NORMMARK: A Weakly Supervised Markov Model for Socio-cultural Norm Discovery

Farhad Moghimifar and **Shilin Qu** and **Tongtong Wu**
Yuan-Fang Li and **Gholamreza Haffari**
Department of Data Science and AI, Monash University, Australia
{first.lastname}@monash.edu

Abstract

Norms, which are culturally accepted guidelines for behaviours, can be integrated into conversational models to generate utterances that are appropriate for the socio-cultural context. Existing methods for norm recognition tend to focus only on surface-level features of dialogues and do not take into account the interactions within a conversation. To address this issue, we propose NORMMARK, a probabilistic generative Markov model to carry the latent features throughout a dialogue. These features are captured by discrete and continuous latent variables conditioned on the conversation history, and improve the model's ability in norm recognition. The model is trainable on weakly annotated data using the variational technique. On a dataset with limited norm annotations, we show that our approach achieves higher F1 score, outperforming current state-of-the-art methods, including GPT3.

1 Introduction

Norms can be thought of as pre-defined socio-culturally acceptable boundaries for human behaviour (Fehr and Fischbacher, 2004), and incorporating them into conversational models helps to produce contextually, socially and culturally appropriate utterances. For instance, identifying the socio-cultural norm of *Greeting* in a negotiation helps to generate responses suitable for the power dynamics and social setting. Whereas, failing to detect and adhere to such norms can negatively impact social interactions (Hovy and Yang, 2021). Recent advances in developing chatbots have also highlighted the necessity of incorporating such implicit socio-cultural information into machine-generated responses, in order to approximate human-like interactions (Huang et al., 2020; Liu et al., 2021).

Norm discovery is a nascent research problem, and current approaches (Hwang et al., 2021) heavily rely on manually constructed sets of rules from available resources such as Reddit (Forbes et al., 2020; Ziems et al., 2022). In addition to the time and cost inefficiency of such approaches, the construction and use of these banks of norms treat each sentence or segment in isolation, and they fail to take the dependencies between norms in the flow of a dialogue into account (Fung et al., 2022; Chen and Yang, 2021). For instance, it is most likely that a dialogue segment containing the norm of *Request* follows a segment that includes *Request* and *Criticism* (Figure 1). Furthermore, such approaches require a large amount of annotated data, limiting their performance on sparsely labelled resources.

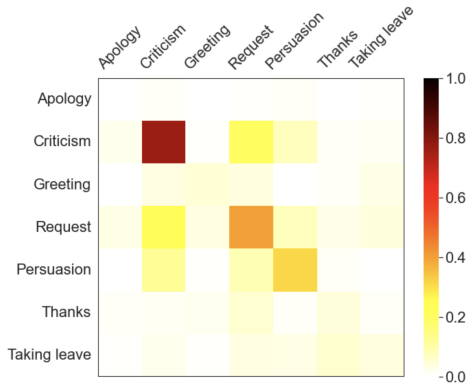

Figure 1: The heatmap of norm distribution based on norm label of the previous segment, constructed from LDC2022E20. Unit in column j of row i shows the probability of norm i following norm j.

To address these limitations, in this paper, we propose a deep generative Markov model that captures the inter-dependencies between turns (segments) of partially-labelled dialogues. The model includes two types of latent variables (LVs): (i) the discrete LVs capture the socio-cultural norms of the dialogue turns, and (ii) the continuous LVs capture other aspects, e.g. related to fluency, topic, and meaning. These latent variables facilitate capturing label- and content-related properties of the previous turns of the conversation, and are conditioned on the previous turns in a Markovian manner. We train the model on weakly annotated data using

Findings of the Association for Computational Linguistics: ACL 2023, pages 5081–5089
July 9-14, 2023

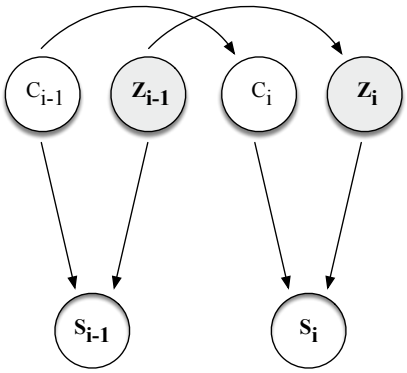

Figure 2: A graphical representation of our probabilistic generative model NORMMARK.

the variational technique, building on variational autoencoders (Kingma and Welling, 2014).

