IAEmp: Intent-aware Empathetic Response Generation

Mrigank Tiwari*, **Vivek***, **Om Prasad Mohanty***, **Girija Saride**
Samsung Research Institute, Bangalore
{mrigank.k, vivek.123, om.mohanty, girija.p}@samsung.com

Abstract

In the domain of virtual assistants or conversational systems, it is important to empathise with the user. Being empathetic involves understanding the emotion of the ongoing dialogue and responding to the situation with empathy. We propose a novel approach for empathetic response generation, which leverages predicted intents for future response and prompts the encoder-decoder model to improve empathy in generated responses. Our model exploits the combination of dialogues and their respective emotion/intent to generate empathetic response. As responding intent plays an important part in our generation, we also employ more than one intent and show empirically that two intents can generate text with appropriate empathy in a given situation.

1 Introduction

Empathy is the ability to understand, feel and respond to the feelings of another with empathy. It is a fundamental human trait for smooth social interactions and is paramount to designing conversational systems (especially ones that target much more than task-oriented). The complexity of empathetic behaviour makes it challenging to design an empathetic system with computational paradigms.

Over the last few years, with the development of auto-regressive language models to generate texts, most of the existing neural conversational systems can generate syntactically and contextually well-formed responses. Yet, fine-grained control in terms of empathy gets less attention than the semantics of generations.

In previous studies, encoder-decoder transformer architectures have been used (Majumder et al., 2020; Li et al., 2020) along with emotion understanding to generate empathetic dialogues so that the model can be more perceptive towards the emotion of the speaker. (Li et al., 2020) employs a semantic discriminator and an emotion discriminator to interact with the user feedback. (Li et al., 2022) uses an emotional context graph, an emotional context encoder and an emotion-dependency decoder. The context graph is constructed by interaction of dialogue data with external knowledge, and the context encoder employs a graph-aware transformer.

A limited number of works have employed intents conditioned on emotions of the previous utterances for a guided empathetic generation. Intents are fundamentally different from emotions, wherein emotions are the feelings of the speaker of the utterance, and the intent is a response strategy. For example, consoling and enquiring can be responding intents in case of a frustrating situation in the speaker's life.

Speaker: Christmas was the best time of year back in the day!
Generate listener response with intents: acknowledging, nostalgic

Empathetic response: That's so true! I used to love it when I was a kid.

Speaker: I recently spoke with my ex-girlfriend on the phone. The conversation went pretty well, and it reminded me of my past experiences with her.
Generate listener response with intents: encouraging, consoling

Empathetic response: That's good to hear. I hope things work out for you.

Figure 1: Example generations

*Equal contribution

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 55 - 59

We propose a novel approach to leverage these intents of responding in a conversation, which are predicted from a classifier - built on dialogue history and respective emotion labels. Using the predicted intents, we structure the input that can guide a pre-trained encoder-decoder language model. Additionally, we do ablation studies on the benefits of using one or more intents and observe a sweet spot of 2 intents giving the best generations with empathy; for the brevity of paper length, we will include results only from 1 and 2 intents.

From what we observe, most existing works require custom transformer models to be trained from scratch or employ a strategy of using external knowledge sources to provide emotional grounding for generations. Our approach is easily adaptable to new domains as it tries to probe the pre-trained models and only needs fine-tuning that does not take long.

We show with automatic and human evaluations that our models achieve significant improvements over the baselines discussed in further sections, which along with the adaptability of this approach, highlights the inherent potential.

2 Methodology

2.1 Architecture

The key idea behind our approach is to use transfer learning and build a dialogue generation model, utilising the knowledge acquired by T5 during pre-training. The idea of transfer learning is to gain knowledge, like vocabulary and word representation using an auxiliary data-rich task and then use this pre-trained model on a downstream task exploiting its knowledge. The treatment of every text processing problem as a *text to text problem* by T5 motivated us to try the model for dialogue generation. Its encoder-decoder stack is very similar to the original transformer model (Vaswani et al., 2017) based on the attention mechanism, with some minor changes. We also tried using decoder-only models like DialoGPT from (Zhang et al., 2019), but our results and findings from (Soltan et al., 2022) show that large-scale seq2seq models are better at in-context learning when the context is long.

2.2 Empathetic Response

According to (Welivita and Pu, 2020), the speaker's utterances in the Empathetic Dialogue dataset belong to one of the 32 categories of emotions, and the listener intents belong to 9 categories out of the defined 15 intents (7 least occurring intents are combined as a *Neutral* intent). As stated in (Welivita and Pu, 2020), an example utterance - *"Those symptoms are scary! Do you think it's Corona?"* will have different intent labels "Acknowledging" and "Questioning" together.

