Question Types and Intents in Human Dialogues

Iuliana Voinea
iuliana.voinea@epfl.ch
(Sciper 308108)

***Abstract*—This project proposes to preprocess the existing EmpatheticDialogues dataset, consisting of 25K conversations, into a new version that comprises dialogues which contain empathetic questions in the final listener's turn. The dialogues were then annotated according to the empathetic question in the final turn, following a fine-grained taxonomy of 9 question types and 12 question intents. The annotation was performed through a set of techniques: manual annotation, pre-trained sentence-similarity classifiers based on siamese and triplet networks, and BERT-based classifiers. This project describes the workflow employed to preprocess the original empathetic dialogues, perform manual annotation on a small subset of dialogues, apply the pre-trained sentence-similarity classifiers to extend the manually-annotated data by a number of additional examples, and train the BERT-based classifiers to annotated the rest of the data points. Although the labels were highly unbalanced for both question types and intents in the extended manually-labelled dataset, the BERT-based classifiers achieved accuracies of 78% for types and 75% for intents, eventually being used to label the rest of the data points. The resulting dataset was confirmed to follow the typical human social interaction patterns after its quality evaluation was performed through visualisation techniques.**

I. Introduction

Strong communication skills have always been an advantage for any individual since Homo Sapiens first developed speech capabilities approximately 150.000 to 200.000 years ago, in Africa, with engravings on red-ochre serving as evidence of early days abstract reasoning [1], [2]. It is well known that these skills have been associated with leadership, persuasion and success throughout the passage of time. Moreover, social conversations have become a fundamental need of the human race as they are required not only for conveying information that is crucial for the community's survival and cultural adaptation to new, potentially hostile habitats [3], but also for receiving emotional relief and support in stressful situations [4].

Consequently, relating to one another is a bonding mechanism that has maintained human societies whole through challenging periods of time or unexplored environments, encouraging them towards collective effort, charity and mutual help which eventually resulted in the prosperity of our species. This mechanism emerged and developed, as many other behavioural and biological features, in parallel with the complexity of human communication, in particular spoken and written language. The better one could explain and share their feelings, views and knowledge with others, the higher their need to feel validated, heard and understood.

Research [4], [5] claims that empathy has evolved to be a key requirement for people's interconnectivity, psychological and physical health. One of the most important skills that an effective communicator can acquire in order to exhibit empathy is *active listening*, be it part of either oral or even written exchanges [6]. It allows the listener to understand the speaker and deliver an appropriate reply [7], hence satisfying the interlocutor's need to be heard. One active listening technique consists of actively asking spontaneous, follow-up questions after comprehending the speaker's response [8]. This tactic encourages the conversation to proceed further and ensures that the speaker feels connected. Unsurprisingly, it was shown that when empathetic listeners want to understand the points of view of the speakers, more than half of their replies during an empathetic dialogue include a question to demonstrate attention and engagement [9].

Incidentally, researchers have been exhibiting an increasing interest in Open-Domain chatbots over the recent years [10]–[12]. With the rise of data-driven techniques, one of their goals is to create empathetic chatbots that can provide natural responses based on the speaker input's semantics. There are a variety of tasks at which they could be helpful such as: interviewing, counseling and even training people on various tasks [10]. One way of providing the chatbots with a "personality" that feels compelling in terms of user experience is active listening [10]. Since asking follow-up questions based on the speaker's input is an active listening technique, it is necessary to extensively comprehend questioning strategies and how they can be employed in the context of chatbots so as to achieve human-like empathetic follow-up questions.

The main approach to understand questioning strategies for creating chatbos capable of providing realistic, engaging responses is defining a taxonomy of question types and intents. This taxonomy can be used to model the dialogue types and intents. For this purpose, the human-computer interaction (HCI) group at EPFL defined one such fine-grained taxonomy, comprising 9 question types and 12 question intents, after thoroughly analysing a number of dialogues extracting from the Empathetic Dialogues (ED) dataset [13]. Then, they manually labelled a small subset of data points, .i.e. questions, resulting a in initial seed dataset of follow-up questions with associated types and intents.

However, the problem that remained was that this seed was too small for any realistic applications, especially regarding those based on data-driven techniques. Moreover, manually labelling the entire dataset is not feasible in terms of labour force, time and cost considerations.

For this reason, the aim of the project was to expand both the type and intent labels to all of the listener questions in the ED dataset given the initial seed. We explored different approaches to address this problem, resulting in a two-fold contribution. First, we conducted an empirical analysis of lexical patterns associated with each question type and intent. We performed both automatic analysis of most frequent n-grams, as well as manual derivation of patterns. While the established dependencies do not uniquely define each given label, they served as a first step towards developing a lexical resource for questions and can be further elaborated to inform the automated labelling and question generation methods. Second, we devised an automatic labelling pipeline for both sets of labels, allowing us to annotate the whole dataset of listener questions with high prediction accuracy.

The remainder of this paper is organised as follows. Section 2 describes the related work on taxonomies of emotions and empathetic responses intents. Section 3 presents the ED dataset, how it was preprocessed and used in order to derive the question types and intents taxonomy, but also finee-tune our models, as well as the initial labelled seed data statistics. Section 4 describes our attempt to build a lexical resource for questions based on n-grams, n-skip-grams and manually extracted lexical patterns which was meant to be used for a heuristic-based annotation procedure. Section 5 presents the models employed for the automated question labelling pipeline and how we combined them to extend the labels form the manually annotated data points to the entire ED question dataset. It also presents the evaluation criteria of the models and the model-specific workflows. The correctness and potential of the results is further discussed in Section 6, whereas the future work is summarised in Section 7. Finally, the project is concluded in Section 8.

II. Related work

The background for this project is provided by presenting an existing dialogue-intent taxonomy and how it revealed the further need for question-specific taxonomies to label listener responses for modelling empathy in social chatbots. Furthermore, the methodology adopted for another similar goal of labelling a huge dataset of 1M movie dialogues is presented, since it was the blueprint for our approach.

Similar work has been conducted to automatically label dialogues based on intent taxonomies. Welivita and Pu [9], the main inspiration for this project, proposed a empathetic response intent taxonomy which can be used to achieve higher interpretability and controllability of the responses produced by social chatbots with the goal of leading healthy, desirable conversations. They created the taxonomy by analysing and deriving the listener intents in responses given to various emotional contexts in the ED dataset. The initial taxonomy and seed dataset containing 521 sentences were obtained by experts who manually annotated listener utterances extracted from 20 randomly selected dialogues for each emotional context. Eventually, they arrived at 15 empathetic response intents. They also showed how to extend the labels from the seed by means of lexical methods and then employed the resulting data to train a BERT transformer-based classifier to eventually annotate all speaker and listener utterances with the 8 most frequent intents and the 32 type of emotion categories given in the ED dataset. The initial expansion of the seed training set was achieved by searching the rest of the data for n-grams characteristic for the intent categories. Subsequently, they used the enhanced training set to fine-tune a pre-trained language model, RoBERTA [14], and automatically labelled the entire dataset with its aid.

Following their analysis of the most commonly associated speaker emotions and listener response intents, they acknowledged that questions are a key element of empathetic responses since the majority of speaker emotions are directly followed by questions.

Regardless of the emotions expressed by the speakers, asking questions allows them to feel heard, understood and comforted. Thus, their work confirmed that there is a need for further analysis of questions in the context of empathetic responses and the creation of a fine-grained taxonomy of empathetic question types and intents. The taxonomy should then be used to label the listener responses in empathetic datasets to allow for the design of social chatbots that can provide healthy, intepretable and controllable empathetic responses based on the speaker's emotional state [15].

