Seeking an Empathy-abled Conversational Agent

Andreea Grosuleac
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
andreea.grosuleac@stud.acs.upb.ro

Ştefania Budulan
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
stefania.budulan@cs.pub.ro

Traian Rebedea
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
traian.rebedea@cs.pub.ro

DOI: 10.37789/rochi.2020.1.1.16

ABSTRACT

A fairly novel area of research, at the conjunction between Artificial Intelligence (AI) and Human-Computer Interaction (HCI), resides in developing conversational agents as more users prefer this type of interaction to conventional interfaces. In this paper, we present an open-domain empathic chatbot, encompassing two of the biggest challenges of dialog systems: understanding emotions and offering appropriate responses. Although these tasks are trivial for a human, it is difficult to create a system that can recognize others' feelings in a discussion. The proposed model is developed based on the Generative Pre-Trained Transformer and it uses two datasets, PersonaChat and Empathetic Dialogues to achieve an empathic chatbot with a cordial personality. The measured performance - 18.20 perplexity, 6.56 BLEU score, and 6.56 accuracy - comes close to the state-of-the-art models, while offering a further refined dialogue persona.

Author Keywords

Empathic dialogue; Open-domain conversational agents; Empathetic chatbot; Conversational interfaces.

INTRODUCTION

As observed by Følstad and Brandtzæg [4] nowadays the interaction between humans and computers seems to turn toward natural-language user interfaces making the development of conversational agents a vital area of research in HCI.

Empathic chatbots are conversational agents that are not only able to generate emotional responses, but can understand the feelings of a user and respond accordingly. While prior work focused on creating conversational systems that can speak coherently and grammatically correct, in the last years the attention of the academic community changed to a more engaging agent that can mimic a real person's skills [14].

This paper tackles the problem of empathic chatbots, a new research focus, that tries to address the lack of empathy in the widely available conversational agents. This is an important issue as there is a current exponential growth in the spread of such agents to solve mundane tasks as website guidance, entertainment, information extraction, or question answering in domains like customer service or education. An important stand-alone application for an empathic chatbot can be made in the healthcare domain as more young people report a decreasing number of friends and personal connections, proving that our generation may be the most connected one, yet the loneliest.

The current work proposes an empathic chat agent that embodies a personality that provides the ability to create more engaging and natural responses. To recognize feelings during the conversation we train a classifier that can predict emotions from a context offered by the user and a small history of the conversation using fastText [8]. To generate responses we apply fine-tuning of a generative pre-trained model [12] using PersonaChat [19] and Empathetic Dialogues [15] datasets.

RELATED WORK

Rashkin et al. [15] proposed a new benchmark for empathic dialogues that is a necessity for further research on this topic. In their work, they designed a novel dataset and tested two types of architectures using this data: retrieval-based, using BERT encoder [2] to find the best match, and generative, using Transformers [15].

MoEL [9] proposed a system that softly combines responses from multiple empathic listeners for each emotion. Despite the occasional confusion created by trying to generate a response from a high variance emotional distribution, the model achieves better results than a generic multi task transformer [17].

CAiRE [10] is the most recent empathic chatbot and it achieves state-of-the-art performance. PersonaChat [19] and Empathetic Dialogues [15] datasets are used to create an open-domain, end-to-end conversational agent.

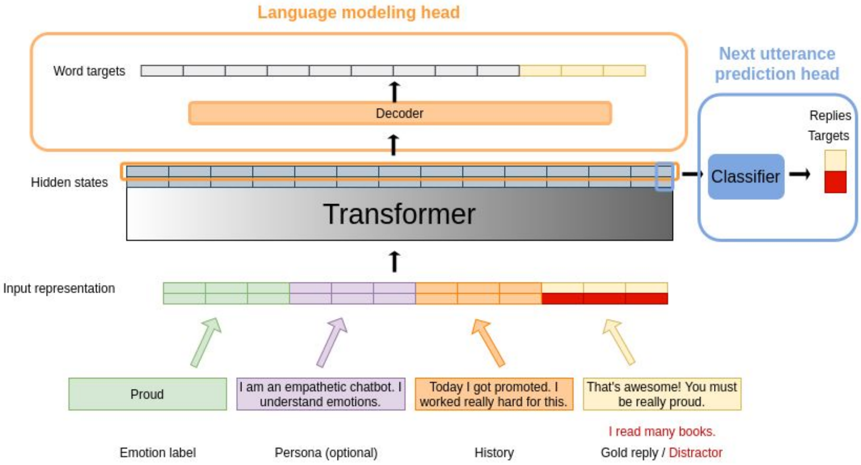

Figure 1. The model architecture used by the proposed empathic agent.

A recent work that focuses on creating a human-like conversational agent is presented by Roller et al. [16]. The authors aim to obtain a recipe for a robust open-domain chatbot that presents not only empathy, but also personal background and knowledge.

