Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes

Hyunwoo Kim **Byeongchang Kim** **Gunhee Kim**
Department of Computer Science and Engineering
Seoul National University, Seoul, Korea
{hyunw.kim, byeongchang.kim}@vl.snu.ac.kr gunhee@snu.ac.kr
https://vl.snu.ac.kr/projects/focused-empathy

Abstract

Empathy is a complex cognitive ability based on the reasoning of others' affective states. In order to better understand others and express stronger empathy in dialogues, we argue that two issues must be tackled at the same time: (i) identifying which word is the cause for the other's emotion from his or her utterance and (ii) reflecting those specific words in the response generation. However, previous approaches for recognizing emotion cause words in text require sub-utterance level annotations, which can be demanding. Taking inspiration from social cognition, we leverage a generative estimator to infer emotion cause words from utterances with no word-level label. Also, we introduce a novel method based on pragmatics to make dialogue models focus on targeted words in the input during generation. Our method is applicable to any dialogue models with no additional training on the fly. We show our approach improves multiple best performing dialogue agents on generating more focused empathetic responses in terms of both automatic and human evaluation.

1 Introduction

Empathy is one of the hallmarks of social cognition. It is an intricate cognitive ability that requires high-level reasoning on other's affective states. The intensity of expressed empathy varies depending on the depth of reasoning. According to Sharma et al. (2020), weak empathy is accompanied by generic expressions such as "*Are you OK?*" or "*It's just terrible, isn't it?*", while stronger empathy reflects the other's specific situation: "*How is your headache, any better?*" or "*You must be worried about the job interview*". In order to respond with stronger empathy, two issues must be tackled: reasoning (i) where to focus on the interlocutor's utterance (for the reason behind the emotion) and (ii) how to generate utterances that focus on such words.

Firstly, which words should we focus on when empathizing with others? As empathy relates to other's emotional states, the reasons behind emotions (*emotion cause*) should be identified. Imagine you are told "*I got a gift from a friend last vacation!*" with a joyful face. The likely words that can be the causes of his/her happiness are "*gift*" and "*friend*". On the other hand, "*vacation*" has less to do with the emotion. If you respond "*How was your vacation?*", the interlocutor may think you are not interested; rather, it is better to say "*Wow, what was the gift?*" or "*Your friend must really like you.*" by focusing on the emotion cause words.

We humans do not rely on word-level supervision for such affective reasoning. Instead, we put ourselves in the other's shoes and simulate what it would be like. *Perspective-taking* is this act of considering an alternative point of view for a given situation. According to cognitive science, *perspective-taking* and *simulation* are key components in empathetic reasoning (Davis, 1983; Batson et al., 1991; Ruby and Decety, 2004). Taking inspiration from these concepts, we propose to train a generative emotion estimator for simulating the other's situation and identifying emotion cause words.

Secondly, after reasoning which words to focus on, the problem of how to generate focused responses still remains. Safe responses that can be adopted to any situations might hurt other's feelings. Generated utterances need to convey the impression that concerns the specific situation of the interlocutor. Such communicative reasoning is studied in the field of computational pragmatics. The Rational Speech Acts (RSA) framework (Frank and Goodman, 2012) formulates communication between speaker and listener as probabilistic reasoning. It has been applied to many tasks to increase the informativeness of generated text grounded on inputs (Andreas and Klein, 2016; Fried et al., 2018; Cohn-Gordon and Goodman, 2019; Shen et al., 2019). That is, RSA allows the input to be more reflected in the generated output.

However, controlling the RSA framework to re-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2227–2240
November 7–11, 2021.

flect specific parts of the input remains understudied. We introduce a novel method for the RSA framework to make models focus on targeted words in the interlocutor's utterance during generation.

In summary, we recognize emotion cause words in dialogue utterances with no word-level labels and generate stronger empathetic responses focused on them without additional training. Our major contributions are as follows:

(1) We identify emotion cause words in dialogue utterances by leveraging a generative estimator. Our approach requires no additional emotion cause labels other than the emotion label on the whole sentence, and outperforms other baselines.

(2) We introduce a new method of controlling the Rational Speech Acts framework (Frank and Goodman, 2012) to make dialogue models better focus on targeted words in the input context to generate more specific empathetic responses.

(3) For evaluation, we annotate emotion cause words in emotional situations from the validation and test set of EmpatheticDialogues dataset (Rashkin et al., 2019). We publicly release our EMOCAUSE evaluation set for future research.

(4) Our approach improves model-based empathy scores (Sharma et al., 2020) of three recent dialogue agents, MIME (Majumder et al., 2020), DodecaTransformer (Shuster et al., 2020), and Blender (Roller et al., 2021) on EmpatheticDialogues. User studies also show that our approach improves human-rated empathy scores and is more preferred in A/B tests.

2 Related Work

Empathetic dialogue modeling. Incorporating user sentiment is one of early attempts for empathetic conversation generation (Siddique et al., 2017; Shi and Yu, 2018). Rashkin et al. (2019) collect a large-scale English empathetic dialogue dataset named EmpatheticDialogues. The dataset is now adopted in other dialogue corpus such as DodecaDialogue (Shuster et al., 2020) and BST (Smith et al., 2020). As a result, pretrained large dialogue agents such as DodecaTransformer (Shuster et al., 2020) and Blender (Roller et al., 2021) now show empathizing capabilities. Empathy-specialized dialogue models are another stream of research. Diverse architectures have been adopted, including emotion recognition (Lin et al., 2020), mixture of experts (Lin et al., 2019), emotion mimicry (Majumder et al., 2020) and persona (Zhong et al., 2020). Li et al. (2020) use lexicon to extract emotion-related words from utterances and feed them to a GAN-based agent.

We aim to improve both pretrained large dialogue agents and empathy-specialized ones by making them focus on emotion cause words in context.

Emotion Cause (Pair) Extraction. The emotion cause extraction (ECE) task predicts causes in text spans, given an emotion. Cause spans have been collected from Chinese microblogs and news (Gui et al., 2014, 2016), English novels (Gao et al., 2017), and English dialogues (Poria et al., 2020). Xia and Ding (2019) propose a task of extracting pairs of both emotion and its cause spans. Previous works tackle these tasks via supervised learning with question-answering (Gui et al., 2017), joint-learning (Chen et al., 2018), co-attention (Li et al., 2018), and regularization (Fan et al., 2019).

