

Affective Decoding for Empathetic Response Generation

Chengkun Zeng[♣], Guanyi Chen[♡], Chenghua Lin^{♣*}, Ruizhe Li[♣], Zhigang Chen[♣]

[♣]Department of Computer Science, University of Sheffield

[♡]Department of Information and Computing Sciences, Utrecht University

[♣]College of Information and Intelligent Engineering, Zhejiang Wanli University

chengkunuzenggo@gmail.com, {c.lin, r.li}@sheffield.ac.uk

g.chen@uu.nl, chenzhigang@zwu.edu.cn

Abstract

Understanding speaker’s feelings and producing appropriate responses with emotion connection is a key communicative skill for empathetic dialogue systems. In this paper, we propose a simple technique called Affective Decoding for empathetic response generation. Our method can effectively incorporate emotion signals during each decoding step, and can additionally be augmented with an auxiliary dual emotion encoder, which learns separate embeddings for the speaker and listener given the emotion base of the dialogue. Extensive empirical studies show that our models are perceived to be more empathetic by human evaluations, in comparison to several strong mainstream methods for empathetic responding.

1 Introduction

Endowing a dialogue system with the ability of empathy responding has attracted a growing body of research (Ma et al., 2020) and is believed to be crucial for many service-oriented applications, such as mental health interventions (Hoermann et al., 2017), assisting medical diagnosis (Xu et al., 2019), and restaurant/hotel booking services (Ghazvininejad et al., 2017; Liu et al., 2018; Wang et al., 2021). Being empathetic requires one to be able to understand the implied feeling of the conversation partner, or in other words, to place oneself in partner’s position. Therefore, to produce proper responses, an Empathetic Dialogue System (EDS) needs to understand not only the situation of the speaker¹ and the causes (Abd Yusof et al., 2017), but also the emotion state of speaker.

In 2019, Rashkin et al. (2019) formally introduced the task for dialogue systems to respond

*Corresponding author

¹We use the term “speaker” referring to the user of the empathetic dialogue system while “listener” inferring to the dialogue system itself.

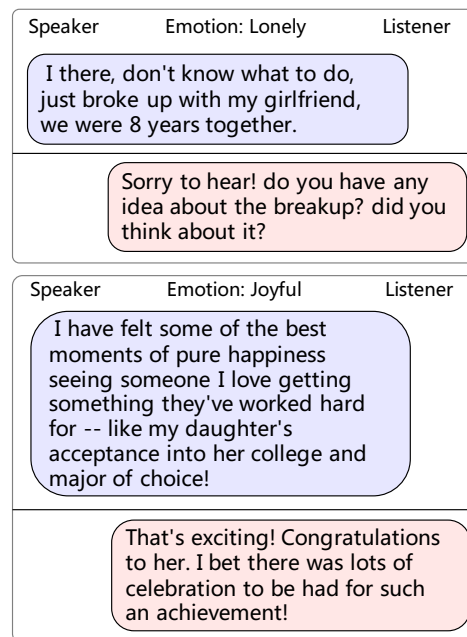


Figure 1: Two example dialogue sessions from the EMPDIAL dataset with the emotion *lonely* and *joyful*, respectively.

to conversations with emotions. They also constructed a benchmark corpus called EMPATHETICDIALOGUES (abbreviated as EMPDIAL), which consists of conversations with a wide range of emotion states for task evaluation. Figure 1 shows an example session of a dialogue from EMPDIAL, where the situation reflects the emotion state of *lonely* and *joyful*. Several approaches have been proposed for modelling emotions, which is a key challenge for building an EDS. These approaches follow two main enterprises: one is multi-task learning (Rashkin et al., 2019; Lin et al., 2020), which trains models for both dialogue generation and predicting the emotion of the dialogue; the other enforces the model to generate empathetic responses conditioning on the emotion state predicted from the dialogue context with a pre-trained

