Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset

Hannah Rashkin[1*], Eric Michael Smith[2], Margaret Li[2], Y-Lan Boureau[2]
[1] Paul G. Allen School of Computer Science & Engineering, University of Washington
[2] Facebook AI Research
hrashkin@cs.washington.edu, {ems,margaretli,ylan}@fb.com

20191018
Yang

Why EMPATHY

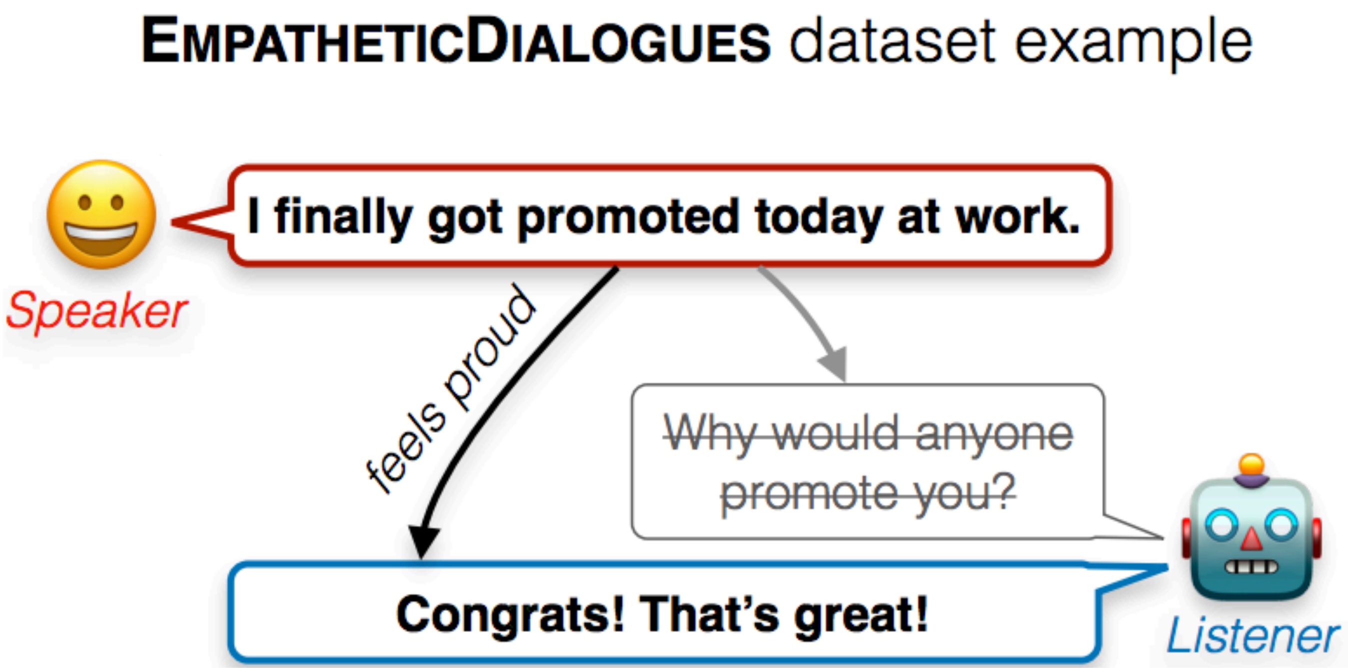


Figure 1: Example where acknowledging an inferred feeling is appropriate

Problem

- Existing chitchat dialogue benchmarks do not capture whether those agents are responding to implicit emotional contexts in an empathetic way

Data Collections

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Figure 2: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own description of a situation when they've felt that way. Then, the speaker tells their story in a conversation with a second worker (the listener).

1. Workers are asked to describe in a 1-3 sentences (19.8 words averagely) a situation based on a feeling label.
2. Each conversation is allowed to be 4-8 utterances long (the average is 4.31 utterances per conversation). The average utterance length was 15.2 words long.
3. 24,850 prompts/conversations from 810 different participants Each conversation is allowed to be 4-8 utterances long

Dataset Statistics

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Figure 2: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own description of a situation when they've felt that way. Then, the speaker tells their story in a conversation with a second worker (the listener).