Another relevant work by Welivita, Xie and Pu [16] served as example in terms of methodology. To address the relatively small size of the ED dataset which is not adequate for training data-hungry neural conversational systems, they curated the OpenSubtitles Emotional Dialogues (OSED) dataset [17] comprising 1M movie dialogues. Each dialogue turn was labelled according to 32 fine-grained emotions and 9 empathetic response intents, reaching a previously unattempted emotional dialogue classification both in terms of number of labels and dataset size. To make the task less challenging, the authors came up with a semi-automated approach using a weak labeler, EmoBERT trained on the ED dataset, to label and filter 1M emotional dialogues based on the prediction confidence. After running the weak classifier, they asked Amazon Mturk workers to select a label from the top 3 predictions or suggest a new one. This approach resulted in an initial seed dataset. To extend the seed, they also employed a Sentence-BERT approach [18] to arrive at semantically meaningful sentence embeddings which can be grouped in therms of cosine-similarity. Thus, they computed dialogue embeddings for the seed and the unlabelled data by employing a decaying weight beginning of at the last turn of the dialogue. Next, they calculated the cosine-similarities between the unlabelled and seed data, ranked them by value, kept only those unlabelled data points whose similarity to a labelled data point was higher than a threshold and annotated them with the same label as the labelled embedding. This way they obtained an additional 3,196 annotated dialogues.

Finally, the authors decided to use the annotated data, crowd-only and crowd and similarity-based, to train a classifier, EmoBERT+ that could then be used to label the entire dataset. EmoBERT+ has a BERT-based architecture, whose weights were initialised from the pre-trained language model RoBERTa [14] and took as input the dialogue turns together with their previous context in reverse order, multiplying the token-embedding of every turn with a decreasing weight factor such that the closer the context to the turn, the higher its weight. Finally, the network input is the sum between the token embedding of each turn in the dialogue multiplied by the weighting scheme and its positional embedding. The best performing model which achieved an accuracy of 65% was identified during the training phase and used to label every turn in the OSED dataset. A similar synergy of labelling techniques and methodology will be employed int this project to label the listener questions in the ED dataset.

III. Dataset

The chosen dataset for this project is the EmpatheticDialogues (ED) dataset created by Rashkin et al. [13] which comprises 24,850 publicly-available dialogues in an open-domain one-to-one setting. The discussions consist of up to 6 turns and take place between 2 individuals conversing about personal past events related to a particular emotion. Each dialogue has an associated emotion annotation selected from a total of 32 emotion labels, aggregated from several other annotation schemes, which span a wide spectrum of emotions with various polarities. The the emotion labels are evenly distributed and richer than other in emotion prediction datasets. Furthermore, the dialogues were crowdscourced from the labour of 810 US workers, using the ParlAI platform [19] to interact with Amazon Mechanical Turk (MTurk).

Criteria	Statistic
Total no. dialogues	24,850
Avg. no. of turns per dialogue	4.31
Total no. of speaker turns	55,984
Total no. of listener turns	51,263
No. of dialogues with at least one question from the listener	15,253 (61.4%)
No. of questions from listeners	20,201

TABLE I: EmpatheticDialogues dataset statistics.

The ED dataset was appropriate for the project, since the workers were explicitly asked to exhibit empathy in their one-to-one exchanges, thus toxic replies are highly unlikely. The creators of the dataset showed that it can be used not only to benchmark the ability to provide empathetic responses in one-to-one dialogue systems, but it also improves their performance when used for training. In addition to that, many of the listener turns in the dialogues contain questions which could be used to derive the initial taxonomy of question types and intents by the EPFL HCI group. Table I presents the several relevant summary statistics of the dataset. The average no. of turns per dialogue is aprox. 4. The total no. of speaker turns is 55,984. The total no. of listener turns is 51,263 . The no. of dialogues which contain at least one question in the listener's turns is 15,253 (61.4%) and the total no. of questions posed by the listeners is 20,201.

Original dialogue	
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 t he trust to loan my friend my vehicle.
Listener:	Ouch... Is it just for a day? Is your friend a safe driver?
Resulting dialogues	
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 t he 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 t he trust to loan my friend my vehicle.
Listener:	Ouch... Is it just for a day? **Is your friend a safe driver?**

TABLE II: Original and resulting dialogues after being preprocessed.

From the original ED dataset, only those dialogues containing questions in at least one listener turn were kept. Since one dialogue could contain several listener questions, each such dialogue was split into several separated dialogues, equal to the number of listener questions. The resulting sub-dialogues were truncated such that they would end with the particular question that they corresponded to so as to allow to label every question in each dialogue, without losing the previous conversational context. The dataset obtained through this procedure could then be used , not only to devise the question types and intents taxonomy and create a manually-annotated seed dataset based on that, but also for the task of labelling each remaining question through advanced NLP methods which will be presented in the following sections. Table II shows an example of a dialogue from the original ED dataset and the resulting dialogues after the split.

Category	Example	Freq.
Request information	Is she having a boy or a girl?	48.21%
Ask about consequence	What did the landlor do about it?	19.66%
Ask about antecedent	What happened for you to feel that way?	10.63%
Suggest a solution	Why didnt' you just reschedule?	9.04%
Ask for confirmation	He does?	5.53%
Suggest a reason	Do you think you were being rowdy?	4.33%
Positive rhetoric	Who needs a calculator with a brain like that, ey?	1.57%
Negative rhetoric	Why let him have all the fun?	1.54%
Irony	Oh my lord, only ten?	0.5%

TABLE III: Taxonomy of empathetic question types with corresponding examples and occurrence frequencies based on the manually annotated 6743 listener questions in the EmpatheticDialogues dataset.

Category	Example	Freq.
Express interest	What is your favourite food?	53.52%
Express concern	Did you get hurt?	21.27%
Sympathize	I bet you were pretty mad?	7.19%
Amplify excitement	Nice, how excited are you for it?	4.5%
Offer relief	Can you get another job?	3.08%
Support	How does that make you feel?	2.1%
Amplify joy	How are you celebrating for him?	1.86%
Amplify pride	I'm so proud of him, what is he taking in university?	1.68%
De-escalate	That was rude, maybe she didn't hear you?	1.66%
Pass judgement	Are you a bit jealous?	1.58%
Moralize speaker	Oh lord, why did you do that?	1.06%
Motivate	Try joining clubs or finding hobbies?	0.5%

TABLE IV: Taxonomy of empathetic question intents with corresponding examples and occurrence frequencies based on the manually annotated 6136 listener questions in the EmpatheticDialogues dataset.

The seed dataset was obtained by manual annotation performed by the EPFL HCI group members and a number of MTurk workers. It consisted of 6743 type annotated and 6136 intent annotated data points, each with a frequency of labels as shown in tables III and IV.

Finally, figures A.1 in the Appendix illustrate the distributions of question types and intents in the seed dataset over the emotional contexts labelled in the original ED dataset. We can see that certain question types or intents tend to occur more often in particular emotional contexts. For example, the emotion labels *embarrassed* tends to co-occur with the 'Ask about consequence' question type and with the *Sympathise* question intent.

IV. Towards a lexical resource for questions

In this section, a first heuristic-based approach that was attempted with the purpose of expanding the seed data collected by manual annotation is presented. The overall idea was to investigate whether it is possible to identify syntactic structures and patterns to build a lexical resource for empathetic question types and intents. Once constructed, the lexical resource would associate the discovered patterns with the labels they are most indicative of, hence being useful when collecting more example utterances for each category. Although int the end the results in this section were not used to label additional questions, the results were interesting and will be subject to future work. It is worth mentioning that at this stage, the entire seed dataset that was later used was not complete as MTurk workers were still annotating dialogues at the time, only 147 questions being annotated with both types and intents.

