social actions and applied it to questions in spontaneous spoken conversations (Stivers and Enfield, 2010). Finally, Huang et al. (2017) developed a taxonomy of 6 question types to describe questions occurring in their dataset of chat-based conversations between strangers instructed to get to know each other.

The described works provide an insightful basis for studying questions in social conversations. However, they do not consider the effect of questions on their addressee's emotional states, neither do they describe specific mechanisms to handle computational modeling. Moreover, most of them apply to spoken dialogs, impeding the extension of their results to chat-based exchanges due to the inherent differences in these modalities. Lastly, they relied mainly on manual annotation, yielding comparatively smaller datasets. In our study, we extended the derived taxonomy to a large corpus using crowd-sourcing and automatic methods and analyzed the emerging patterns on a large scale. We summarize the comparison of our question taxonomy with the existing schemes in Table 1.

3 Dataset

For taxonomy derivation, we sought a dataset that contains social dialogs with diverse emotional expressions and could be applicable to train a chat-

Taxonomy	# labels	social function	emotional function	dataset
(Graesser et al., 1994)	18	✗	✗	✗
(Freed, 1994)	16	✓	✗	✗
(Enfield et al., 2010)	7	✓	✗	✗
(Huang et al., 2017)	6	✓	✗	✗
EQT	21	✓	✓	✓

Table 1: Comparison of question taxonomies.

bot with advanced question-generating abilities. We avoided datasets featuring multi-modal dialogs (IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019)) as well as transcribed spoken conversations (Emotionlines (Hsu et al., 2018), Switchboard (Jurafsky et al., 1997)). Such dialogs contain back-channel communication and other sensory signals that are not present in chat-based conversations and, therefore, are not well-suited for the modeling task. Similarly, we rejected datasets that assist other tasks than social conversation modeling, such as SQuAD (Rajpurkar et al., 2016) (reading comprehension) or QoQA (Reddy et al., 2019) (information gathering). Finally, we did not consider datasets from social media as they can contain toxic and aggressive responses (Zhang et al., 2018).

We opted for the EmpatheticDialogues (ED) dataset (Rashkin et al., 2019), a benchmark dataset for empathetic dialog generation containing 24,850 conversations grounded in emotional contexts. Each dialog is initiated by a speaker describing a feeling or experience and continued by a listener who was instructed to respond empathetically. The dialogs are evenly distributed over the 32 emotional contexts, covering various speaker sentiments (e.g., *sad*, *joyful*, *proud*). We found the ED dataset to be a rich source of question-asking as over 60% of all dialogs contain a question in one of the listeners' turns, resulting in a total of 20K listener questions. Basic statistics of the dataset are given in Table 2.

Descriptor	Value
# dialogs in total	24,850
# turns per dialog on avg.	4.31
# dialogs with at least one question from listener	15,253 (61.4%)
# questions from listeners	20,201

Table 2: Statistics of the EmpatheticDialogues dataset.

4 Defining Empathetic Question Taxonomy

Given the community's interest in question-asking functionality for chatbots and its significance for empathetic response generation, we aimed at developing a taxonomy of listeners' questions asked in response to speakers' emotional inputs. For this purpose, being guided by prior literature review, we employed a qualitative coding method, which is an established approach for such tasks (Stivers and Enfield, 2010; Huang et al., 2017; Zeinert et al., 2021). Qualitative coding is a process of grouping and labeling similar types of data and iteratively validating the labels.

To cover a diverse range of speakers' emotions, we sampled several hundred dialogs uniformly from the 32 emotional contexts in the ED corpus. The sample size was chosen to balance the need for the diversity of questions with researchers' ability to consider each question carefully and was consistent with prior practice. The coding process was informed by previous question classification schemes (Table 1) and knowledge about general principles of emotional regulation (Gross, 2013). Iterative adjustments were applied resulting from discussions of the concrete data. Specifically, the first author made several iterations of coding trials to develop an initial set of labels. Throughout the process, a number of review sessions were held with the last author to merge the labels into more focused classes. As a result, we developed the Empathetic Question Taxonomy (EQT) with two distinguished branches: ***question acts*** describe semantic-driven features of questions (e.g., *ask for confirmation*, *positive rhetoric*), whereas ***question intents*** characterize their emotion-regulation functions targeted at the interlocutor's emotional state (e.g., *sympathize*, *amplify excitement*). As it will be revealed further (§7), an empathetic listener can use different question acts to deliver the same intent, justifying the proposed branching.

Overall, more than 310 questions were annotated. EQT consists of 9 labels for question acts and 12 labels for question intents. The granularity of the taxonomy was driven by earlier linguistic findings and empirical observations about the interplay of the labels in two branches. For example, question acts *request information* (Enfield et al., 2010), *ask about consequence* (Graesser et al., 1994), and *ask about antecedent* (Graesser et al., 1994) are related and could possibly be grouped. However, we decided to keep them separately as listeners use them with unequal frequencies in positive and negative emotional contexts and combine them with different question intents (§7). Similarly, the initial set of labels for question intents was created based on the variety of emotions present in the dataset. We further reduced it to a manageable size to make it more applicable for an annotation task, while still preserving sufficient expressiveness of labels to represent subtleties of the data (Zeinert et al., 2021). We present the labels with their definitions below and provide several examples in Figure 1. Examples for each act and intent label are given correspondingly in Tables 4 and 5 from Appendix A.

Question acts

Request information (38.7%): Ask for new factual information.
Ask about consequence (21.0%): Ask about the result of the described action or situation.
Ask about antecedent (17.1%): Ask about the reason or cause of the described state or event.
Suggest a solution (8.7%): Provide a specific solution to a problem in a form of a question.
Ask for confirmation (5.8%): Ask a question to confirm or verify the listener's understanding of something that has been described by the speaker.
Suggest a reason (5.2%): Suggest a specific reason or cause of the event or state described by the speaker in a form of a question.
Irony (1.3%): Ask a question that suggests the opposite of what the speaker may expect, usually to be humorous or pass judgement.
Negative rhetoric (1.3%): Ask a question to express a critical opinion or validate a speaker's negative point without expecting an answer.
Positive rhetoric (1.0%): Ask a question to make an encouraging statement or demonstrate agreement with the speaker about a positive point without expecting an answer.

Question intents

Express interest (57.1%): Express the willingness to learn or hear more about the subject brought up by the speaker; demonstrate curiosity.
Express concern (20.3%): Express anxiety or worry about the subject brought up by the speaker.
Offer relief (4.8%): Reassure the speaker who is anxious or distressed.
Sympathize (3.9%): Express feelings of pity and sorrow for the speaker's misfortune.
Support (2.6%): Offer approval, comfort, or encouragement to the speaker, demonstrate an interest in and concern for the speaker's success.
Amplify pride (2.6%): Reinforce the speaker's feeling of pride.
Amplify excitement (1.9%): Reinforce the speaker's feeling of excitement.
Amplify joy (1.6%): Reinforce the speaker's glad feeling such as pleasure, enjoyment, or happiness.
De-escalate (1.6%): Calm down the speaker who is agitated, angry, or temporarily out of control.
Pass judgement (1.6%): Express a (critical) opinion about the subject brought up by the speaker.
Motivate (1.0%): Encourage the speaker to move onward.
Moralize speaker (1.0%): Judge the speaker.

To validate the interpretability of the labels and efficacy of the instructions for the crowd-sourcing task, we invited two other members from our research group and asked them to annotate questions in 20 randomly selected dialogs, containing 25 questions. The annotators were instructed to consider the preceding dialog turns while assigning the labels as the same question might fall into different categories based on the context. For example, the question "What happened!?" can be classified as *Express interest* or *Express concern*, depending on the valence of the speaker's emotion. We computed both the Fleiss kappa (Fleiss, 1971) and the observed agreement among the first author and two annotators. The observed agreement was calculated as a percentage of questions with at least two agreed labels (Endriss and Fernández, 2013). We considered it as a reliable measure of inter-rater

– *My cat vomited on my shoes today* (Negative)

– ***Is your cat ill?*** (Suggest a reason, Sympathize) ***or does cat always do that?*** (Request info, Express concern)

– *no he just ate too much* (Neutral)

– *I got approved to adopt a dog!* (Positive)

– *Yay! I love dogs!* ***Do you have any you want to get specifically or are you just going to look until you find one that clicks?*** (Ask about consequence, Amplify excitement)

– *Oh I already picked one! I'll be picking her up this weekend.* (Positive)

Figure 1: Examples of dialogs grounded in negative (top) and positive (bottom) emotional contexts. Listeners' questions are shown in bold with the assigned *(act, intent)* labels given in parenthesis. The valence of speaker's emotions in each turn is also indicated.

agreement as the number of coding categories was large (9 for acts and 12 for intents), yielding relatively low chance agreement (11.1% and 8.3% respectively). The agreement resulted in 92% for acts ($\kappa = 0.52$) and 80% for intents ($\kappa = 0.31$), supporting the satisfactory interpretability of EQT.

5 Crowd-Sourced Annotation

For further analysis, we annotated a larger subsample of the ED dataset with the EQT labels by designing and launching a crowd-sourcing task on Amazon Mechanical Turk (Mturk). The design was refined based on three pilot studies: one internal and two Mturk-based. For the annotation, we sampled about 40% of dialogs from each of the original 32 emotional contexts. We only sampled the dialogs with at least one question in one of the listener's turns. The dialogs were then pre-processed so that each dialog ended with a question requiring a label. Further, we distributed the dialogs into individual human intelligent tasks (HITs) and launched them on Mturk in a sequence of batches. For each HIT we collected the annotations from three workers. The incentive for one HIT varied from $0.4 to $0.9 depending on the worker's performance and task configuration. We describe the details about the task design and the annotation procedure below; exhaustive explanations about dialog pre-processing and the task user interface are provided in Appendix B.