- Training, val, testing set are respectively 19533 / 2770 / 2547 conversations

Emotion label statistics

Emotion	Most-used speaker words	Most-used listener words	Training set emotion distrib
Surprised	got,shocked,really	that's,good,nice	5.1%
Excited	going,wait,i'm	that's,fun,like	3.8%
Angry	mad,someone,got	oh,would,that's	3.6%
Proud	got,happy,really	that's,great,good	3.5%
Sad	really,away,get	sorry,oh,hear	3.4%
Annoyed	get,work,really	that's,oh,get	3.4%
Grateful	really,thankful,i'm	that's,good,nice	3.3%
Lonely	alone,friends,i'm	i'm,sorry,that's	3.3%
Afraid	scared,i'm,night	oh,scary,that's	3.2%
Terrified	scared,night,i'm	oh,that's,would	3.2%
Guilty	bad,feel,felt	oh,that's,feel	3.2%
Impressed	really,good,got	that's,good,like	3.2%
Disgusted	gross,really,saw	oh,that's,would	3.2%
Hopeful	i'm,get,really	hope,good,that's	3.2%
Confident	going,i'm,really	good,that's,great	3.2%
Furious	mad,car,someone	oh,that's,get	3.1%
Anxious	i'm,nervous,going	oh,good,hope	3.1%
Anticipating	wait,i'm,going	sounds,good,hope	3.1%
Joyful	happy,got,i'm	that's,good,great	3.1%
Nostalgic	old,back,really	good,like,time	3.1%
Disappointed	get,really,work	oh,that's,sorry	3.1%
Prepared	ready,i'm,going	good,that's,like	3%
Jealous	friend,got,get	get,that's,oh	3%
Content	i'm,life,happy	good,that's,great	2.9%
Devastated	got,really,sad	sorry,oh,hear	2.9%
Embarrassed	day,work,got	oh,that's,i'm	2.9%
Caring	care,really,taking	that's,good,nice	2.7%
Sentimental	old,really,time	that's,oh,like	2.7%
Trusting	friend,trust,know	good,that's,like	2.6%
Ashamed	feel,bad,felt	oh,that's,i'm	2.5%
Apprehensive	i'm,nervous,really	oh,good,well	2.4%
Faithful	i'm,would,years	good,that's,like	1.9%

Figure 3: Distribution of conversation labels within EmpatheticDialogues training set and top 3 content words used by speaker/listener per category.

- The distribution is also evenly (sample distribution)

Modeling

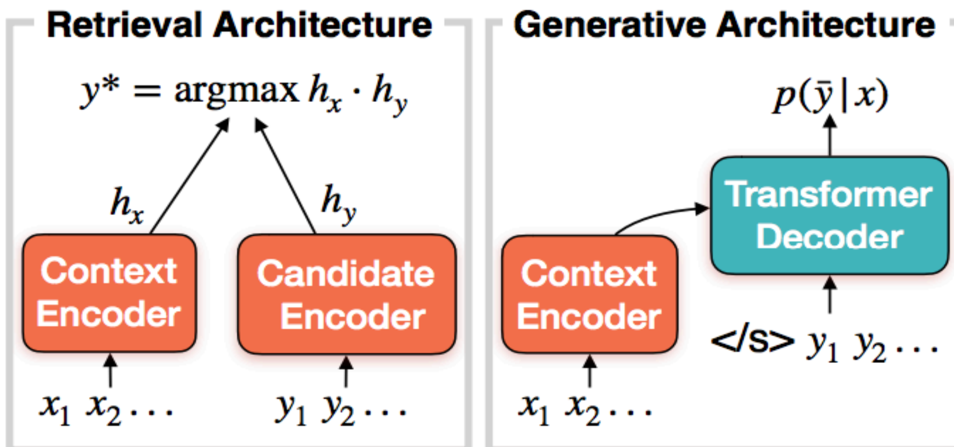


Figure 4: Dialogue generation architectures used in our experiments. The context of concatenated previous utterances is tokenized into $x_1, x_2, \cdots$, and encoded into vector h_x by the context encoder. *Left:* In the retrieval set-up, each candidate y is tokenized into $y_1, y_2, \cdots$ and encoded into vector h_y by the candidate encoder. The system outputs the candidate y^* that maximizes dot product $h_x \cdot h_y$. *Right:* In the generative set-up, the encoded context h_x is used as input to the decoder to generate start symbol $</s>$ and tokens $y_1, y_2, \cdots$. The model is trained to minimize the negative log-likelihood of target sequence $\bar{y}$ conditioned on context.