Datasets

PersonaChat [19] dataset makes a step towards a more engaging dialogue. This is a very useful dataset because it proposes a novel turn-based dialogue that can improve a conversational agent with a consistent personality. This dataset is made over 160,000 utterances between crowdworkers from Amazon Mechanical Turk, who received a different persona to mimic during the conversation. Each dialogue has a minimum 6 or 8 utterances, each with a limit of 15 words.

Empathetic Dialogues [15] is a novel dataset based on sentimental conversations between a speaker, who describes a situation when they felt in a certain way, and a listener that has to respond empathically. This dataset provides 32 emotions distributed almost evenly over 24,850 conversations from 810 Amazon Mechanical Turk workers. Each pair was asked to choose an emotion and every participant had to provide a description of a scene when they felt that way.

PROPOSED SOLUTION

The main steps followed by our solution are preprocessing the data, transforming the data to match with the pre-trained model input representation, converting the inputs into tensors, and finally, fine-tuning the model as presented in Figure 1. For the emotion classifier, this work uses an external trained model for text classification, fastText [3].

The base architecture we used in our work is a unidirectional GPT (Generative Pre-trained Transformer). The model follows standard fine-tuning [18] and it is based on the open-source implementation published on GitHub [5] which was extended for the Empathetic Dialogues dataset.

Input Representation

The GPT model receives as input a sequence of words. To fine-tune this model for a dialogue task, all the main features of a conversation must be represented in the resulting embedding space. The current work follows the same design presented by Wolf et al. [18] for the PersonaChat dataset, adapted to the Empathetic Dialogues.

To speed up training, other than the gold reply, represented by the correct next utterance, one distractor was used. The distractor was picked randomly from a pool of candidates that includes all possible individual utterances from the training dataset.

The input fed to the network must contain the concatenation of the emotion label, the custom persona (”i am an empathetic chatbot.”, ”i understand emotions.”, ”i am friendly.”, ”i want to help humans.”), the history of the conversation and the correct reply or the distractor. Apart from the word embeddings, the model needs more information about the input, such as the position for each token used by the attention mechanism and delimitation between the segments.

The delimitation is made using special tokens such as the start of a sequence, the end of a sequence and the indexes of the two speakers. A padding token is used to fill the remaining positions up to 512, which represents the length of the input sequence.

Training

In our work, we start from the GPT Double Heads [6] model and the corresponding tokenizer [7] released by OpenAI. We use a Double Heads model because during the fine-tuning we optimize a multi-task loss function that combines both language model loss and next-sentence prediction loss. For the transfer learning part, there are used 5 epochs, a batch size of 4, a learning rate equal to *6.25e-5* that was linearly decayed to zero during training, a max history for each speaker of 2 and 4 steps of gradient accumulation. The current work uses AdamW, an improved version of Adam optimizer.

To train the classifier in order to predict the emotion, the current work utilizes the fastText [3] library. The text from which the model learns is assembled from the context and the full conversation. The model is trained for 50 epochs and with a learning rate equal to *0.7*.

Decoding

During inference, a decoder is used to predict the next utterance as a sequence of words, based on the current input. The algorithm used is the combination of the top-k and top-p sampling. Top-*k* sampling reduces the candidates at the best *k* possibilities with the highest probability and top-*p* sampling keeps only the candidates for which the sum is greater than the *p* parameter. Therefore there are kept only the tokens with the higher probability acquired from the two different methods.

The parameters used for decoding are *k=0.8* and *p=0.9*. Other parameters used are *max_length=20* that limits the number of generated words and *min_length=1*. The max number of utterances in the history is 4.

EXPERIMENTS AND RESULTS

In this section, there will be presented two experiments that aim to explore different possible approaches and to compare their effectiveness in the context of dialogue systems that show empathy. The evaluation metrics used are perplexity and BLEU score to assess the current proposal performance relative to other models used for the same goal, and accuracy to measure the emotion classification results. Perplexity refers to the uncertainty from the model prediction and it measures how well it predicts the next golden token for the test data and BLEU score measures the quality of a text by calculating the distance between the generated text and the golden reply.

The first experiment is fine-tuning the GPT network using the Empathetic Dialogues dataset (denoted as **GPT+ED**). In this case, the input is built using the emotion label, the history, and the reply.

The second experiment is using transfer learning from the Empathetic Dialogues to the pre-trained GPT model using PersonaChat (denoted as **GPT+PC+ED**). For this attempt, we started from a pre-trained model [11]. In this case, the model already performs well on a general dialogue task and the experiment aims to adapt it for a more empathic version by prepending the emotion at the top of the input.

