Multi-Task Learning of Generation and Classification for Emotion-Aware Dialogue Response Generation

Tatsuya Ide and **Daisuke Kawahara**
Waseda University
{t-ide@toki., dkw@}waseda.jp

Abstract

For a computer to naturally interact with a human, it needs to be human-like. In this paper, we propose a neural response generation model with multi-task learning of generation and classification, focusing on emotion. Our model based on BART (Lewis et al., 2020), a pre-trained transformer encoder-decoder model, is trained to generate responses and recognize emotions simultaneously. Furthermore, we weight the losses for the tasks to control the update of parameters. Automatic evaluations and crowdsourced manual evaluations show that the proposed model makes generated responses more emotionally aware.

1 Introduction

The performance of machine translation and summarization has been approaching a near-human level in virtue of pre-trained encoder-decoder models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). The same technology has been applied to dialogue systems, which are now expected to be put to practical use.

To interact naturally with a human, the computer needs to be human-like. Several methods have been proposed to build such dialogue systems. They include a system interacting based on knowledge and common sense (Dinan et al., 2019) and that interacting by considering one's own and the other's personality (Zhang et al., 2018). In particular, we focus on the viewpoint of emotion as targeted in Rashkin et al. (2019).

In this paper, we propose a multi-task learning method for building a dialogue system that takes the speaker's emotions into account. Also, we focus on the hierarchy of emotions (Kumar et al., 2019) and simultaneously train multiple emotion recognition tasks with different granularity. Our multi-task learning model is not expected to share complementary information among similar tasks as previous work (Liu et al., 2019), and we do not aim at improving the accuracy of emotion recognition. Instead, we focus on generating emotion-aware responses. Also, concerned that the ratio of emotion recognition in multi-task learning is too large, we explore further quality improvement by weighting each loss. We build a model based on BART (Lewis et al., 2020), a pre-trained Transformer (Vaswani et al., 2017) model, to implement multi-task learning of response generation and emotion recognition.

Experiments are performed using a dialogue corpus without context. The effectiveness of the proposed method in generating responses is confirmed by automatic and manual evaluations. Multi-task learning of response generation and emotion recognition makes generated responses more emotionally aware of utterances. The improvement is not only on the emotional aspect but also on the quality of fluency, informativeness, and relevance. We also found that controlling the parameters by weighting the losses improved the performance of the model.

2 Related Work

One of the previous studies on emotion-based response generation is the Emotional Chatting Machine (ECM) (Zhou et al., 2018). ECM is used together with an emotion classifier to generate a response based on a given emotion. EmpTransfo (Zandie and Mahoor, 2020) is a similar model to ours. Given an utterance, a model based on GPT (Radford et al., 2018) learns an emotion and an action simultaneously in addition to a response, which improves the quality of generated responses. These models focus on the emotion of a response so that they do not generate a response based on the emotion of an utterance.

Lubis et al. (2018) incorporate an emotion encoder into a hierarchical seq2seq architecture, enabling a system to understand the emotional context on a user. TG-EACM (Wei et al., 2020), the suc-

Proceedings of NAACL-HLT 2021: Student Research Workshop, pages 119–125
June 6–11, 2021. ©2021 Association for Computational Linguistics