Compared to those tasks, we recognize emotion cause words with no word-level labels using a generative estimator. Our method does not require word-level labels other than the emotion labels of the whole sentences. We then generate more specific empathetic responses focused on them.

Rational Speech Acts (RSA) framework. The RSA framework (Frank and Goodman, 2012) has been applied to many NLP tasks including referencing (Andreas and Klein, 2016; Zarrieß and Schlangen, 2019), captioning (Vedantam et al., 2017; Cohn-Gordon et al., 2018), navigating (Fried et al., 2018), translation (Cohn-Gordon and Goodman, 2019), summarization (Shen et al., 2019), and dialogue (Kim et al., 2020). It can improve informativeness of generated utterances better grounded on inputs (*e.g.* images, texts).

Compared to previous use of RSA, we propose an approach that can control the models to focus on targeted words from the given input.

3 Identifying Emotion Cause Words with Generative Emotion Estimation

Our approach consists of two steps: (i) recognizing emotion cause words from utterances with no word-level labels (§3), and (ii) generating empathetic responses focused on those words (§4). In this section, we first train a generative emotion estimator to identify emotion cause words.

3.1 Why Generative Emotion Estimator?

We leverage a *generative* model by taking inspiration from *perspective-taking* (*i.e. simulating* one-

self in other's shoes) to reason emotion causes; not requiring word-level labels. Our idea is to estimate the emotion cause weight of each word in the utterance while satisfying the following three desiderata.

(1) Do not require word-level supervision for learning to identify emotion cause words in the utterances. Humans do not need word-level labels to infer the probable causes associated with the other's emotion during conversation.

(2) Simulate the observed interlocutor's situation within the model. *Simulation theory* (ST) from cognitive science explains that this mental imitation helps understanding the internal mental states of others (Gallese et al., 2004). Much evidence for ST is found from neuroscience including mirror neurons (Rizzolatti and Craighero, 2004), action-perception coupling (Decety and Chaminade, 2003), and empathetic perspective-taking (Ruby and Decety, 2004).

(3) Reason other's internal emotional states in Bayesian fashion. Studies from cognitive science argue that human reasoning of other's affective states and minds can be described via Bayesian inference (Griffiths et al., 2008; Ong et al., 2015; Saxe and Houlihan, 2017; Ong et al., 2019).

Interestingly, a generative emotion estimator (GEE), which models $P(C, E) = P(E)P(C|E)$ with text sequence (*e.g.* context) C and emotion E, satisfies all the above conditions. First, the generative estimator computes the likelihood of C by *generating* C given E, which can be viewed as a *simulation* of C. Second, it estimates $P(E|C)$ via Bayes' rule. Finally, the association between the emotion estimate and each word comes for free by using the likelihood of each words; without using any word-level supervision. We use BART (Lewis et al., 2020) to implement a GEE.

3.2 Training to Model Emotional Situations

Dataset. To train our GEE, we leverage the EmpatheticDialogues (Rashkin et al., 2019), a multi-turn English dialogue dataset where the speaker talks about an emotional situation and the listener expresses empathy. An example is shown in Table 1. The emotion and the situation sentence are only visible to the speaker. Situations are collected beforehand by asking annotators to recall related experiences for a given emotion label. The dataset includes a rich suite of 32 emotion labels that are evenly distributed.

Emotion: Grateful **Situation**: I was grateful when my mother visited me for my birthday.
Speaker: It was my birthday, my mom came to surprise me. **Listener**: Aw that's so nice, how did she surprise you? **Speaker**: She showed up to my house and brought me a cake. **Listener**: Cakes! yessss winning. :)

Table 1: A dialogue example in EmpatheticDialogues.

Emotion: Joyful **GEE**: • I got accepted into a masters program in neuroscience.
Emotion: Angry **GEE**: • I was so mad at my cousin. He stole my daughters stuff.
Emotion: Grateful **GEE**: • The night my dad got me a new car was a magical time.

Table 2: Example of sampled outputs from our generative emotion estimator (GEE) using Nucleus sampling.

Training. Given an emotion label E, GEE is trained to generate its corresponding emotional situation $C = \{w_1, ..., w_T\}$, where w_i is a word. As a result, our GEE learns the joint probability $P(C, E)$. The trained GEE shows perplexity of 13.6 on the test situations of EmpatheticDialogues.

3.3 Recognizing Emotions

Once trained, GEE can predict $P(E|C = c)$ for a word sequence c (*e.g.* utterance) using Bayes' rule:

$$P(E|C = c) \propto P(C = c|E)P(E). \quad (1)$$

We compute the likelihood $P(C = c|E)$ by GEE's generative ability as described in §3.1. Since emotions in EmpatheticDialogues are almost evenly distributed, we set the prior $P(E)$ to a uniform distribution. Finally, we find the emotion with the highest likelihood of the given sequence c.

We comparatively report the emotion classification accuracy of GEE in Appendix.

3.4 Weakly Supervised Emotion Cause Word Recognition

We introduce how GEE can recognize emotion cause words solely based on emotion labels without word-level annotations. For a given word sequence $c = \{w_1, w_2, ..., w_T\}$ (*e.g.* utterance), GEE can reason the association $P(W|E = \hat{e})$ of each word w_t in the sequence c to the recognized emotion $\hat{e}$ in Bayesian fashion:

$$P(W|E = \hat{e}) \propto P(E = \hat{e}|W)P(W). \quad (2)$$

The emotion likelihood is computed as

$$P(\hat{e}|W = w_t) = \mathbb{E}_{w_{<t}}[P(\hat{e}|w_t, w_{<t})] \approx \frac{P(w_t|\hat{e}, w_{<t})P(\hat{e})}{\sum_{e' \in \mathcal{E}} P(w_t|e', w_{<t})P(e')}, \quad (3)$$

where $w_{<t}$ is the partial utterance up to time step $t-1$. Since computing the expectation over all possible partial utterance $w_{<t}$ is intractable, we approximate it by a single sample. We build set $\mathcal{E}$ to include $\hat{e}$ and emotions with the two lowest probability of $P(E|C = c)$ when recognizing emotion in Eq.(1). We assume the marginal $P(W)$ is uniform. We choose the top-k words reasoned by GEE as emotion cause words, and focus on them during empathetic response generation.

