I KNOW THE FEELING: LEARNING TO CONVERSE WITH EMPATHY

Yang Zhao
AI Lab, NLP Center

Why EMPATHY (Empathy:共情)

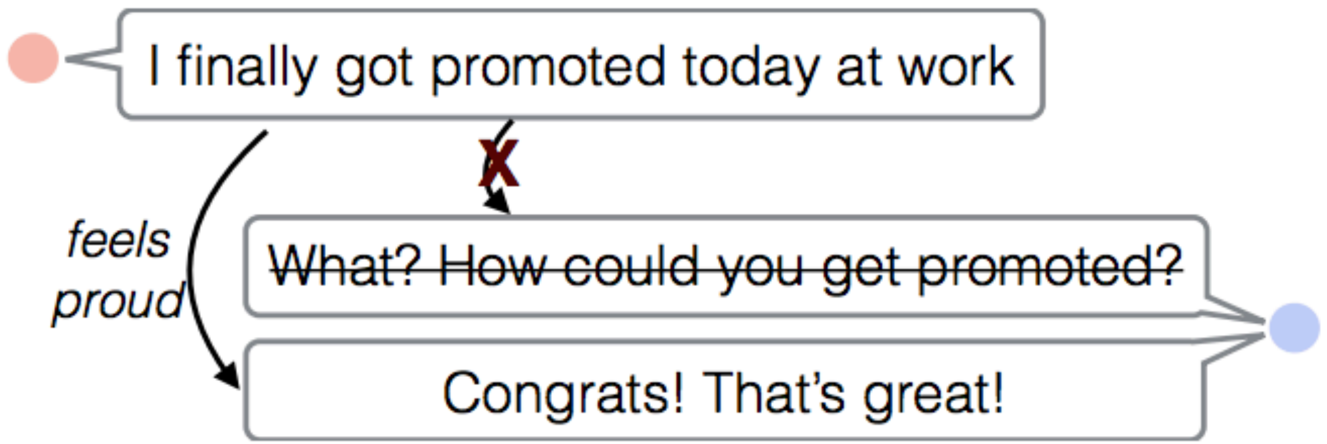


Figure 1: Example conversation where acknowledging an inferred feeling might be appropriate

Problem

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

Data Collection

Table 1: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own prompt based on a situation when they've felt that way. Then, the speaker tells their story in a conversation with a second worker (the listener).

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!

Distribution of 32 Labels

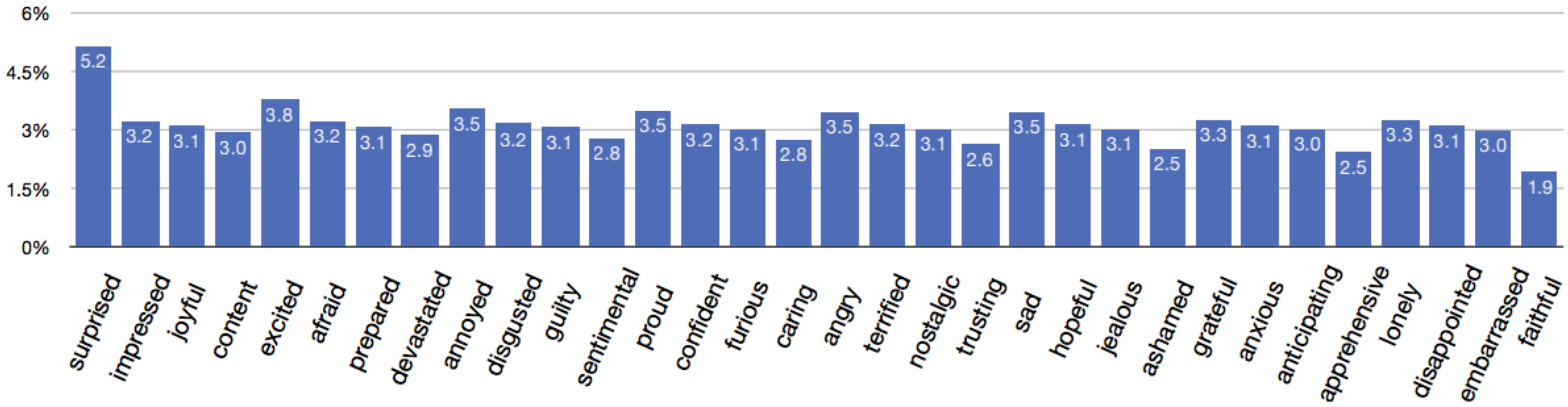


Figure 2: Distribution of situation/conversation labels within EMPATHETICDIALOGUES. Percentages per class are also listed in the appendix.

