Evaluating Significance of Emotion Classification in Emotion-Aware Empathetic Dialogue Systems

Akshen Kadakia akshenhe@usc.edu
Mrinal Kadam mrkadam@usc.edu
Swetha Sivakumar swethasi@usc.edu
Yashwanth Gondi gondi@usc.edu
Tanmey Saraiya tsaraiya@usc.edu

1 Project Domain & Goals

A desirable trait in a human-facing dialogue agent is to appropriately respond to a conversation partner that is describing personal experiences, by understanding and acknowledging any implied feelings - a skill we refer to as empathetic responding.

Our project aims to facilitate evaluating models' ability to produce empathetic responses. We are carrying out two experiments that are different than that proposed by the original paper-

1. Changing the pre-trained emotion classifier used to train the 'Emotional Awareness' quotient of the dialogue system
2. Changing the EMPATHETICDIALOGUES (ED) dataset, a novel dataset with about 25k personal dialogues, given in the original paper, by reducing the number of categories of the target class(emotion) from 32 to a smaller subset of emotions,

following which we will check whether there has been any significant increase/decrease in our model's performance or not. Some examples of dialogues in the ED dataset grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding have been shown in Fig. 1.

While it is straightforward for humans to recognize and acknowledge others' feelings in a conversation, this is a significant challenge for AI systems due to the paucity of suitable publicly available datasets for training and evaluation. How to integrate empathetic responding into more general dialogue when, for example, the needs for empathy have to be balanced with staying on topic or providing the information is also very important. These are the broad problems that we want to address.

The dialogue models that use the ED dataset will be perceived to be more empathetic by human evaluators, compared to models merely trained on large-scale Internet conversation data. It is important to note that engaging in social talk, reacting to emotional cues, and displaying a caring attitude have been associated with better task outcomes in many domains.

We need NLP to solve this problem because it will help us resolve the ambiguity in language and will also aid us in adding useful numeric structure to unstructured data for our application which is text and sentiment analytics.

At the end of this, we will be able to build a good dialogue system that can provide empathetic responses by judging the conversation at hand. This will make the conversation flow more naturally and will encourage the user to correspond with the dialogue system further.

2 Related Work

Recent research has focused on finding new ways of encoding emotion in text data. One way is inspired by the concept of Word2Wec word embeddings (Mikolov et al. 2013), using word embeddings to not only encode word's meaning but also encode the emotion it represents (Fung et al. 2019), to achieve promising results.

While EMPATHETICDIALOGUES represent emotion as one of 32 discrete values, there have been other attempts at more fine-grained discrete representations, with 64 emojis (X. Zhou and Wang 2018) which generate high-quality conversational responses which are emotion-aware. Some have chosen a less granular path with as less as 2 classes in case of (Kong et al. 2019), 6 classes by (H. Zhou et al. 2018) and (Peng et al. 2019) and 9 classes by (Huang et al. 2018).

Some techniques have been developed to blend several conversational skills into one dialogue system to give a more well-rounded experience to the consumer (Smith et al. 2020). These skills include knowledge, using the Wizard of Wikipedia dataset

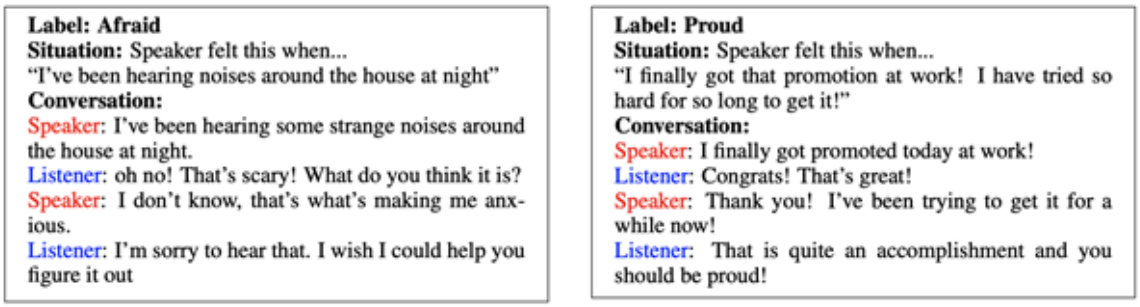

Figure 1 Examples of dialogues in the ED dataset. Taken from (Rashkin et al. 2019)

(Dinan et al. 2019), empathy using the Empathetic Dialogues, and talking about one's own personal background (L. Zhou et al. 2020).

Our work seeks to expand upon previous research and the research of the original paper by exploring if any improvements can be made on the pre-trained model's structure to extract better performance. We will also explore ways to improve the performance by changing the number of classes of emotions we consider and quantifying the change in performance.

3 Datasets

We plan on using the EMPATHETICDIALOGUES dataset as proposed by (Rashkin et al. 2019). The EMPATHETICDIALOGUES dataset is composed of 24,850 conversations. Each conversation is 4-8 utterances long. A sample from the dataset is shown in Fig. 1, in which a situation describes the topic the speaker starts the conversation with.

The ED dataset uses a granular 32 emotions classifier, which we intend to change based on the emotion hourglass as proposed by (Cambria, Livingstone, and Hussain 2011). We intend to reduce the granularity of emotions used, by mapping the labels to new sets of labels of sizes like 6 and 9 as tested by other related work and quantify the results.

4 Technical Challenges

4.1 Training large-capacity models require heavy computational investment

In the experiments of our proposed project, we will be working with two types of models. Each model has its own large-capacity architectures which are as follows:

- **Retrieval-Based Models:** In retrieval models, we use a variety of Transformer-based architectures and BERT-based models. The Transformer architecture consists of two encoders- context encoder and candidate encoder and shares the same base architecture of four layers and 6 Transformer heads. The experiment is also repeated with "Large" Transformers with five layers instead of four.
- **Generative Models:** Uses Full Transformer Architecture ie. Encoder and Decoder.