Model details

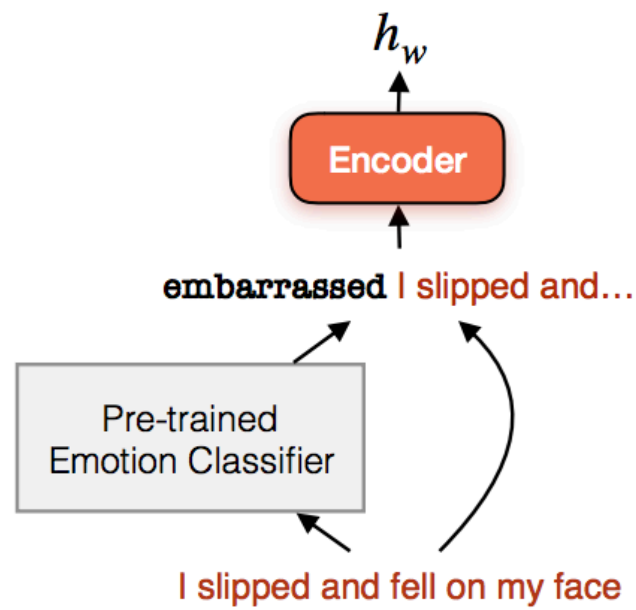
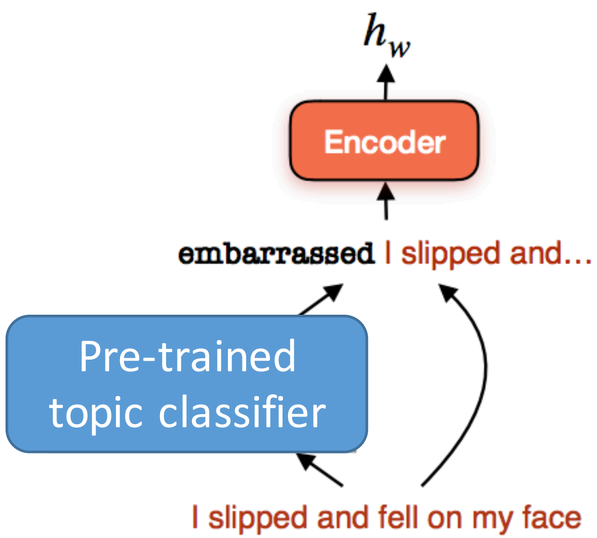


Figure 5: Incorporating additional supervised information, here from an emotion classification task. An input sequence (either a dialogue context or a candidate) is run through a pre-trained classifier, and the top k output labels are prepended to the sequence, which is then run through the corresponding (context or candidate) encoder to output a hidden representation h_w (either h_x or h_y) as in the base setting.

- Train a classifier to predict the emotion label from the description of the situation written by the Speaker before the dialogue for the training set dialogues of

Supervision from a more distant task would be help as well?



- also experiment with a classifier trained on the 20-Newsgroup dataset (Joachims, 1996), for topic classification (TOPICPREPEND-1).

Evaluation

• For the retrieval systems, we additionally compute p@1,100, the accuracy of the model at choosing the correct response out of a hundred randomly selected examples in the test set.

•Evaluate Relevance, Fluency, Empathy: did the responses show understanding of the feelings of the person talking about their experience? (1: not at all, 3: somewhat, 5: very much)

•Source candidate during inference: in addition to EMPATHETICDIALOGUES, the DailyDialog (Li et al., 2017) training set and up to a million utterances from a dump of 1.7 billion Reddit conversations are included

Quantitative Results

		Retrieval		Retrieval w/ BERT		Generative	
Model	Candidate Source	P@1,100	AVG BLEU	P@1,100	AVG BLEU	PPL	AVG BLEU
Pretrained	R	-	4.10	-	4.26	27.96	5.01
	ED	43.25	5.51	49.94	5.97	-	-
Fine-Tuned	ED	**56.90**	5.88	65.92	**6.21**	**21.24**	**6.27**
	ED+DD	-	5.61	-	-	-	-
	ED+DD+R	-	4.74	-	-	-	-
EmoPrepend-1	ED	56.31	5.93	**66.04**	6.20	24.30	4.36
TopicPrepend-1	ED	56.38	**6.00**	65.96	6.18	25.40	4.17

Table 1: Automatic evaluation metrics on the test set. Pretrained: model pretrained on a dump of 1.7 billion REDDIT conversations (4-layer Transformer architecture, except when specified BERT). Fine-Tuned: model fine-tuned over the EMPATHETICDIALOGUES training data (Sec. 4.2). EmoPrepend-1, Topic-Prepend1: model incorporating supervised information from an external classifiers, as described in Sec. 4.3. Candidates come from REDDIT (R), EMPATHETICDIALOGUES (ED), or DAILYDIALOG (DD). P@1,100: precision retrieving the correct test candidate out of 100 test candidates. AVG BLEU: average of BLEU-1,-2,-3,-4. PPL: perplexity. All automatic metrics clearly improve with in-domain training on utterances (Fine-Tuned vs. Pretrained), other metrics are inconsistent. *Bold: best performance for that architecture.*