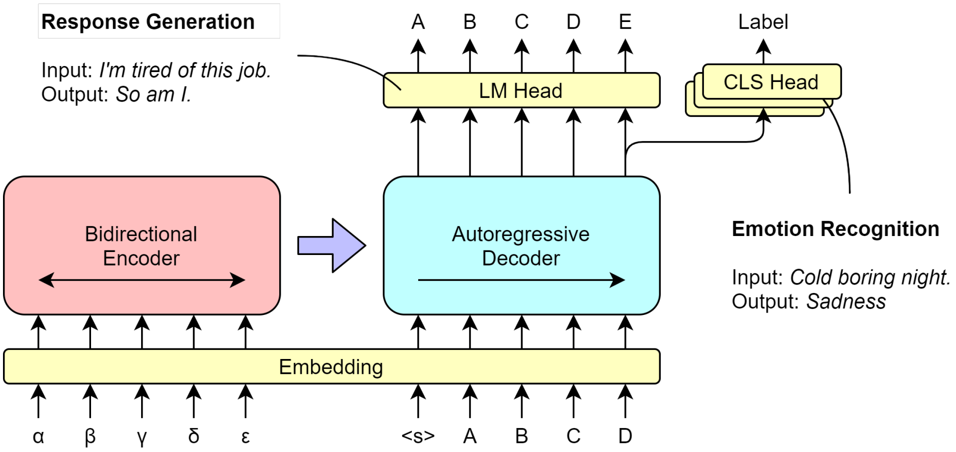


Figure 1: The architecture of our model, based on BART (Lewis et al., 2020). It contains one LM head and several CLS heads, which solve generation and classification, respectively. In our experiments, three CLS heads are used for the emotion recognition tasks with different granularity.

cessor of EACM (Wei et al., 2019), is a model that considers not only the emotion in an utterance but also the emotion that a response should have. The model learns a distribution to infer both the emotion of the utterance and the response from a given utterance. CARE (Zhong et al., 2021) uses some commonsense to generate a response with both rationality and emotion. Through latent concepts obtained from an emotionally aware knowledge graph, predicted responses can be emotional and rational.

Actually, the above models require separate units or special architecture for understanding emotion in a dialogue. In contrast, our proposed model achieves that with a single structure, inherited from Transformer (Vaswani et al., 2017) and BART (Lewis et al., 2020). In other words, our model does not need an extra unit. Therefore, the proposed method consequently reduces the redundancy of Transformer parameters (Kovaleva et al., 2019) and realizes more efficient understanding of emotion to generate a response.

3 Emotion-Aware Response Generation by Multi-Task Learning

3.1 Overview

Our model learns response generation as a generation task and emotion recognition as a classification task. By learning response generation and emotion recognition simultaneously through multi-task learning, it is possible to generate a response by considering the emotion of a given utterance.

Multi-task learning often involves several similar tasks because they can share information and thus the performance of each task can be improved. However, the purpose of our multi-task learning method is to improve the quality of response generation, not to improve the performance of emotion recognition. This is different from general multi-task learning.

Our model is based on BART (Lewis et al., 2020). Its architecture is shown in Figure 1. The model has several output layers, or heads, for the tasks to be trained, which include an LM head for generating words in response generation and CLS heads for solving classification tasks. Given a sentence, the CLS head predicts its label such as `positive` or `negative`. One CLS head is set for each classification task.

The input/output format of each task is the same as that in BART. In the generation task, we put an utterance and a right-shifted response into the encoder and decoder, respectively. In the classification task, we put an utterance and a right-shifted utterance into the encoder and decoder, respectively. Following the learning algorithm of MT-DNN (Liu et al., 2019), each task that the model learns is selected for each mini-batch. A different loss is calculated for each task, and the parameters are updated for each mini-batch.

3.2 Losses of Generation and Classification Tasks

Let $\boldsymbol{x} = (x_1, \ldots, x_M)$ be the given utterance and $\boldsymbol{\theta}$ be the parameters of the model. Our model is trained by updating $\boldsymbol{\theta}$ based on the loss for each task.

Dataset	Train	Validation	Test
DailyDialog	76,052	7,069	6,740
TEC	16,841	2,105	2,105
SST-2	16,837	872	1,822
CrowdFlower	15,670	1,958	1,958

Table 1: The statistics of the datasets for our experiments, where TEC stands for Twitter Emotion Corpus. Because TEC and CrowdFlower have no split of train, validation, and test, we split them into three at 8:1:1.

Generation The response to $\boldsymbol{x}$ is defined as $\boldsymbol{y} = (y_1, \ldots, y_N)$. The model infers an appropriate $\boldsymbol{y}$ from $\boldsymbol{x}$. The generation loss $\mathcal{L}_{\text{gen}}$ is calculated as the negative log-likelihood loss.

$$\mathcal{L}_{\text{gen}} = -\sum_{j=1}^{N} \log p(y_j | \boldsymbol{x}, y_1, \ldots, y_{j-1}; \boldsymbol{\theta}) \quad (1)$$

Classification If the correct label of $\boldsymbol{x}$ is c, the model infers c from $\boldsymbol{x}$. The negative log-likelihood loss is also used for the classification loss $\mathcal{L}_{\text{cls}}$.

$$\mathcal{L}_{\text{cls}} = -\log p(c | \boldsymbol{x}; \boldsymbol{\theta}) \quad (2)$$

3.3 Loss Weighting

Although the proposed multi-task learning model learns the generation and classification tasks simultaneously, there is a possibility that the ratio of learning for the classification task is too large. When solving a general classification task, the end of learning is often determined by the convergence of the loss in the validation data. On the other hand, the target of our model is a generation task, and the number of epochs required for generation is larger than that of the classification task.