4 Controlling the RSA framework for Focused Empathetic Responses

We introduce how to control the Bayesian Rational Speech Acts (RSA) framework (Frank and Goodman, 2012) to focus on targeted words in the context during response generation. We first preview the basics of RSA for dialogues (§4.1). We then present how to control the RSA with word-level focus (§4.2), where our major contribution lies. Figure 1 is the overview of our method.

4.1 The Rational Speech Acts Framework

Applying the RSA framework is computing the posterior of the dialogue agent's output distribution over words each time step. Hence, it is applicable to any existing pretrained dialogue agents on the fly, with no additional training.

The RSA framework formulates communication as a reference game between speaker and listener. Based on recursive Bayesian formulation, the speaker (*i.e.* dialogue model) reasons about the listener's belief of what the speaker is referring to. We follow the approach of Kim et al. (2020) for adopting RSA to dialogues. Our goal here is to update a base speaker S_0 to a pragmatic speaker S_1 that focuses more on the emotion cause words in dialogue context c (*i.e.* dialogue history).

Base Speaker S_0. Let c and u_t denote dialogue context and the output word of the model at time step t, respectively. The base speaker S_0 is a dialogue agent that outputs u_t for a dialogue context and partial utterance $u_{<t}$: $S_0(u_t|c, u_{<t})$. As described, one can use any dialogue models for S_0.

Pragmatic Listener L_0. The pragmatic listener is a posterior distribution over which dialogue context the speaker is referring to. It is defined in terms of the base speaker S_0 and a prior distribution $p_t(C)$ over the context in Bayesian fashion:

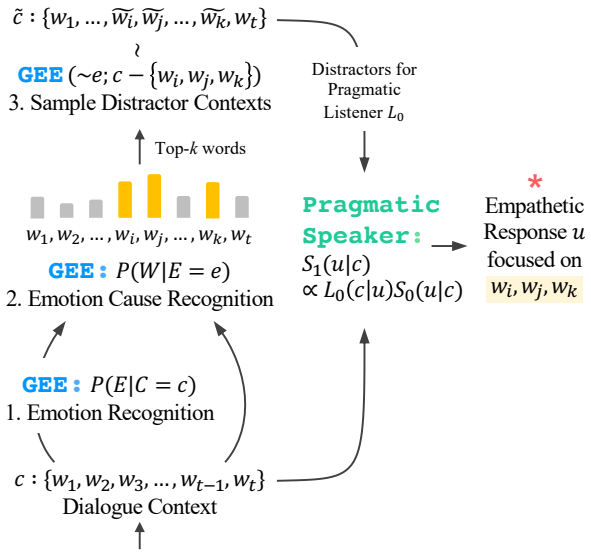

Figure 1: Overview of our method, consisting of emotion recognition (§3.3), emotion cause word recognition (§3.4), distractor context sampling (§4.2), and pragmatic generation (§4.1). GEE denotes our generative emotion estimator.

$$L_0(c|u_{\leq t}, p_t) \propto \frac{S_0(u_t|c, u_{<t})^\beta \times p_t(c)}{\sum_{c' \in \mathcal{C}} S_0(u_t|c', u_{<t})^\beta \times p_t(c')}. \quad (4)$$

The *shared world* $\mathcal{C}$ is a finite set comprising the given dialogue context c and other contexts (coined as *distractors*) different from c. Our contribution lies in how to build world $\mathcal{C}$ to endow the dialogue agent with controllability to better focus on targeted words, which we discuss in §4.2. We update prior $p_{t+1}(C)$ with L_0 from time step t as follows: $p_{t+1}(C) = L_0(C|u_{\leq t}, p_t)$. β is the rationality parameter which controls how much the base speaker's distribution is taken into account. We note that L_0 is simply a distribution computed in Bayesian fashion, not another separate model.

Pragmatic Speaker S_1. Integrating L_0 with S_0, we obtain the pragmatic speaker S_1:

$$S_1(u_t|c, u_{<t}) \propto L_0(c|u_{\leq t}, p_t)^\alpha \times S_0(u_t|c, u_{<t}). \quad (5)$$

Since the pragmatic speaker S_1 is forced to consider how its utterance is perceived by the listener (via L_0), it favors words that have high likelihood of the given context c over other contexts in shared world C. Similar to Eq. 4, α is the rationality parameter for S_1.

4.2 Endowing Word-level Control for RSA to Focus on Targeted Words in Context

We aim to make dialogue models focus on targeted words from the input (*i.e.* dialogue context) during generation via shared world $\mathcal{C}$. The shared world C consists of the given dialogue context c and other distractor contexts. It is used for computing the likelihood of the given context c in Eq. 4.

Previous works of RSA in NLP manually (or randomly) select pieces of text (*e.g.* sentences) entirely different from the given input (Cohn-Gordon and Goodman, 2019; Shen et al., 2019; Kim et al., 2020). In our context, it means distractors will be totally different contexts from c in the dataset. For example, when given a context "*I got a gift from my friend.*", a distractor might be "*Today, I have an exam at school.*". Although such type of distractors helps improve the specificity of the model's generated outputs, it is difficult to finely control which words the models should be specific about.

Our core idea is to build distractors by replacing the emotion cause words in c with different words via sampling with GEE. It can enhance the controllability of the RSA by making models focus on targeted words (*e.g.* emotion cause words recognized by GEE) from the dialogue context.