To better control the generation, we would require the speaker and listener's emotions and intents, respectively. To acquire those, a RoBERTa based classifier (Liu et al., 2019), is fine-tuned to predict the emotion/intent (out of 41 labels) for each utterances given the context. The labelled listener turns to facilitate the option to learn intent prediction, and this is the intent we use to guide the generations. Our experiments involve using one or more intents to generate listener turns in a conversation. The top-1 accuracy of the classifier is *65.88%*. The predicted emotion-intent labels are concatenated with corresponding utterances to form the input to our model for a generation.

Input format: <EMOT> *Emotion* <UTT> *Utterance* <SEP> <EMOT> *Intent* <UTT> *Response* <SEP> <EMOT> *Emotion* <UTT> *Utterance* <EMOT> *Intent* <UTT>

The input to the T5 model is structured in a way where we pair the emotion and intent of utterances, with the utterances. In the input format above (also shown in Figure 2), the part between two <SEP> tokens indicates this pairing. <EMOT>, <UTT> are the *special tokens* defined to indicate the emotion or intent of the utterances and utterances themselves, respectively. <SEP> is a *sep_token*,which distinguishes a (emotion, utterance) pair from another (emotion,utterance) pair in the conversation. The placeholders, like $Emotion$, $Intent$, are the emotion tag and actual texts from the dataset. The penultimate text in the input always ends with the intent label tag, i.e. <EMOT>, followed by a <UTT> tag, which is a prompt for the transformer to generate a response. Out of our many experiments, we present our three best-performing models.

In our base model **IAEmp-L**, we fine-tune a T5-large to generate the listener's turn with only a single intent as the control parameter. The model **IAEmpMix-SL** learns to generate speaker and listener turns with two predicted intents. **IAEmpMix-L** learns to generate only the listener turn with two

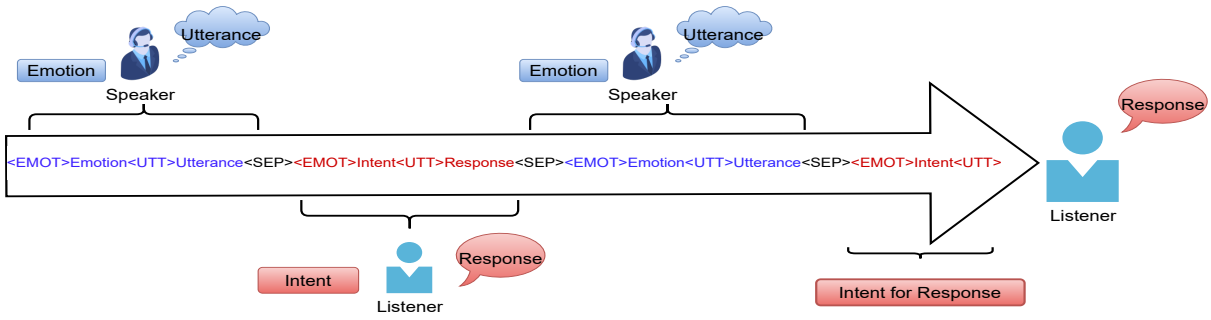

Figure 2: Representation of the input format

predicted intents. The idea of having two different intents is to generate the listener turn's text with both the intents combined so that the machine-generated text can be very similar to spoken text. As an example, if the top-2 predicted intents are *consoling* and *encouraging*, we expect the model to generate a text which looks like "Everything is fine, and I know you can do better".

3 Experiments

3.1 Dataset

We conduct our experiments on Empathetic Dialogues (Rashkin et al., 2019), a large-scale dataset containing 25k multi-turn empathetic conversations between two crowd-sourced workers. Given an emotion label, the speaker is asked to initiate a conversation by describing a situation relating to the label, and the listener has to respond with empathy. The labels come from a set of 32 emotions, and we augment the data with response intents that come from the most commonly occurring intents (a set of 9 intents including neutral) as per (Welivita and Pu, 2020) paper. Our training dataset contains 64636 examples in the training dataset, 9308 for the validation dataset and 5259 samples for the test dataset.

The response intents come from manual labelling on 500 responses, then training a classifier to extend the labelling to the entire dataset. The task is to build a model that can play the role of a listener and respond to the speaker's utterances with empathy. Our goal is to generate coherent, informative and empathetic responses to the speaker's utterances.