Therefore, we consider weighting the loss functions. While the weight for response generation is fixed at 1, the weight for emotion recognition is varied between 0 and 1. This makes the contribution of the classification task reduced in updating the parameters.

4 Experiments

4.1 Datasets

We train a model with three tasks of emotion recognition in addition to response generation using multi-task learning. Each emotion recognition task is a classification task with 6, 2, and 12 labels, and we call them emotion recognition, coarse-grained emotion recognition, and fine-grained emotion recognition, respectively. The datasets for such emotion recognition were selected according to Bostan and Klinger (2018). The numbers of instances are summarized in Table 1.

Response Generation DailyDialog (Li et al., 2017) is used for response generation. The dataset is a multi-turn dialogue corpus, and we obtain pairs of an utterance and a response by extracting two turns at a time. Each utterance in the corpus has an emotion label, but we do not use these labels in the experiment. This is because almost all of the emotion labels are `other`, which is not suitable for our method.

Emotion Recognition For the core emotion recognition dataset, we use the Twitter Emotion Corpus (Mohammad, 2012). It was constructed based on Twitter hashtags and consists of six labels: {`anger`, `disgust`, `fear`, `joy`, `sadness`, `surprise`}. Because there is no distinction between train, validation, and test in the dataset, 80% of the total samples is assigned to train, and the remaining 10% each is assigned to validation and test.

Coarse-Grained Emotion Recognition For coarse-grained emotion recognition, we use SST-2 (Socher et al., 2013). This is a dataset of movie comments labeled with {`positive`, `negative`}. To maintain a balance with the number of instances for the other emotion recognition tasks, we reduce the number of instances for training to 25%.

Fine-Grained Emotion Recognition For fine-grained emotion recognition, we use the emotionally-tagged corpus provided by CrowdFlower.[1] We exclude the label `empty` and adopt this corpus for a classification task with 12 labels: {`anger`, `boredom`, `enthusiasm`, `fun`, `happiness`, `hate`, `love`, `neutral`, `relief`, `sadness`, `surprise`, `worry`}. As with the Twitter Emotion Corpus, this corpus does not have a split of train, validation, and test, and thus the whole data is divided into 8:1:1. Furthermore, for the same reason as in SST-2, only 50% of the total data is used.

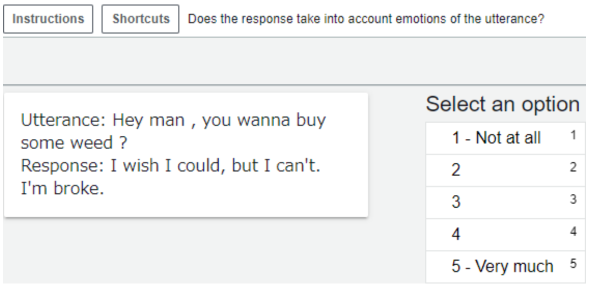
4.2 Training

The hyperparameters are set based on BART (Lewis et al., 2020) and the Fairseq

[1] The original link is no longer available. An alternative is https://data.world/crowdflower/sentiment-analysis-in-text.

Model	Auto Eval				Manual Eval			
	BLEU	*dist*-1	*dist*-2	Avg Len	*Emo*	*Flu*	*Info*	*Relv*
R	32.35	5.87	30.48	14.12	3.44	3.48	3.63	3.55
R+E6	32.29	5.93	30.48	14.12	**3.59**	**3.82**	3.62	**3.96**
R+E6+E2	32.39	**6.00**	**30.77**	14.11	3.58	3.75	**3.74**	3.70
R+E6+E12	**32.55**	5.89	30.57	**14.14**	3.52	3.48	3.55	3.58
R+E6+E2+E12	32.29	5.91	30.47	14.12	**3.59**	3.75	3.57	3.64

Table 2: Evaluation results of our models by multi-task learning. R stands for response generation, and E• is emotion recognition with • labels. *Emo*, *flu*, *info*, and *relv* are the four aspects for the manual evaluation by crowdsourcing.