To evaluate the performance of our model in the task of socio-cultural norm discovery, we conducted experiments on an existing dataset. Experimental results show superiority of our model, by 4 points in F1 score, over the state-of-the-art approaches, in which each segment of a dialogue is modelled independently of the others. Furthermore, by evaluating our model on low amounts of training data, we show the capability of our proposed approach in capturing socio-cultural norms on partially-labeled data.

2 Related Works

Recent approaches have tried to develop models with the human psychological and behavioural capabilities (Jiang et al., 2021; Botzer et al., 2022; Lourie et al., 2021). Other approaches targeted identifying implicit social paradigms by developing sequence generation models (Moghimifar et al., 2020; Bosselut et al., 2019). However, the task of socio-cultural norm discovery has been overlooked, mostly due to the lack of proper annotated data (Fung et al., 2022). Forbes et al. (2020) present a dataset of social norms, collected from Reddit and propose a generative model to expand this collection. Zhan et al. (2022) also showed how social norms can be useful in conducting better negotiation dialogues. In a similar approach, Ziems et al. (2022) present a corpus of moral norms. Zhan et al. (2023) and Fung et al. (2022) use a prompt-based large-scale language model to generate rules from dialogues. More similar to our approach, existing models identify labels associated with utterances of dialogues (Chen and Yang, 2021; Yang et al., 2019; Yu et al., 2020). However, these approaches fail to take into account the flow of contextual information throughout a dialogue. In contrast to these studies, our approach addresses this task by considering the inter-dependencies between turns of dialogues.

3 A Generative Markov Model for Socio-Cultural Norm Discovery

We are given a set of dialogues $\mathcal{D} = \{\boldsymbol{d}^i\}_{i=1}^n$, where each dialogue consists of a set of turns (or segments) $\boldsymbol{d}^i = \{\boldsymbol{s}_j^i\}_{j=1}^m$. Each turn consists of a sequence of tokens from a vocabulary set $\mathcal{V}$. The dialogue set $\mathcal{D}$ consists of two subsets of labeled ($\mathcal{D}_L$) and unlabeled ($\mathcal{D}_U$) dialogues, where each turn $\boldsymbol{s}_m^i \in \mathcal{D}_L$ is annotated with a socio-cultural norm label $c_i \in \mathcal{C}$ with a total of K norm classes. The turns in the unlabeled dataset lack socio-cultural norm labels. Our goal is to develop a model that, by using contextual information carried from previous turns of the dialogue, discovers the socio-cultural norm associated with the turns of a dialogue.

Probabilistic Generative Model. Our model (shown in Fig. 2) assumes a directed generative model, in which a turn is generated by a factor capturing the socio-cultural norms and another factor capturing other aspects, e.g. topic and syntax. For each turn, the socio-cultural norm factor is captured by a discrete latent variable c_i, and the other aspects are captured by a continuous latent variable $\boldsymbol{z}_i$. As our aim is to leverage the contextual information, the latent variables of each turn of the dialogue are conditioned on those from the previous turn in Markovian manner. As such, our proposed generative model for each turn is as follows:

$$p_\theta(\boldsymbol{s}_i, \boldsymbol{z}_i, c_i | \boldsymbol{z}_{i-1}, c_{i-1}) = p_\theta(\boldsymbol{s}_i | c_i, \boldsymbol{z}_i) p_\theta(\boldsymbol{z}_i | \boldsymbol{z}_{i-1}) p_\theta(c_i | c_{i-1})$$

where $p_\theta(c_i|c_{i-1})$ and $p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i-1})$ capture the dependency of the causal factors on the previous turn, and $p_\theta(\boldsymbol{s}_i|c_i, \boldsymbol{z}_i)$ is a sequence generation model conditioned on the causal factors.

Training. To train the model, the likelihood function for a dialogue in D_U is:

$$p_\theta(\boldsymbol{s}_1..\boldsymbol{s}_n) = \sum_{c_1..c_n} \int d(\boldsymbol{z}_1)..d(\boldsymbol{z}_n) \times \prod_{i=1}^{n} p_\theta(\boldsymbol{s}_i|c_i, \boldsymbol{z}_i) p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i-1}) p_\theta(c_i|c_{i-1}).$$

