Appendix

Anonymous EMNLP submission

A Empathy Expression Bias

For more comprehensive recognization of the severe emathy expression bias between existing empathetic dialogue models and humans, we further quantify the bias of Multitask-Transformer (Rashkin et al., 2019) in Figure 1 and MoEL (Lin et al., 2019) in Figure 2. The results are similar with MIME (Majumder et al., 2020), we can see the large intent distribution bias and the monotony of empathy expression of existing models.

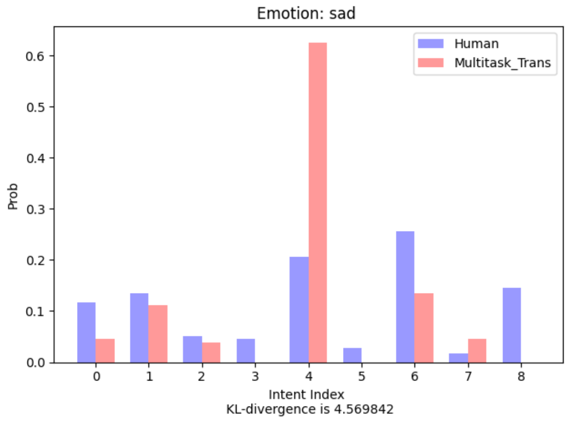

Figure 1: Empathetic intent distribution of human and Multitask-Transformer (Rashkin et al., 2019)

B Intent Keywords Collection

The keywords are retrieved from the training set of **EmpatheticIntents** dataset (Welivita and Pu, 2020) by using TF-IDF method. **EmpatheticIntents** has 5490 responses labeled with intents, where each intent group has 610 responses. Based on the intent label for each response in the training set, we concatenate all the responses which are in the same group and remove all the stop words. Finally, we apply TF-IDF to obtain the top k keywords for each intent group, we set k to 30 in our experiments. See Table 1 for top ten keywords for each intent.

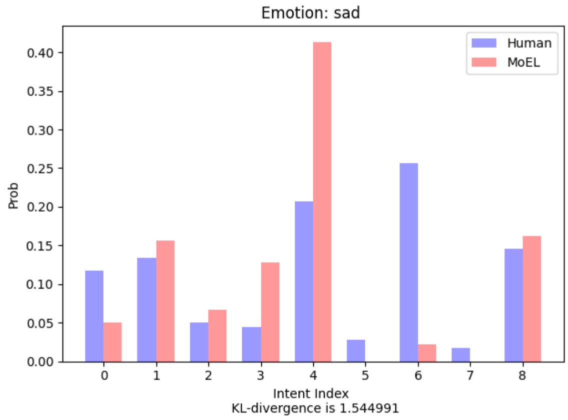

Figure 2: Empathetic intent distribution of human and MoEL (Lin et al., 2019)

C Error Analysis

Although EmpHi achieves huge improvement in terms of empathy, relevance, and diversity in empathetic dialogue generation, there is still some flaws. At first, the generation task of EmpHi is far difficult than existing models, because it needs to generate response condition on both context and the predicted intent, while other models generate response only condition on the context, therefor the exposure bias of EmpHi is more severe. See in Table 2, although the predicted most possible intent of EmpHi is the same as reference and its corresponding response is great, EmpHi also gives high probability for *Questioning* intent and the corresponding response is not very contextually relevant, EmpHi knows it is suitable for asking more details to show its caring, but it does not know how to ask under this context, thus EmpHi needs better understanding for context information. We believe this issue could be mitigated when using more dialogue data for pretraining.

Intent	Keywords
Agreeing	know, understand, agree, definitely, feel, feeling, exactly, mean, oh, right
Acknowledging	sounds, nice, awesome, like, great, cool, would, oh, must, really
Encouraging	hope, well, hopefully, get, good, time, bet, great, goes, soon
Consoling	hope, hopefully, get, better, well, soon, time, find, good, things
Sympathizing	sorry, hear, oh, am, happened, loss, feel, hope, really, aw
Suggesting	maybe, get, time, could, think, well, next, something, try, go
Questioning	oh, get, go, going, long, kind, like, work, good, maybe
Wishing	congratulations, luck, good, wish, best, well, happy, oh, birthday, wow
Neutral	good, like, sure, well, time, one, have, people, never, get

Table 1: Keywords for each intent.

Emotion		Proud
Context		My wife was able to get a job with microsoft without going to a four-year college!
Reference		That' s pretty amazing, she must be skillful regardless.
Intent	Prob	Generated Response
Agreeing	2.91%	Oh, wow ! That's amazing ! I know how you feel.
Acknowledging	**40.9%**	That's awesome ! I'm glad you're going to be proud of her.
Encouraging	9.98%	That's awesome ! I hope you guys have a great job!
Consoling	3.66%	Wow, That' s good, i hope you are able to get it .
Sympathizing	1.43%	Oh wow ! I am sorry to hear that.
Suggesting	0.62%	That's good ! Maybe she will be fine.
Questioning	28.6%	That's great ! How long have you been working ?
Wishing	9.19%	Congratulations ! That' s awesome ! congratulations !
Neutral	2.75%	That is a good thing to do, I wish i was a lot of work .

Table 2: Error analysis of EmpHi, although most responses are reasonable, there are still some flaws.

References

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.