For a dialogue context $c = \{w_1, ..., w_T\}$ where w_i is a word, GEE outputs top-k emotion cause words regarding the recognized emotion $\hat{e}_1$ from context c, denoted by $\mathcal{W}_{gee}$. Next, we concatenate the least likely n emotions from GEE with the context c removing the top-k emotion cause words: $[\hat{e}_{-1}, ..., \hat{e}_{-n}; c - \mathcal{W}_{gee}]$, which is input to GEE. We then sample different words $(\tilde{w}_i, \tilde{w}_j, \ldots, \tilde{w}_k)$ from GEE's output in place of $\mathcal{W}_{gee}$ to construct a distractor $\tilde{c}$. For example, given a context c "*I was sick from the flu*" and "*sick, flu*" as the top-2 emotion cause words, a sampled distractor $\tilde{c}$ can be "*I was laughing from the relief*". We use these altered contexts $\{\tilde{c}_1, ..., \tilde{c}_i\}$ as distractors for the shared world $\mathcal{C}$ in the pragmatic listener L_0 (Eq. 4). We set n and cardinality of world $\mathcal{C}$ to 3 (*i.e.* $\mathcal{C} = \{c, \tilde{c}_1, \tilde{c}_2\}$). We run experiments and find the best k $(= 5)$ (see Appendix).

The only difference between the original context c and the sampled distractor $\tilde{c}$ is those emotion cause words. The pragmatic speaker S_1 (Eq. 5) prefers to generate words that have a higher likelihood of the given context c (including the original emotion cause words $\mathcal{W}_{gee}$) than the distractor context $\tilde{c}$. As a result, the pragmatic agent can generate utterances more focused on those original emotion cause words.

	#Emotion	Label	#Label/Utt	#Utt
RECCON	8	Span	2.0	6.3K
EMOCAUSE (Ours)	32	Word	2.3	4.6K

Table 3: Statistics of the EMOCAUSE evaluation set compared to RECCON (Poria et al., 2020). Utt denotes utterance.

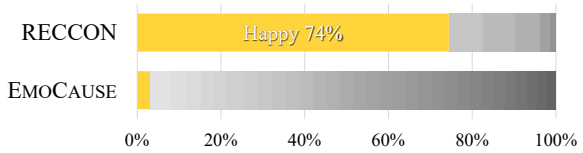

Figure 2: Emotion ratio of RECCON and our EMOCAUSE evaluation set.

Emotion	Situation
Surprised	Man, I did not expect to see a bear on the road today.
Afraid	I have to take a business trip next week, I'm not looking forward to flying.
Sad	I feel sad that I am spending so much time this late on the internet.
Joyful	I'm excited I get to go to Disney in October!

Table 4: Examples of annotated emotion cause words.

Embarrassed	pant, fell, dropped, people, tripped, toilet
Nostalgic	old, childhood, memory, friend, back
Trusting	friend, gave, best, daughter, money, phone
Anxious	job, interview, exam, new, presentation
Proud	graduated, daughter, college, son, school
Disappointed	not, son, car, failed, get, job, hard, friend

Table 5: The most frequent cause words for each emotion. Other emotions can be found in Appendix.

5 EMOCAUSE: Emotion Cause Words Evaluation Set

5.1 Collecting Annotations

To evaluate the performance of GEE, we annotate emotion cause words[1] in the situations of validation and test set in EmpatheticDialogues (Rashkin et al., 2019) (§3.2). Using Amazon Mechanical Turk, we ask three workers to vote which words (*e.g.* object, action, event, concept) in the situation

[1]As existing works annotate emotion cause spans for a given emotion label, we also coin our annotations as emotion cause words. However, in terms of "*causality*", we note that the *true* cause of the given emotion can be annotated only by the original annotator of the emotion label.

sentence are the cause words to the given emotion. Since explicit emotion words in the text (*e.g.* happy, disappointed) are not cause words of emotion, we discourage workers from selecting them.

Annotators are required to have a minimum of 1000 HITs, 95% HIT approval rate, and be located at one of [AU, CA, GB, NZ, US]. We pay the annotators $0.15 per description. To further ensure quality, only annotators who pass the qualification test are invited to annotate. Nevertheless, speculations for emotion causes are subjective and can vary among annotators. Therefore, we use *only* unanimously selected words (*i.e.* earning *all* three votes) to ensure maximum objectivity.

5.2 Analysis

We analyze the characteristics of our emotion cause words in the EMOCAUSE evaluation set. In Table 3 and Figure 2, we compare the basic statistics of our annotation set and RECCON (Poria et al., 2020), which is an English dialogue dataset annotating emotion cause spans on the DailyDialog (Li et al., 2017) and IEMOCAP (Busso et al., 2008) with a total of 8 emotions. Since our EMOCAUSE is based on emotional situations from an empathetic dialogue dataset (Rashkin et al., 2019), emotion causes play a more important role than in casual conversations from RECCON. While 74% of RECCON's labels belong to a single emotion *happy*, EMOCAUSE provides a balanced range of 32 emotions labels. Therefore, our evaluation set presents a wider variety than RECCON. Table 4 shows some examples of the annotated emotion cause words.

Table 5 reports the most frequent cause words for some emotions. We find "embarrassing" events happen frequently in *toilets* and in front of *people*. "Proud" and "disappointed" are closely related to *children*. Interestingly, *phones* are associated with "trusting", which may be due to smartphones containing sensitive personal information. More examples and results can be found in Appendix.

6 Experiments

We first evaluate our generative emotion estimator (GEE) on weakly-supervised emotion cause word recognition (§6.2). We then show our new controlling method for the RSA framework can improve best performing dialogue agents to generate more empathetic responses by better focusing on targeted emotion cause words (§6.3).

6.1 Datasets and Experiment Setting

EmpatheticDialogues (ED) (Rashkin et al., 2019). This dataset is an English empathetic dialogue dataset with 32 diverse emotion types (§3.2). The task is to generate empathetic responses (*i.e.* responses from the listener's side in Table 1) when only given the dialogue context (*i.e.* history) without emotion labels and situation descriptions. It contains 24,850 conversations partitioned into training, validation, and test set by 80%, 10%, 10%, respectively. We additionally annotate cause words for the given emotion for all situations in the validation and test set of EmpatheticDialogues (§5).