A. N-grams and n-skip-grams

The first method employed to construct the lexical resource relied on finding the most common n-grams and n-skip-grams for each question type and intent categories. The most common n-grams and n-skip-grams were computed and then manually analysed for all values of n between and including 1 and 4, but no clear type- or intent-specific n-grams emerged. Although Welivita and Pu [9] successfully used this approach to collect additional utterances for each of their labels, their categories are more disjoint, including a variety of sentence types i.e. questions, exclamations, enunciations. Furthermore, all questions were mapped to a single class in their work, *Questioning*. It is true that empathetic listeners' questions can belong to different type and intent classes, however it is extremely difficult to distinguish them without considering the previous utterances and conversational context. This is due to the fact that most questions are formulated in similar ways and contain the same question-specific words and word orders e.g. *why... ?, how...?, when... ?* etc. Since there were no significant findings with regard to n(-skip)-grams, the results will not be included in this report.

B. Manually specified patterns

The second method employed to construct the lexical resource involved the manual analysis of a subset of the 147 listener questions utterances in order to extract lexical patters that seem to frequently occur with particular question types and intents. The difficulty in this case stemmed from finding the right balance between the generalisation and specificity of the patterns such that there would be enough matching questions, yet not too many overlapping matches between a single pattern and several question types or intents. Following their extraction, the patterns were converted to regular expressions and applied to a all the 147 questions. Tables A.1 and A.2 in the Appendix show the final selection of patterns together with the most frequent question types and intents for which the patterns occur and the percentage of the questions that contain the patterns from the most frequent type and intent classes.

V. Automatic labelling of questions

In this section, two automatic labelling procedures are presented. The first one, distant learning using dialogue embeddings, successfully expanded the seed dataset of manually-labelled data points, resulting in a lager training set for the second classifier-based technique which was successfully employed to label the rest of the data.

A. Distant learning using dialogue embeddings

The first technique addressed in order to extend the initial annotated seed dataset, obtined by both the EPFL HCI group and the Amazon MTurk workers, is distance learning using dialogue embeddings. This procedure is based on Sentence-BERT (SBERT) [18] which results in semantically meaningful embeddings computed by siamese and triplet networks. Once the embeddings are generated, cosine-similarity can be used as metric to identify the most similar empathetic questions. We followed almost the same approach as Welivita and Pu [16], with some additional modifications. The same *roberta-base-nli-stsb-mean-tokens* model as suggested by Welivita and Pu [16] was employed, fine-tuned on NLI [20] and the STS benchmark (STSb) datasets [21], as they concluded that it performs better in terms of efficiency than *roberta-large* and scores well on the STS benchmark.

As such, we created 2 classifiers, one for question types and one for question intents. The same methodologies were applied to both. To test their performance, we performed 5-split stratified cross validation such that the data folds would be balanced across all class labels. For both classifiers, the final test performance was calculated by taking the average of the accuracies computed for the 5 folds. At this stage, only the labelled data points were used in order to identify the similarity thresholds above which two empathetic questions could be classified with high confidence as belonging to the same type or intent category. Once the threshold was identified, the classifiers could be used further to extend the seed dataset. The two classifier types which we investigated are described in the following two subsections.

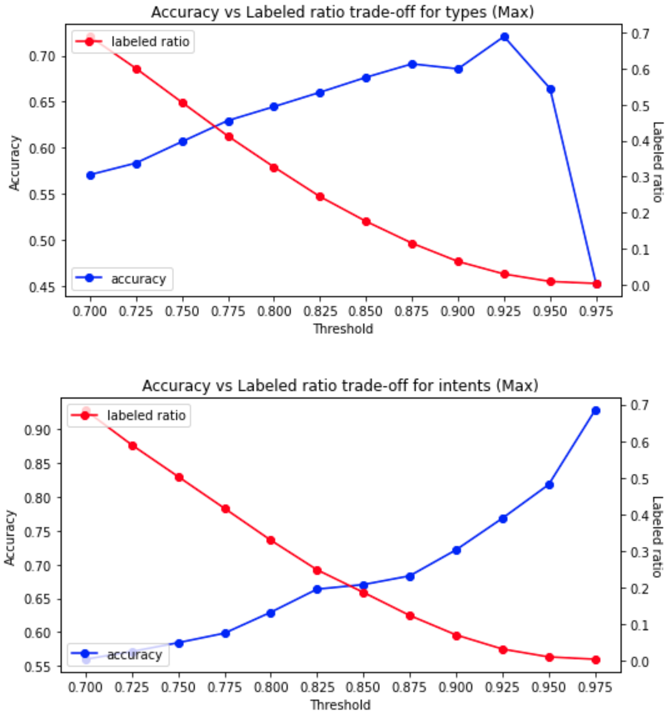


Fig. 1: Accuracy and ratio of labelled data points per threshold during training for the max. similarity apporach.

1) Maximum similarity approach: The first type of classifier which we evaluated used a maximum similarity approach. It involved the following steps. 1) Dialog turn embeddings of 768 dimensions were computed by the SBERT model for all utterances in the entire preprocessed dataset. 2) Next, the dialogue-wise embeddings were obtained for both labelled and unlabelled conversations by applying a weight decay, beginning from the last turn towards the first one in the dialogue. Then, a weighted average of the utterance embeddings in each dialogue was taken to arrive at the final dialogue-level embedding. The weighting scheme was based on half decaying [16] which is explained in Appendix A.2. 3) Next, the cosine-similarity was computed between the test and train sets of annotated data points for each fold and the results ranked. 4) The data points in the test folds were annotated with the label of the embedding in the train folds with the highest cosine-similarity above a selected threshold. 5) Finally, after computing the accuracies for each test fold, the final accuracy for the current threshold was calculated by taking their average. The same procedure was repeated for a number of thresholds in order to identify the one with the highest accuracy. 6) After identifying the best threshold, the classifier was applied to the unlabelled data as well. The unannotated question dialogues received the label of the annotated embedding with the highest cosine-similarity value above the selected threshold. If no embedding with a similarity higher than the threshold was found, the unlabelled dialogue remained unannotated. The same procedure was employed for both the question types and intents classifiers. Figure 1 illustrates the accuracies and ratio of labelled data points per threshold.

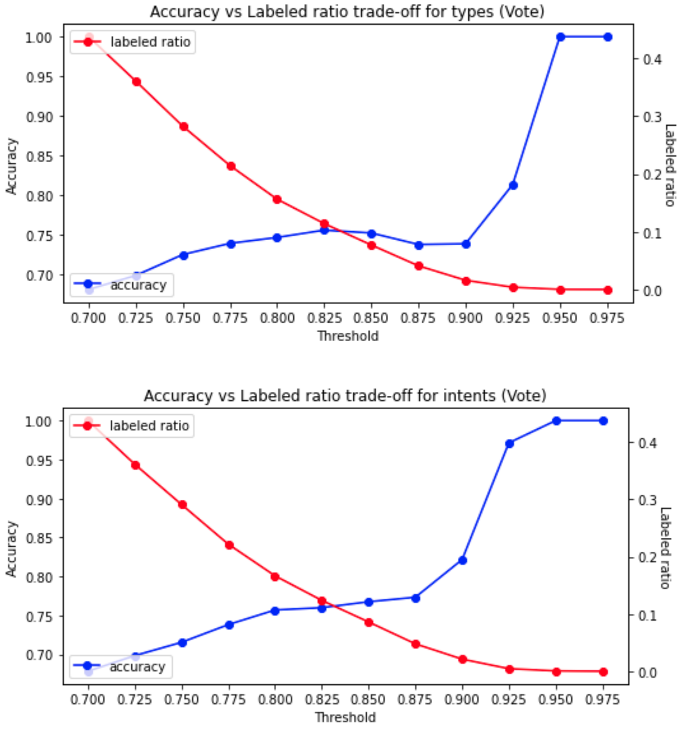


Fig. 2: Accuracy and ratio of labelled data points per threshold during training for the top 3 max. similarity vote approach.