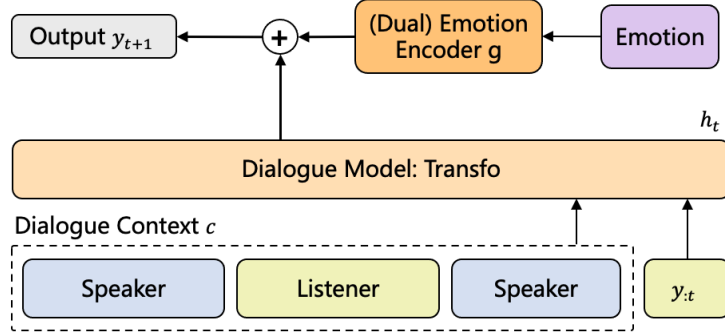


Figure 2: The overall architecture of our empathetic dialogue system.

emotion classifier (Rashkin et al., 2019).

Our work follows a similar vein to the second enterprise, where we propose a simple yet efficient technique coined *Affective Decoding* (AD), which can effectively incorporate emotion signals into model training and generate more empathetic responses. Our method can work in two different modes. The first mode injects emotion embedding in each decoding step. This is different to Rashkin et al. (2019) who only applied prepending at the first time-step on the encoder. For the second mode, we introduce an additional auxiliary *Dual Emotion Encoder*, which learns separate embeddings for the speaker and listener given the emotion base of the dialogue. In addition, we systematically evaluate and compare AD with the existing mainstream emotion modelling methods for empathetic responding, including both prepending emotion embeddings (Rashkin et al., 2019) and multi-task learning (Lin et al., 2020).

Based on the EMPDIAL dataset, we conducted comprehensive empirical studies, which include automatic (e.g., BLEU, BOW) and two human evaluations. For human evaluation, we assess both model-level performance and finer-grained level aspects concerning empathy, relevance, and fluency of the generated responses. Empirical results show the effectiveness of our affective decoding and that our model with the auxiliary dual emotion encoder works the best. While Rashkin et al. (2019) reported that multi-task learning did not provide consistent improvements for the task, we actually found multi-task learning performs even worse than a pre-trained language model (Wolf et al., 2019) fine-tuned on the EMPDIAL dataset.

To summarise, our contributions are 3-fold: (1) we introduce a simple yet efficient decoding method called *affective decoding* to the task of em-

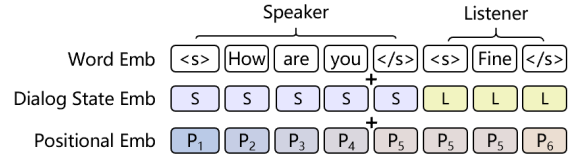


Figure 3: Illustration of the input format of our model.

pathetic response generation; (2) we conduct a comprehensive comparison between various emotion modelling methods in empathetic dialogue modelling by means of automatic evaluation and 2 human evaluations; (3) empirical results show the effectiveness of our affective decoding method and that with the auxiliary dual emotion encoder, our model can further support the analysis of listeners and speakers’ behaviours in terms of how they utter with respected to the same emotion.

The rest of the paper is organised as follows. §2 presents our model for empathetic response generation. We show the experimental setup and results in §3. §4 presents some case studies and finally §5 concludes the paper. The code is available at: <https://github.com/zenggo/affective-decoding-4-empathetic-dialog>.

2 Methodology

In this section, we describe our Affective Decoding model, which consists of two key components, namely, a pre-trained response generator, *Transfo* (Wolf et al., 2019), and the *affective decoder* for enhancing empathetic responding.

2.1 Dialogue Modelling

We use *Transfo*, which is built upon the Generative Pre-trained Transformer (Radford, 2018, GPT) pre-trained on the BOOKSCORPUS dataset (Zhu et al., 2015), and which gives good performance on

building conversational agents (Dinan et al., 2020). When fine-tuning on the EMPDIAL dataset, a response is generated given the dialogue context c , which contains single or multi-turn conversations. For each input token, it is represented as a summation of its **word embedding, positional embedding and dialogue state embedding**, as illustrated in Figure 3. We model two possible dialogue states, where state S corresponds to the speaker and state L to the listener.