Human Results

	Model	Candidate	Empathy	Relevance	Fluency
Retrieval	*Pre-trained*	R	2.82 ± 0.12	3.03 ± 0.13	4.14 ± 0.10
		R+ED	3.16 ± 0.14	3.35 ± 0.13	4.16 ± 0.11
		ED	3.45 ± 0.12	3.55 ± 0.13	4.47 ± 0.08
	Fine-tuned	ED	$\mathbf{3.76 \pm 0.11}$	3.76 ± 0.12	4.37 ± 0.09
	EmoPrepend-1	ED	3.44 ± 0.11	3.70 ± 0.11	4.40 ± 0.08
	TopicPrepend-1	ED	3.72 ± 0.12	$\mathbf{3.91 \pm 0.11}$	$\mathbf{4.57 \pm 0.07}$
Retrieval w/ BERT	*Pre-trained*	R	3.06 ± 0.13	3.29 ± 0.13	4.20 ± 0.10
		R+ED	3.49 ± 0.12	3.62 ± 0.12	4.41 ± 0.09
		ED	3.43 ± 0.13	3.49 ± 0.14	4.37 ± 0.10
	Fine-tuned	ED	3.71 ± 0.12	3.76 ± 0.12	4.58 ± 0.06
	EmoPrepend-1	ED	3.93 ± 0.12	3.96 ± 0.13	4.54 ± 0.09
	TopicPrepend-1	ED	$\mathbf{4.03 \pm 0.10}$	$\mathbf{3.98 \pm 0.11}$	$\mathbf{4.65 \pm 0.07}$
Generative	*Pre-trained*	–	2.31 ± 0.12	2.21 ± 0.11	3.89 ± 0.12
	Fine-Tuned	–	$\mathbf{3.25 \pm 0.12}$	$\mathbf{3.33 \pm 0.12}$	4.30 ± 0.09
	EmoPrepend-1	–	3.16 ± 0.12	3.19 ± 0.13	4.36 ± 0.09
	TopicPrepend-1	–	3.09 ± 0.13	3.12 ± 0.13	$\mathbf{4.41 \pm 0.08}$
Gold Response	–	–	4.19 ± 0.10	4.55 ± 0.07	4.68 ± 0.06

Table 2: Human ratings. Fine-tuning on ED and using ED candidates generally improves scores, especially on Empathy, with minimal retraining. Additional external supervision (Prepend) improves the Empathy and Relevance scores for BERT-based models. Bold: best score for that group. Italics: reference model for the group.

Fine-tuning using their dataset

	P @1,100		BLEU	
Model	DD	R	DD	R
Pretrained	39.04	58.95	6.65	1.43
Fine-Tuned	44.58	56.25	7.14	1.64
Pretrained-Large	42.28	61.60	6.94	1.42
Fine-Tuned-Large	48.96	58.71	7.42	1.73

Table 6: Performance of the retrieval-based pretrained model and retrieval-based models fine-tuned on ED data for next utterance prediction in other datasets, with both context and candidates from the same dataset (R=Reddit, DD=DailyDialog).

Case study

Context	Setup	Model	Cand	Generated Response
Someone came to my door with a gun the other day.	Gen	Pretrained	-	I'm not sure if you're joking or not.
	Gen	Fine-Tuned	-	Oh no! What did you do?
	Gen	EmoPrepend	-	Oh no! Did you have a weapon?
	Gen	TopicPrepend	-	I'm glad to hear that.
My son failed his exams! Can you believe it! I don't know what to do!	Ret-BERT	Pretrained	R	Hopefully his mistakes can help you know what not to do on the tests.
	Ret-BERT	Pretrained	ED	Oh no! I ' m so sorry! I hope you could pull your grade up!
	Ret-BERT	Fine-Tuned	ED	Oh no! Did he not study?
	Ret-BERT	EmoPrepend	ED	Im sorry about that! Did your son start school already?
	Ret-BERT	TopicPrepend	ED	Oh no! Does he have the option of taking it again?

Table 4: Examples of model responses. Shading is provided for better legibility.

Conclusion

- Introduce a new dataset of 25k dialogues with emotion label
- Experiments show that using this dataset to provide retrieval candidates or fine-tune conversation models leads to responses that are evaluated as more empathetic

End