5.1 Task design

The interface consisted of four main components: instructions, terminology, terminology quiz, and the annotation task. The instructions informed the workers about the purposes of the task. Next, the terminology page outlined the description of the EQT, listing the definition of each label with examples. The terminology quiz contained six dialogs from the terminology page and invited the worker to select correct labels for questions in each dialog. Finally, the annotation task included 25 dialogs, each ending with a listener turn with one or multiple questions. Under each question, labels from two EQT branches were presented, and the worker had to select one most suitable label within each of the sets.[2] Twenty out of the 25 dialogs were treated as points for annotation, and the other 5 were bonus dialogs. For the bonus questions, we identified the gold labels during the manual annotation phase and used them to control workers' quality: a worker had to select the correct labels to score the points counting towards additional incentive ($0.2).

We required all workers who accepted one of our tasks for the first time to take the terminology quiz. Workers who assigned the correct labels to at least three questions could proceed to the annotation task and were granted bonus payment for passing the quiz ($0.1). The quiz was not required for the workers who had successfully passed it once.

5.2 Quality control

In addition to the terminology quiz, we used several mechanisms to control the annotation quality. First, following Mturk recommendations, we only allowed the workers with a 98% approval rate to access our tasks. Second, we rejected assignments whose completion time significantly deviated from the expected average. Further, we ran additional checks for the workers who accepted several of our assignments simultaneously. Lastly, we computed the inter-rater agreement for each batch and discarded the submissions that harmed the agreement.

5.3 Results

Overall, we launched 556 HITs and 465 of them were completed. The rejection rate after the quality control was 4.7%. Upon obtaining the results, we first computed the Fleiss kappa scores for acts ($\kappa = 0.34$) and for intents ($\kappa = 0.27$) to validate that the agreement between the workers is acceptable. Then, we identified the final labels using the majority vote: if at least two workers agreed on a label, we chose it as a final label. This resulted in an 83.6% observed agreement score for acts and 75.8% observed agreement for intents. The majority vote approach was shown to be able to filter noisy judgments of amateurs, producing the labeled set of comparable quality to the annotations of experts (Nowak and Rüger, 2010). As a final check, we computed the kappa agreement between the crowd-sourced labels and the first author annotations for the subset of 450 randomly sampled questions. The scores equaled 0.57 for acts (71.6% observed agreement) and 0.50 for intents (68.0% observed agreement), indicating moderate agreement, which we treat as satisfactory for this type of task. As a result, an act label was assigned to 6,433 questions and an intent label – to 5,826 questions, with an intersection of 4,962 questions.

[2] In our task design, we chose to ask for a single most suitable label to facilitate further data analysis, however allowing the selection of multiple applicable labels is also possible. We discuss this possibility further at the end of the paper (§8).

6 Automatic Labeling

To show how EQT can be operationalized, we demonstrate the use of the taxonomy for annotating the reminder of the ED dataset. We first formulate the question act and intent prediction problems and then build two classification models to address them. Before training, we augmented the labeled set using k-Nearest-Neighbors (k-NN) method. We also tried training the classifiers without data augmentation, but their performance was weaker (see Appendix D for details).

6.1 Data Augmentation

We employed the Sentence-BERT (SBERT) framework (Reimers and Gurevych, 2019) to obtain embeddings for all questions with their contexts. Then we used the cosine similarity measure to find k labeled NNs for each question in the unlabeled set and assign the same labels to them. For the first step, we computed the embeddings of each dialog turn using the *roberta-base-nli-stsb-mean-tokens* SBERT model and then combined them into a single embedding per question with the weighted average. We opted for weighed average instead of concatenation to keep manageable size of the embedding vector. We used a half-decaying weighting scheme, providing the highest weight to the final question to indicate its importance. The usage of this weighting scheme is guided by our previous experiments of similar nature, where we observed that the models with decaying weights performed better than the ones without them (Welivita et al., 2021). Next, we tested several approaches for identifying semantically similar dialogs to propagate the labels. One strategy was to take the same label as the top-1 NN, given that the similarity was higher than a predefined threshold. The other strategy was to use the label identified with the majority vote from the top-3 NNs. We did not experiment with higher values of k due to resource considerations. We ran several cross-validation experiments on the labeled set with grid search over various cosine-similarity thresholds. Top-3 majority vote strategy was shown to produce higher accuracy with a 0.825 cosine similarity threshold value resulting in the acceptable trade-off between the accuracy (∼76% for both label sets) and the number of labeled questions. Therefore, we applied this strategy for the whole dataset, which produced additional 1,911 labels for question acts and 1,886 labels for question intents. More details are provided in Appendix C.

6.2 Classifier Models

Using the human-annotated and augmented labels, we trained two classifiers, which we collectively call QBERT. QBERT models have identical architecture and vary only in the number of output categories in the final layer. Each model consists of a BERT-based representation network, an attention layer, one hidden layer, and a softmax layer. For the representation network, we used the architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters. We initialized it with the weights of RoBERTa language model pre-trained by Liu et al. (2019) and for training used the same hyper-parameters as the authors. As input, we fed a listener question and preceding dialog turns in the reverse order. To prioritize the question, the half-decaying weighting scheme as described above was applied to the token embeddings of each turn.

Before training, we took out a stratified random sample of 20% of the questions (1,500) as a test set. The test set contained respectively 1156 human- and 344 SBERT-annotated questions. We separately trained each model on 80% of the remaining datapoints (5,475 acts, 4,969 intents), keeping the rest as a validation set (1,369 acts, 1,243 intents). We trained each model for 15 epochs and for prediction retained the ones with the lowest validation loss (see Appendix D for details). The classifiers achieved 74.7% accuracy for intents and 79.1% accuracy for acts on the test set. Further breakdown accuracies for human- and SBERT-annotated test samples are given in Table 3. According to previous work, human-human agreement can be used as a proxy for human accuracy (Kumar, 2014; Somasundaran and Chodorow, 2014). Given the agreement in our Mturk experiment (∼75-85%), QBERT exhibited reasonable predictive accuracy and validated applicability and usefulness of EQT for language modeling tasks.

Label source	Question intents	Question acts
human	71.0%	77.1%
SBERT	86.9%	87.5%
both	74.7%	79.1%

Table 3: Accuracy of QBERT classifiers on different slices of test data based on the source of annotations (human, SBERT, or both).

7 Analysis of Questioning Strategies

In this section we present the analysis of questioning strategies adopted by the empathetic listeners. We base our examination on human-annotated questions instead of the whole ED dataset to avoid any potential noise which might have been introduced by automatic classification. Visualizations for the whole dataset are included in Appendix E. Here, by a *questioning strategy*, we imply a combination of act and intent labels assigned to each question. We first analyzed which labels from the two EQT branches form such strategies by plotting the co-occurrences of each pair (Figure 2). Larger circles represent more frequent strategies, while an empty cell indicates that people do not use the given act to deliver the corresponding intent. For example, to amplify partner's joy, one may request information for more details or ask about consequences of the event, but will unlikely raise a negative rhetorical question. Several strategies are much more frequent than others. Act *Request Information* and intent *Express interest* dominate in our dataset, occurring together for 39% of questions. They define the most general type of questions, which are probably easy to ask, providing a reason why listeners use them often. At the same time, dialogs in the ED dataset are relatively short, and it can be difficult for listeners to fully understand the ideas and feelings of speakers in a couple of turns. In this case, requesting information and expressing interest demonstrates listener's attentive curiosity about the situation. Once listeners feel more confident about the speakers' sentiments and contexts, they employ more specific question-asking strategies.

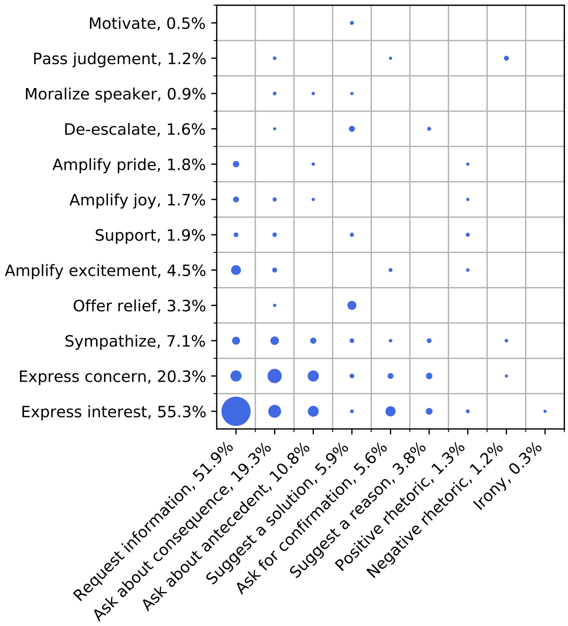


Figure 2: Joint distribution of question intents and acts for 5,272 human-labeled questions. Blue circles are proportional to the frequency of each pair's co-occurrence.