2) Top 3 maximum similarity vote approach: The second type of classifier which we evaluated used a top 3 maximum similarity vote approach. The first 3 steps are the same as in the case of the maximum similarity approach, but the subsequent steps include some modifications. 4) The data points in the test folds were annotated with the label decided on by a majority vote between the top 3 embeddings in the train folds with the highest cosine-similarity above a selected threshold. 5) Finally, after computing the accuracies for each test fold, the final accuracy for the current threshold was calculated by taking their average. The same procedure was repeated for a number of thresholds in order to identify the one with the highest accuracy. 5) After identifying the best threshold, the classifier was applied to the unlabelled data as well. The unannotated question dialogues received the label decided on by a majority vote between the top 3 annotated embeddings with the highest cosine-similarity values above the selected threshold. If 3 embeddings with a similarity higher than the threshold were not found, the unlabelled dialogue remained unannotated. The same procedure was employed for both the question types and intents classifiers. Figure 2 illustrates the accuracies and ratio of labelled data points per threshold.

3) Results: Following the experiments, it was decided that the top 3 maximum similarity vote approach was superior. For both the type and intent classifiers, a threshold of 0.825 was deemed as providing the optimal trade-off between accuracy and amount of labelled data points. As a result of this choice, the seed dataset was extended by 1911 type-labelled and 1886 intent-labelled dialogues. Out of these, 1874 were annotated with both types and intents.

Finally, it is worth mentioning that an alternative method of computing the dialogue-level embeddings was attempted for the distance learning classifiers. This procedure was meant to take into consideration the emotional context of the dialogues as well by calculating the one-hot-encoding vectors of the emotional context labels (provided by the ED dataset) and appending them to the final embeddings described in the above subsections. However, it led to no significant improvements and it was not employed in the end due to the unnecessary computational overhead and increased dimensionality of the embeddings.

B. Classification models for questions types and intents

To label the rest of the dataset, two final classifiers, one for question intents and one for question types, were fine-tuned using the seed dataset combining the data points manually-labelled by the EPFL HCI group members and the Amazon MTurk workers with those found using the the previously mentioned SBERT models. Aside from the final softmax layer which was modified to accommodate the number of types and intents in the question taxonomy, both classifiers follow the EmoBERT+ architecture introduced by Welivita and Pu [16], illustrated in fig. 3. These BERT-based architectures are comprised of an attention layer, a hidden layer, and a softmax layer. The representation networks consist of 12 layers, 768 dimensions, 12 heads and 110M parameters. Following Wlivita and Pu's [16] approach, the network is initialised with weights from RoBERTa [14], a pre-trained language model, to later be fine-tuned using the annotated data points available in the extended seed.

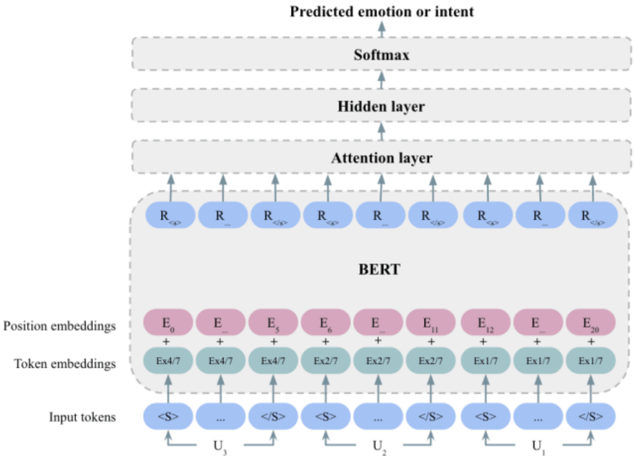


Fig. 3: EmoBERT+ architecture [16].

The input of the networks comprises the listener's question turn together with the previous dialogue utterances in reverse order. The token embeddings representing each turn were multiplied by a decreasing weighting scheme so as to give increasingly high importance to the utterances as they get closer to the final question-containing turn. The embedding corresponding to the question is the final one in the dialogue and is assigned the highest importance weight. The final input representations were obtained by summing the importance weight-multiplied token embeddings with the BERT-specific positional token embeddings. The maximum length of the input tokens was set to 100.

Both models were fine-tuned for 10 epochs with a learning rate of 2e-5, using a batch size of 50 due to the limitations of the available hardware. After 4 epochs for the type classifier and 3 epochs for the intent classifier, both the models achieved the lowest validation loss before starting to overfit as shown in fig. 4. Thus, we selected the parameters available at those epochs. Since the models were not trained from scratch, but only

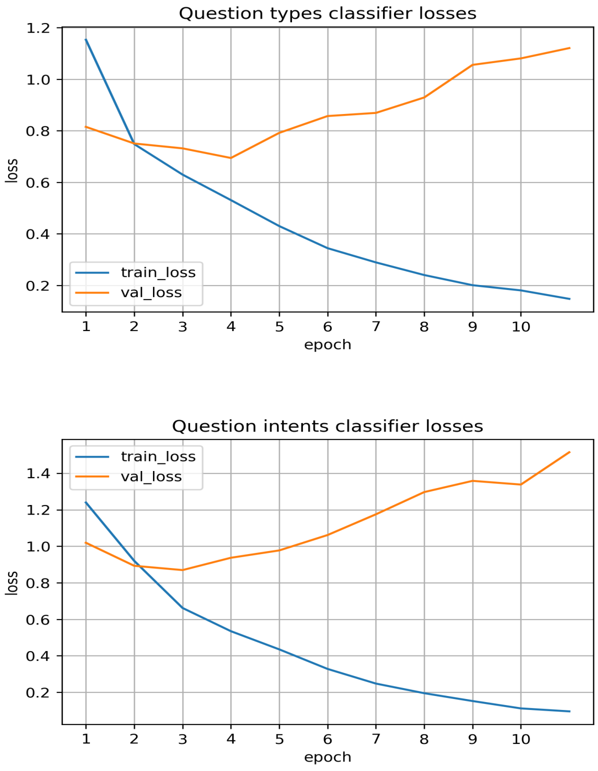


Fig. 4: Question types and intents BERT-based classifiers losses.

fine-tuned, it is reasonable that the best models emerged after a small number of iterations. Before starting to fine-tune the models, the seed dataset contained 9,220 examples for which at least one of the type or intent label was available, 8,344 type-labelled examples, 7,712 intent-labelled examples and 6,836 examples annotated with both labels. To test the final models, 1,500 of the examples annotated with both labels were set aside and used as test set.

To fine-tune the question types classifier, 6,844 type-annotated examples remained after removing the ones belonging to the test set. These examples were further split into 80% train set and 20% validation set i.e. 5,475 data points for training and 1,369 data points for validation. Similarly, when fine-tuning the question intents model, we were left with 6,212 intent-annotated examples after removing the ones belonging to the test set. They were also split into 80% train set and 20% validation set i.e. 4,969 training examples and 1,243 validation examples.