- 24,850 prompts/conversations from 810 different participants
- Each conversation is allowed to be 4-8 utterances long
- The average utterance length is 15.2 words long

Modeling

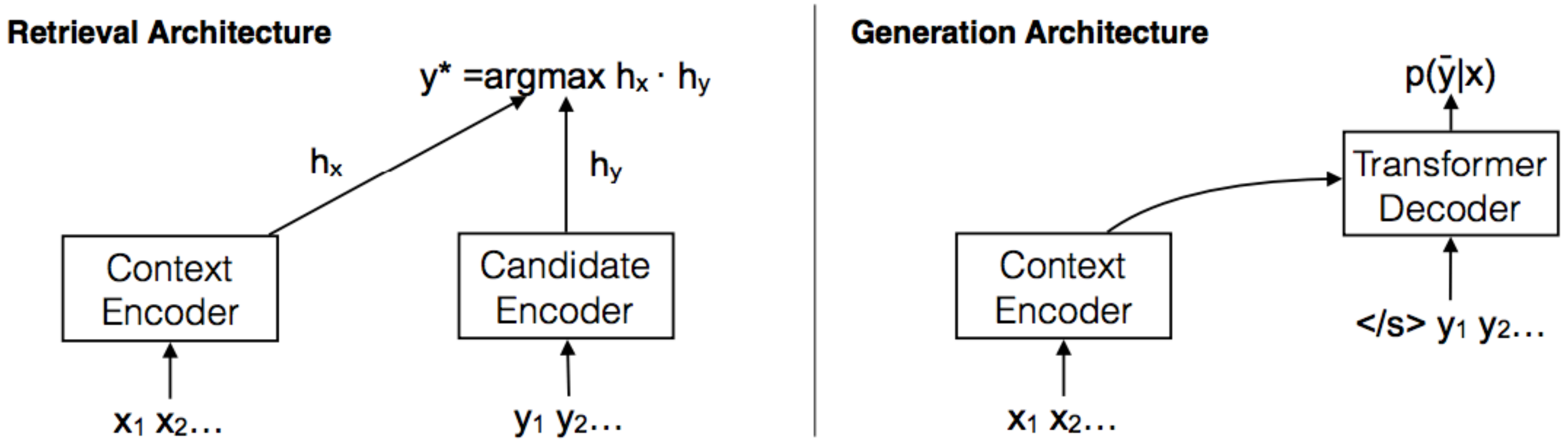
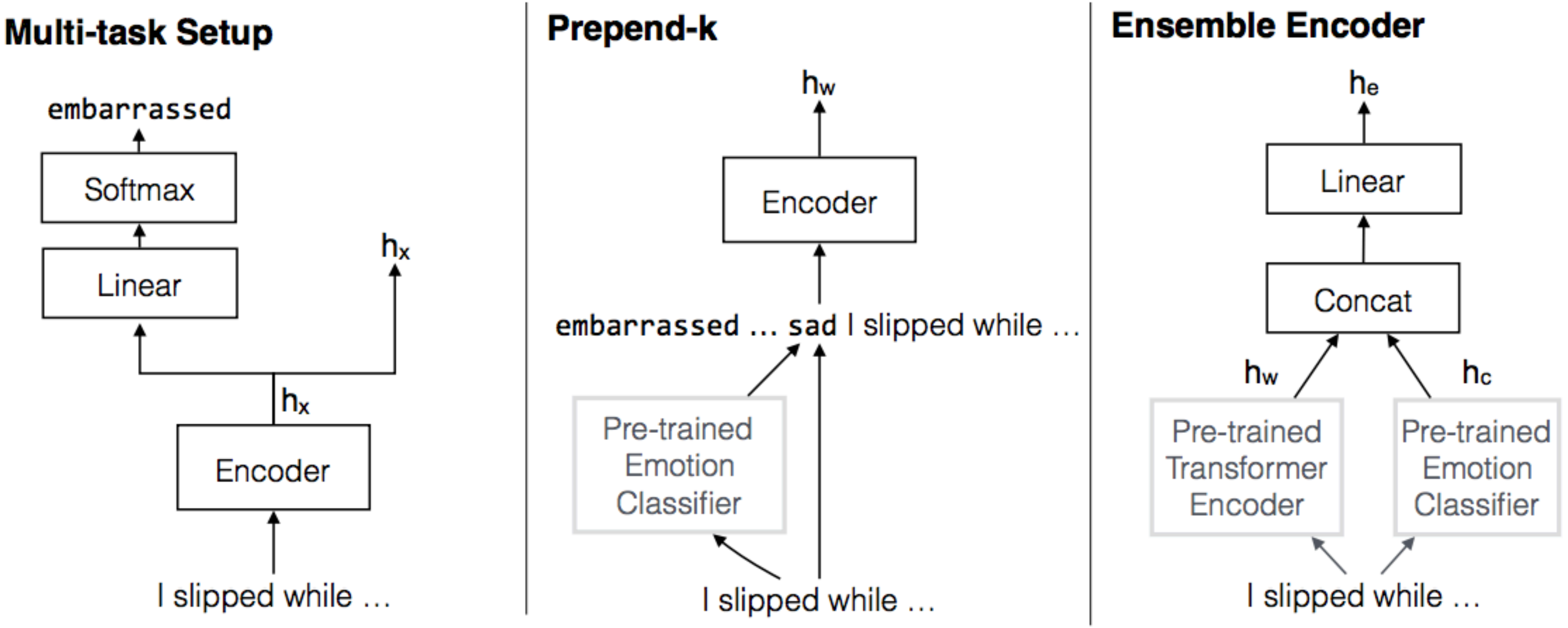
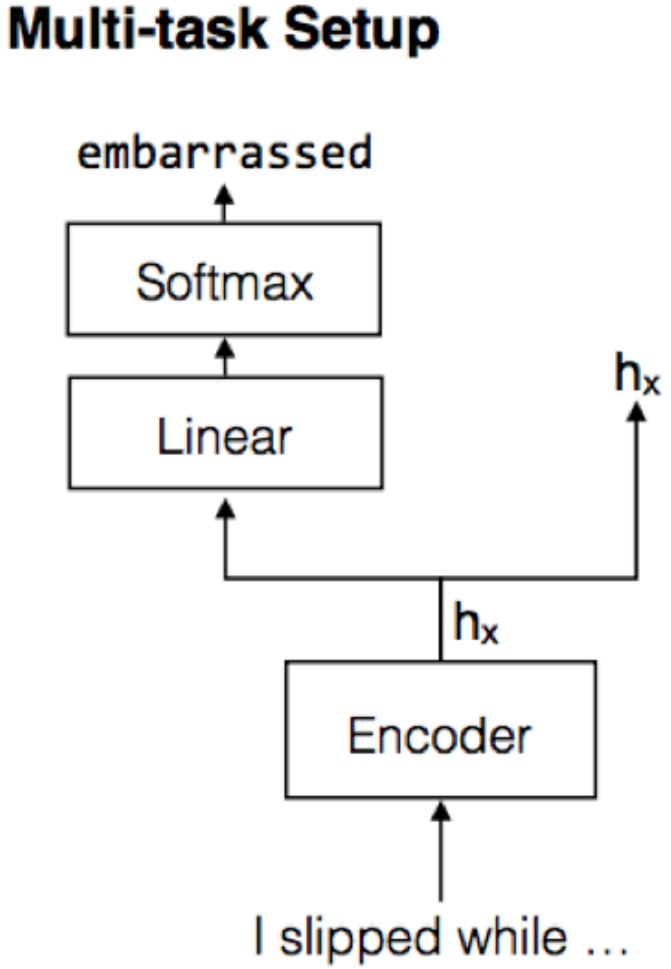


Figure 3: Dialogue generation architectures used in our experiments. The context of concatenated previous utterances is tokenized into $x_1, x_2, \cdots$, and encoded into vector h_x by the context encoder. *Left:* In the retrieval set-up, each candidate y is tokenized into $y_1, y_2, \cdots$ and encoded 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 x.

