A Review of Text Generation Models for Empathic Responses

Kim Ngan Phan[1], Hyung-Jeong Yang[2], Jieun Shin[3], Seungwon Kim[4], Soo-Hyung Kim*[,5]
[1]PhD's Student, Department of AI Convergence, Chonnam National University
[2,3,4,5]Professor, Department of AI Convergence, Chonnam National University
*Corresponding Author

kimngan260997@gmail.com, hjyang@jnu.ac.kr, jieunshin@jnu.ac.kr, seungwon.kim@jnu.ac.kr, shkim@jnu.ac.kr

ABSTRACT

Empathetic response generation is a challenging concern in the field of natural language processing. Recent studies are trying to generate empathic responses to humans in dialogues. The published EmpatheticDialogues dataset is a solid foundation for the task of generating empathic responses. Many researchers have experimented with the EmpatheticDialogues dataset, which has many potential variations of transformer architectures. In this paper, we survey several previous approaches to the task of generating empathic responses with the aim of indicating the potential of future deep-learning models.

1. Introduction & Related Wor

The content of daily human dialogues is very rich and diverse. Not only asking and answering, empathy plays a very important role in conversations. The utterance in human dialogue carries various features such as knowledge, emotion, connotation, keywords, etc. However, those features constitute the implicit emotional information that the listener wants the speaker to be able to grasp. In other ways, the empathic response is the listener's expectation factor in direct dialogue. Empathy plays an important role in human dialogues because it has the ability to strengthen their emotional bond. Generating empathic responses is also important to improve the user experience in human-computer interaction. Improving empathy is a factor for computers to integrate into human society. Indeed, a non-empathetic response can frustrate users in interacting with computers because the responses are mechanical, and incoherent, and thus lead to limitations in human-computer interaction.

Research on generating empathetic responses in dialogues has been continuously improving over the years. Rashkin recently published the EmpatheticDialogues dataset [1] for models that generate empathetic responses. This dataset is popular and widely used for related research. The dataset consists of dialogues between listeners and speakers, where the listeners actively talk about their concerns based on the given situation and emotions. The model functions as a listener and provides empathetic responses to the speaker.

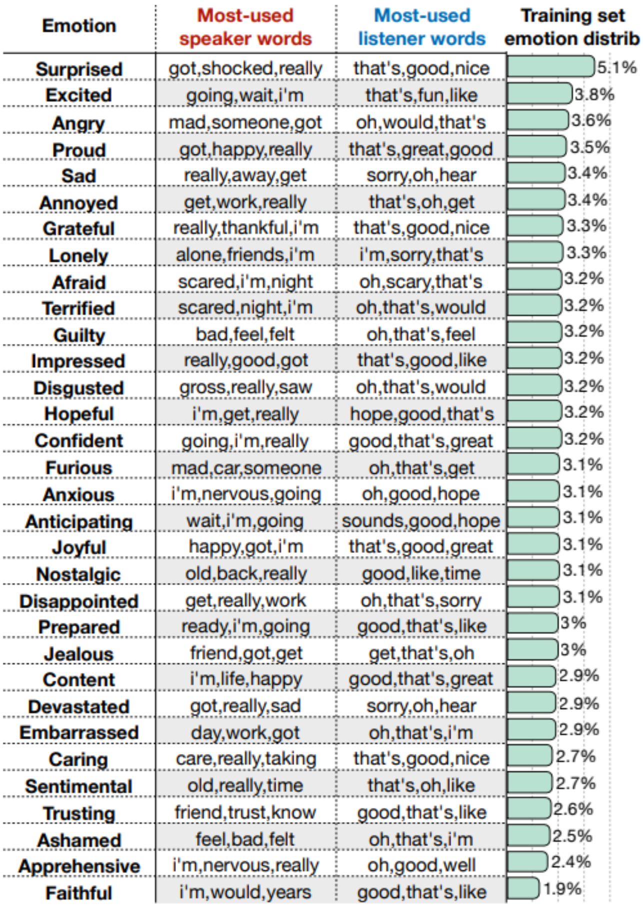
Studies are continuously being published with experiment results based on the EmpatheticDialogues dataset. Their results also showed that the Transformer-based experiments showed the ability to make the model more empathetic. Therefore, the transformer architecture serves as a foundation for subsequent studies based on the EmpatheticDialogues dataset. The authors have proposed transformer variants combined with other powerful architectures. From this, we notice the potential and effectiveness of combined architectures in generating empathetic responses. In this paper, we survey methods that have used the EmpatheticDialogues dataset, with the aim of indicating potential future directions for the task of generating empathetic responses.