Model	Precis.	Recall	F1 - score	Accuracy
Q. types classifier	0.6	0.55	0.55	0.78
Q. intents classifier	0.6	0.3	0.33	0.75

TABLE V: Performance results of the BERT-based question types and intents final classifiers on the test set.

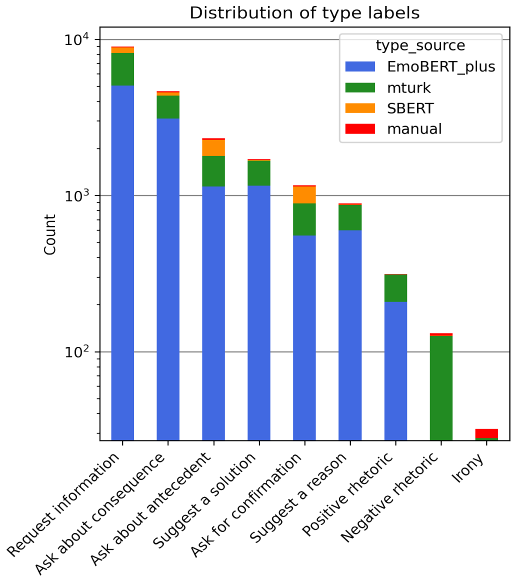


Fig. 5: Type labels distribution for the fully annotated dataset.

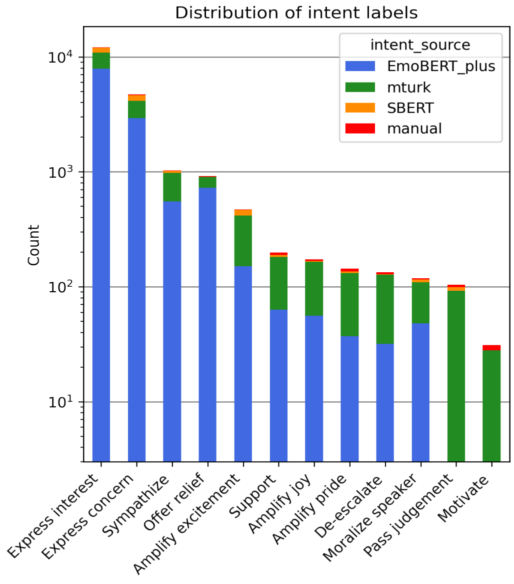


Fig. 6: Intent labels distribution for the fully annotated dataset.

C. Results

The optimal question types and intents classifiers' performance results are shown in table V. Their precision, recall, macro-F1 metrics and accuracy scores are presented as computed on the common test set of 1,500 examples annotated with both type and intent labels. Since the achieved accuracies were satisfactory, the models were ran on the remaining unlabelled data points until the entire dataset was tagged with question types and intents categories. Figs. 5 and 6 depict the type and intent labels distributions of the fully annotated dataset, as well as the amount of examples annotated by each different method, on a logarithmic scale. The joint distribution of the types and intents categories is also available in Appendix A.3.

VI. Discussion

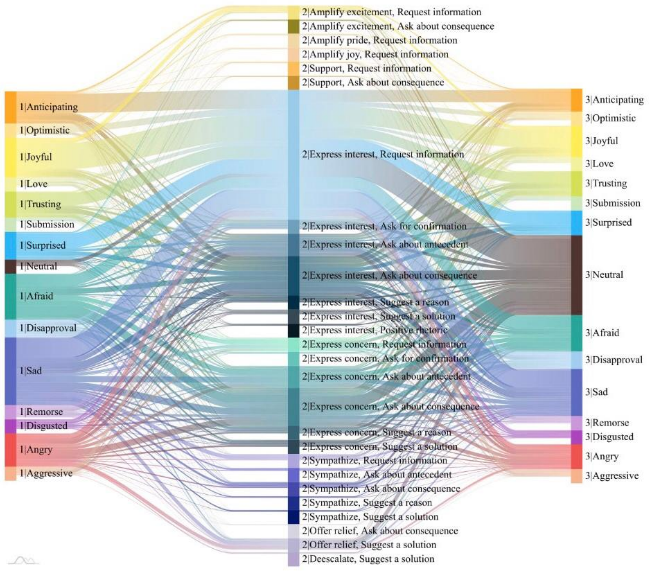


Fig. 7: Emotion-type/intent-emotion flow visualisation of the resulting ED questions dataset.

This project proposed a listener's questions annotation pipeline for empathetic dialogues. The two question types and intents classifiers, that were created to extend the labels from the manually-annotated seed to the entire dataset, were inspired from Welivita and Pu [16] who designed an architecture that can be applied as a general-purpose classifier for one-to-one social dialogues. Although the models achieved high accuracies compared to the model of inspiration, 78% for question types and 75% for question intents, and a comparable precision of 60%, the F1-scores indicate that there is a significant imbalance between the precision and recall capabilities of the models. In particular, the F1-score of the intents classifier is only 33%. These results, however, were expected since the models were subject to highly unbalanced training datasets, 53.62% of the intent-labelled data points belonging to the *Express interest* category and 48.21% of the type-labelled dataset belonging to the *Request information* category. Since the available training data was far from ideal in terms of balance, the reported F1-scores, especially that of 55% achieved by the types classifier, are comparable to the state-of-the-art performance and could be improved should additional training data points for the underrepresented classes be acquired.

The emotion-type/intent-emotion exchange patterns in fig. 7 indicate that the final ED question dataset reflects the typical emotion-type/intent-emotion flows encountered in social dialogues according to existing work [9], [22]. As a result, the fully labelled question dataset that we obtained can be used to endow social chatbots with the ability to exhibit empathy by asking situation-appropriate empathetic questions.

Speaker: (Guilty)	I wish I could work less... I feel so bad that I don't get to spend as much time with my daughter.
Listener: (Suggest a solution)	I hear ya. Any way to get a new job?
Speaker: (Hopeful)	I've been trying, what I really need is a higher paid one so that way I don't have to work 2 jobs.

TABLE VI: Positive emotional state shift in speaker after being asked a *Suggest a solution* type empathetic question.

Speaker: (Devastated)	I can't believe the Falcons blew the Super Bowl. I was there.
Listener: (Express interest)	Oh you were? I really am not a huge sports fan, so what was so big about this?
Speaker: (Neutral)	It had never happened before.

TABLE VII: Positive emotional state shift in speaker after being asked an *Express interest* intent empathetic question.

Among the benefits that this project introduces, we can enumerate 1) a better quality assessment of empathetic chatbots's question-based responses, 2) the ability to nudge social chatbots in the direction of appropriate empathetic question-based responses given a particular emotional context, and 3) additional aid in the creation of neural chatbots that are more interpretable and controllable [15]. To showcase the importance of delivering appropriate empathetic responses, tables VI and VII show the effect of certain question types and intents on the speaker's emotion from the fully labelled dataset. Table VI demonstrates how asking an empathetic question of type *Suggest a solution* to a person that feels guilty, can shift their emotional state to a hopeful one. Likewise, table VII presents how the impact of responding with an *Express interest* intent question can bring an individual from a devastated emotional state back to a more neutral one as they shift their attention on briefing the listener on the subject.

VII. Limitations and future work

Although the project led to many promising results, there is still a fair amount of work that can be done in the future. The first area which left significant room for exploration is the pattern derivation for building a lexical resource for questions. The work done on this topic was not conclusive, thus we did not manage to engineer any additional pattern-related features for the classification task. In the future, we could try to analyse the obtained patterns more extensively and, based on the findings, engineer new features that could improve the performance of our classifiers. One interesting direction to follow in this case is designing features based on grammatical forms of questions since linguistic knowledge is known to add a lot of value to rule-driven NLP [23], even for the difficult case of question utterances. This suggest that features based on syntactic rules could also add value to neural classifiers.