A major challenge in achieving the expected results lies in training large-capacity models especially like BERT-based models which have long training times.

4.2 Longer training times despite powerful computational resources

Fig. 2., an extract from our base paper shows the Training time and computational resources obtained during the experiments. As we can clearly see from this table, large-capacity models, specifically BERT-based models like the Pretrained-BERT-R model which is a retrieval model trained on 1.7 Billion data points from Reddit using 8GPUs still take 13.5 days to train. In our proposed model, we will majorly be working on fine-tuning BERT and Transformer models using ED dataset and Transfer learning Techniques if required.

4.3 Emotion Aggregation to enable multiple data sources

In emotion-aware dialogue systems, a crucial role lies in curating our dataset and deciding what emotions the model must be able to capture and react to. In our base paper, the models capture 32

	Model	Params, resources, train examples	Emp	Rel	Fluent
Retrieval	Pretrained-R	84.3M, 2.5 days, 8GPUs, 1.7B	2.8	3.0	4.1
	Pretrained-ED	same , same, same	3.5	3.6	4.5
	Fine-Tuned	same , + 0.5 hour, 1 GPU, +22.3k	3.8	3.8	4.4
	Pretrained-Bert-R	217M, 13.5 days, 8GPUs , 1.7B	3.1	3.3	4.2
	Pretrained-Bert-ED	same, same, same	3.4	3.5	4.4
	Fine-Tuned-Bert	same, +1hour, 8GPUs, +22.3k	3.7	3.8	4.6
Generative	Pretrained	85.1M, 2 days, 32 GPUs, 1.7B	2.3	2.2	3.9
	Fine-Tuned	same , +1 hour, 1 GPU, +22.3k	3.3	3.3	4.3
	Pretrained-Large	86.2M, 2.5 days, 32 GPUs, 1.7B	2.8	3.0	4.0
	Fine-Tuned-Large	same , +0.5 hour, 1 GPU, +22.3k	3.6	3.6	4.5

Figure 2 Training time and resources of experimental models. Taken from (Rashkin et al. 2019)

emotions. But for our experiments, we will be aggregating the 32 emotions into 'K' major groups and testing the model performance. The challenge here is in identifying and grouping the 32-emotions into 'K' groups while maintaining a balance between the emotions.

4.4 Relevant, Accurate, and Labelled Data Paucity

A major practical challenge in achieving emotion-awareness in dialogue systems is due to the lack of human annotations since it is a time-consuming process. In the base paper, the EmpatheticDilaogues (ED) data is collected through MTurk with 810 Mturk workers. And the samples thus collected are grounded to a situation and are based on one-on-one conversations between a Speaker and a Listener. For the purposes of our experimentation, we will be creating, circulating, and using online surveys to collect data grounded to emotions.

4.5 Ensuring Domain-Consistency

The base paper uses additional information from external predictors to define EmoPrepend and TopicPrepend Variations of the models. As a part of our proposal, we aim to modularly replace the current external predictors with state of the art Emotion-classification models. To ensure there is no degradation in performance, it is required that the new replacement emotion-classification model that we will introduce has been pre-trained on a similar domain with balanced data which is grounded in emotions and stresses on empathetic aspects of the dialogue.

References

Cambria, E., Andrew G. Livingstone, and Amir Hussain (2011). "The Hourglass of Emotions". In: *COST 2102 Training School*.

Dinan, Emily et al. (2019). "Wizard of Wikipedia: Knowledge-Powered Conversational Agents". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1l73iRqKm.

Fung, P. et al. (2019). "Empathetic dialog systems". English. In: *LREC 2018 - 11th International Conference on Language Resources and Evaluation*. Cited By :3. URL: www.scopus.com.

Huang, Chenyang et al. (June 2018). "Automatic Dialogue Generation with Expressed Emotions". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 49–54. DOI: 10.18653/v1/N18-2008. URL: https://aclanthology.org/N18-2008.

Kong, Xiang et al. (2019). "An Adversarial Approach to High-Quality, Sentiment-Controlled Neural Dialogue Generation". In: *CoRR* abs/1901.07129. arXiv: 1901.07129. URL: http://arxiv.org/abs/1901.07129.

Mikolov, Tomás et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *CoRR* abs/1310.4546. arXiv:

1310.4546. URL: http://arxiv.org/abs/1310.4546.

Peng, Yehong et al. (2019). "Topic-enhanced emotional conversation generation with attention mechanism". In: *Knowledge-Based Systems* 163, pp. 429–437. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2018.09.006. URL: https://www.sciencedirect.com/science/article/pii/S095070511830457X.

Rashkin, Hannah et al. (July 2019). *Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset*. Florence, Italy: Association for Computational Linguistics, pp. 5370–5381. DOI: 10.18653/v1/P19-1534. URL: https://aclanthology.org/P19-1534.

Smith, Eric Michael et al. (2020). "Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills". In: *CoRR* abs/2004.08449. arXiv: 2004.08449. URL: https://arxiv.org/abs/2004.08449.

Zhou, Hao et al. (2018). *Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory*. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455/15753.

Zhou, Li et al. (Mar. 2020). "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot". In: *Computational Linguistics* 46.1, pp. 53–93. ISSN: 0891-2017. DOI: 10.1162/coli_a_00368. eprint: https://direct.mit.edu/coli/article-pdf/46/1/53/1847834/coli_a_00368.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00368.

Zhou, Xianda and William Yang Wang (July 2018). "MojiTalk: Generating Emotional Responses at Scale". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1128–1137. DOI: 10.18653/v1/P18-1104. URL: https://aclanthology.org/P18-1104.