We further analyzed this phenomenon temporally across dialog turns (Figure 3). Primarily, we studied how listeners' questioning strategies affect speakers' emotions by visualizing the mappings between them. For this visualization, we used 41 emotion and intent labels describing each turn in the ED dataset produced by Welivita and Pu (2020). To avoid clutter, we mapped the original 41 labels to 3 coarser categories: positive, negative, and neutral using our best judgement (see Appendix E for details). Then, for the dialogs containing a question in the second turn, we plotted how speakers' emotions and listeners' questioning strategies shift over the first three turns. We computed the frequencies of all questioning strategies and, for the ones occurring in more than 0.5% of cases, we plotted the flow patterns. We restricted our analysis to the first three turns because over 70% of dialogs in the ED dataset have only four of them, excluding the possibility to study the influence of questioning strategies on further speakers' turns. In order to still get an intuition how listeners' question-asking behavior changes in the consecutive turns, we plotted the dynamics of the ratios of question act and intent labels across the dialog depth.

Figure 3a shows the flow rates between speakers' emotions and listeners' questioning strategies. As observed before, listeners most likely use follow-up questions to elicit more details about the situation by expressing interest and requesting information. In most of such cases, the speaker's emotion remains preserved in their consecutive utterance as the speaker elaborates on the first turn, maintaining the sentiment. When speakers explain themselves with sufficient clarity already in the first turn, listeners raise more precise questions, adapting the strategy to the affective context. If speakers share a positive experience, listeners try to amplify their emotions by requesting more information or asking about the consequences of the situation. On the contrary, when speakers disclose a negative sentiment, listeners try to validate and alleviate their feelings. They typically intend to express concern, sympathize, offer relief, or de-escalate the issue, and achieve it by asking about what preceded or fol-

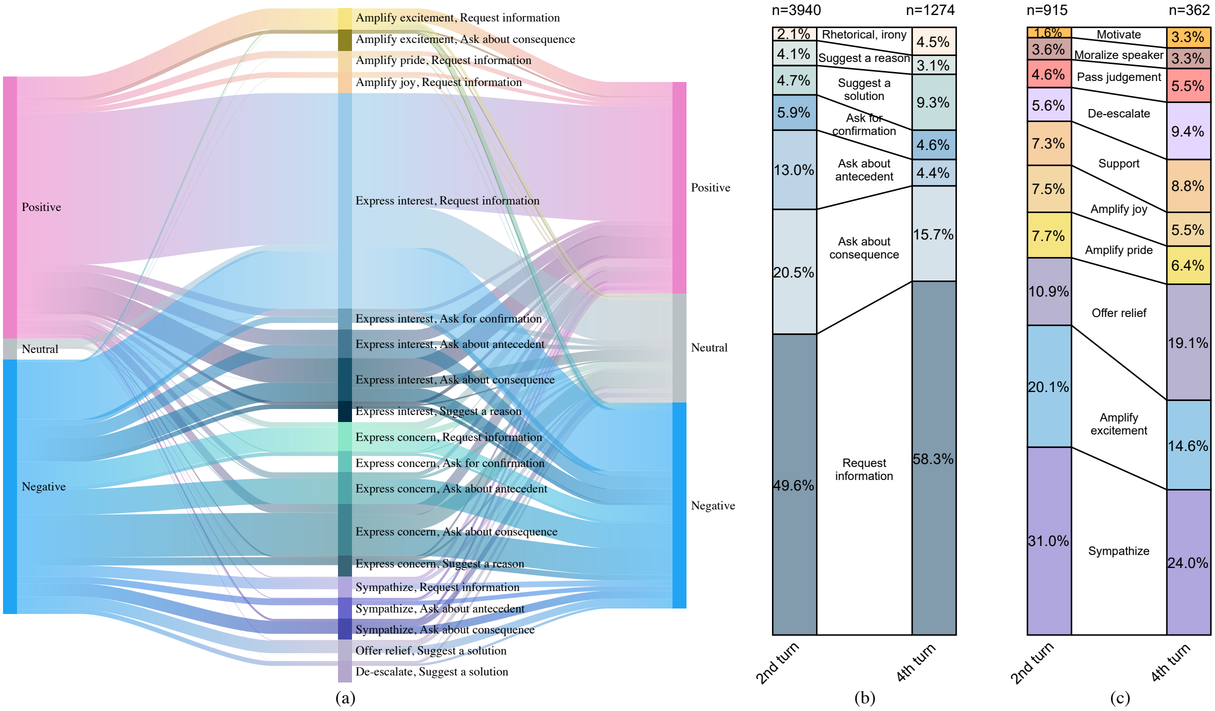


Figure 3: a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs (human-labeled ED subset). b) Frequency distribution of question acts across dialog turns (human-labeled ED subset). c) Frequency distribution of question intents across dialog turns (human-labeled ED subset). Two prevalent intents were excluded for visual clarity; their percentage rates computed for all questions (n=3940 and n=1274) are: *Express interest*: 54.3% $\rightarrow$ 57.9%, *Express concern*: 22.5% $\rightarrow$ 13.7%

lowed the situation and politely suggesting possible solutions or potential reasons for the issue. These specific strategies demonstrate their effectiveness as almost a half of negative speakers' emotions gets mitigated after the question intervention, while two thirds of positive emotions keep up in the following speaker's turn. The examples of dialogs showing how listeners use questions to treat both positive and negative speakers' sentiments are given in Figure 1. Additional examples are also available in Figure 9 of Appendix D.

Figures 3b and 3c demonstrate how ratios of different acts and intents evolve over two successive listeners' responses. Even though the horizon of four dialog turns might be too short to trace all the patterns, a few observations can be made. With increasing depth of the dialog, the overall number of questions decreases, while two types get more prominent: general questions (*Request Information*, *Express interest*) and questions aiming at suppressing speakers' negative emotions (e.g., *Suggest a solution*, *Offer relief*). It may indicate that listeners employ specific strategies to react to positive speakers' emotions immediately after their disclosure, but in case of negative contexts they tend to ask for extra clarifications in the first place and deliver targeted emotional treatment only in the next turn. As dialogs converge to more neutral exchanges, reducing the need to manage speakers' feelings, the ratio of questions demonstrating listeners' general curiously about the subject increases.

Finally, we reflected on the scarcely represented labels. Among acts, *Positive* and *Negative rhetoric* and *Irony* appear least frequently. These labels can be broadly classified as rhetorical questions. They typically serve for self-expression than conversational engagement and, therefore, are less common than other forms of questions (Huang et al., 2017). Moreover, negative rhetorical prompts may harm the conversation quality (Zhang et al., 2018), which could also explain why listeners avoided them in empathetic dialogs. The same reasoning applies to the two infrequent intents, *Pass judgement* and *Moralize speaker*. Another surprisingly rare intent is *Motivate*. We believe that motivation might be difficult to express in the form of a question. Moreover, for people who did not undergo special training, expressing motivation might be more challenging than other intents as it suggests a more thorough approach to solving one's problems.

8 Limitations and Future Work

Due to the nature of the ED dataset, some EQT labels are less represented than others. We kept them under consideration as we observed their distinctive role in managing speaker's emotions. Their further analysis is crucial for further identifying and designing effective questioning strategies for empathetic conversations, such as promoting motivational questions and avoiding judgmental ones. Eliciting additional samples for these categories could be possible by applying QBERT classifiers to other datasets capturing social dialogs.

Our taxonomy does not cover the phatic role of questions typically occurring during greetings, e.g., "What's up?" or "How's it going?" Such questions were very rare in the ED dataset. We chose not to analyze them, since these routine questions are the most superficial (Huang et al., 2017) and unlikely to serve any emotion-regulation function.

In the design of our annotation task, we opted for asking the crowd workers to choose a single most specific label from each of the two EQT branches. This was done with the aim of facilitating further analysis of questioning strategies withing the scope of this study. Nevertheless, according to Graesser et al. (1994), most adequate classification schemes in the social sciences allow assigning an observation to multiple rather than only one category. This also applies to our case. For example, for the question "Did you go through a breakup recently?" both *Suggest a reason* and *Request information* can be relevant. Future work can explore the possibilities of using multiple applicable labels in addition to the most specific one. Additional labels can be obtained either by tagging the samples manually or by taking top-N most confident predictions from the classifiers.

The results of this paper can facilitate the development of question-asking mechanisms of conversational chatbots. One can employ conditional training (See et al., 2019) to train an end-to-end neural model on a subset of most effective questioning strategies as defined by the co-occurrences of the EQT labels and their mappings with speakers' emotions (cf. Figure 3). To achieve even greater interpretability and controllability, researchers can devise architectures that dynamically model the selection of appropriate questioning strategy before generating a question. The strategy can be selected based on the conversational history and speaker's emotion and further passed into the question generation module. The main purpose of such modeling approaches is to lead an engaging empathetic conversation by raising meaningful questions, which deliver desirable effect on user's emotional state. Moreover, EQT along with QBERT models can be used to label questions originating from other corpora or chat logs and evaluate their effectiveness for regulating speaker's emotions, as described above.

9 Conclusion

In this paper we introduced EQT, an Empathetic Question Taxonomy depicting acts and intents of questions in social dialogs. We used crowdsourcing and automatic methods to tag all listeners' questions from the ED dataset with the EQT labels, which validated their interpretability and produced useful annotations for future research. Further analysis of the dataset with the visualization techniques shed light on various question-asking strategies employed by listeners in response to speakers' emotionally-ridden inputs. We identified several useful question-asking behaviors for favorable emotional regulation. We expect that our findings will enable the development of more controllable and effective question-generation models.