Three models

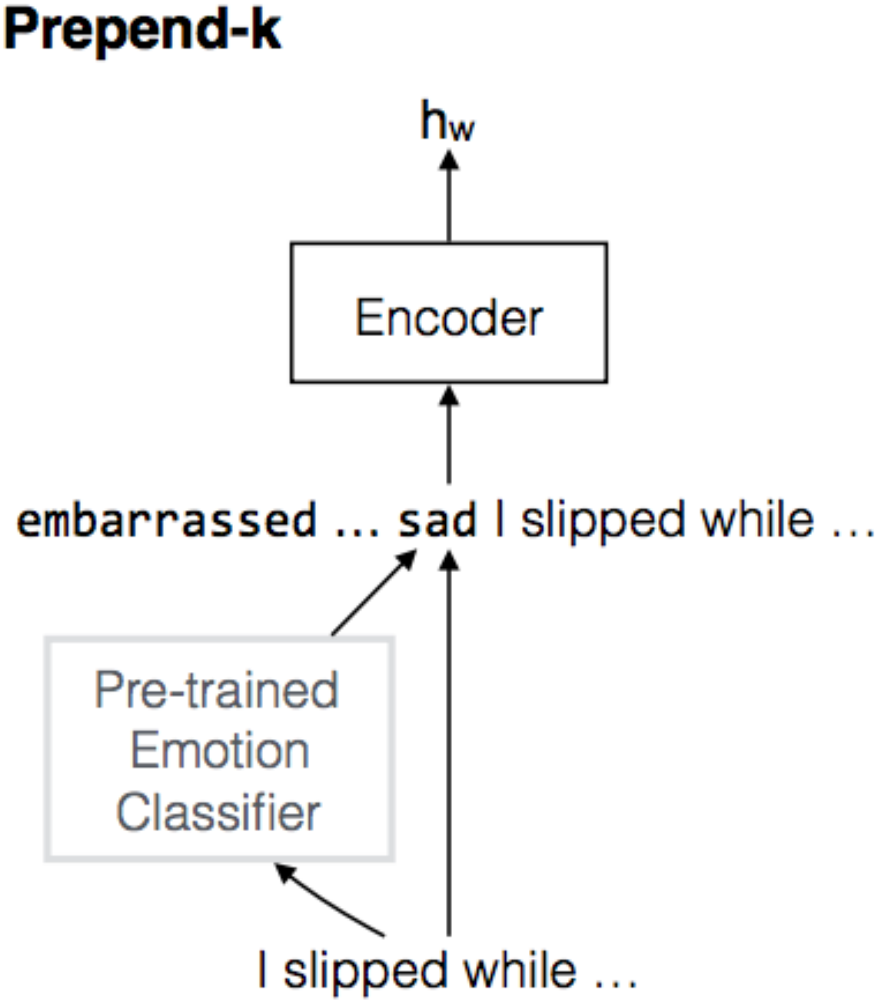


Multi-Task Objective



- alter the objective function to also optimize for predicting the given emotion label.