2.2 Affective Decoding

One of the key challenges of building an effective EDS is recognising and understanding the emotion of the speaker. Inspired by the affective language model of Ghosh et al. (2017), we tackle the problem by proposing *Affective Decoding* (AD), a simple strategy **which injects emotion embeddings into each decoding step**. Such a strategy allows our model to encode dialogue’s emotion base effectively, and to distribute more probability mass towards the words in the utterance that are highly correlated with the dialogue emotion, leading to enhanced empathetic responding.

Concretely, at each time step t , we first encode the emotional label with a one-hot embedding, which is then mapped into a dense vector $g(e)$ by the emotion encoder $g(\cdot)$ (see Figure 2 for details). $g(e)$ is then used for predicting the next word y_{t+1} jointly with the dialogue context c and the decoded outputs for all previous time steps $y_{:t}$. Formally, the probability of $P(y_{t+1})$ is given as

$$P(y_{t+1}|y_{:t}, c, e) = \text{softmax}(Wh_t + Vg(e)), \quad (1)$$

where h_t is the representation of c and $y_{:t}$ encoded by *Transfo*; W and V are weights in the output layer. Similar to prior studies (Rashkin et al., 2019; Lin et al., 2020), our AD model maintains one emotion embedding for the whole dialogue session. **Dual Emotion encoder.** We observe that in the dialogues with emotional situations, speaker and listener tend to utter with distinctive styles. **That is, the speaker normally describes his/her own experience with personal emotions, while the listener tries to respond in the way which can establish an emotional connection with the speaker based on speaker’s emotional needs** (e.g., encouraging and motivating). For example, in the dialogue with a emotional base of *joyful* in Figure 1, the speaker used phrases like “*happiness*” and “*love*” while listeners used “*exciting*” and “*congratulation*”. Based on this observation, we introduce a

mechanism **so called Dual Emotion (DE) encoder**, which learns separate embeddings for the speaker and listener given the emotion base of the dialogue. We coin our model augmented with the auxiliary DE component **AD+DE**, and its generation process becomes:

$$P(y_{t+1}|y_{:t}, c, e) = \begin{cases} \text{softmax}(Wh_t + V_S g_S(e)) & s_t = S \\ \text{softmax}(Wh_t + V_L g_L(e)) & s_t = L, \end{cases} \quad (2)$$

where $s_t \in \{S, L\}$ is the dialogue state at the step t . With the dual embedding space, we hypothesise that the interpretability of our model’s behaviour will also be enhanced, as it makes possible to identify the differences of the language use between speakers and listeners.

3 Experiment

We evaluate our models on EMPDIAL, using the original split of Rashkin et al. (2019) and their emotion classifier based on FastText (Joulin et al., 2017). Table 1 shows the statistics of the EMPDIAL dataset, which contains 32 emotion labels. We compare our model against four competitive baselines in the experiment, including

- **MoEL**: a transformer model with multiple independent transformer decoders for generating different contextual responses (Lin et al., 2019);
- **Transfo**: a pre-trained transformer model which is fine-tuned using multi-task learning in language modelling and next-utterance classification tasks (Wolf et al., 2019);
- **PRE**: a Transfo model with an emotion embedding prepended to the dialogue context (Rashkin et al., 2019);
- **MTL**: a Transfo model with multi-task learning, where the main task is dialogue response generation and the secondary task uses the encoded dialogue context for predicting the emotion for the whole session (Lin et al., 2020).

In terms of our own models, apart from **AD** and **AD+DE**, we also further tested a model variant (**ADM**), which considers multi-task learning. We detail each of our model variants below.