EmoCause (§5). We compare our GEE with four methods that can recognize emotion cause words with no word-level annotations: random, RAKE (Rose et al., 2010), EmpDG (Li et al., 2020), and BERT (Devlin et al., 2019). For random, we randomly choose words as emotion causes. RAKE is an automatic keyword extraction algorithm based on the word frequency and degree of co-occurrences. EmpDG leverages a rule-based method for capturing emotion cause words using EmoLex (Mohammad and Turney, 2013), a large-scale lexicon of emotion-relevant words. Finally, we train BERT for emotion classification with the emotion labels in ED. For BERT, we select the words with the largest averaged weight of BERT's last attention heads for the classification token (*i.e.* `[CLS]`). More details can be found in Appendix.

Dialogue models for base speakers. We experiment our approach on three recent dialogue agents: MIME (Majumder et al., 2020), DodecaTransformer (Shuster et al., 2020), and Blender (Roller et al., 2021). MIME is a dialogue model explicitly targeting empathetic conversation by leveraging emotion mimicry. We select MIME, since it reportedly performs better than other recent empathy-specialized models (Rashkin et al., 2019; Lin et al., 2019) on EmpatheticDialogues. DodecaTransformer is a multi-task model trained on all DodecaDialogue tasks (Shuster et al., 2020) (*i.e.* 12 dialogue tasks including ED, image and knowledge grounded ones) and finetuned on ED. Blender is one of the state-of-the-art open domain dialogue agent (Roller et al., 2021) trained on Blended-SkillTalk dataset (Smith et al., 2020) which adopts contexts from ED. We also finetune Blender on ED. For all models, we use the default hyperparameters from the official implementations. More details are in Appendix.

Model	Top-1 Recall	Top-3 Recall	Top-5 Recall
Human	41.3	81.1	95.0
Random	10.7	30.6	48.5
EmpDG	13.4	36.2	49.3
RAKE	12.7	35.8	55.0
BERT-Attention	13.8	40.6	61.2
GEE (Ours)	**17.3**	**48.1**	**68.4**

Table 6: Comparison of emotion cause word recognition performance between our generative emotion estimator (GEE), random, RAKE (Rose et al., 2010), EmpDG (Li et al., 2020), and BERT on our EMOCAUSE evaluation set (§5).

Automatic evaluation metrics. For weakly-supervised emotion cause word recognition, we report the Top-$1, 3, 5$ recall scores.

For EmpatheticDialogues, we report coverage and two scores for specific empathy expressions (Exploration, Interpretation) measured by pretrained empathy identification models (Sharma et al., 2020). The coverage score refers to the average number of emotion cause words included in the model's generated response.

The (i) Exploration and (ii) Interpretation are metrics for expressed empathy in text, introduced by Sharma et al. (2020). They both require responses to focus on the interlocutor's utterances and to be specific. (i) Explorations are expressions of active interest in the interlocutor's situation, such as "*What happened?*" or "*So, did you pass the chemistry exam?*". The latter is rated as a stronger empathetic response since it asks specifically about the interlocutor's situation. (ii) Interpretations are expressions of acknowledgments or understanding of the interlocutor's emotion or situation, such as "*I know your feeling.*" or "*I also had to speak in front of such audience, made me nervous.*" Expressions of specific understanding are considered to be more empathetic. RoBERTa models (Liu et al., 2019) that are separately pretrained for each metric rate each agent's response by returning values of 0, 1, or 2. Higher scores indicate stronger empathy.

6.2 Weakly-Supervised Emotion Cause Word Recognition

Table 6 compares the recall of different methods on our EMOCAUSE evaluation set (§5). Our GEE outperforms all other alternative methods. RAKE performs better than EmpDG that uses a fixed lexicon of emotion-relevant words. Compared to RAKE, methods leveraging dense word representations (*i.e.* BERT, GEE) perform better. Selecting words by BERT's attention weights does not attain better performance on capturing emotion cause words than GEE. The gap between GEE and other methods widens when the number of returned words from models is more than one (*i.e.* Top-$3, 5$).

We also evaluate human performance to measure the difficulty of the task. We randomly sample 100 examples from the test set and ask a human evaluator to select five best guesses for the emotion causes. As the performance gap between GEE and human is significantly large, there is much room for further improvement in weakly-supervised emotion cause recognition.

Model	Coverage	Exploration ↑	Interpretation ↑
MIME (Majumder et al., 2020)			
S_0	0.22	0.12	0.05
Plain S_1	0.22	0.23	0.10
Focused S_1	**0.24**	**0.24**	**0.13**
DodecaTransformer (Shuster et al., 2020)			
S_0	0.34	0.25	0.24
S_0+Emotion	0.34	0.21	0.20
Plain S_1	0.43	0.30	0.23
Focused S_1	**0.49**	**0.32**	**0.30**
Blender (Roller et al., 2021)			
S_0	0.35	0.28	0.22
S_0+Emotion	0.34	0.31	0.20
Plain S_1	0.43	0.37	0.21
Focused S_1	**0.54**	**0.38**	**0.26**

Table 7: Comparison of our approach (Focused S_1) with other speakers on EmpatheticDialogues (Rashkin et al., 2019). Exploration, and Interpretation scores are evaluated by pretrained RoBERTa models from Sharma et al. (2020).

6.3 Empathetic Response Generation

Results on Automatic Evaluation. Table 7 reports the performance of different dialogue agents on EmpatheticDialogues (Rashkin et al., 2019) with automatic evaluation metrics. Our *Focused* S_1 significantly outperforms the base model S_0 in terms of Interpretation and Exploration scores that measure more focused and specific empathetic expression. We also test the plain pragmatic method (*Plain* S_1) that use random distractors as in previous works (Cohn-Gordon et al., 2018; Kim et al., 2020). The *Focused* S_1 consistently outperforms *Plain* S_1 on Interpretation score with similar or better Exploration scores. The *Focused* S_1 models show higher coverage scores than other mod-

Model	Empathy ↑	Relevance ↑	Fluency ↑
MIME (Majumder et al., 2020)			
S_0	2.94	3.17	2.75
Focused S_1	**3.09**	**3.21**	**2.83**
DodecaTransformer (Shuster et al., 2020)			
S_0	2.53	3.47	2.56
Focused S_1	**2.71**	**3.57**	**2.75**
Blender (Roller et al., 2021)			
S_0	2.91	3.12	3.46
Focused S_1	**3.00**	**3.25**	**3.57**

Table 8: Comparison of our approach (Focused S_1) with base speakers (S_0) on human rating.