Does the response take into account emotions of the utterance?

Utterance: Hey man , you wanna buy some weed ?
Response: I wish I could, but I can't. I'm broke.

Figure 2: An example of the manual evaluation by crowdsourcing on Amazon Mechanical Turk. Workers are supposed to answer such questions by rating the given dialogue on a five-point scale.

example.[2] The learning rate is set to 3e-5, and the parameters are optimized by Adam with weight decay. For response generation, we apply label smoothing of 0.1 to the negative log-likelihood loss. The number of input and output tokens is set to 64, and training is performed for 64 epochs. We use beam search with 5 beams to select words and eliminate cases where there are more than three repeated n-grams. Training and generation are performed on NVIDIA Tesla V100.

4.3 Evaluation Metrics

We evaluate the trained models automatically and manually.

Automatic Evaluation First, we evaluate how much the output responses are related to the correct response using BLEU (Papineni et al., 2002). Second, we evaluate whether the output responses are lexically diverse using *distinct* (Li et al., 2016). For *distinct*, *distinct*-1 and *distinct*-2 are calculated, which focus on unigrams and bigrams, respectively. We also compare the average number of words in output responses, which is based on the assumption that the longer a response is, the less common it is. The large average number indicates that generated responses tend to be not dull.

Manual Evaluation Actually, the lack of correlation between automatic and manual evaluation (Liu et al., 2016) has been indicated especially in regards to generation tasks. Thus, we perform manual evaluation by crowdsourcing, where Amazon Mechanical Turk is used as the platform. We use four metrics mainly following Rashkin et al. (2019): *emotion*, *fluency*, *informativeness*, and *relevance*. Each of the questions asks whether the generated response takes into account the emotion of the utterance, whether the generated response is syntactically correct, whether a generated response provides some information for the utterance, and whether the content of the response is appropriately related to the utterance. A total of 100 randomly selected responses for the test data are asked to rate the above four metrics on a five-point scale. US residents are designated as workers, and seven workers are requested for each metric of each sample. The final score is obtained as the average of the values obtained from the seven workers. An example of the questions asked to the workers is shown in Figure 2.

4.4 Results

Multi-Task Learning The evaluation results are shown in Table 2. The response generation is denoted by R, and the emotion recognition for the Twitter Emotion Corpus, SST-2, and CrowdFlower datasets is denoted by E6, E2, and E12, respectively. In terms of automatic evaluation, R+E6+E2 and R+E6+E12 maximized the *distinct* and BLEU, respectively. In the proposed multi-task learning model, therefore, emotion recognition of different granularity is effective in relevance and diversity. For manual evaluation, all models that include emo-

[2] https://github.com/pytorch/fairseq/blob/master/examples/bart/README.summarization.md.

Utterance	Gold Response	R	R+E6
Yes, my parents, and soon my brother as well.	I didn't know you had a brother.	I see. In a few months, you'll be moving to a new place.	I hope I get to meet you soon.
It falls on the fifth day of the fifth lunar month.	Could you tell me the origin of the two festivals?	Oh, then on that day?	Oh? Why is that day?
How smart of you to book the tickets before hand!	Oh, do you really think so?	Sorry, I didn't think of it that way.	Yeah, yeah, I know. It was the last thing I wanted to do.
I can't. I'm waiting for Paul, but he's late.	I'll say the movie's starting in the minute.	Would you like me to call him?	Where is he?
thank you. I think I should also have a look at the Internet to see if anyone's got a sublet.	how much do are you looking to spend?	if you need a flat, you can always look online. Do you have any questions?	good idea. You can ask around at the office. If you need a flat, they'll probably have one.

Table 3: Examples generated by the baseline model (R) and the proposed model (R+E6). We chose examples where the *emotion* score in manual evaluation is quite improved by adding E6 to R.