- **AD**: a simple model by injecting emotion embeddings into each decoding step;

Dataset	Emotion Labels	Train	Validation	Test
EMPDIAL	32	19,533	2,770	2,547

Table 1: The statistics of EMPDIAL dataset.

Model	PPL	BLEU	BOW _e	BOW _a	BOW _g	DIST-1	DIST-2
MoEL	38.19	2.84	0.502	0.878	0.679	0.005	0.023
Transfo	13.94	1.75	0.729	0.747	0.559	0.015	0.070
Prepend	13.90	1.75	0.731	0.753	0.565	0.016	0.079
MTL	14.95	1.49	0.733	0.757	0.566	0.015	0.067
ADM	14.50	1.81	0.733	0.756	0.564	0.015	0.068
AD	14.04	1.69	0.734	0.756	0.570	0.016	0.072
AD+DE	14.03	1.71	0.736	0.757	0.571	0.016	0.074

Table 2: Automatic evaluation results.

- **AD+DE**: a variant of **AD** by introducing the Dual Emotion encoder to separately model embeddings for the speaker and listener;
- **ADM**: a variant of our model by combining **AD+DE** with multi-task learning, adopting a similar strategy to the **MTL** baseline.

3.1 Automatic Evaluation

For automatic evaluation, we evaluate the models in three aspects, i.e., *fluency*, *adequacy*, and *diversity*. Particularly, fluency is measured by perplexity, adequacy by BLEU and BOW embedding metrics, and diversity by DIST. We describe each of the metrics in detail below.

- **Perplexity (PPL)**: measures how well a language model is (lower the better);
- **BLEU (Papineni et al., 2002)**: n -gram overlap between the system output and the reference;
- **BOW embedding (Liu et al., 2016b)**: the cosine similarity between the bag-of-words embeddings of the output and the reference. Specifically, there are three matching strategies:
 - Greedy (**BOW_g**): the average cosine similarities between word embeddings of the two utterances which are greedily matched (Rus and Lintean, 2012);
 - Average (**BOW_a**): the cosine similarity between the averaged word embeddings in the two utterances (Mitchell and Lapata, 2008);
 - Extreme (**BOW_e**): the cosine similarity between the largest extreme values in the word embeddings of the two utterances (Pennington et al., 2014);
- **DIST (Li et al., 2016)**: measures the corpus level diversity of the outputs by calculating the ratio of unique n -grams ($n = 1, 2$) over all n -grams in the outputs.

3.1.1 Experimental Results

Table 2 shows the automatic evaluation results for the tested models. Overall, the results do not seem to provide strong evidence in terms of which models perform best. Among the baselines, MoEL achieves the highest BLEU, BOW_a and BOW_g, while it has the worst PPL and diversity (i.e., DIST-1 and DIST-2). Prepend in contrast, performs the best in terms of PPL and diversity, and gives similar performance in the BOW metrics when compared to other baselines (except MoEL) and our models.

Our AD+DE model gives similar performance to Prepend, i.e., it achieves fairly balanced performance across all types of metrics and gives the highest scores in BOW_e and DIST-1. AD+DE also appears to slightly outperform AD, but the difference is somewhat minimal. In addition, it is surprising to see no significant difference between Transfo and other models for all metrics, where the latter explicit account for the emotional signals of the dialogue. We also see that MTL has a lower BLEU score even than Transfo. Conversely, comparing AD+DE and ADM, multi-task learning helps to yield better BLEU but yields worse performance

on other metrics. In summary, we are not able to establish a clear winner based on automatic metrics, although Prepend seems to slightly outperform other baseline models overall.

3.2 Human Evaluation

To assess the performance of the tested models more robustly and comprehensively, we conducted two forms of human evaluation: *ranking* for evaluating the overall performance of each system (Duh, 2008), and *multi-item rating* (Diamantopoulos et al., 2012) for evaluating the system performance against more fine-grained aspects (e.g. whether the response is relevant or not).