Model	Win	Lose	Tie
MIME (Majumder et al., 2020)			
Focused S_1 vs S_0	**46.7**%	20.0%	33.3%
DodecaTransformer (Shuster et al., 2020)			
Focused S_1 vs S_0	**42.1**%	28.8%	29.1%
Blender (Roller et al., 2021)			
Focused S_1 vs S_0	**44.6**%	37.4%	18.0%

Table 9: Comparison of our approach (*Focused* S_1) with base speakers (S_0) on A/B test for empathetic response generation. The win and lose rates are based on *Focused* S_1.

Model	Win	Lose	Tie
Blender (Roller et al., 2021)			
Focused S_1 vs Plain S_1	**39.6**%	32.3%	28.0%

Table 10: Comparison of our approach (*Focused* S_1) with plain pragmatic approach (*Plain* S_1) on A/B test for reflecting the given emotion cause words. The win and lose rates are based on *Focused* S_1.

Emotion: Angry **Context**: I just moved to this neighborhood and some dumb criminals shot my neighbors and ran into the woods!
Emotion cause words from GEE: dumb, neighbors, shot, moved, ran
S_0: That's terrible. Did you get a chance to talk to them? **Plain S_1**: Oh my god, that is horrible! Where are they? **Focused S_1**: Oh my god, that is horrible! Did they get *caught*?
Emotion: Trusting **Context**: I'm a big believer in listening to your gut. Sometimes I might wonder where my intuition is leading me, but it always seems to be spot on.
Emotion cause words from GEE: believer, leading, intuition, wonder, spot
S_0: That's great. I'm glad you have a good intuition. **Plain S_1**: Good for you. I have a gut feeling about a lot of things. **Focused S_1**: That's great! I'm a big *believer* in *intuition* too.
Emotion: Proud **Context**: We were living overseas at the times – the most impressive part was crossing the Grand Canal over a pontoon of boats to finish the marathon ;-)
Emotion cause words from GEE: impressive, Canal, Grand, part, finish
S_0: Wow, that sounds like a lot of fun! I'm glad you had a good time! **Plain S_1**: Wow, that sounds amazing! I bet you were so excited! **Focused S_1**: Wow, that sounds amazing! I've never been to *Europe*, but I've heard of it.

Table 11: Examples of recognized emotion cause words from our GEE and responses from S_0 and Focused S_1. We underline words where our Focused S_1 reflects the emotion cause words returned by GEE.[2]

els, indicating they more reflect the context's emotion cause words in responses. As MIME is only trained on EmpatheticDialogues, its Exploration and Interpretations scores are lower than models pretrained on other larger corpus. As a result, we find our approach is effective in both large pretrained open domain dialogue models and empathy-specialized one.

We also finetune DodecaTransformer and Blender with explicit emotion information (S_0+Emotion). Following Rashkin et al. (2019), we concatenate the ground-truth emotion label to the dialogue context during training. At inference, the top predicted emotion from GEE is used. We find the Interpretation or Exploration scores of S_0+Emotion models drop. Thus, simply adding emotion information is insufficient to make models focus more on the interlocutor's emotional event.

Results on Human Evaluation. We conduct user study and A/B test via Amazon Mechanical Turk. We randomly sample 100 test examples, each rated by three unique human evaluators. Following previous works (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020), we rate empathy, relevance, and fluency of generated responses.

[2]Since Grand Canal is a famous tourist attraction in Venice, Italy, the word 'Europe' is closely related to it. We note that there is another famous Grand Canal in China. This might be a bias in BART, since it is trained on English datasets.

Given the dialogue context and model's generated response, evaluators are asked to rate each criterion in a 4-point Likert scale, where higher scores are better. We also run human A/B test to directly compare the *Focused* S_1 and base S_0. We ask three unique human evaluators to vote which response is more empathetic. They can select *tie* if both responses are thought to be equal.

Table 8 and 9 summarizes the averaged human rating and A/B test results on MIME (Majumder et al., 2020), DodecaTransformer (Shuster et al., 2020), and Blender (Roller et al., 2021). Our *Focused* S_1 agents are rated more empathetic and relevant to the dialogue context than the base agent S_0, with better fluency. Also, users prefer responses from our *Focused* S_1 agent over those from the base agent S_0. The inter-rater agreement (Krippendorff's α) for human rating and A/B test are 0.26 and 0.27, respectively; implying fair agreement.

In addition to the coverage score in Table 7, we run A/B test on Blender (Roller et al., 2021) to compare the *Focused* S_1 and *Plain* S_1 for reflecting the given emotion cause words in the responses. We random sample 200 test examples and ask three unique human evaluators to vote which response is more focused on the given emotion cause words from the context.

Table 10 is the result of A/B test for focused response generation on Blender (Roller et al., 2021). Users rate that responses from *Focused* S_1 more reflect the emotion cause words than those from the *Plain* S_1 approach. Thus, both quantitative and qualitative results show that our *Focused* S_1 approach helps dialogue agents to effectively generate responses focused on given target words.

Examples of the recognized emotion cause words from GEE and generated responses are in Table 11. Our *Focused* S_1 agent's responses reflect the context's emotion cause words returned from our GEE, implicitly or explicitly.

7 Conclusion

We studied how to use a generative estimator for identifying emotion cause words from utterances based solely on emotion labels without word-level labels (*i.e.* weakly-supervised emotion cause word recognition). To evaluate our approach, we introduce EMOCAUSE evaluation set where we manually annotated emotion cause words on situations in EmpatheticDialogues (Rashkin et al., 2019). We release the evaluation set to the public for future research. We also proposed a novel method for controlling the Rational Speech Acts (RSA) framework (Frank and Goodman, 2012) to make models generate empathetic responses focused on targeted words in the dialogue context. Since the RSA framework requires no additional training, our approach is orthogonally applicable to any pretrained dialogue agents on the fly. An interesting direction for future work will be reasoning how the interlocutor would react to the model's empathetic response. Such reasoning is an essential part for expressing empathy.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research was supported by Samsung Research Funding Center of Samsung Electronics under project number SRFCIT2101-01. The compute resource and human study are supported by Brain Research Program by National Research Foundation of Korea (NRF) (2017M3C7A1047860). Gunhee Kim is the corresponding author.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. In *EMNLP*.