10 Ethical Considerations

In this work, we used Mturk platform to collect annotations for the dataset. Crowd workers on Mturk are known to be underpaid according to western standards, earning a median hourly wage of only ∼\$2/h (Kaufmann et al., 2011). At the same time, monetary remuneration is not the only factor defining people's motivation to work on such crowdsourcing platforms (Hara et al., 2018). For example, workers might also engage with HITs to learn new or train existing skills, pass free time, or meet new people. Taking these factors into account, we designed our annotation experiments so that workers received ∼\$6/h on average to achieve reasonable trade-off between the number of HITs we could launch with the available budget and the offered payment. While being slightly lower than the US minimum wage (\$7.25), it was deemed a fair compensation given that it is three times higher than the reported median wage and workers could have other reasons to complete the tasks than purely monetary reward. Nevertheless, we encourage future works of similar nature to offer higher compensation to the workers if possible.

References

Peter Bregman. 2020. Validation. In M. Goldsmith and S. Osman, editors, *Leadership in a Time of Crisis: The Way Forward in a Changed World*, 100 Coaches. RosettaBooks.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Ulle Endriss and Raquel Fernández. 2013. Collective annotation of linguistic resources: Basic principles and a formal model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 539–549, Sofia, Bulgaria. Association for Computational Linguistics.

N.J. Enfield, Tanya Stivers, and Stephen C. Levinson. 2010. Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42(10):2615–2619. Question-Response Sequences in Conversation across Ten Languages.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Alice F Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of Pragmatics*, 21(6):621–644.

Arthur C Graesser, Cathy L McMahen, and Brenda K Johnson. 1994. Question asking and answering. In Morton Ann Gernsbacher, editor, *Handbook of Psycholinguistics*. Academic Press.

James J Gross. 2013. *Handbook of emotion regulation*. Guilford publications.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. *A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. 2017. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3):430.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money: Worker motivation in crowdsourcing-a study on mechanical turk.

Ritesh Kumar. 2014. Developing politeness annotated corpus of Hindi blogs. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1275–1280, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

P. McEvoy and R. Plant. 2014. Dementia care: using empathic curiosity to establish the common ground that is necessary for meaningful communication. *Journal of Psychiatric and Mental Health Nursing*, 21(6):477–482.

César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A. Lozano, and María Inés Torres. 2019. A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction*, 3(3).

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, page 557–566, New York, NY, USA. Association for Computing Machinery.

Amber Paukert, Brian Stagner, and Kerry Hope. 2004. The assessment of active listening skills in helpline volunteers. *Stress, Trauma, and Crisis*, 7(1):61–76.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jeffrey D. Robinson and John Heritage. 2006. Physicians' opening questions and patients' satisfaction. *Patient Education and Counseling*, 60(3):279–285. EACH Conference 2004.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses ('use these words to write a sentence based on this picture'). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Baltimore, Maryland. Association for Computational Linguistics.

Tanya Stivers and N.J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626. Question-Response Sequences in Conversation across Ten Languages.

Weichao Wang, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Answer-guided and semantic coherent question generation in open-domain conversation. 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 5066–5076, Hong Kong, China. Association for Computational Linguistics.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2193–2203, Melbourne, Australia. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If i hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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)*, pages 3181–3197, Online. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Examples from Empathetic Question Taxonomy

Tables 4 (acts) and 5 (intents) present the two EQT branches with examples for each label. Examples are selected from the initial manually annotated subset. For each label we include its frequency for the three corresponding sets: manually-labeled, Mturk-labeled, and overall (both manually-, Mturk-, and automatically-labeled). The frequencies are approximately the same across each label, which validates that our annotation methods produced credible results. Examples of automatically assigned labels are given in Appendix D.

Question Act	Definition and Example
Request information 38.7%, 52.5%, 51.4%	Ask for new factual information. *- when i left my family to study in another city i got upset.* *- I'm sorry to hear that.* ***What are you studying?***
Ask about consequence 21.0%, 19.2%, 17.9%	Ask about the result of the action or situation described by the speaker. *- Our home was broken into* *- Oh no!* ***Did they steal a lot?***
Ask about antecedent 17.1%, 10.5%, 11.3%	Ask about the reason or cause of the event or state described by the speaker. *- Hi, I had a great vacation but something went wrong* *- Oh no, I'm sorry to hear that.* ***What happened?***
Suggest a solution 8.7%, 5.7%, 8.0%	Provide a specific solution to a problem in a form of a question. *- I lost my favorite jacket and I can't find it* - ***did you try redoing your steps of the last day?***
Ask for confirmation 5.8%, 5.6%, 5.2%	Ask a question to confirm or verify the listener's understanding about something that has been described by the speaker. *- I applied for a job last week.* - ***Oh did you?***
Suggest a reason 5.2%, 3.7%, 4.1%	Suggest a specific reason or cause of the event or state described by the speaker in a form of a question. *- i felt scared walking home alone the other day.* *- That's terrible!* ***Were you in a bad part of town or anything?***
Positive rhetoric 1.0%, 1.3%, 1.1%	Ask a question in order to make an encouraging statement or demonstrate agreement with the speaker about a positive point without expecting an answer. *- I couldn't pay for all my groceries and someone came up from the line behind and paid for the rest. I was so touched!* - ***Wow, how amazing is that!?***
Negative rhetoric 1.3%, 1.1%, 0.8%	Ask a question in order to express a critical opinion or validate a speaker's negative point without expecting an answer. *- I swear my friend is always using me* - ***that sucks is she really your friend then?***
Irony 1.3%, 0.3%, 0.2%	Ask a question using words that suggest the opposite of what the listener intends, usually to be humorous or pass judgement. *- I ate 10 Big Macs the other day.* *- oh my lord!* ***only ten?***

Table 4: Classification of question acts with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall.

Question Intent	Definition and Example
Express interest 57.1%, 55.2%, 60.2%	Express the willingness to learn or hear more about the subject brought up by the speaker; demonstrate curiosity. - *I just applied for a higher paying position within my company.* - ***That's cool, what is the position?***
Express concern 20.3%, 20.3%, 23.4%	Express anxiety or worry about the subject brought up by the speaker. - *I cry every time I think of my sister.* - *Why??* ***what happened to her!?***
Sympathize 3.9%, 7.3%, 5.1%	Express feelings of pity and sorrow for the speaker's misfortune. - *my girlfriend cheated on me* - *Oh no!* ***How did you find out?***
Offer relief 4.8%, 3.2%, 4.5%	Reassure the speaker who is anxious or distressed. - *They stopped making donuts at my favorite bakery.* - *Oh no!* ***Can you get donuts somewhere else?***
Amplify excitement 1.9%, 4.7%, 2.3%	Reinforce the speaker's feeling of excitement. - *lol. Going on vacation to Florida in a couple weeks!* - *Wow that's awesome!* ***To the beach?***
Support 2.6%, 1.8%, 1.0%	Offer approval, comfort or encouragement to the speaker, demonstrate interest in and concern for the speaker's success. - *I studied so hard for my test.* - ***I hope you did well?***
Amplify joy 1.6%, 1.7%, 0.9%	Reinforce the speaker's glad feeling such as pleasure, enjoyment, or happiness. - *I just received my certification to teach english as a second language!* - *Congrats!!!* ***Do you already have a job lined up?***
Amplify pride 2.6%, 1.7%, 0.7%	Reinforce the speaker's feeling of pride. - *My nephew caught a huge bass this weekend!* - ***That is cool, did you teach him how to fish?***
De-escalate 1.6%, 1.6%, 0.7%	Calm down the speaker who is agitated, angry or temporarily out of control. - *My neighbor threw their nasty trash all over their yard and won't clean it up! It's sooo gross!* - *Oh, that's disgusting!* ***Have you tried to talk to them about it?***
Moralize speaker 1%, 0.6%, 0.6%	Judge the speaker. - *I broke my TV remote and i blamed it on my kid* - *That's kinda terrible.* ***Did you apologize to him?***
Pass judgement 1.6%, 1.2%, 0.5%	Express an opinion (especially critical) about the subject brought up by the speaker. - *I hope the government can give some free course about the benefit of staying calm and healthy* - ***Government?*** *No way, it is interested in quite the opposite my friend.*
Motivate 1%, 0.5%, 0.2%	Encourage the speaker to move onward. - *This weekend is so boring so far* - *yeah? nothing interesting whatsoever?* ***why not make it exciting yourself?***

Table 5: Classification of question intents with corresponding definitions and examples. Under each label its frequency is given for the three corresponding sets: manually labeled, Mturk labeled, and overall.

B Details about Mturk Annotation Task

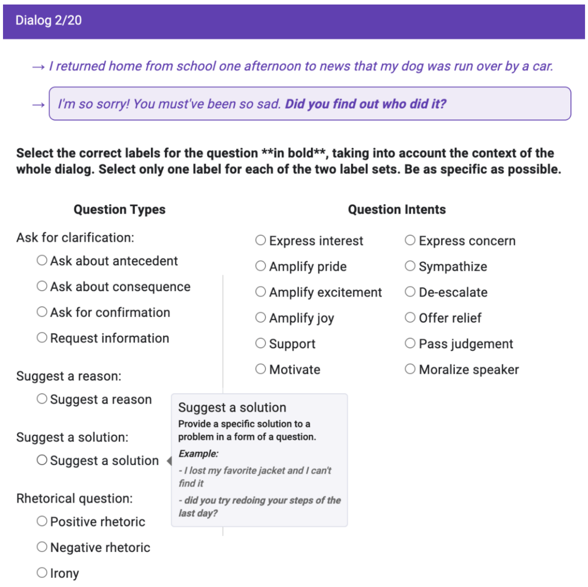
B.1 Dialog Pre-processing

Throughout our study, we only used those ED dialogs that contained questions in at least one listener turn. Since one dialog could contain several listener questions, for all downstream annotation tasks each such dialog was split into several separated dialogs, equal to the number of listener questions. The resulting sub-dialogs were truncated such that they would end with the particular question to which they corresponded to allow labeling every question in each dialog, without losing the previous conversational context. Figure 4 shows an example of a dialog from the original ED dataset and the resulting dialogs after the split.