Model	Acc [%]	F1 [%]
R	17.29	12.25
R+E6	64.32	56.53
R+E6+E2	63.23	54.88
R+E6+E12	66.27	57.86
R+E6+E2+E12	64.85	56.09

Table 4: Emotion recognition (E6) performance of our models in Table 2. The values for R, trained only on response generation, are very low, while R+E6+E12 marks the best score among these models.

tion recognition outperformed the model with only response generation. Moreover, R+E6 scores were particularly high for all four metrics. The proposed multi-task learning model not only makes the generated responses more emotionally aware but can also improve the quality of other metrics, such as fluency and informativeness.

Several examples of responses generated by the obtained model are shown in Table 3. We compare the given utterances and their responses of R and R+E6. We can see that R+E6 generated more emotion-sensitive sentences, such as "Yeah, yeah, I know" and "good idea."

In addition, we show the results of emotion recognition in Table 4, which is especially on a six-label classification task. We calculate accuracy and F1-score as metrics for evaluation. The result shows that, on emotion recognition, increasing the number of tasks to train does not necessarily lead to improvement of the scores. We can see that models with training of emotion recognition on fine-grained labels tend to outperform the other models. However, the goal of our model is not improvement of classification but that of generation, so that those score variation is not essential in this work.

Loss Weighting The evaluation results for different loss weighting are shown in Table 5. The weight for the loss of E• is denoted as $\lambda_{\mathrm{E}\bullet}$. In automatic evaluation, we can see the improvement of the scores by weighting, especially in the model with E12. On the other hand, the manual evaluation shows that weighting improves some scores, with the case (.5, .5, 0) producing the highest score. Therefore, weighting each loss can improve the quality of generated responses, and in the condition of our experiment, it is most effective to reduce the weights of E6 and E2 by half.

5 Conclusion

We worked on improving the quality of neural network-based response generation. Focusing on the aspect of emotion, we proposed a multi-task learning response generation model that includes the tasks of generation and classification. Through automatic and manual evaluations, we confirmed that the proposed model improved several metrics of performance. Moreover, we further improved the quality of the model by weighting losses. As a result, we found that such weighting improved

$(\lambda_{\mathrm{E6}}, \lambda_{\mathrm{E2}}, \lambda_{\mathrm{E12}})$	Auto Eval				Manual Eval			
	BLEU	*dist*-1	*dist*-2	Avg Len	*Emo*	*Flu*	*Info*	*Relv*
(1, 0, 0)	32.29	5.93	30.48	14.12	3.59	3.82	3.62	**3.96**
(.5, .5, 0)	32.48	5.86	30.54	**14.15**	**4.00**	**4.16**	**4.01**	**3.96**
(.5, 0, .5)	**32.52**	5.93	30.62	14.04	3.37	3.60	3.37	3.36
(.33, .33, .33)	32.43	**5.97**	**30.81**	14.01	3.63	3.37	3.49	3.66

Table 5: Evaluation results for differed loss. $\lambda_{\mathrm{E}\bullet}$ indicates the weight for the loss of E•, and the metrics are the same as those of Table 2. The weight for the response generation loss (λ_{R}) is fixed at 1 throughout the experiments. Note that (1, 0, 0) is equivalent to R+E6 in Table 2.

several scores and the balance of parameter updates was also an important factor.

This paper focused on the emotion of the dialogue and generated responses that take into account the emotion of an utterance. On the other hand, we did not focus on the emotion of a response, which is a subject for our future work. We plan to work on estimating the emotions that a response should have and generating a response based on a specified emotion. In the experiments of this paper, we omitted the context of a dialogue. However, it is also necessary to consider past utterances and their effects on emotions for generating responses, which is also an issue to be addressed in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP18H03286.

References

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

A. Kumar, A. Ekbal, D. Kawahra, and S. Kurohashi. 2019. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining 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.

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. *Proceedings* *of the AAAI Conference on Artificial Intelligence*, 32(1).

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wei Wei, Jiayi Liu, Xianling Mao, Guibin Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. 2020. Target guided emotion aware chat machine. *arXiv preprint arXiv:2011.07432*.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1401–1410, New York, NY, USA. Association for Computing Machinery.

Rohola Zandie and Mohammad H. Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Care: Commonsense-aware emotional response generation with latent concepts. *arXiv preprint arXiv:2012.08377*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).