C Daniel Batson, Judy G Batson, Jacqueline K Slingsby, Kevin L Harrell, Heli M Peekna, and R Matthew Todd. 1991. Empathic Joy and the Empathy-Altruism Hypothesis. *Journal of Personality and Social Psychology*, 61(3):413.

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.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint Learning for Emotion Classification and Emotion Cause Detection. In *EMNLP*.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in Machine Translation: A Method to Reduce Meaning Loss. In *NAACL-HLT*.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically Informative Image Captioning With Character-level Inference. In *NAACL-HLT*.

Mark H Davis. 1983. Measuring Individual Differences in Empathy: Evidence for a Multidimensional

Approach. *Journal of Personality and Social Psychology*, 44(1):113.

Jean Decety and Thierry Chaminade. 2003. Neural Correlates of Feeling Sympathy. *Neuropsychologia*, 41(2):127–138.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. 2019. A Knowledge Regularized Hierarchical Approach for Emotion Cause Analysis. In *EMNLP*.

Michael C Frank and Noah D Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *NeurIPS*.

Vittorio Gallese, Christian Keysers, and Giacomo Rizzolatti. 2004. A Unifying View of the Basis of Social Cognition. *Trends in Cognitive Sciences*, 8(9):396–403.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of NTCIR-13 ECA Task. In *NTCIR-13*.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian Models of Cognition. *Cambridge Handbook of Computational Cognitive Modeling*, pages 115–126.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A Question Answering Approach for Emotion Cause Extraction. In *EMNLP*.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-Driven Emotion Cause Extraction with Corpus Construction. In *EMNLP*.

Lin Gui, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou. 2014. Emotion Cause Detection with Linguistic Construction in Chinese Weibo Text. In *NLPCC*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness. In *EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *COLING*.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A Co-Attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness. In *EMNLP*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of Empathetic Listeners. In *EMNLP*.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2020. CAiRE: An End-to-End Empathetic Chatbot. In *AAAI*.

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:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *EMNLP*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv:1705.06476*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2015. Affective Cognition: Exploring Lay Theories of Emotion. *Cognition*, 143:141–162.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science*, 11(2):338–357.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2020. Recognizing Emotion Cause in Conversations. *arXiv:2012.11820*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*.

Giacomo Rizzolatti and Laila Craighero. 2004. The Mirror-Neuron System. *Annual Review of Neuroscience*, 27:169–192.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2021. Recipes for Building an Open-Domain Chatbot. In *EACL*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic Keyword Extraction from Individual Documents. *Text mining: applications and theory*, 1:1–20.

Perrine Ruby and Jean Decety. 2004. How would You Feel Versus How do You Think She would Feel? A Neuroimaging Study of Perspective-taking with Social Emotions. *Journal of Cognitive Neuroscience*, 16(6):988–999.

Rebecca Saxe and Sean Dae Houlihan. 2017. Formalizing Emotion Concepts within a Bayesian Model of Theory of Mind. *Current Opinion in Psychology*, 17:15–21.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP*.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically Informative Text Generation. In *NAACL-HLT*.

Weiyan Shi and Zhou Yu. 2018. Sentiment Adaptive End-to-End Dialog Systems. In *ACL*.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The Dialogue Dodecathlon: Open-domain Knowledge and Image Grounded Conversational Agents. In *ACL*.

Farhad Bin Siddique, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Zara Returns: Improved Personality Induction and Adaptation by an Empathetic Virtual Agent. In *ACL*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In *ACL*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-Aware Captions from Context-Agnostic Supervision. In *CVPR*.

Rui Xia and Zixiang Ding. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In *ACL*.

Sina Zarrieß and David Schlangen. 2019. Know What You Don't Know: Modeling a Pragmatic Speaker that Refers to Objects of Unknown Categories. In *ACL*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *EMNLP*.

A Implementation Details

Weakly-supervised emotion cause word recognition. We use *rake-nltk*[3] to implement RAKE (Rose et al., 2010), and the official code of EmpDG[4] from the authors (Li et al., 2020). We respectively finetune BERT-based-uncased (Devlin et al., 2019) for BERT-Attention and BART-large (Lewis et al., 2020) for our generative emotion estimator (GEE). We set a learning rate to 3e-5 for BERT-Attention and 1e-5 for GEE. Other than the learning rate, we follow the default hyperparameters in ParlAI framework[5] (Miller et al., 2017). We select the best performing checkpoint using the Top-1 recall for emotion cause word recognition on the validation set. We run experiments 5 times with different random seeds and report averaged scores on Table 6.

Dialogue models. We use MIME (Majumder et al., 2020), DodecaTransformer (Shuster et al., 2020), and Blender 90M (Roller et al., 2021) as dialogue models for base speakers. For MIME, we use the codes and pretrained weights of the authors' official implementation[6] as is. For DodecaTransformer and Blender, we use the ParlAI framework with the default hyperparameters and finetune them on EmpatheticDialogues (Rashkin et al., 2019). We select the best performing checkpoint via perplexity on the validation set.

During inference, we use greedy decoding and set RSA parameter α and β to 2.0 and 0.9 for MIME, 3.0 and 0.9 for DodecaTransformer, and 4.0 and 0.9 for Blender. We select the best performing α and β from the candidates of $[1.0, 2.0, 3.0, 4.0]$ and $[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ with one trial for each. Inference on the test set of EmpatheticDialogues takes 0.4 hours with Blender 90M base speaker.

Evaluation metrics. To compute Exploration and Interpretation scores (Sharma et al., 2020), we separately finetune RoBERTa-base for each score using the author's official code[7].

Sensitivity to k of top-k emotion cause words. In all experiments, we use $k = 5$, which is found by validation with $k = 1, 2, 4, 8$ using Blender (Roller et al., 2021) on EmpatheticDialogues (Rashkin et al., 2019). Table 12 summarizes the results.