3.2.1 Ranking based Human Evaluation

We use pairwise binary ranking (i.e., preference test) (Vilar et al., 2007), which has been shown reliable for comparing the performance of multiple models. We randomly sample 100 dialogue context from the test set for both single-turn and multi-turn dialogues (i.e., 50 samples each type). We then generate a response with each tested model given a sampled context. Given two responses generated by two models, two raters (PhD students in computer science) were asked to decide *which model is better in terms of empathetic responding or there is no difference*.

We report the results of this pairwise preference test in Table 3, and the corresponding break down results of the single-turn and multi-turn dialogues in Table 4. Take the number 48 corresponding to Transfo and MoEL in Table 3 as an example. It means that 48% of the judges prefer Transfer over MoEL by considering both single-turn and multi-turn dialogues in the test set. Table 4 gives the break down results for single-turn (i.e., 46%) and multi-turn dialogue (i.e., 50%), respectively. By taking the average, we can derive 48% as the overall result. Clearly, human evaluation (i.e., Table 3) shows very different observations compared to the automatic evaluation. On the one hand, AD and AD+DE are clear winners this time, which significantly outperform all other models including the best performed baseline *PRE*. It can also be observed that AD+DE slightly outperformed AD but the difference is insignificant. On the other hand, multi-task learning shows a negative effect on empathetic dialogue modelling, i.e., by comparing MTL with Transfo and by comparing ADM with AD+DE. We give more discussions regarding this phenomenon in the Rating experiment section.

In addition, it can be observed that MoEL gives the worst performance compared to all other models by a large margin, but one might argue that the results are not directly comparable because the non pre-trained MoEL has less capacity than other pre-trained baseline models (e.g., the parameters of Transfo are 5 times as many as that of MoEL). The inconsistency on the results of automatic evaluation and preference test somewhat resemble the observation of prior studies that automatic metrics show low validity for evaluating empathetic dialogue systems (Liu et al., 2016a). To further investigate the underlying issue, we interview our raters as to which factor influences most on their decisions. It turns out that small errors in the responses that cannot be detected by the automatic measures (e.g. BLEU or BOW) can have a great impact. For instance, wrong reference (e.g., responding “*I’m happy for you.*” when the speaker is actually describing an experience of his/her sister) or wrong tense (e.g., responding “*I hope you will be fine.*” when the speaker is describing an experience happened in the past).

3.2.2 Rating based Human Evaluation

Likert Scale Rating (LSR) and Magnitude Estimation (ME) are two popular rating based methods. It is reported that ME performs better for evaluating goal-oriented dialogue systems (Santhanam and Shaikh, 2019) and language generation systems (Novikova et al., 2018) while LSR works better for measuring acceptability of text (Langsford et al., 2018). Considering the degree of empathetic is tied to the acceptability of the generated responses and that multi-item LSR is on a par with ME (van der Lee et al., 2019), we opt for LSR with three dimensions listed below. Model responses (the same set used in the ranking study) were scored by same two raters. The rating score ranges from 0 to 3.

- Empathy: *Does the listener understand the speaker’s feelings, and responds appropriately?*
- Relevance: *Is the content of the reply relevant to the topic mentioned by the speaker? Is it informative?*
- Fluency: *Does the response look fluent?*

The rating results in Table 5 and Table 6 (break down results) show a similar tendency with the

Model	MoEL	Transfo	PRE	MTL	ADM	AD	AD+DE
MoEL	-	12	11	26	18	11	12
Transfo	48 [†]	-	23	28	29	13	17
PRE	52 [†]	34 [†]	-	49 [†]	34	19	22
MTL	48 [†]	25	19	-	27	16	15
ADM	56 [†]	45 [†]	26	46 [†]	-	22	21
AD	55 [†]	39 [†]	45 [†]	44 [†]	40 [†]	-	8
AD+DE	57 [†]	42 [†]	35 [†]	45 [†]	43 [†]	12	-

Table 3: Results of the pairwise preference test: the number indicates the percentage (%) of responses generated by system A (row) is favoured by raters comparing to B (column). [†] means the preference is significant with $p < .05$ (two-proportion z-test).