In the Mturk interface, if the given listener turn contained multiple questions, we showed the resulting sub-dialogs in the same page one after another for contextual consistency. But if the original dialog contained listener questions in several turns, we showed the resulting dialogs in the two separate pages. Using the example from Figure 4, we would show the first resulting dialog in one page and the last two resulting dialogs together in another page.

B.2 Task User Interface

The user interface for the annotation task is illustrated in Figure 5.



Dialog 2/20

→ *I returned home from school one afternoon to news that my dog was run over by a car.*

→ *I'm so sorry! You must've been so sad.* ***Did you find out who did it?***

Select the correct labels for the question **in bold, taking into account the context of the whole dialog. Select only one label for each of the two label sets. Be as specific as possible.**

Question Types

Ask for clarification:
○ Ask about antecedent
○ Ask about consequence
○ Ask for confirmation
○ Request information

Suggest a reason:
○ Suggest a reason

Suggest a solution:
○ Suggest a solution

Rhetorical question:
○ Positive rhetoric
○ Negative rhetoric
○ Irony

Suggest a solution
Provide a specific solution to a problem in a form of a question.
Example:
- I lost my favorite jacket and I can't find it
- did you try redoing your steps of the last day?

Question Intents

○ Express interest
○ Amplify pride
○ Amplify excitement
○ Amplify joy
○ Support
○ Motivate

○ Express concern
○ Sympathize
○ De-escalate
○ Offer relief
○ Pass judgement
○ Moralize speaker

Figure 5: The user interface of the Mturk crowd-sourcing annotation task.

Original dialog

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch... Is it just for a day? Is your friend a safe driver?

Resulting dialogs

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch... Is it just for a day?

Speaker:	– You are never going to believe what I did!
Listener:	– What did you do?
Speaker:	– Well, I normally do not feel comfortable lending things to my friends, but recently I mustered up the trust to loan my friend my vehicle.
Listener:	– Ouch... Is it just for a day? Is your friend a safe driver?

Figure 4: Original and resulting dialogs after pre-processing.

C Details about Data Augmentation with Lexical Similarity

C.1 Setup and Results

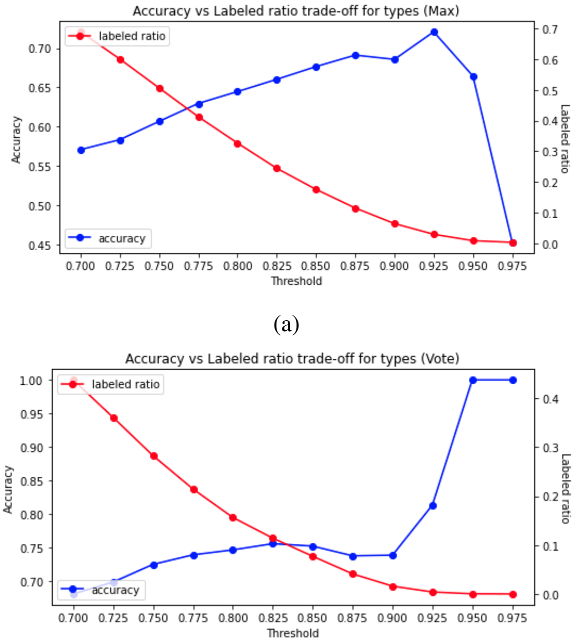
We used a half-decaying weighting scheme to encode questions with preceding context for the data augmentation process. The highest weight was always assigned to the final question to give it a higher preference. For example, if the dialog context consisted of three turns with embeddings e_1, e_2, e_3 and the fourth turn was a listener's question with embedding e_4^*, the final dialog embedding was $(8/15)e_4^* + (4/15)e_3 + (2/15)e_2 + (1/15)e_1$.

Figures 6 and 7 demonstrate the results of cross-validation runs for question acts and question intents for the Nearest-Neighbor label propagation approach. For each label set, we experimented with two similarity strategies: taking the same label as the top-1 most similar dialog according to the cosine similarity (*Max*, included in sub-figures 6a and 7a) and identifying the label with the majority vote from the top-3 most similar dialogs (*Vote*, included in sub-figures 6b and 7b). For each cross-validation launch we conducted a grid-search over cosine-similarity thresholds in a range between 0.7 and 1.

We also tried concatenating one-hot-encoded emotional context vectors with the dialog embeddings before running the cross-validation, but it did not result in any improvement in the accuracy and the resulting plots were almost identical to Figures 6 and 7, so we decided not to proceed with this approach.

C.2 Examples of Annotated Questions

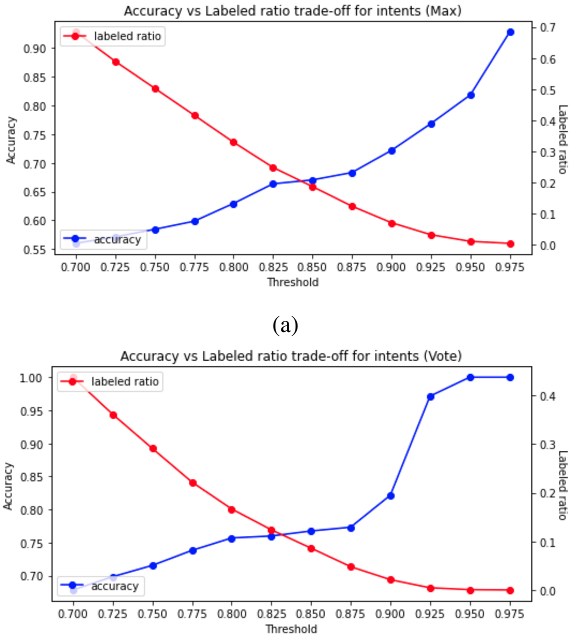
Table 6 presents several examples of propagated labels obtained using the outlined data augmentation process to give a better idea on the accuracy of this approach.



(a)

(b)

Figure 6: Cross-validation results for question acts for the two considered strategies: *Max* in sub-figure 6a and *Vote* in sub-figure 6b.



(a)

(b)

Figure 7: Cross-validation results for question intents for the two considered strategies: *Max* in sub-figure 6a and *Vote* in sub-figure 6b.