Secondly, we only ran one iteration of supervised training which already resulted in fairly satisfactory models in terms of accuracy. However, a semi-supervised learning approach can be attempted in order to improve even further. In addition to that, the F1-scores that we obtained were not very good due to the imbalanced between the models' precision and recall. This issue was most likely caused by the significant imbalance of labels in the manually-annotated seed dataset. For this reason, in the future, the BERT-based classifiers could be retrained on a higher quality seed dataset obtained after collecting more data points belonging to the underrepresented categories of question types and intents.

Lastly, due to the unavailability of specialised hardware (GPUs), we trained the models with the default hyperparameters proven to perform well by Wlivita and Pu [16] on related tasks. The possibility remains that better hyperparameter values can be found such that the models may achieve superior performance results. Hence, it is worth exploring different configurations in the future.

VIII. Conclusion

In this project, we preprocessed and annotated a dataset containing 25K empathetic dialogues based on a fine-grained taxonomy of 9 question types and 12 question intents. The resulting dataset is focused on dialogues containing empathetic questions in listeners' turns, unlike other similar ones that are focused on more general types of empathetic utterances, not specifically questions. At first, the creation of a lexical resource for questions was attempted to serve as a heuristic-based annotation method, however it did not yield conclusive results. To perform the annotation, two families of question intents and types classifiers were developed. A classifier from both families was created for each one of the cases of identifying question types or intents, resulting in 4 models in total. The first type of classifiers are based on sentence-similarity approaches using cosine-similarity as metric. They were used to extend the available manually-labelled seed dataset with several high confidence predictions whose similarity to annotated examples exceeded a designated threshold. The second type of classifiers (BERT-based) were trained on the extended seed to annotate the rest of the data points with significant accuracy. As future work, the results from the pattern analysis should be explored more extensively in order to engineer supplementary features related to grammatical structures . Furthermore, in the future, the BERT-based classifiers are intended to be trained on a higher quality manually-annotated seed dataset with a better balance of labels across classes. In addition to that, further model selection should be performed in order to identify the optimal hyperparametrs of the BERT-based models should the necessary hardware resources become available. Lastly, the models were trained for one iteration in a supervised manner, hence additional improvement could be achieved following a semi-supervised learning approach.

References

[1] C. S. Henshilwood, F. d'Errico, R. Yates, Z. Jacobs, C. Tribolo, G. A. Duller, N. Mercier, J. C. Sealy, H. Valladas, I. Watts *et al.*, "Emergence of modern human behavior: Middle stone age engravings from south africa," *Science*, vol. 295, no. 5558, pp. 1278–1280, 2002.

[2] C. S. Henshilwood and B. Dubreuil, "Reading the artefacts: gleaning language skills from the middle stone age in southern africa," *The cradle of language*, vol. 2, pp. 61–92, 2009.

[3] M. Pagel and R. Mace, "The cultural wealth of nations," *Nature*, vol. 428, no. 6980, pp. 275–278, 2004.

[4] M. Reblin and B. N. Uchino, “Social and emotional support and its implication for health,” *Current opinion in psychiatry*, vol. 21, no. 2, p. 201, 2008.

[5] N. Eisenberg and N. D. Eggum, “Empathic responding: Sympathy and personal distress,” *The social neuroscience of empathy*, vol. 6, no. 2009, pp. 71–830, 2009.

[6] C. Bauer, K. Figl, and R. Motschnig-Pitrik, “Introducing’active listening’to instant messaging and e-mail: Benefits and limitations,” *IADIS International Journal on WWW/Internet*, vol. 7, no. 2, pp. 1–17, 2010.

[7] D. A. Kolb, I. M. Rubin, and J. M. McIntyre, *Organizational psychology: readings on human behavior in organizations*. Prentice Hall, 1984.

[8] S. Louw, R. W. Todd, and P. Jimarkon, “Active listening in qualitative research interviews,” in *Proceedings of the International Conference: Research in Applied Linguistics, April*, 2011.

[9] A. Welivita and P. Pu, “A taxonomy of empathetic response intents in human social conversations,” in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4886–4899. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.429

[10] Z. Xiao, M. X. Zhou, W. Chen, H. Yang, and C. Chi, “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*, 2020, pp. 1–14.

[11] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, “The design and implementation of xiaoice, an empathetic social chatbot,” *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[12] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, “Caire: An end-to-end empathetic chatbot,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 622–13 623.

[13] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381. [Online]. Available: https://www.aclweb.org/anthology/P19-1534

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[15] C. Xu, W. Wu, and Y. Wu, “Towards explainable and controllable open domain dialogue generation with dialogue acts,” *arXiv preprint arXiv:1807.07255*, 2018.

[16] A. Welivita, Y. Xie, and P. Pu, “Fine-grained emotion and intent learning in movie dialogues,” *arXiv preprint arXiv:2012.13624*, 2020.

[17] P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” 2016.

[18] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[19] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston, “Parlai: A dialog research software platform,” *arXiv preprint arXiv:1705.06476*, 2017.

[20] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” *arXiv preprint arXiv:1508.05326*, 2015.

[21] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, “Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation,” *arXiv preprint arXiv:1708.00055*, 2017.

[22] 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.

[23] “Question type extraction using spacy,” https://towardsdatascience.com/linguistic-rule-writing-for-nlp-ml-64d9af824ee8#8a27, accessed: 2021-06-11.