2. EmpatheticDialogues Dataset

The dataset has been shown to be a solid foundation for exploring the empathetic response generation model in dialogues. In the experiments of the empathetic response generation model, EmpatheticDialogues data is commonly used to evaluate the proposed models. The dataset has dialogues between listeners and speakers based on different contexts. It has 32 negative and positive emotion labels, each emotion and a situation description are assigned to each dialogue. The speaker based on the given emotional label and situation description initiates a dialogue and the listener perceives and responds. The speaker and listener then exchange up to 6 more turns.

The dataset consists of 24,850 dialogues collected from 810 different participants. The datasets are divided into 19533, 2770, and 2547 dialogues for training, validation, and testing respectively.

Figure 1 depicts the label distribution of the training dataset and the top 3 most used content words by speakers and listeners in each category.



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%

Fig. 1. Distribution of conversation labels within the EMPATHETICDIALOGUES training set, along with the top 3 content words used by the speaker and listener in each category. [1]

3. Generation Model Model

3.1 Transformer Model

The transformer model [2] consists of encoder-decoder architectures, transformers have become a rapidly growing but challenging research potential in modeling conditional conversational response generation. Several published studies have used the transformer as a foundation to reformulate their models in related tasks. The model takes as input a dialogue context sequence and outputs a response sequence. The words of context sequence are embedded in a higher-dimensional space of arbitrary dimensionality. The encoder architecture performs spatial encoding with a combination of a multi-head attention module and a feed-forward layer. After encoding, the tensors are fed into the decoder architecture to perform decoding. The decoder architecture also has a multi-head attention module and a feed-forward layer as its main modules. The decoder then generates an output sequence. At each step, it takes a word from the vocabulary based on the probability of the output. From there it forms the output response. The transformer model is illustrated in Figure 2.

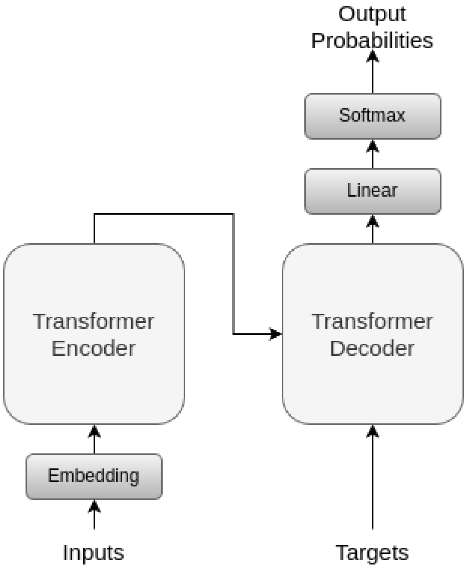


Fig. 2. Overview architecture of Transformer.

3.2 Empathetic response generation models

Based on the EmpatheticDialogues dataset, we found a wide variety of previous methods. However, our survey was only conducted on a few typical survey was only conducted on a few typical methods of the model reform trend.

MIME (Majumder et al., 2020) [3]: The model relies on 32 emotion labels to divide into two clusters of positive and negative emotions. Emotional mimicry leads to improved empathy. The model initially implements a transformer encoder to encode the context. They then obtain response-emotion presentation by sampling response-emotion distributions to form two groups mimicking and non-mimicking. They are then also refined to accommodate both positive and negative contextual cases. Finally, they are fed into a transformer decoder to generate an empathic response.

CEM (Sabour et al., 2021) [4]: Not only user emotions, CEM suggests that cognitive understanding can also be considered for the empathic response generation model. In detail, the model uses COMET [5] to generate five commonsense knowledge. The xReact knowledge is encoded and combined with the encoding information for the affective relation. The remaining four pieces of knowledge are also encoded and combined with the encoding information for the cognitive relation. The affection-refined and cognition-refined are performed in separate encoders again. The model performs the knowledge selection and finally performs the decoder.

SEEK (Wang et al., 2022) [6]: SEEK is proposed architecture with the desire to select accurate information from combining common knowledge sources. The model also uses COMET to extract five commonsense. The authors use a transformer encoder to encode utterances and common knowledge. Bi-LSTM is implemented to extract related sequences between utterances and common knowledge. Then, it models the emotions and fine-grained emotions at the utterance levels. Before decoding, the representation needs to be filtered by the cross-attention knowledge attention module. This attention receives the utterance representation sequence as a query vector, the generated knowledge text from the COMET model is vector and values keys. This helps the model generate an empathic response in the dialogue.