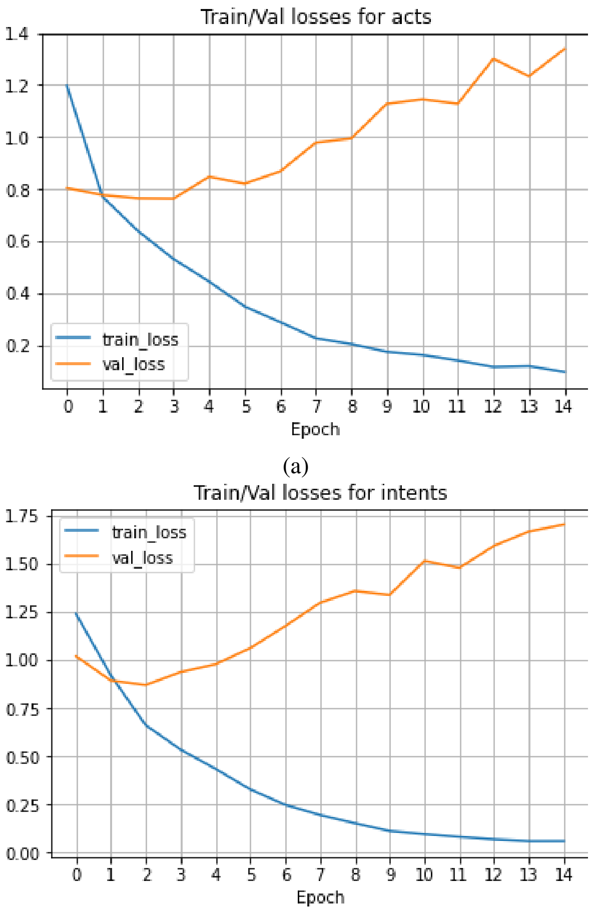
Head of Table 6

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– I get a good feeling when I think back to a birthday I had when I was a kid and all of my friends and I got to see a really funny movie at the mall. *– Awww!* ***What movie did you go to see?*** (Request information, Express interest)	*– I went to the movies by myself yesterday. I have no friends.* *–* ***what movie did you see?*** (0.87: Request information, Express interest)	*– I was happy when we were going to a new movie last weekend. I had waited all summer for it* *–* ***What movie was it?*** (0.87: Request information, Express interest)	*– I'm going to see a film tonight at the cinema.* *– oh really?* ***what movie?*** (0.86: Request information, Express interest)
– It really sucked, since a month ago I was dating this girl and she dumped me so early on. *– I'm so sorry.* ***Are you okay?*** (Request information, Express concern)	*– I hurt me when my parents got divorced. I never thought that would happen* *–* ***I'm so sorry, are you okay?*** (0.92: Request information, Express concern)	*– I am really feeling bad* *– I'm so sorry!* ***Is everything ok?*** (0.90: Request information, Express concern)	*– I just found out that my girlfriend has been cheating on me. God this is the worst week of my life.* *– I feel really sorry for you.* ***Will you be okay?*** (0.84: Request information, Express concern)
– One time my mom bought an ice cream from Mcdonalds! *–* ***Really?*** (Ask for confirmation, Express interest)	*– I saw someone putting mayo on their ice cream.* *–* ***Really?*** (0.92: Ask for confirmation, Express interest)	*– I accidentally ate someone else's cake at work* *–* ***Really?*** (0.91: Ask for confirmation, Express interest)	*– I just ate 5 donuts by myself* *–* ***Really?*** (0.86: Negative rhetoric, Express interest)
– i was scared walking home last night *–* ***Why was you scared was it too dark?*** (Suggest a reason, Express concern)	*– I used to be so scared to go to sleep as a kid.* *– How come?* ***Were you scared of the dark?*** (0.92: Suggest a reason, Express concern)	*– I stay away from the dark.* *– Why do you do that?* ***Are you scared of the dark?*** (0.86: Suggest a reason, Sympathize)	*– i was scared walking home the other day* *–* ***Why were you scared?*** (0.83: Ask about antecedent, Express concern)
– I one time lost my trunks in the pool! People saw me in a way I didn't want! *– Oh no! That must have been super embarrassing!* ***How did you react to that?*** (Ask about consequence, Sympathize)	*– a girl i like at school told me today she doesn't like me in front of everyone* *– Oh no! That must have been really embarrassing!* ***How did you respond?*** (0.85: Ask about consequence, Sympathize)	*– I fell down on stage while dancing, I felt so bad.* *–* ***oh dear, that must've been embarrassing, are you okay though?*** (0.84: Ask about consequence, Sympathize)	*– Once at a swimming competition, I had a wardrobe malfunction in front of a lot of people* *– Oh my goodness, that must have been humiliating.* ***What did you do?*** (0.83: Ask about consequence, Sympathize)
– My neighbor died in a car crash. *– Oh my. I'm so sorry to hear that.* ***What happened?*** (Ask about antecedent, Sympathize)	*– My nephew died yesterday.* *– I am so sorry to hear that.* ***What happened?*** (0.89: Ask about antecedent, Sympathize)	*– My pet ferret Fuzzy died the other day. I was so heartbroken.* *– I'm so sorry to hear that.* ***What happened?*** (0.88: Request information, Sympathize)	*– When my pet died I felt liek I lost my family member, My best friend.* *– Im sorry to hear that.* ***What happened?*** (0.88: Ask about antecedent, Sympathize)
– My brother just turned 16 and he's about to get his first car! I'm so excited for him. *– Whoa that's exciting!* ***What kind of car we looking at?*** (Request information, Amplify excitement)	*– I can't wait! We just bought a car today! Going to pick it up soon!* *– Oh nice! That is exciting!* ***What kind of car did you get?*** (0.89: Request information, Amplify excitement)	*– I just bought a brand new car* *– How exciting!* ***What kind of car is it?*** (0.86: Request information, Amplify excitement)	*– I was surprised when my dad got me my first car. I was not expecting it* *– That must have been exciting for you.* ***What car was it?*** (0.85: Request information, Amplify excitement)

Continuation of Table 6

Annotated question	Top-1 NN	Top-2 NN	Top-3 NN
– I spent hours reviewing notes and course content to prepare myself for a few trials that a company wanted me to go through. *– Good job!* ***Do you feel pretty prepared?*** (Request information, Support)	*– I have an important job interview this week* ***– Have you prepared well for it?*** (0.85: Request information, Express interest)	*– I have been studying for my final math exam all week long.* *– I hope you do well on it!* ***Do you feel prepared?*** (0.83: Ask for confirmation, Support)	*– Ive got a big interview on Friday. It for a job I really want.* *– I hope it goes well!* ***are you prepared?*** (0.83: Request information, Support)
– Friends threw me a surprise party yesterday. *– thats awesome, and happy birthday !!!* *– Thanks! I got so many cool gifts! I was so happy.* ***– what kind of gifts did you get?*** (Ask about consequence, Amplify excitement)	*– I was happy to find that at work my coworker prepared a birthday party for me. I was not expecting it.* *– Wow. I bet that was a nice surprise.* ***Did you get a lot of presents?*** (0.84: Ask about consequence, Amplify excitement)	*– My friends threw me a surprise birthday party last year!* *– That is very nice* *– It was! I was shocked and I felt very loved.* ***– Did they brought any special gift?*** (0.84: Request information, Express interest)	*– My friends planned a surprise party for my birthday.* *– Exciting!* ***Did you get any neat gifts?*** (0.84: Ask about consequence, Amplify excitement)
– I'm living my best life. I could'not be any happier. *– good to know.* ***and what makes your life so good, huh?*** (Request information, Amplify joy)	*– I am so happy with my life right now.* *– You sound very content.* ***What makes you happy?*** (0.86: Request information, Express interest)	*– I feel good. Everything finally seems to be working out.* *– That's great!* ***What are some things you're enjoying about life right now?*** (0.86: Request information, Amplify joy)	*– I've been happy with the way things have been going in my life lately.* *–* ***That's awesome, glad to hear, what are you most happy with?*** (0.86: Ask about antecedent, Amplify joy)
– I was happy when my brother finished school. I was proud of him *– That is awesome.* ***Was it high school or college?*** (Request information, Amplify pride)	*– It felt great to see my son graduate. Like I succeeded as a parent.* *– That's awesome.* ***high school?*** (0.88: Request information, Amplify pride)	*– I use to be the number one tennis player in the state.* *– That is an awesome achievement!* ***Was it for high school or college?*** (0.86: Request information, Amplify pride)	*– I'm a Phd student and I'm taking a really hard class. I have to do well so I was really happy when I got an A on a test!* *– thats awesome!* ***what college you go to?*** (0.84: Request information, Express interest)
– I cheated at cards. *–* ***Did you feel bad about it?*** (Ask about consequence, Moralize speaker)	*– I cut someone off in traffic today* *–* ***Do you feel bad about it?*** (0.85: Ask about consequence, Moralize speaker)	*– Yesterday, i had a night out with my friends, but i lied to partner that i will be staying late for work. I did not want to see her nagging* *– That's really not good.* ***Did you feel bad about it?*** (0.85: Negative rhetoric, Moralize speaker)	*– I was really hungry today and ate my roomates' leftovers.* *–* ***Do you feel bad about it?*** (0.85: Ask about consequence, Moralize speaker)
– I stole money from my friend. *– oh..* ***why did you do that?*** (Ask about antecedent, Pass judgement)	*– I stole money from my son's piggy bank.* *–* ***Why did you do that?*** (0.94: Ask about antecedent, Pass judgement)	*– I stole money from someone at a party years ago and I still feel bad about it.* *–* ***Why did you do that?*** (0.91: Ask about antecedent, Pass judgement)	*– I told my best friends secret to another one of our friends.* *–* ***Why did you do it?*** (0.89: Ask about antecedent, Pass judgement)

Table 6: Examples of propagated labels obtained using majority vote from the top-3 Nearest-Neighbor (NN) dialogs according to cosine similarity. The first column includes the newly annotated question, and the other three show the top-3 NN dialogs with respective question labels and a similarity value. Spelling and punctuation of the original source have been preserved.



(a)

(b)

Figure 8: Train and validation losses over the course of approximately 15 training epochs for question acts (sub-figure 8a) and question intents (sub-figure 8b).

D Details about training automatic classifiers

For our automatic classifiers, we used GELU as a hidden activation function and applied a 0.1 dropout to all layers and attention weights. For training, we used Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-6}$, and the peak learning rate of 2×10^{-5}. The maximum number of input tokens was set to 100, and we used the batch size of 50. The evolution of train and validation losses over the course of 15 training epochs is shown in Figure 8. We used Google Colab environment for the training.

The performance of classifiers trained only on a human-annotated subset was several percent lower than training on augmented data (see Section 6.2), resulting in 75% accuracy for acts and 70% for intents on the same (human-annotated) test set. Therefore, in this paper, we focus on the results obtained with the augmented data.

Figure 9 demonstrates several examples of automatically labeled questions in the ED dialogs. We specify both the predicted act and intent labels for each listeners' question and emotions expressed by speakers in each turn to observe how they are influenced by listeners' questions. Here we combine the pre-processed dialogs (cf. Section B.1) back to their original format, which explains why some labeled questions appear in the middle of the dialogs.

E Extended Analysis of Questioning Strategies

E.1 Mapping of Emotions and Empathetic Intents

Table 7 presents the mapping of 32 emotions (Rashkin et al., 2019) and 9 empathetic intents (Welivita and Pu, 2020) to three coarser emotion categories of different valence, which we used to produce visualizations for the analysis.

E.2 Additional plots for Human-Labeled Subset

Figures 10 and 11 show the breakdown of flow rates between speakers' emotions and listeners' questioning strategies (Figure 3) into separate mappings for acts and for intents, respectively.

E.3 Analysis of Questioning Strategies on the whole Dataset

For completeness, we include the same analytical visualizations as presented in Section 7 for the whole ED dataset (Figures 12, 13, 14, and 15). From these Figures, one can observe higher presence of more "general" categories (*Request information*, *Express interest*), which presumably originates from the fact that QBERT classifiers are slightly biased towards these classes due to the class imbalance in the training data.[3] Nevertheless, despite this remark, other major patterns revealed by the analysis of human-annotated subset (cf. Section 7), preserve in the Figures produced for the whole ED dataset (including automatically-annotated questions).