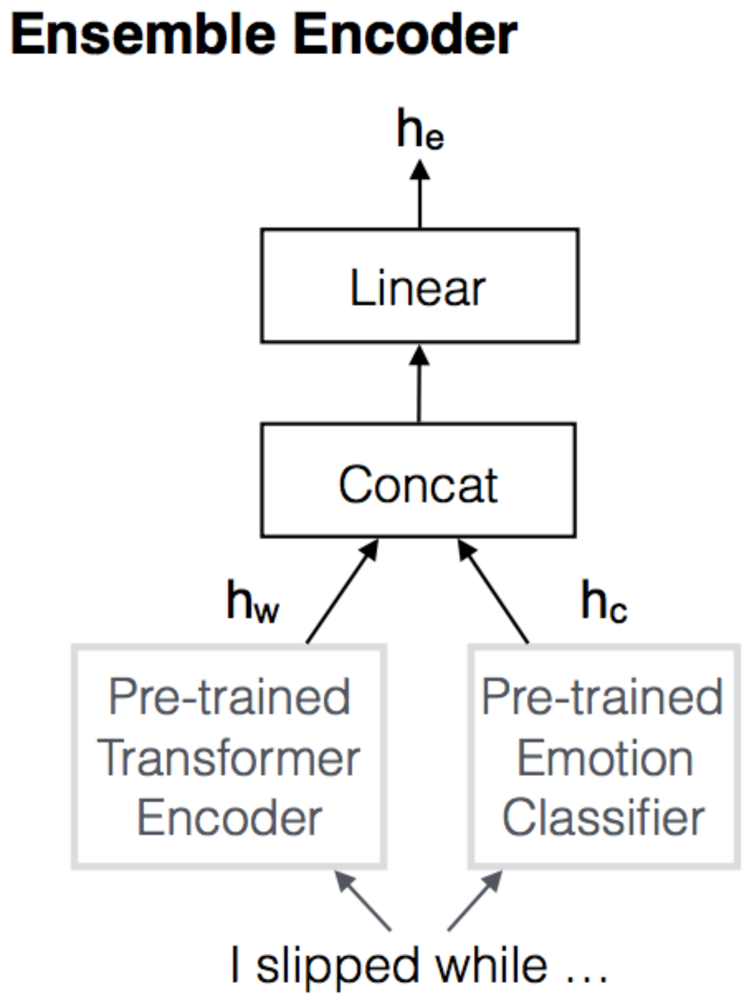
Prepending Top-K Emotion Predictions



- explicitly add the best emotion predictions from a simple emotion classifier to the input text.
- use a fastText model trained to predict the emotion label from the description of the situation written by the Speaker before the dialogue for the training set dialogues.

I finally got promoted! $\longrightarrow$ `proud excited joyful` I finally got promoted!

Ensemble of Encoders



- Take an off-the-shelf classifier for emotion prediction, DeepMoji from Felbo et al. (2017) with the weights as released by the authors, ENSEM-DM

- Use a version of the same DeepMoji architecture that is first re-trained on the situation descriptions from our training data, ENSEM-DM+.

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

Experimental Results

Table 2: Automatic evaluation metrics on the test set. Pretrained: basic transformer model pre-trained on a dump of 1.7 billion REDDIT conversations. Base: model fine-tuned over the EMPATHETICDIALOGUES training data. Remaining rows: models incorporating emotion supervised information, as described in Sec. 4.2. Candidates come from REDDIT (R), EMPATHETICDIALOGUES (ED), or DAILYDIALOGUES (DD). All automatic metrics clearly improve with in-domain training (Base vs. Pretrained), but the effects of adding supervised information are inconsistent on the automated metrics, although ensembling with a deep emotion classifier consistently improves generation.

	Retrieval			Generation	
Model	P @1,100	Candidate Source	AVG BLEU	PPL	AVG BLEU
Pretrained	43.25	R	4.1	27.96	5.01
	-	ED	5.51	-	-
Base	**56.90**	ED	5.88	21.24	6.27
	-	ED+DD	5.61	-	-
	-	ED+DD+R	4.74	-	-
MULTITASK	55.73	ED	6.18	24.07	5.42
PREPEND-1	56.31	ED	5.93	24.30	4.36
PREPEND-3	55.75	ED	**6.23**	23.96	2.69
PREPEND-5	56.35	ED	6.18	25.40	5.56
ENSEM-DM	52.71	ED	6.03	**19.05**	**6.83**
ENSEM-DM+	52.35	ED	6.04	19.1	6.77
ENSEM-TRAN	51.69	ED	5.88	19.21	6.41

	Retrieval			Generation	
Model	P @1,100	Candidate Source	AVG BLEU	PPL	AVG BLEU
Pretrained	43.25	R	4.1	27.96	5.01
	-	ED	5.51	-	-
Base	**56.90**	ED	5.88	21.24	6.27
	-	ED+DD	5.61	-	-
	-	ED+DD+R	4.74	-	-
MULTITASK	55.73	ED	6.18	24.07	5.42
PREPEND-1	56.31	ED	5.93	24.30	4.36
PREPEND-3	55.75	ED	**6.23**	23.96	2.69
PREPEND-5	56.35	ED	6.18	25.40	5.56
ENSEM-DM	52.71	ED	6.03	**19.05**	**6.83**
ENSEM-DM+	52.35	ED	6.04	19.1	6.77
ENSEM-TRAN	51.69	ED	5.88	19.21	6.41

- Using only in-domain candidates leads to slightly higher BLEU scores
- For retrieval systems, adding emotion supervision explicitly decreases the accuracy of the rankings, p@1,100, but generally improves the average BLEU scores
- The ensemble encoders improve the generation models in perplexity and BLEU