[3] https://github.com/csurfer/rake-nltk
[4] https://github.com/qtli/EmpDG
[5] https://parl.ai
[6] https://github.com/declare-lab/MIME
[7] https://github.com/behavioral-data/Empathy-Mental-Health

k	Exploration ↑	Interpretation ↑
1	0.32	0.27
2	0.34	0.29
4	0.35	0.30
8	0.36	0.29

Table 12: Comparison of different k values for top-k emotion cause words on generating empathetic responses in EmpatheticDialogues (Rashkin et al., 2019). Exploration and Interpretation scores are evaluated by pretrained RoBERTa models from Sharma et al. (2020).

Experiments for emotion cause word recognition and emotion classification are run on one NVIDIA Quadro RTX 6000 GPU. Experiments for empathetic response generation are run on two GPUs.

B Emotion Classification

We report the classification performance of emotion classifiers used in empathetic response generation. Table 13 shows the Top-1, 5 emotion classification accuracy for each model. For reference, BERT (Devlin et al., 2019) shows 0.55 and 0.88 for Top-1 and 5 accuracy.

Model	Top-1	Top-5
MoEL (Lin et al., 2019)	0.38	0.74
MIME (Majumder et al., 2020)	0.34	0.77
GEE (Ours)	0.40	0.77

Table 13: Comparison of emotion classification accuracy from different models trained on EmpatheticDialogues (Rashkin et al., 2019).

C Details of EMOCAUSE Evaluation Set

Table 14 shows some selected examples of emotion cause words with given emotion and situation. Table 15 shows Top-10 frequent cause words per emotion. Interestingly, same words can be seen in both positive and negative emotions. For example, we can find the word *interview* on both "Anxious" and "Confident". "Anticipating" and "Disappointed" are closely related to *vacation*. This result shows that understanding the context is one of key prerequisites for emotion cause word recognition.

Emotion: Surprised We just got a new puppy . My older dog knew to let that one out first when I get home from work .
Emotion: Faithful My boyfriend is going out with a bunch of people I do n't know tonight . But I trust him that he will be a good boy .
Emotion: Anticipating I am really waiting on getting my tax returns this year I could use new carpet
Emotion: Trusting I trust my own intuitions when it comes to my health .
Emotion: Embarrassed i was super late for my meeting on tuesday
Emotion: Sad My girlfriend 's cat is sick with Cancer . I do n't think she 's going to make it for much longer and I 'm really shaken up by it .
Emotion: Proud I put in a lot of effort and energy and I found a new job . It 's an online teaching position and I feel so good about myself .
Emotion: Terrified Driving down the highway during a heavy thunderstorm and a car crash happens in front of me where a car flips over .
Emotion: Confident I studied all night for my final exam
Emotion: Guilty I made a really inappropriate joke about someone I work with to other coworkers and it got back to them . I feel really bad about it .

Table 14: Examples of our annotated emotion cause words. Words with background color are selected as emotion cause words by annotators.

Emotion	#Label/Utt	Top-10 frequent emotion cause words
Afraid	2.12	alone, night, spider, house, noise, movie, dark, storm, hurricane, heard
Angry	2.62	car, dog, neighbor, friend, husband, brother, not, stole, hit, kid
Annoyed	2.59	dog, people, cat, work, loud, late, night, sister, neighbor, friend
Anticipating	2.04	new, waiting, vacation, coming, son, job, forward, next, friend, back
Anxious	2.05	interview, job, exam, presentation, big, dentist, going, test, girlfriend, back
Apprehensive	2.11	job, nervous, new, first, interview, driving, moving, car, day, night
Ashamed	2.48	stole, ate, friend, forgot, girlfriend, missed, drunk, bad, money, mistake
Caring	2.49	dog, sick, care, wife, friend, home, helped, puppy, girlfriend, baby
Confident	1.95	exam, studied, job, interview, win, test, well, prepared, good, answer
Content	2.04	life, good, happy, relaxing, watching, weekend, back, breakfast, family, live
Devastated	2.42	dog, passed, died, away, lost, friend, father, job, cancer, cat
Disappointed	2.59	not, son, car, failed, get, hard, job, n't, birthday, vacation
Disgusted	2.47	dog, poop, threw, friend, dead, food, roach, puked, eat, animal
Embarrassed	2.73	pant, fell, dropped, people, tripped, stuck, slipped, toilet, front, friend
Excited	1.95	vacation, new, friend, first, trip, car, puppy, see, won, coming
Faithful	2.09	loyal, girlfriend, husband, year, relationship, boyfriend, family, friend, married, good
Furious	2.58	car, dog, neighbor, hit, broke, without, son, room, accident, cheated
Grateful	2.42	friend, helped, life, job, family, good, help, husband, work, parent
Guilty	2.64	ate, stole, friend, forgot, money, candy, eating, cake, bar, girlfriend
Hopeful	1.91	job, promotion, future, new, better, get, interview, ticket, college, well
Impressed	2.30	friend, daughter, guy, car, new, well, man, brother, world, backflip
Jealous	2.66	friend, car, new, husband, girl, girlfriend, bought, got, boyfriend, won
Joyful	2.18	first, child, wife, friend, family, together, daughter, baby, birthday, trip
Lonely	2.18	friend, alone, moved, husband, family, myself, away, wife, went, left
Nostalgic	2.59	old, childhood, friend, memory, game, school, child, family, back, comic
Prepared	2.00	ready, packed, studied, exam, everything, supply, ingredient, studying, set, all
Proud	2.40	graduated, college, daughter, job, first, son, school, brother, won, new
Sad	2.39	dog, died, passed, away, cat, sick, friend, not, lost, put
Sentimental	2.40	old, picture, passed, photo, dog, childhood, school, away, toy, found
Surprised	2.29	friend, party, birthday, found, baby, car, gift, home, pregnant, won
Terrified	2.28	night, dog, tornado, car, bad, chased, someone, storm, fly, crash
Trusting	2.17	friend, best, daughter, drive, car, brother, sister, card, dog, phone

Table 15: Number of emotion cause words per utterance and Top-10 frequent emotion cause words for each emotion.