Model	Single-turn							Multi-turn						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1 MoEL	-	15	12	29	19	14	16	-	9	10	23	17	8	8
2 Transfo	46	-	14	28	36	14	22	50	-	32	28	22	12	12
3 Prepend	47	42	-	46	44	24	20	57	26	-	52	24	14	24
4 MTL	42	32	16	-	26	16	16	54	18	22	-	28	16	14
5 MTL+AD+DE	52	44	22	44	-	28	28	60	46	30	48	-	16	14
6 AD	49	52	40	48	46	-	16	61	26	50	40	34	-	8
7 AD+DE	51	52	32	42	48	10	-	63	32	38	48	38	6	-

Table 4: Results of the pairwise preference test for single-turn EDS (left) and multi-turn EDS (right): the number indicates the percentage (%) of responses generated by system A (row) is favoured by raters comparing to B (column).

Model	Empathy	Relevance	Fluency
MoEL	1.629 ^{†‡}	1.525 ^{†‡}	2.031 ^{†‡}
Transfo	1.987 [†]	2.043 [†]	2.307
PRE	1.940 [†]	2.043 [†]	2.270
MTL	1.850 [†]	1.917 [†]	2.313
ADM (ours)	2.020 [†]	2.012 [†]	2.330
AD (ours)	2.177 [‡]	2.233 [‡]	2.380
AD+DE (ours)	2.187 [‡]	2.237 [‡]	2.387

Table 5: LSR Results. [†] means AD+DE significantly outperforms the corresponding model with $p < .05$ and [‡] means the corresponding model outperforms Transfo significantly with $p < .05$ (paired t -test).

ranking experiment and give some additional insights. Regarding *fluency*, how or whether a model incorporates emotion information seems to have no impact to the fluency of the generated responses. In terms of the other two aspects, we have the following findings: (1) similar to the ranking experiment, MoEL gives the worst performance, regardless the rating aspect. Comparing to *transfo*, the PRE, MTL, and ADM models cannot improved either *empathy* or *relevance* of the generated responses, significantly. The remarkable well performance of the vanilla *transfo* model embodies that by fine-tuning the model on EMPDIAL, this GPT

based model is a decent baseline for understanding emotion and responding empathetically; (2) in line with the ranking experiment, AD and AD+DE give the best performance. Although AD+DE performs slightly better than AD, the difference between them is not significant. Joining with other results, it seems that learning separate embeddings for the speaker and listener does bring some benefit but it is not as strong as expected. Nonetheless, we found that introducing DE can help analyse the behaviours of listeners and speakers in terms of how they utter with respected to the same emotion situation, which will be discussed in detail in §4.2; (3) comparing the results of MTL with Transfo and that of ADM with AD+DE reveals that MTL decreases both empathy and relevance of the responses. One possible reason behind why MTL does not yield positive effect in EDS (based on the results of both ranking and rating experiments) is that there might exist trade-off between the optimisation of the dialogue generator’s objective and that of the emotion classifier’s objective (Sener and Koltun, 2018). As a result, the overall performance is harmed by the naive linear combination of the two objectives.

Model	Single-turn			Multi-turn		
	Empathy	Relevance	Fluency	Empathy	Relevance	Fluency
MoEL	1.714	1.652	2.050	1.543	1.397	2.012
Transfo	2.073	2.193	2.373	1.900	1.893	2.240
Prepend	1.920	2.047	2.327	1.960	2.040	2.213
MTL	1.920	2.027	2.353	1.780	1.807	2.273
MTL+AD+DE	2.067	2.093	2.340	1.973	1.933	2.320
AD	2.187	2.260	2.373	2.187	2.213	2.400
AD+DE	2.180	2.240	2.367	2.173	2.227	2.393

Table 6: LSR results for single-turn and multi-turn dialogues respectively.