[3] One possible way to overcome the class imbalance issue in future work is to use the weighted loss function for training.

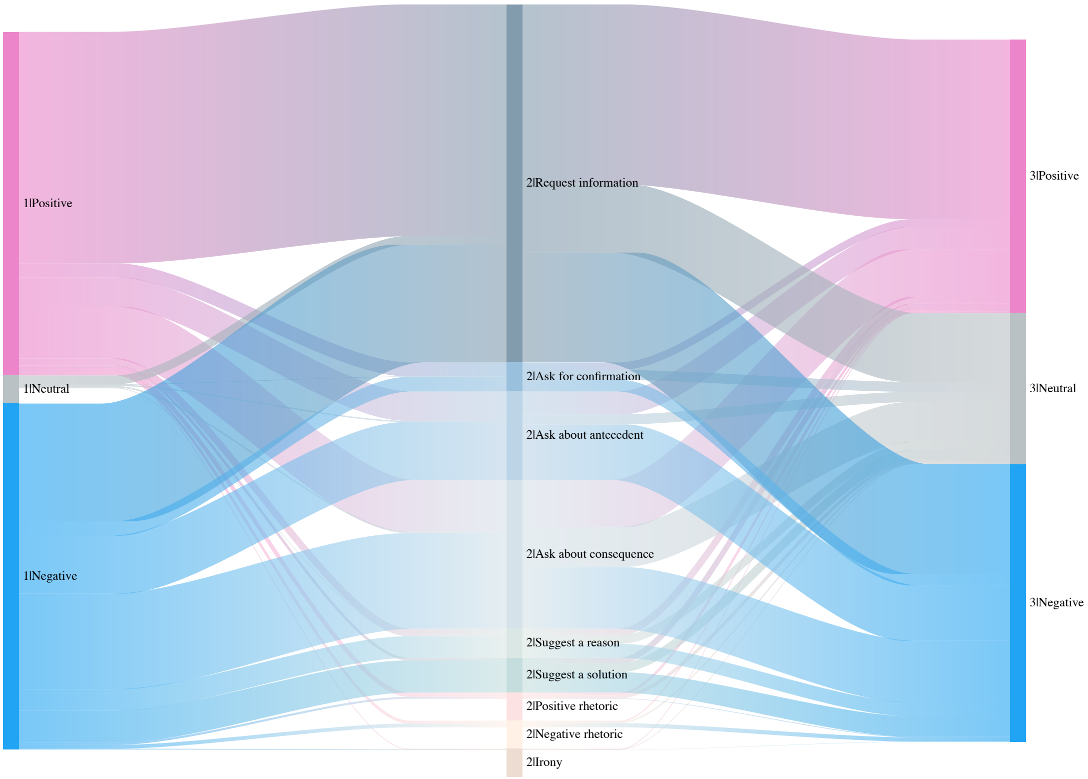


Figure 10: Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (human-labeled ED subset).

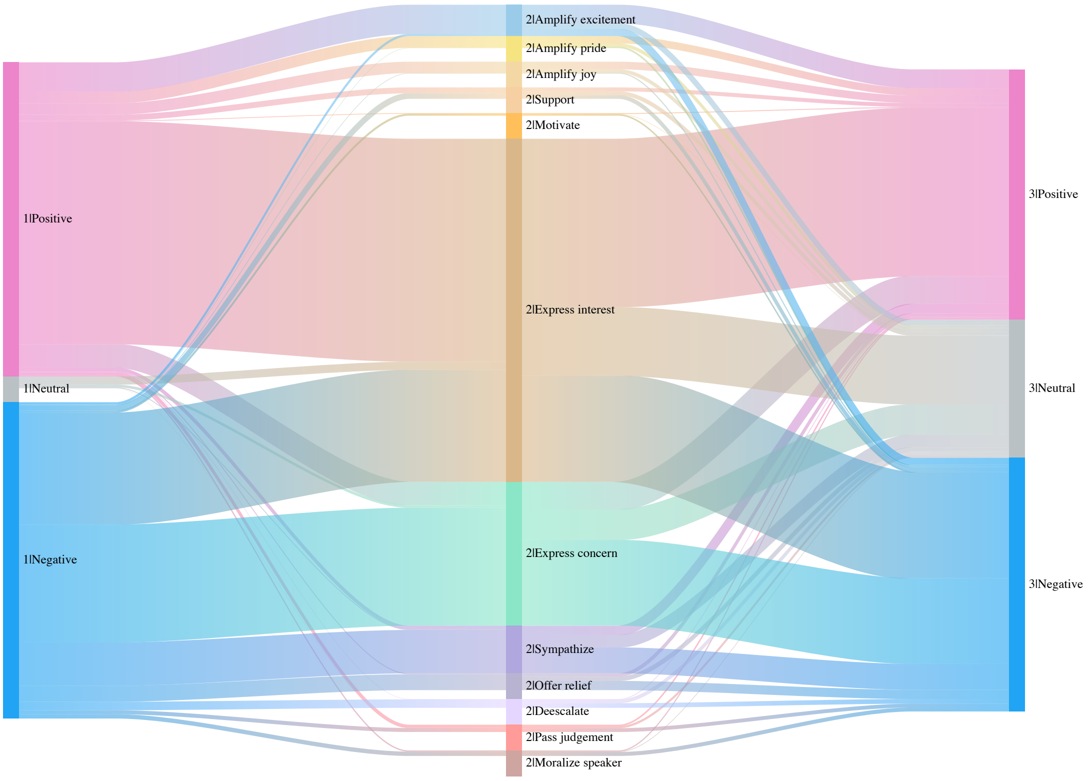


Figure 11: Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (human-labeled ED subset).

– *I am proud of my girlfriend for getting a full time job, I am sure she will do great!* (Positive)

– *That's awesome i bet she will too!* ***when does she start?*** (Request information, Express interest)

– *She starts in exactly a week* (Positive)

– ***woo hoo so you guys going out to celebrate?*** (Ask about consequence, Amplify excitement)

– *I am so happy to be having a boy* (Positive)

– *That's great! Congratulations!* ***Is this your first child?*** (Request information, Amplify joy)

– *Thanks. Yes it is. I already got a crib and baby bath.* (Neutral)

– *My daughter scored the winning goal at her last soccer game. I was so happy that all her hard work paid off!* (Positive)

– *That's great.* ***Does she practice a lot?*** (Request information*, Amplify pride)

– *Yes, she practices almost every day after school with her friends and also with her team. She says she will be a professional player one day!* (Positive)

– *Man.....my cat died:(I feel horrible.* (Negative)

– ***That's awful, how did your cat die?*** (Ask about antecedent, Sympathize)

– *Old age. she had a good life but it's still tearing me up.* (Neutral)

– *I took a test last week that I had studied very hard for. I know I got most of the answers right, but I got a failing grade* (Negative)

– *Must've been a really difficult exam.* ***Will there be other exams to balance it out?*** (Ask about consequence*, Offer relief*)

– *The person sitting next to me copied my answers, so the teacher failed both of us.* (Negative)

– ***I guess the teacher wasn't going to listen to you?*** (Suggest a reason, De-escalate) *That sucks.*

– *I ordered a gift for a friend and it says it was delivered but I never received it. Now the company says it takes 14 days for a refund.* (Negative)

– ***Don't you hate how "customer service" has no service anymore?*** (Negative rhetoric, Sympathize) ***Did you get the refund at least?*** (Suggest a solution, Offer relief)

– *Still waiting..... That's the most upsetting. Because they waste no time taking your money* (Negative)

– *I didn't realize that stealing was bad until I realized how it made me feel afterwards* (Negative)

– *So you probably felt pretty guilty huh.* ***Did you return what you stole?*** (Ask about consequence, Moralize speaker)

– *No, I was scared to get charged, but I stopped after that* (Neurtal)

Figure 9: Examples of questions labeled automatically with QBERT. Question acts and intents marked with a star* were annotated by Mturk workers.

Category	Mapped emotions and intents
Positive:	trusting, surprised, caring, content, joyful, excited, anticipating, hopeful, prepared, nostalgic, impressed, faithful, confident, proud, grateful
Neutral:	neutral, encouraging, agreeing, suggesting, acknowledging, sympathizing, wishing, consoling, questioning
Negative:	devastated, afraid, apprehensive, terrified, disappointed, disgusted, lonely, anxious, sad, embarrassed, annoyed, furious, ashamed, angry, sentimental, guilty, jealous

Table 7: Mapping of 32 emotions and 9 empathetic intents describing the EmpatheticDialogues dataset to three emotion categories of different valence.

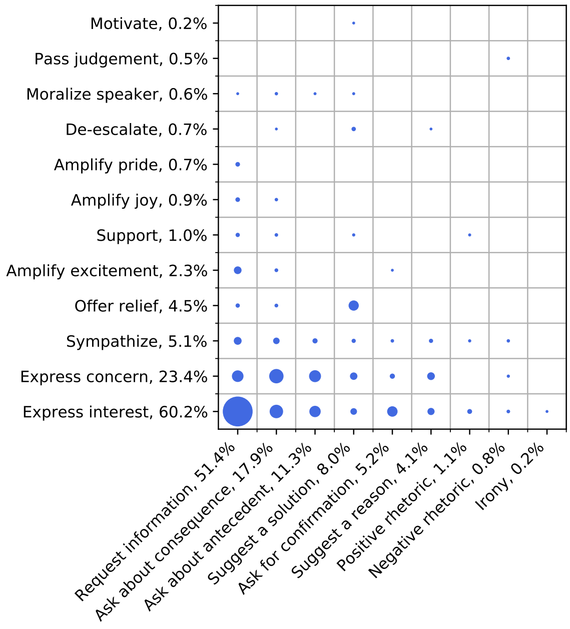


Figure 12: Joint distribution of question intents and acts for 20,201 labeled questions (whole ED dataset). Blue circles are proportional to the frequency of each pair's co-occurrence.