APPENDIX

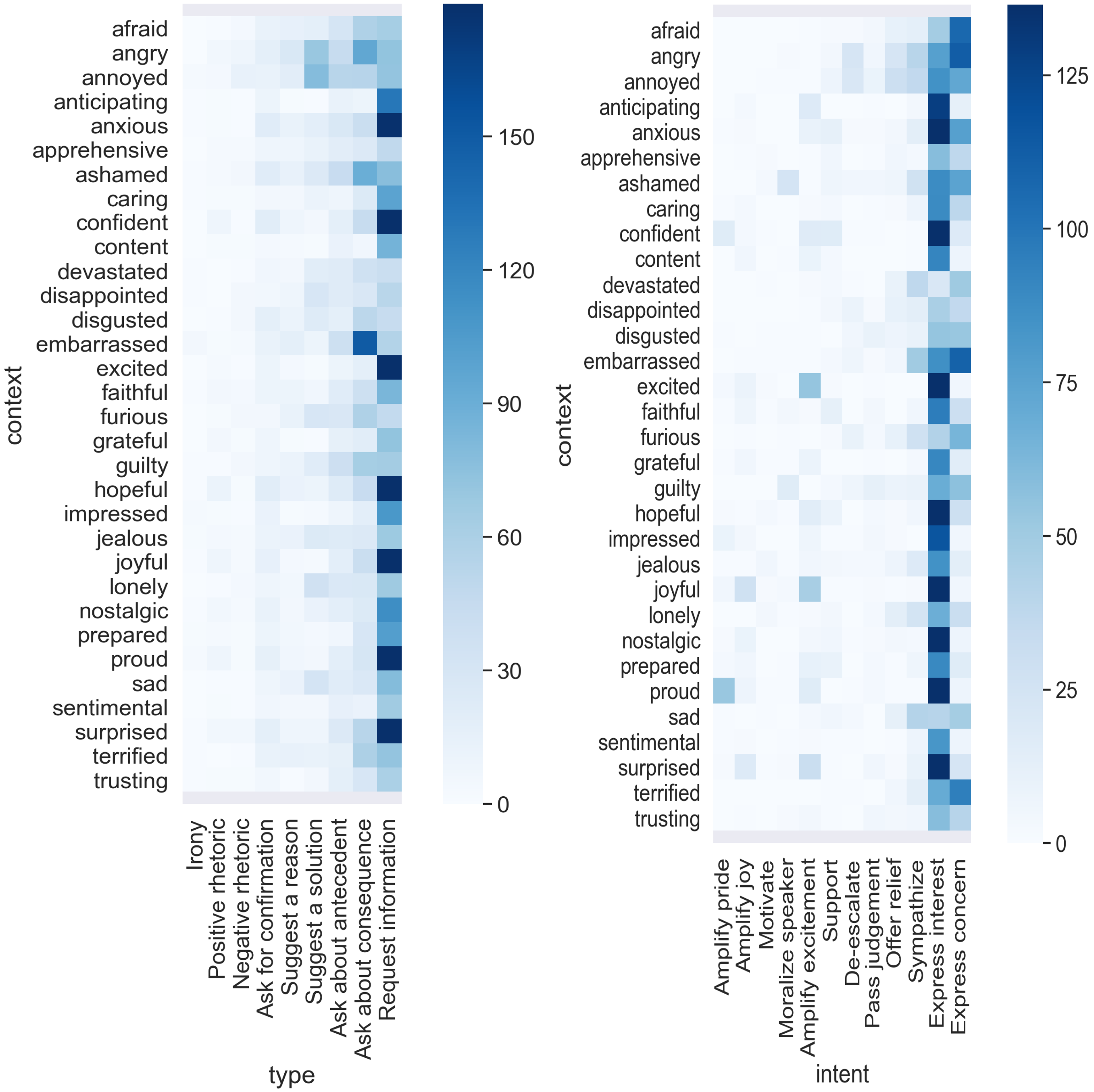


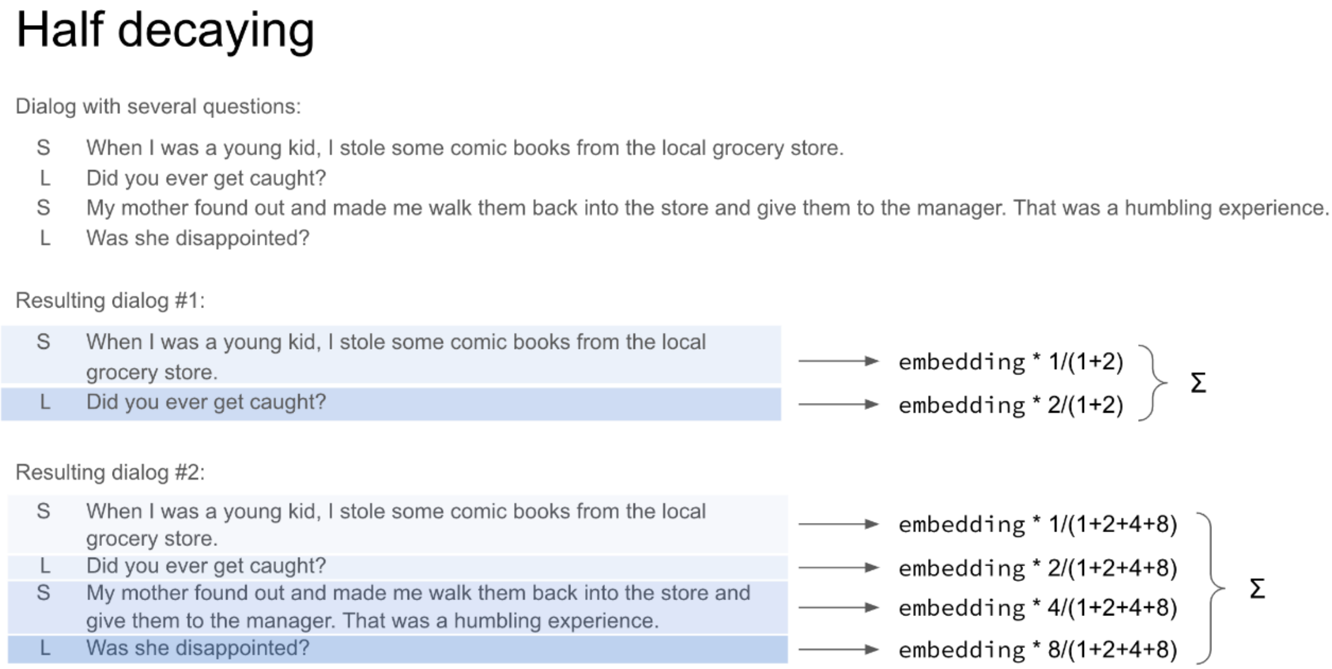
Fig. A.1: Question types and intents distributions over emotional context.

Lexical Pattern	Most frequent question type	Most frequent question type %
[max. 3-word nominal group/phrase]?	Ask for confirmation	48.15%
... right?	Positive rhetoric	17.86%
... really?	Ask for confirmation	14.81%
Why/When/Where/How did you/they/he/she/it ...?	Ask about antecedent	13.66%
Maybe/Perhaps ...?	Suggest a solution	9.52%
... proud ...?	Positive rhetoric	7.14%
Have/has you/they/he/she tried ...?	Suggest a solution	7.14%
... isn’t/doesn’t he/she/it?	Positive rhetoric	7.14%
How long/many/old/often ...?	Request information	6.18%
What kind of ....?	Request information	5.38%
... do you think ...?	Suggest a reason	4.94%
... speak/spoken/spoke to ...?	Suggest a solution	3.97%
... hard ...?	Negative rhetoric	3.85%
... , huh?	Positive rhetoric	3.57%
... , you know?	Positive rhetoric	3.57%
.... , ok?	Positive rhetoric	3.57%
Maybe ... could/shoul/couldn’t/shouldn’t ...?	Suggest a solution	3.17%
Were/Was you/they/he/she/it able to ...?	Suggest a solution	3.17%
... what made you/them/her/him/it ...?	Ask about antecedent	3.08%
A/An [max. 6-word nominal group/phrase]?	Ask for confirmation	2.78%
And/So, do/does you/they/he/she/it?	Ask for confirmation	2.78%
And/So, did you/he/she/they?	Ask for confirmation	1.85%
Did you/they/he/she/it ... at least ...?	Suggest a solution	1.59%
Did anyone/anybody ...?	Ask about consequence	1.4%
... i’m guessing/i guess ...?	Suggest a reason	1.23%
... prepared ...?	Suggest a reason	1.23%
And/So, were/was you/they/he/she/it?	Ask for confirmation	0.93%
And/So, you/they/he/she/it do/does/did?	Ask for confirmation	0.93%
And/So, are you/they?	Ask for confirmation	0.93%
And/So, do you/they?	Ask for confirmation	0.93%
... or what?	Ask for confirmation	0.93%
Have/Has you/they/he/she (ever) tried/thought o...	Suggest a solution	0.79%
How about ...?	Suggest a solution	0.79%
... you might want to ...?	Suggest a solution	0.79%
Have you ever/always/already ...?	Suggest a solution	0.79%
... terrible ...?	Suggest a solution	0.79%
... sorry ...?	Ask about consequence	0.47%
... awful ...?	Ask about consequence	0.47%
... sad ...?	Ask about antecedent	0.44%
... jealous ...?	Ask about consequence	0.23%
... do you know?	Request information	0.09%