Generated Response	
Input	S: I got a free pizza from Domino’s but felt bad about how it happened.
Transfo	oh no ! what did you do ?
PRE	oh no , what happened ?
MTL	i love domino ’s pizza ! i love domino ’s !
ADM	oh no , what happened ?
AD	why did you feel bad about it ?
AD+DE	why did you feel bad about it ?
Input	S: Do any of your kids play any sports? L: No, I do not have kids. Yours? S: My daughter skateboards. She went down the big ramp the other day. It was so cool.
Transfo	that sounds like a lot of fun !
PRE	wow , that must have been a lot of fun .
MTL	that ’s cool . i ’ve never been on a rollercoaster . i ’ve never been on a rollercoaster .
ADM	that is so cool ! i bet you are so proud of her !
AD	that ’s awesome ! i bet you were so proud !
AD+DE	that ’s awesome ! i bet you were so proud of her !

Table 7: Example outputs generated by the baselines and our models, where S stands for speaker and L stands for Listener.

4 Case Studies

4.1 Sample Outputs of Different Empathetic Dialogue Systems

Table 7 lists a number of sample responses generated by the baselines and our models. It can be observed that our AD and AD+DE models produce high quality empathetic responses. For the first example, our models can follow the context and ask the reason why the speaker felt bad about getting a free pizza, whereas some of the baseline models produce uninformative responses (e.g., *what did you do?*) and some of them respond with incorrect emotion (e.g., *I love domino’s pizza!*). In the second sample, our models can generate more em-

pathetic responses (i.e., providing more approval and praise) compared to other baselines. In contrast, method like MTL generate irrelevant content (i.e., *rollercoaster*). Another observation is that the responses generated by AD and AD+DE are quite similar to each other, which is in line with the evaluation results.

4.2 Interpreting Dual Emotional Embeddings

We also conducted an experiment to assess how the learnt emotion embeddings by AD+DE differ with respect to speakers and listeners. Given an emotion label, we listed the label’s top-10 nearest neighbours in the speaker space and listener

Emotion State	Nearest neighbour words of the emotion label
Proud	S: son, graduated, proud, honour, daughter, happy, pleased, nephew, musicians, said L: celebrate, bet, con, proud, keep, parent, started, moment, congratulations
Sad	S: sad, cried, upset, bummed, died, passed, cry, depressed L: sorry, retrace, memories, sleazy, lose, toll, alive, sudden

Table 8: The 10 nearest neighbour words of the emotion label **PROUD** and **SAD** in the speaker (S) and listener (L) space, respectively.

space (see Table 8), respectively, based on the label embedding. Take the emotion label “proud” as an example, words like *proud*, *happy*, *honour* in the speaker space are very close semantically and are highly relevant to the emotion label. Also words like *son*, *daughter* are often be mentioned in parents’ expression of pride. In the listener space, words like *congratulations*, *proud*, *celebrate* are commonly used for responding to the speaker’s emotion of proud and the corresponding experience. These examples not only show consistency with people’s conversation habits, but also illustrate the difference between the speaker’s and the listener’s diction.

4.3 Generating Empathetic Dialogues from Scratch

Since we jointly model the speakers and listeners in the empathetic dialogues, our system is capable to generate a multi-turn conversation given an emotional situation and a prompt. Figure 4 provides some example dialogues generated by AD+DE in such a way. After given a specific emotion label (e.g., *joyful* and *disappointed* from the predefined label set of the EMPDIAL dataset), our model can generate relevant and empathetic responses conditioned on the initial prompt such as “my mother”. It can be observed that the generated multi-turn conversations are coherent and respect the given emotion labels.