Intuitively, the training objective for each dialogue turn corresponds an extension of the variational autoencoder (VAE) which involves: (i) both discrete and continuous latent variables, and (ii) conditioning on the latent variables of the previous turn. As such, we resort to the following variational evidence lowerbound (ELBO) for the unlabeled turns:

$$\log p(\boldsymbol{s}_i|c_{i-1}, \boldsymbol{z}_{i-1}) \geq \mathbb{E}_{q_\phi(c_i|\boldsymbol{s}_i,c_{i-1})}\{ \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{s}_i,\boldsymbol{z}_{i-1})}\big[\log p_\theta(\boldsymbol{s}_i|\boldsymbol{z}_i, c_i)\big]\} - KL[q_\phi(\boldsymbol{z}_i|\boldsymbol{s}_i, \boldsymbol{z}_{i-1})||p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i-1})] - KL[q_\phi(c_i|\boldsymbol{s}_i, c_{i-1})||p_\theta(c_i|c_{i-1})]$$

where q_ϕ's are variational distributions. We have nested ELBOs, each of which corresponds to a turn in the dialogue. We refer to the collection of these ELBOs for all dialogues in D_U by $\mathcal{L}(D_U)$. For the labeled turns, the ELBO for a dialogue turn is,

$$\log p(\boldsymbol{s}_i, c_i|c_{i-1}, \boldsymbol{z}_{i-1}) \geq \log p_\theta(c_i|c_{i-1}) + \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{s}_i,\boldsymbol{z}_{i-1})}\big[\log p_\theta(\boldsymbol{s}_i|\boldsymbol{z}_i, c_i)\big] - KL[q_\phi(\boldsymbol{z}_i|\boldsymbol{s}_i, \boldsymbol{z}_{i-1})||p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i-1})]$$

where we also add the term $\log q_\phi(c_i|\boldsymbol{s}_i, c_{i-1})$ to the training objective. We refer to the collection of ELBOs for all dialogues in the labeled data as $\mathcal{L}(D_L)$. Finally, the training objective based on the labeled and unlabeled dialogues is $\mathcal{L} = \mathcal{D}_\mathcal{U} + \lambda\mathcal{D}_\mathcal{L}$, where λ trades off the effect of the labeled and unlabeled data. We resort to the reparametrisation trick for continuous and discrete (Gumble-softmax (Jang et al., 2017)) latent variables when optimising the training objective.

Architectures. We use a transformer-based encoder to encode the turns $\boldsymbol{s}_i$ with a hidden representation $\boldsymbol{h}^s_i$. The *classifier* $q_\phi(c_i|\boldsymbol{s}_i, c_{i-1})$ is a 2-layer MLP with `tanh` non-linearity whose inputs are $\boldsymbol{h}^s_i$ and the embedding of c_{i-1}. For $q_\phi(\boldsymbol{z}_i|\boldsymbol{s}_i, \boldsymbol{z}_{i-1})$, we use a a multivariate Gaussian distribution, whose parameters are produced by MLPs from $\boldsymbol{h}^s_i$ and $\boldsymbol{z}_{i-1}$. For $p_\theta(\boldsymbol{s}_i|\boldsymbol{z}_i, c_i)$, we use an LSTM decoder, where this is performed by replacing pre-defined special tokens in the embedding space with $\boldsymbol{z}_i$ and c_i. For $p_\theta(c_t|c_{t-1})$, we use MLP with a softmax on top.

4 Experiments

In this section we report the performance of our model on the task of socio-cultural norm discovery in comparison to the current state-of-the-art models.

Dataset In our experiments, we use LDC2022E20. This dataset consists of 13,074 segments of dialogues in Mandarin Chinese. The dialogues are from text, audio, and video documents, where we transcribed the audio and video files using Whisper (Radford et al., 2022). The segments have been labelled from the set of socio-cultural norm labels of *none*, *Apology*, *Criticism*, *Greeting*, *Request*, *Persuasion*, *Thanks*, and *Taking leave*. We split the data into train/test/development sets with the ratio of 60:20:20. Each dialogue is divided into sequences of segments of length 5, where on average each segment consists of 8 sentences. We report the performance of our model, in comparison to the baselines, when using the maximum number of labeled data in the training set (Max). In addition, to evaluate the effect of the amount of training data on the performance of our model, we randomly select 50 and 100 of these sequences of dialogues for training, and report the results on the test set.

Baselines. We compare our model with LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2019), where each turn of a dialogue is encoded separately. We use WS-VAE-BERT (Chen and Yang, 2021) as another baseline, which encodes the contextual representation of a segment via a latent variable. However, WS-VAE-BERT does not capture the connections between segments. To experiment with the performance of our model on limited labeled data, we compare it to SetFit (Tunstall et al., 2022), which has proven to be a strong few-shot learning model. Similar to our model, we use `'bert-base-chinese'` as the backbone of BERT and WS-VAE-BERT, and `'sbert-base-chinese-nli'` has been used in SetFit. Additionally, we compare our model with a prompt-base large-scale language model GPT-3 `text-davinci-003` (Brown et al., 2020) and ChatGLM (Du et al., 2022), where the norm labels are given to the model with segments of dialogue, and the model is asked to generate a socio-cultural norm label from the list.