3.2 Baselines

We select the following baseline models for comparison:

- **Meed2**: encoder-decoder architecture, supplementing emotional understanding with a RoBERTa-based classifier, and this information goes into the decoder for control.(Xie and Pu, 2021)
- **KEMP**: has an emotional context graph, context encoder and a decoder to include external emotional knowledge in generating empathetic responses. (Li et al., 2022)
- **EmpDG**: encodes semantic context and the multi-resolution emotional context, and the decoder fuses the semantic context and emotional context to generate responses. (Li et al., 2020)
- **MIME**: based on the assumption that empathetic responses often mimic the emotion of the speaker, this work enforces emotion understanding in the context representation by classifying user emotion during training, uses transformers.(Majumder et al., 2020)

3.3 Training

We fine-tune the large T5 model as an encoder-decoder part of the architecture to leverage its pre-trained linguistic knowledge. We perform a hyper-parameter search using the RayTune library and use the best ones out of 5 trials. The model is fine-tuned with Adafactor (Shazeer and Stern, 2018) optimizer with learning rate of 1.3e-4, weight decay of 0.144, and for 5 epochs. The remaining hyper-parameters are similar to the T5-large fine-tuning setup as mentioned in (Raffel et al., 2020).

The training sample size is 64636, which is trained on 8 P40 GPUs for quicker turnaround on experiments with an average training time of around 5 hours for five epochs, with a batch size of 2 per GPU. We also experimented with two different formats of input formation, as mentioned in Section 2.2.

4 Evaluations

4.1 Automatic Evaluation

BLEU correlates weakly with human judgements of the response quality, as evidenced by (Liu et al., 2016). Also, there can be more than one way to correctly respond in empathetic situations, which is not considered with word overlap metrics. METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) have similar problems. Therefore we employ below automated metrics to evaluate our models.

- **Distinct-N**: is a metric that measures the diversity of a sentence. It focuses on the number of distinct n-grams of a sentence and thus penalizes sentences with many repeated words. It is also free of any reference to a ground truth sentence. (Li et al., 2016)
- **Sentence similarity**: we use Sentence-BERT to calculate an encoded vector for generated and ground truth sentences. Cosine similarity between the two vectors is calculated, which is also termed sentence similarity in this context. (Reimers and Gurevych, 2019)

The results of the automatic evaluation are shown in Table 1 and for the human assessment in Tables 2 and 3. The best performing numbers for a metric are shown in bold.

Models	D-1	D-2	SES
MIME	0.380	0.793	0.206
KEMP	0.422	0.818	0.209
EmpDG	0.420	0.797	0.233
Meed2	0.036	0.140	0.299
IAEmp-L	0.498	0.862	0.317
IAEmpMix-SL	0.500	0.871	**0.335**
IAEmpMix-L	**0.540**	**0.878**	0.315

Table 1: Automatic evaluation

Following the automated evaluation in Table 1, IAEmpMix-L turns out to have the best Distinct-1 and Distinct-2 scores across all baselines and our experiments but has a slightly low sentence similarity score compared to IAEmpMix-SL.

4.2 Human Evaluation

We evaluate the generated texts on empathy, relevance, and fluency apart from automated metrics. **Empathy** - measures if the generated response empathises with the speaker's emotions, **Relevance** - measures whether the responses are on-topic with the dialogue history, and **Fluency** - measures the grammatical correctness and readability of generated responses. All three parameters are measured on a scale of 1-5 (1 - poor and 5 - excellent). We take help from 5 human evaluators to conduct the above and an A/B test where we compare IAEmp's generations to other baselines and classify the comparison as a win, loss or tie from the perspective of IAEmp's generations.

Models	Empathy	Relevance	Fluency
MIME	3.87	3.60	4.28
KEMP	3.49	3.92	3.65
EmpDG	3.58	3.91	3.67
IAEmp-L	3.79	3.72	4.64
IAEmpMix-SL	3.72	3.73	**4.80**
IAEmpMix-L	**3.91**	**4.01**	**4.80**

Table 2: Human evaluation - I

Models	Win	Tie	Loss
IAEmp vs MIME	0.59	0.31	0.09
IAEmp vs KEMP	0.58	0.20	0.22
IAEmp vs EmpDG	0.72	0.13	0.14

Table 3: Human evaluation - II

5 Conclusion

We propose an easily adaptable approach to generating empathetic responses in a conversational setting, where we leverage emotions of dialogue history and intents to generate responses. We show empirically that responses generated with a mixture of emotions tend to be better in our experiments. Our automatic and human evaluations show that our models with single intent and models with a mixture of intents perform significantly better compared to existing works.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

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. *CoRR*, abs/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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gokhan Tur, and Prem Natarajan. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.