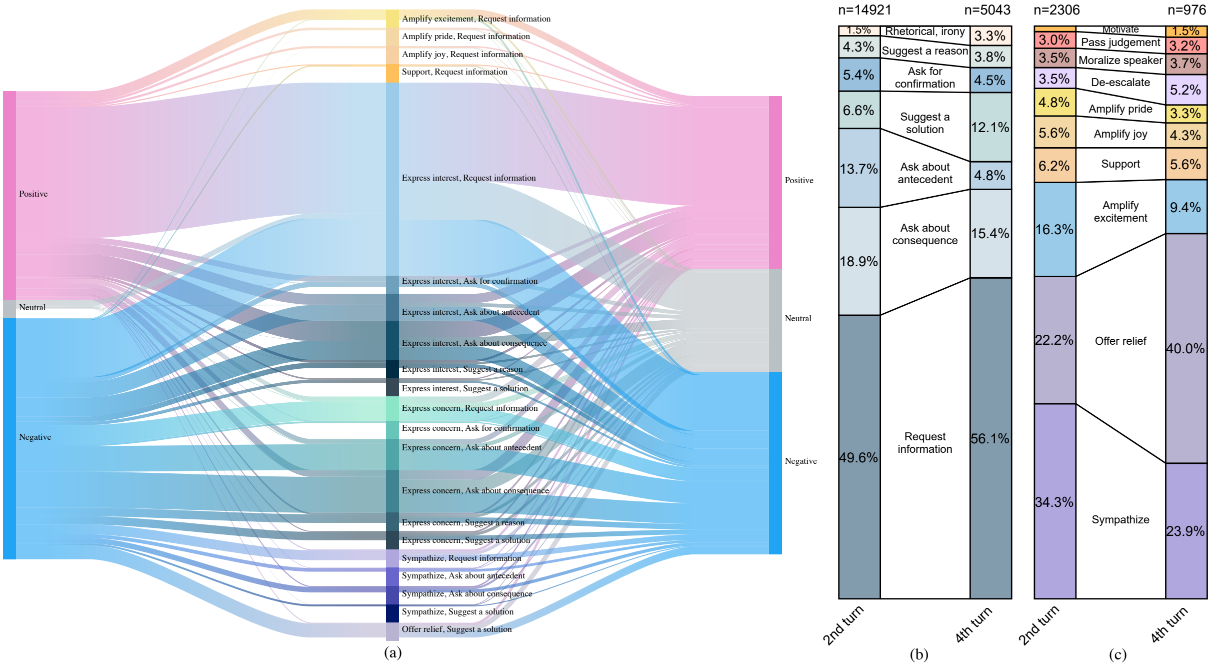


Figure 13: a) Mappings between emotions disclosed by the speakers and listeners' questioning strategies in the first three turns of the ED dialogs (whole ED dataset). b) Frequency distribution of question acts across dialog turns (whole ED dataset). c) Frequency distribution of question intents across dialog turns. Two prevalent intents were excluded for visual clarity; their percentage rates computed for all questions (n=14921 and n=5043) are: *Express interest*: 59.7% → 61.1%, *Express concern*: 24.9% → 19.3%

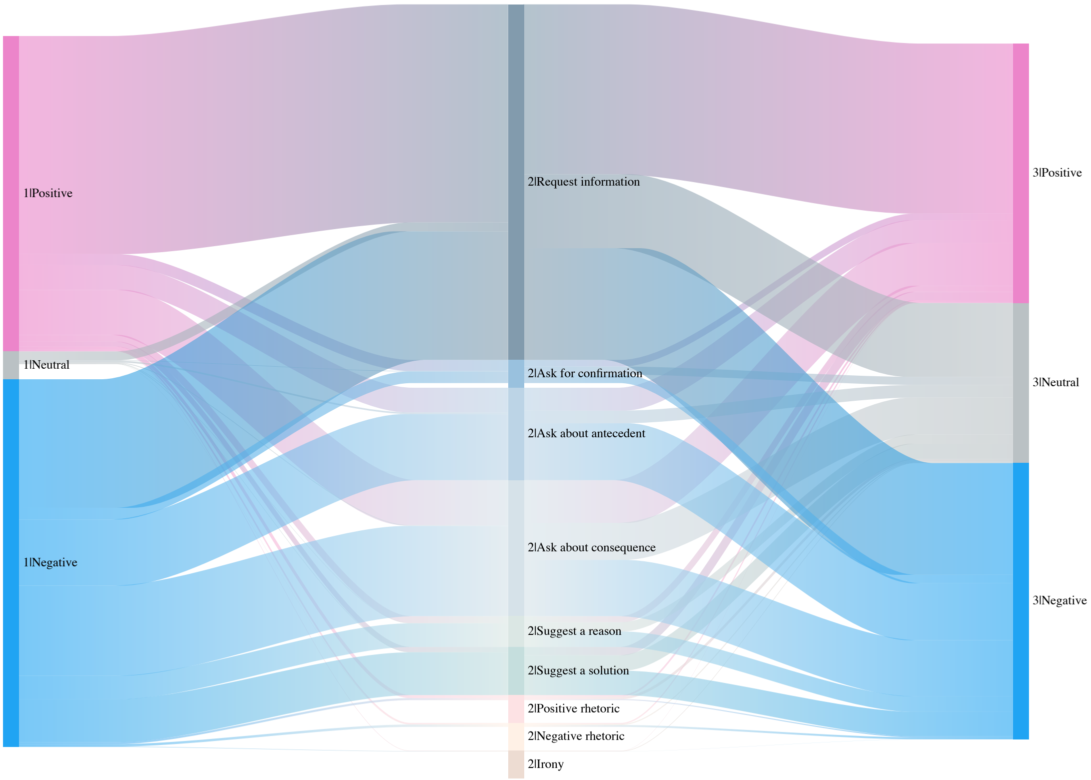


Figure 14: Mappings between emotions disclosed by the speakers and question acts used by listeners in the first three turns of the ED dialogs (whole ED dataset).

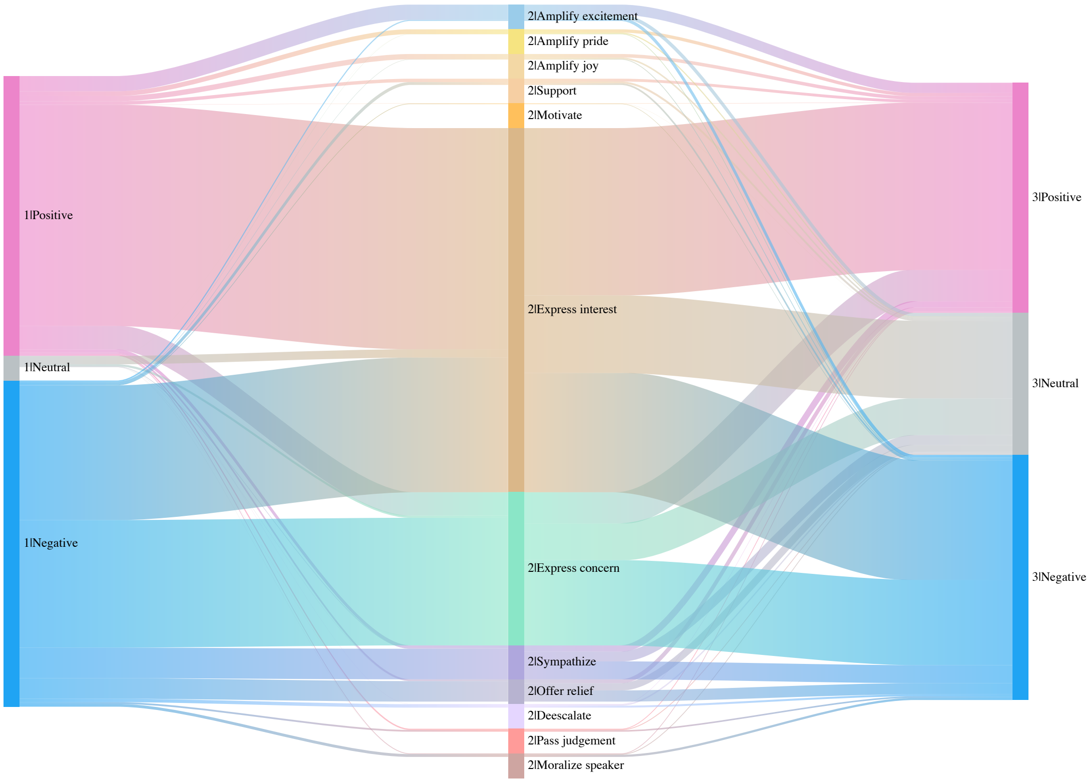


Figure 15: Mappings between emotions disclosed by the speakers and question intents used by listeners in the first three turns of the ED dialogs (whole ED dataset).