Evaluation Metrics. Following previous works in classification tasks, we report the macro averaged precision, recall, and F1 score of the models in predicting the socio-cultural norm label of each segment of a dialogue.

	Max			
Model	**P**	**R**	**F1**	Size
LSTM	9.13	12.57	10.08	0.4M
BERT	38.42	32.33	33.34	109M
WS-VAE-BERT	42.03	**40.74**	39.01	132M
SetFit	41.42	40.23	40.54	102M
ChatGLM	17.64	20.55	17.19	6B
GPT-3	39.86	35.05	33.61	175B
NORMMARK$_{\texttt{zero}}$	44.14	36.97	39.49	136M
NORMMARK	**47.92**	38.67	**44.20**	131M

Table 1: Segment-level socio-cultural norm prediction performance (precision, recall and F1 score). The results are reported by training the models on maximum number of labelled sequences of dialogues.

4.1 Results

Table 1 summarises the main results of the conducted experiment on LDC2022E20 data. On Max setting, where the model uses the maximum number of datapoints in the training set, our model outperforms all of the baselines with a margin of 4 and 6 points on F1 and precision, respectively, and achieves a comparable result in recall. This gap between our model and WS-VAE-BERT indicates the effect of carrying contextual information from previous turns of conversation. In addition, lower results of GPT-3 suggest that discovering socio-cultural norms is a challenging task, which needs higher-level reasoning.

Amount of Labeled Data. To evaluate the performance of our model with less amount of training data, we report the results on using only 50 and 100 datapoints during training, in Table 3. When using 100 sequences of turns, our model achieves the highest score in F1, and improves the precision and recall by more than 3 points over non-prompt based models. However, GPT-3 outperforms our proposed model in these two metrics. Similarly, on a more limited number of training data (setting 50), GPT-3 shows its dominance. Nevertheless, our model performs the best amongst the other baselines, by improving the F1 score by 3 points.

Model	**50**	**100**	**Max**
NORMMARK$_{\texttt{zero-extended}}$	13.41	14.33	20.33
NORMMARK$_{\texttt{extended}}$	13.48	14.76	20.25
NORMMARK	32.46	34.43	44.2

Table 2: Segment-level socio-cultural norm prediction of two variation of our approach, in comparison to our model. The results are macro-averaged F1 score.

	50			100		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
LSTM	7.92	12.5	9.69	11.06	12.58	9.90
BERT	10.85	15.74	12.58	10.46	15.50	11.82
WS-VAE-BERT	20.43	17.21	16.60	36.38	22.48	23.37
SetFit	30.42	26.25	27.86	32.12	30.54	31.32
ChatGLM	17.64	20.55	17.19	17.64	20.55	17.19
GPT-3	**39.86**	**35.05**	**33.61**	**39.86**	**35.05**	33.61
NORMMARK$_{\texttt{zero}}$	25.94	19.71	20.04	35.41	27.73	28.33
NORMMARK	**32.46**	**30.02**	**30.72**	**36.41**	**33.48**	**34.43**

Table 3: Segment-level socio-cultural norm prediction performance (precision, recall and F1 score). The results are reported by training the models on 50, 100 number of labelled sequences of dialogues.

Conditioning on the Context. To analyse the effect of carrying contextual information from previous turns of dialogue, we report the performance of the simplified version of our model (NORMMARK$_{\texttt{zero}}$), where the connections from previous turn are omitted. As can be seen in Table 1, in all of the settings NORMMARK outperforms the simplified version, indicating the importance of inter-dependencies between turns. Furthermore, we developed two variations of NORMMARK and NORMMARK$_{\texttt{zero}}$ where the contextual information from previous turns is carried directly through the previous segment (Figure 4). In Table 2, the lower performance of these models suggests that the contextual information from the previous turn overshadows the representation of the latent variable as well as the norm label, and consequently the norm classifier is profoundly biased towards the previous turn of dialogue.

Markov Order. We further analysed the effect of carrying contextual meaning from previous turns of dialogues, by varying the size of the Markov conditioning context l from 1 to 9, i.e. each of our

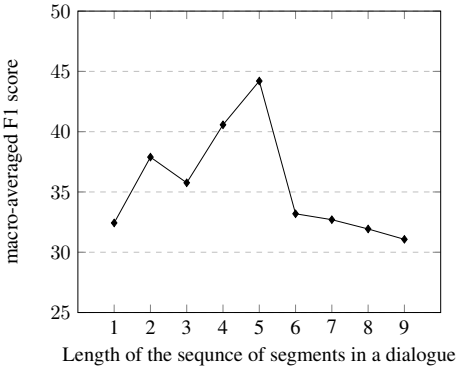

Figure 3: The performance of NORMMARK with different length of sequence of segments.

proposed latent variables is conditioned on previous l turns of dialogue.

