Empathy for Emotional Support

Sahand Sabour
2020280401

Marcin Malinowski
202038003

❖ Introduction

- Dialogue Systems (DS)
  - AI programs that can simulate human conversations.
  - Used to help users accomplish tasks (task-oriented).
  - Can also engage in normal conversations (open-domain).
- Empathy
  - A complex multidimensional construct.
  - Enables people to understand the situation and the implicit emotions of other people.
- Empathetic DS (EDS)
  - Aims to demonstrate an understanding of user utterances.
  - Responds appropriately in an empathetic manner.
  - Builds a trusting relationship with the user.
- EPITOME
  - Three mechanisms for expressing empathy
    - Emotional Reactions (ER)
    - Explorations (EX)
    - Interpretations (IR)

I am 18 years old and I was recently rejected by my dream school. My parents say it's ok but this makes me really disappointed in myself.

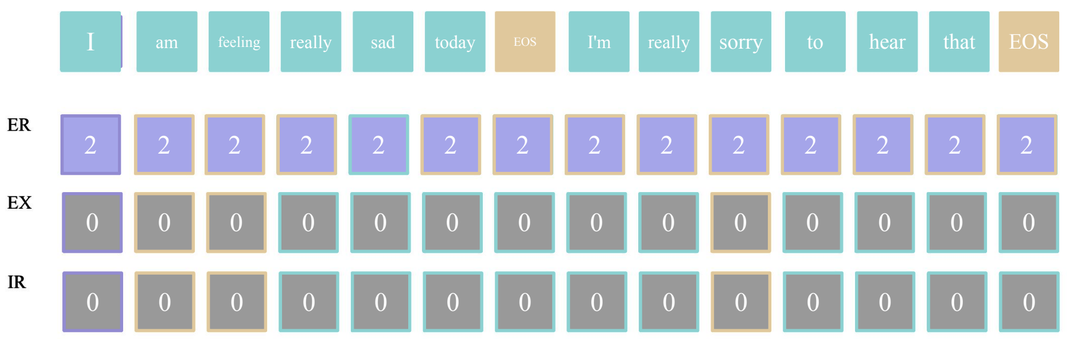
<ER> I'm sorry to hear that. <IR> I understand how that would make you dissapointed in yourself. <EX> Have you talked to any of your friends about this?

❖ Method

- Let $S_i = (s_{i1}, s_{i2}, ..., s_{in})$ be a post and $R_i = (r_{i1}, r_{i2}, ..., r_{im})$ be one of its corresponding responses.
- For general language modelling we have

$$P(r_{im+1} | S_i, r_{i1}, r_{i2}, ..., r_{im})$$

- With Embeddings



- With Label Tokens
  - Let T_{ER}, T_{EX}, and T_{IR} be the label tokens

$$P(r_{im+1} | S_i, T_{EX}, T_{ER}, T_{IR}, r_{i1}, r_{i2}, ..., r_{im})$$

- Top-P sampling

$$P(r_i = V_L | r_1, r_2, \ldots, r_{i-1}) = Softmax(u/T)$$

❖ Experiments

- Dataset
  - Collected posts and responses of 55 subreddits from Reddit.
  - Filtered and cleaned data based on best practices.
  - Annotated the dataset using EPITOME.

Dataset	Examples	avg tokens
posts	1,048,675	199
all responses	8,436,372	63
post responses	3,479,249	73

Table 1: Gathered data statistics.

Dataset	Examples	avg post tokens	avg response tokens
2-2-2	5,865	261	121
ones and up	26,123	233	116
at least one 1	2,683,009	190	80

Table 2: Data statistics of annotated data

- Models
  - Base: pretrained only.
  - Finetuned: trained on the raw data.
  - W/ Embeddings: trained with mechanisms as embeddings.
  - W/ Label tokens: trained with mechanisms as tokens.
- Variations
  - GPT-2
  - DialoGPT
  - BART
- Automatic Evaluation

Model		PPL	BLEU (%)				ROUGE-L			DISTINCT	
			B-1	B-2	B-3	B-4	Prec	Recall	F1	D-1	D-2
Base	DialoGPT	18.354	41.26	24.11	13.98	8.09	0.076	0.112	0.082	0.035	**0.227**
Fine-tuned	GPT-2	8.623	47.88	29.21	17.34	10.22	0.158	0.157	0.136	0.014	0.108
	DialoGPT	8.232	49.14	29.31	17.36	10.21	0.155	0.157	0.134	0.014	0.121
	BART	9.364	47.23	28.08	17.14	10.09	0.162	0.158	0.140	0.009	0.097
w/ Embeddings	GPT-2	8.643	47.26	28.99	17.29	10.24	**0.177**	0.161	**0.146**	**0.015**	0.119
	DialoGPT	8.311	**48.09**	**29.43**	**17.52**	**10.37**	0.174	0.161	0.144	**0.015**	0.131
	BART	9.614	47.51	29.02	17.12	10.34	0.163	0.160	0.139	0.010	0.105
w/ Label tokens	GPT-2	8.472	47.38	28.84	17.10	10.08	0.150	0.161	0.136	0.014	0.107
	DialoGPT	**8.295**	47.67	28.98	17.16	10.09	0.148	**0.162**	0.136	0.014	0.115
	BART	9.297	46.93	27.92	17.02	10.02	0.141	0.137	0.142	0.012	0.107

Table 2: Results of automatic evaluation on our collected Reddit dataset. The best results are in **bold**.

- Human Evaluation

Model		Fluency	Relevancy	Empathy
Fine-tuned	GPT-2	3.72	3.47	3.51
	DialoGPT	**3.76**	3.51	3.56
	BART	3.62	3.49	3.23
w/ Embeddings	GPT-2	3.59	3.59	3.84
	DialoGPT	3.61	3.62	3.86
	BART	3.55	3.49	3.64
w/ Label tokens	GPT-2	3.62	3.69	3.86
	DialoGPT	3.60	**3.71**	**3.91**
	BART	3.57	3.52	3.52

Table 4: Results of human evaluation. The best results are in **bold**.

❖ Discussions and Conclusions

- Human evaluation results demonstrate that using empathy mechanisms improves the empathy of generated response.
- Overall, DialoGPT pairs best with these mechanisms.
- Using these mechanisms as label tokens provides the most empathetic results. However, the results obtained from adding embeddings are also considerable.