TABLE A.1: **Question lexical patterns extracted by manual analysis of the EmpatheticDialogues dataset, the most frequent question types that contain them and the percentage of the questions belonging to the most frequent question types that contain them.**

Lexical Pattern	Most frequent question intent	Most frequent question intent %
[max. 3-word nominal group/phrase]?	Express concern	21.66%
Maybe/Perhaps ...?	Motivate	20.0%
How long/many/old/often ...?	Amplify pride	19.51%
Why/When/Where/How did you/they/he/she/it ...?	Amplify joy	10.0%
... you might want to ...?	Motivate	10.0%
Have/has you/they/he/she tried ...?	Motivate	10.0%
... proud ...?	Amplify pride	9.76%
What kind of ....?	Amplify excitement	8.08%
Maybe ... could/shoul/couldn't/shouldn't ...?	Moralize speaker	7.14%
... do you think ...?	Moralize speaker	7.14%
... speak/spoken/spoke to ...?	De-escalate	6.45%
... prepared ...?	Support	6.25%
... right?	Support	6.25%
.... , ok?	Express concern	3.69%
... what made you/them/her/him/it ...?	Pass judgement	3.45%
... really?	Pass judgement	3.45%
... isn't/doesn't he/she/it?	Amplify joy	3.33%
How about ...?	De-escalate	3.23%
Were/Was you/they/he/she/it able to ...?	De-escalate	3.23%
And/So, do/does you/they/he/she/it?	De-escalate	3.23%
... hard ...?	Support	3.12%
... i'm guessing/i guess ...?	Support	3.12%
... , huh?	Amplify pride	2.44%
... sorry ...?	Sympathize	2.03%
... awful ...?	Sympathize	1.35%
Did anyone/anybody ...?	Sympathize	1.35%
... or what?	Offer relief	1.33%
... , you know?	Offer relief	1.33%
Have you ever/always/already ...?	Offer relief	1.33%
Have/Has you/they/he/she (ever) tried/thought o...	Offer relief	1.33%
... terrible ...?	Offer relief	1.33%
Did you/they/he/she/it ... at least ...?	Express concern	0.46%
A/An [max. 6-word nominal group/phrase]?	Express interest	0.33%
... sad ...?	Express concern	0.23%
... jealous ...?	Express concern	0.23%
And/So, did you/he/she/they?	Express interest	0.17%
... do you know?	Express interest	0.08%
And/So, were/was you/they/he/she/it?	Express interest	0.08%
And/So, are you/they?	Express interest	0.08%
And/So, you/they/he/she/it do/does/did?	Express interest	0.08%
And/So, do you/they?	Express interest	0.08%

TABLE A.2: **Question lexical patterns extracted by manual analysis of the EmpatheticDialogues dataset, the most frequent question intents that contain them and the percentage of the questions belonging to the most frequent question intents that contain them.**



Half decaying

Dialog with several questions:

S When I was a young kid, I stole some comic books from the local grocery store.
L Did you ever get caught?
S My mother found out and made me walk them back into the store and give them to the manager. That was a humbling experience.
L Was she disappointed?

Resulting dialog #1:

S When I was a young kid, I stole some comic books from the local grocery store. → embedding * 1/(1+2)
L Did you ever get caught? → embedding * 2/(1+2)
Σ

Resulting dialog #2:

S When I was a young kid, I stole some comic books from the local grocery store. → embedding * 1/(1+2+4+8)
L Did you ever get caught? → embedding * 2/(1+2+4+8)
S My mother found out and made me walk them back into the store and give them to the manager. That was a humbling experience. → embedding * 4/(1+2+4+8)
L Was she disappointed? → embedding * 8/(1+2+4+8)
Σ

Fig. A.2: Half-decay weighting scheme [16].

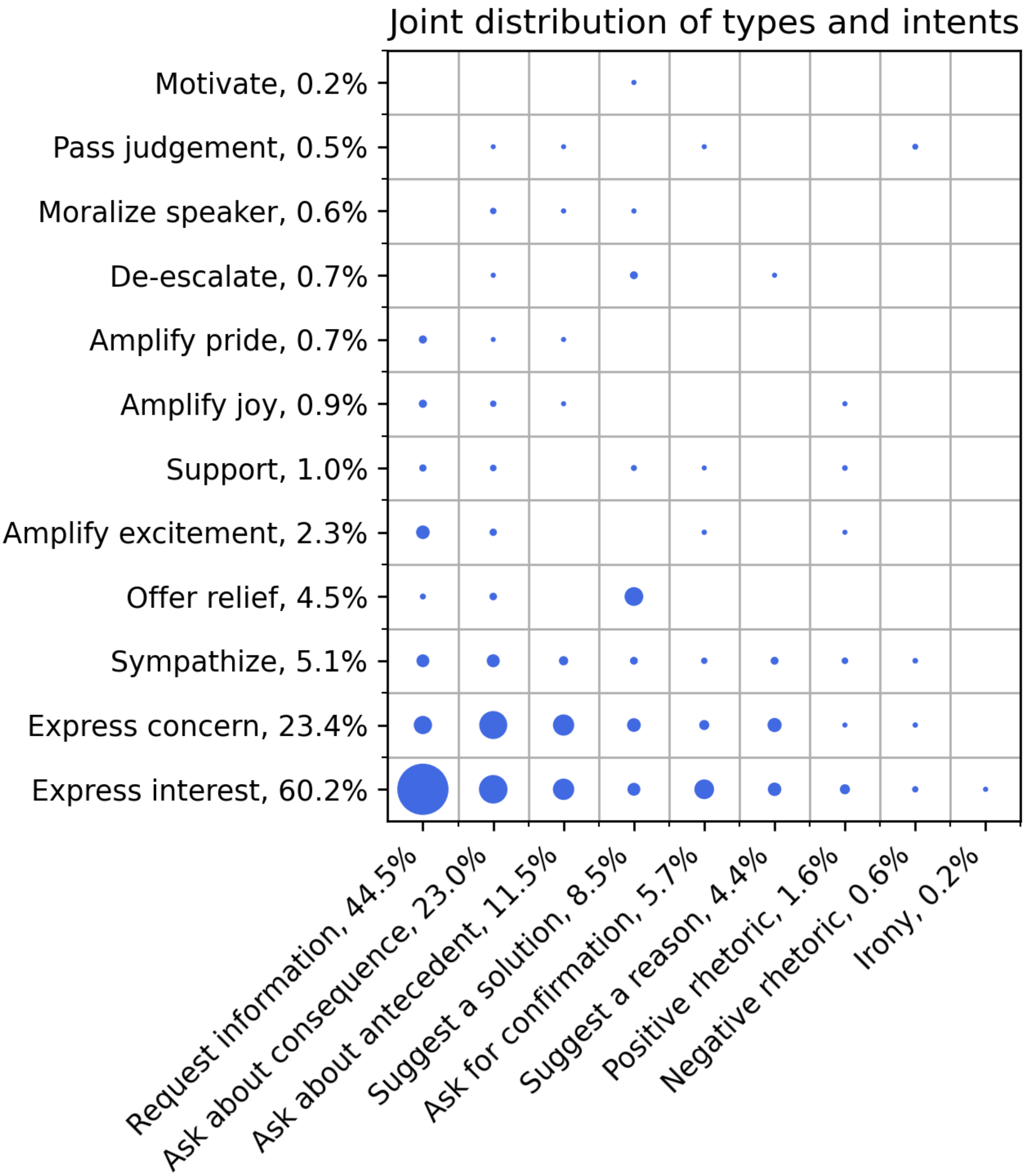


Fig. A.3: Question types and intents joint distribution for the fully annotated dataset.