5 Conclusion

In this paper, we propose a simple and effective technique called Affective Decoding for empathetic response generation. Empirical results based on extensive human evaluation show that our models (AD and AD+DE) outperform several strong baselines. Simply fine-tune the pre-trained *Transfo* on EMPDIAL achieves decent performance. MTL,

[joyful]

S: my mother just got a promotion at her job !
L: that ’s great ! what kind of job is it ?
S: it ’s a financial analyst job !
L: that ’s great ! i ’m sure you ’re proud of her !

[disappointed]

S: my mother was diagnosed with pancreatic cancer a few weeks ago .
L: oh no ! i ’m sorry to hear that . is she going to be okay ?
S: i think so , but i was n’t expecting it at all .
L: i ’m sorry to hear that . i hope everything works out for you .

Figure 4: Given the initial word *my mother*, two example dialogues are generated conditioning on the given emotion “joyful” and “disappointed”.

which has been used in some EDS, shows negative effects on the overall performance. As a side outcome, we also confirm the low validity of the mainstream automatic metrics for evaluating empathetic dialogue systems.

It was noted that empathetic dialogue systems tend to generate generic responses such as “*I’m sorry to hear that.*”. Therefore, one important future work is to improve the diversity and informativeness of the empathetic responses generated by an EDS. One possible technical direction is to employ variational autoencoders (Zhao et al., 2017; Li et al., 2019, 2020), which have been shown effective in improving the diversity in response generation.

Acknowledgement

This work is supported by the awards made by the UK Engineering and Physical Sciences Research Council (EP/P011829/1) and Ningbo Natural Science Foundation (202003N4320, 202003N4321). We thank anonymous reviewers for their insightful comments.

References

- Noor Fazilla Abd Yusof, Chenghua Lin, and Frank Guerin. 2017. [Analysing the causes of depressed mood from depression vulnerable individuals](#). In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 9–17, Taipei, Taiwan. Association for Computational Linguistics.
- Adamantios Diamantopoulos, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski, and Sebastian Kaiser. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3):434–449.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.
- Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194.
- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.
- Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis Philippe Morency, and Stefan Scherer. 2017. [Affect-LM: A neural language model for customizable affective text generation](#). *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:634–642.
- Simon Hoermann, Kathryn L McCabe, David N Milne, and Rafael A Calvo. 2017. Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *JMIR*, 19(8):e267.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- Steven Langsford, Amy Perfors, Andrew T Hendrickson, Lauren A Kennedy, and Danielle J Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1).
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. 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.
- Ruizhe Li, Xiao Li, Guanyi Chen, and Chenghua Lin. 2020. [Improving variational autoencoder for text modelling with timestep-wise regularisation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2381–2397, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ruizhe Li, Xiao Li, Chenghua Lin, Matthew Collinson, and Rui Mao. 2019. [A stable variational autoencoder for text modelling](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 594–599, Tokyo, Japan. Association for Computational Linguistics.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. 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 121–132.
- Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *AAAI*, pages 13622–13623.
- Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. [Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

- Chia Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016a. [How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2122–2132.
- Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016b. 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.
- Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. *arXiv preprint arXiv:1803.05928*.
- 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Alec Radford. 2018. Improving Language Understanding by Generative Pre-Training.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset](#). pages 5370–5381.
- Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, pages 675–676. Springer.
- Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. *arXiv preprint arXiv:1909.10122*.
- Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *arXiv preprint arXiv:1810.04650*.
- David Vilar, Gregor Leusch, Hermann Ney, and Rafael E Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103.
- Dingmin Wang, Chenghua Lin, Qi Liu, and Kam-Fai Wong. 2021. [Fast and scalable dialogue state tracking with explicit modular decomposition](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–295, Online. Association for Computational Linguistics.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. [TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents](#). *arXiv preprint arXiv:1901.08149*.
- Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. [Learning discourse-level diversity for neural dialog models using conditional variational autoencoders](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.