In Table 1 are presented the scores obtained by our models. Both of them are evaluated on the Empathetic Dialogues dataset. The best results in the domain of empathic chatbots are currently obtained by CAiRE [10], the state of the art model. The improvements that CAiRE brings are the addition of a persona that supports the empathic component as well as using a pre-trained model such as GPT. These aspects, that differentiated CAiRE from other models, were followed by the current work and the results presented in Table 1 show that the **GPT+PC+ED** version has only slightly lower scores.

Table 1. Comparison between existing models in the domain of empathic chatbots and the current work. The first five models were introduced by Rashkin et al. [15], based on a Transformer, pretrained on 1.7 billion conversations from Reddit. (The first six scores are taken from [15] and [10])

Model	PPL	BLEU	Accuracy
Pretrained	27.96	5.01	-
Fine-Tuned	21.24	6.27	-
MULTITASK	24.07	5.42	-
EmoPrepend-1	24.30	4.36	-
ENSEM-DM	19.05	6.83	-
CAiRE	**13.32**	**7.03**	**0.516**
GPT+ED	21.34	5.82	0.483
GPT+PC+ED	*18.20*	*6.56*	*0.483*

Table 2. Examples of generated answers and the predicted emotion based on an initial context.

Context	Real / Predicted	GPT+ED	GPT+PC+ED
I am so impressed by my current boss. He is a terrific example of leadership in every way.	Impressed / Impressed	that is awesome! i bet you are so happy to have him as your boss.	that is good to hear. what makes you think that?
I recently got very upset with my sister when said some very mean things to me. I let her know that she was out of line by yelling at her. I	Angry / Guilty	it's okay. you have to be honest with her.	i'm sorry, what did she say?

feel bad for both of us.			
I hate it when my husband goes away for work. There is no one else here but me.	Lonely / Lonely	why do you hate it when he is gone?	that's a bummer. have you tried talking to him about it?

One main difference perceived between the two experiments is the type of sentence the model generates as shown in Table 2. **GPT+ED** tends to generate statements that interrupt the flow of the conversation and that focus just on being empathic towards the speaker. **GPT+PC+ED** usually contributes to the discussion by asking questions while still revealing empathy. This dissimilarity may be caused by the datasets the models use. While Empathetic Dialogues focuses on expressing the emotion recognized from the text, the PersonaChat dataset has conversations where speakers want to find out more about each other.

By analyzing the confusion matrix for the emotion classifier, it shows that it performs well and predicts in most of the cases the correct label. For some sentiments such as “angry” and “furious” or “terrified” and “afraid”, the classifier is not capable of distinguishing properly, as they are close in signification.

Failure cases

Some of the failure cases met during the evaluation process consist in repetition of words and ideas, the inability to refer to past information or shallow understanding of concepts and ideas. However, these are recurring problems in conversational agents and may require fundamental novel approaches to mitigate the occasional poor performances.

CONCLUSION

In this work, we introduced an empathic conversational agent that can understand emotions during a discussion and respond properly. During the development of the dialogue system, we provided two experiments that show that the best approach is to endow the chatbot with a friendly personality that can help it generate more engaging and empathic answers. Using the PersonaChat dataset for fine-tuning the model increases the chit-chat capability and entertains longer conversations.

The main contribution of our paper is the proposal to not integrate the emotion classification into the training pipeline, and to independently learn to classify emotions to ease the training process and to make this prediction more robust. One of the immediate improvements that can greatly improve the performance of the model is to replace the transformer architecture with improved versions of GPT, such as GPT 2 [13] or GPT 3 [1].

The current results are promising and with further improvements, we will be able to build a truly empathic chatbot that displays more skills from humans’ social behavior and that can be used at large scale to nourish the emotional and social needs of people.

REFERENCES

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
3. FastText library https://fasttext.cc/ last accessed on 22 June 2020.
4. Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. Interactions, 24(4), 38-42.
5. GitHub https://github.com/huggingface/transfer-learning-conv-ai last accessed on 8 July 2020.
6. GPT Double Head model https://huggingface.co/transformers/model_doc/gpt.html#openaigptdoubleheadsmodel last accessed on 8 July 2020.
7. GPT Tokanizer https://huggingface.co/transformers/model_doc/gpt.html#openaigpttokenizer last accessed at 8 July 2020.
8. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
9. Lin, Z., Madotto, A., Shin, J., Xu, P., & Fung, P. (2019). Moel: Mixture of empathetic listeners. arXiv preprint arXiv:1908.07687.
10. Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., & Fung, P. (2020). CAiRE: An End-to-End Empathetic Chatbot. In AAAI (pp. 13622-13623).
11. Pretrained GPT with PersonaChat https://s3.amazonaws.com/models.huggingface.co/transfer-learning-chatbot/gpt_personachat_cache.tar.gz last accessed at 17 June 2020.
12. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
14. Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579.
15. Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. arXiv preprint arXiv:1811.00207.

16. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Boureau, Y. L. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
18. Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149.
19. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. arXiv preprint arXiv:1801.07243.