Human Evaluation Results

Table 3: Human evaluation metrics from rating task. Training on EMPATHETICDIALOGUES improves all scores. Encoding supervised emotion information improves the empathy score (and sometimes the relevance and fluency by a smaller margin). *Bold: results within 1 SEM of best model.*

	Model	Candidates	Empathy	Relevance	Fluency
Retrieval	Pretrained	R	2.58±0.14	2.97±0.14	4.11±0.12
	Base	ED	3.27±0.13	3.42±0.14	4.44±0.08
	Multitask	ED	**3.58±0.12**	**3.58±0.14**	**4.46±0.09**
	Prepend-1	ED	**3.51±0.13**	**3.61±0.15**	4.45±0.10
	Prepend-3	ED	**3.62±0.14**	**3.50±0.15**	**4.54±0.08**
	Prepend-5	ED	**3.52±0.14**	**3.64±0.14**	**4.47±0.09**
	Ensem-DM+	ED	3.36±0.14	3.33±0.14	4.13±0.11
Generation	Pretrained	-	2.26±0.13	2.37±0.13	4.08±0.12
	Base	-	2.95±0.15	3.10±0.14	4.37±0.10
	Multitask	-	3.17±0.14	**3.23±0.14**	4.29±0.11
	Prepend-1	-	2.66±0.15	2.63±0.15	4.22±0.12
	Prepend-3	-	**3.34±0.13**	**3.31±0.15**	**4.58±0.09**
	Prepend-5	-	**3.35±0.15**	**3.20±0.15**	4.41±0.10
	Ensem-DM+	-	3.17±0.14	**3.19±0.14**	4.31±0.11
Gold Response	-	-	*4.19±0.06*	*4.48±0.06*	*4.67±0.04*

	Model	Candidates	Empathy	Relevance	Fluency
Retrieval	Pretrained	R	2.58±0.14	2.97±0.14	4.11±0.12
	Base	ED	3.27±0.13	3.42±0.14	4.44±0.08
	Multitask	ED	**3.58±0.12**	**3.58±0.14**	**4.46±0.09**
	Prepend-1	ED	**3.51±0.13**	**3.61±0.15**	4.45±0.10
	Prepend-3	ED	**3.62±0.14**	**3.50±0.15**	**4.54±0.08**
	Prepend-5	ED	**3.52±0.14**	**3.64±0.14**	**4.47±0.09**
	Ensem-DM+	ED	3.36±0.14	3.33±0.14	4.13±0.11
Generation	Pretrained	-	2.26±0.13	2.37±0.13	4.08±0.12
	Base	-	2.95±0.15	3.10±0.14	4.37±0.10
	Multitask	-	3.17±0.14	**3.23±0.14**	4.29±0.11
	Prepend-1	-	2.66±0.15	2.63±0.15	4.22±0.12
	Prepend-3	-	**3.34±0.13**	**3.31±0.15**	**4.58±0.09**
	Prepend-5	-	**3.35±0.15**	**3.20±0.15**	4.41±0.10
	Ensem-DM+	-	3.17±0.14	**3.19±0.14**	4.31±0.11
Gold Response	-	-	*4.19±0.06*	*4.48±0.06*	*4.67±0.04*

- All of the models with explicit emotion improved, meaning that the more explicit emotion supervision does allow models to better condition responses for the tone of the conversation

- most of them also maintain fluency scores (which are all above 4 on average)

Cases

Table 5: Examples of model responses from the validation set.

Context	Model	Generated Utterance
Tomorrow is my first university day and I'm very nervous	Gen-base	What are you studying?
	Gen-multitask	Good luck!
	Gen-prepend-5	I'm sure you will be fine.
	Gen-ensemb-DM+	Are you nervous ?
I had to go out of town by myself for three weeks. It was hard not being able to talk to anyone I knew.	Ret-base	Do you still talk to the people you went with?
	Ret-multitask	I'm sorry to hear that. Do you think you struggle to communicate with people?
	Ret-prepend-5	That sounds difficult. Did you try talking to people in your neighborhood?
	Ret-ensemb-DM+	Did you start to feel a bit lonely?

This work v.s. Ours

1. Artificial data v.s. real user interaction

2. Reddit: dialogue among strangers, how empathy; but the dialogue among strangers works for personality

3. Emotion might change, while personality is consistent

END