CARE (Wang et al., 2022) [7]: The idea-based model combines all the interdependent and simultaneous cause-and-effect relationships, based on emotions, and past and future dialogue contents. The model performs the prior and posterior causal graph. The prior causal graph includes causal relationships explicitly mentioned in the user's previous utterances, whereas the posterior graph incorporates additional causalities from the user's subsequent utterances. These causalities are fed into multi-source attention at the decoder to generate a response.

InfRa (Li et al., 2024) [8]: InfRa incorporates discourse features to enhance structural dialogue understanding, typically utilizing a novel edge pruning and mutual information learning module to refine the representation. In this study, the model performs BART [9] encoder and decoder architecture.

Table 1. Summary of performance results of several published articles on the task of generating empathic responses.

Model	**PPL**	**Dist-1**	**Dist-2**	**BLEU-3**	**BLEU-4**
MIME [3]	37.33 [8]	0.26 [8]	0.87 [8]	-	-
CEM [4]	36.11	0.66	2.99	-	-
SEEK [6]	37.09	0.73	3.23	-	-
CARE [7]	32.84	-	-	4.88	2.95
InfRa [8]	**25.87**	-	-	**3.70**	**22.41**

Table 1 is referenced from the performance results of a number of published papers. Although the papers employ several additional evaluation metrics, we only present the evaluation metrics Dist-1, Dist-2, and BLEU-3, BLEU-4. Since some of the required MIME [3] performances were unavailable, we gathered additional performances from InfRa [8]. The trend of the reported results of the previously proposed methods shows that the COMET combination does not help the model improve much. COMET is a pre-train model of the GPT language model. COMET combined with the encoder in the transformer causes a lot of noisy information in the context of the situation. We need a module with a noise-filtering function. SEEK [5] has improved the weakness of the model combination with COMET but the efficiency has not improved much. Another idea was more groundbreaking when the authors tried to combine graphs and encoder/decoder architecture in many potential and powerful ideas. When combined with transformer encoders and decoders, CARE has demonstrated improved effectiveness over most other methods. However, the effectiveness of the model is not a superior indicator. In InfRa, the author used the BART encoder and decoder. Along with the integration of discourse features by graph, it has shown superior results than previous methods. This shows that graphs have the potential for the task of generating empathetic responses. Because graph extracts potential and robust representations within the data used, these representations can closely reflect the context of the dialogue. The study using the transformer and graph combination needs more improvements in many other directions to produce better performance results.

3.3 Evaluation metrics

In the previously proposed methods, models are generally evaluated by the Perplexity (PPL) index. Additionally, the authors can evaluate the model in detail through their scientific reasoning. The PPL evaluates the model's understanding of language structure, with lower perplexity reflecting higher fluency and greater prediction accuracy. Besides, indexes such as Dist-n [10], BLEU [11], and Accuracy are also commonly used by authors to evaluate their models. Specifically, Dist-n is a diversity-n score that calculates the number of n-grams in the responses, reflecting their informativeness. BLEU evaluates the quality of machine-generated text with one or more reference responses. A strongly evaluated model mostly satisfies these evaluation indicators.

4. Conclusion

In this survey, we have shown a typical view of these various research efforts under the task of empathic response generation. The researchers explored many contexts related to empathy and improved the results over the previous method. We also present the potential through typical methods. However, improving empathic response generation models still faces many difficulties in the field of human-computer interaction. Research needs many experiences for new directions to explore the hidden context of dialogue.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT), the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437718) supervised by IITP, and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS- 2023-00219107).

References

[1] 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 Annual Meeting of the Association for Computational Linguistics, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

[2] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[3] 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.

[4] Sabour, Sahand, Chujie Zheng, and Minlie Huang. "Cem: Commonsense-aware empathetic response generation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 10. 2022.

[5]Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

[6] Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4634–4645, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[7] Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022. CARE: Causality Reasoning for Empathetic Responses by Conditional Graph Generation. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 729–741, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[8] Li, Bobo, et al. "Integrating discourse features and response assessment for advancing empathetic dialogue." Information Processing & Management 61.5 (2024): 103803.Li, Bobo, et al. "Integrating discourse features and response assessment for advancing empathetic dialogue." Information Processing & Management 61.5 (2024): 103803.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

[10] 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.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.