Figure 3 summarises the results. It shows that shorter context results in lower performance, due to passing less contextual information to the next turns. On the other hand, too long context results in lower performance as well, due to extra complexity of modelling longer dependencies in latent variables and norm labels. As shown in the figure, our model performs best with a context size of 5 on this dataset.

5 Conclusion

In this work, we address the task of socio-cultural norm discovery from open-domain conversations. We present a probabilistic generative model that captures the contextual information from previous turns of dialogues. Through empirical results, we show that our model outperforms state-of-the-art models in addressing this task.

6 Limitations

We have studied the task of socio-cultural norm discovery based LDC2022E20 dataset, which consists of everyday situational interactions in Mandarin Chinese. Although we believe that our approach can used in other cultural settings, the current state of the model might not be generalisable to other cultures, unless further tuning is possible. Our model's ability in discovering such norms can help to improve conversational agents, however, real-world scenarios involving duplicitous or ambiguous terms might confuse our proposed approach. In addition, our model is limited to the textual modality, and we believe incorporating audio and visual features into the model can improve identifying socio-cultural norms. Nonetheless, the reliance of our model on large-scale pre-trained language models might result in some deployment challenges in situations with limited resources. Besides, all the reported results are by fixing a random seed running all experiments once.

7 Ethics Statement

Our work leverages pre-trained language models (BERT), therefore similar potential risks of this model is inherited by our work.

8 Acknowledgements

This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The authors are grateful to the anonymous reviewers for their helpful comments.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. Analysis of moral judgment on reddit. *IEEE Transactions on Computational Social Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*.

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Ernst Fehr and Urs Fischbacher. 2004. Social norms and human cooperation. *Trends in cognitive sciences*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2022. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *arXiv preprint arXiv:2210.08604*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Jimin Hong, Jungsoo Park, Daeyoung Kim, Seongjae Choi, Bokyung Son, and Jaewook Kang. 2022. Tess: Zero-shot classification via textual similarity comparison with prompting using sentence encoder. *arXiv preprint arXiv:2212.10391*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations*.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Farhad Moghimifar, Lizhen Qu, Yue Zhuo, Mahsa Baktashmotlagh, and Gholamreza Haffari. 2020. CosMo: Conditional Seq2Seq-based mixture model for zero-shot commonsense question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5347–5359, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, et al. 2023. Socialdial: A benchmark for socially-aware dialogue systems. *arXiv preprint arXiv:2304.12026*.

Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2022. Let's negotiate! a survey of negotiation dialogue systems. *arXiv preprint arXiv:2212.09072*.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

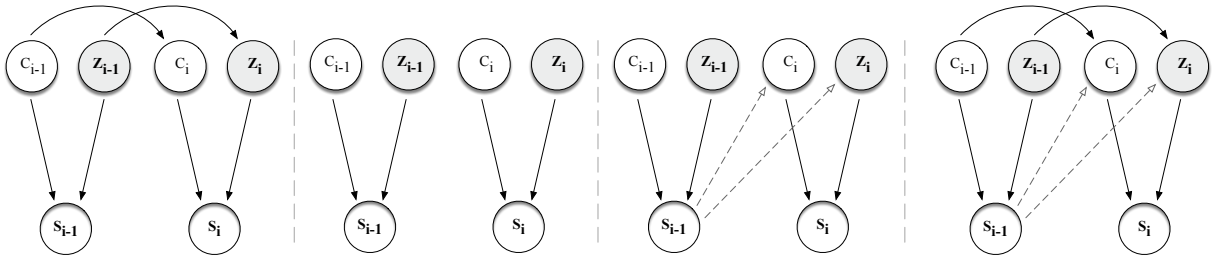

Figure 4: A graphical representation of our probabilistic generative model NORMMARK. The second model from left is a simplified version of our proposed approach where the contextual information from previous turns of dialogue is not carried through the current step (NORMMARK$_{\texttt{zero}}$). The next two models are `extended` versions of NORMMARK$_{\texttt{zero}}$ and NORMMARK, respectively, where direct contextual information from previous segment is carried to the current turn.

A Experimental Details

To train our model, we have used the pre-trained `'bert-base-chinese'`, which is licensed free to use for research purposes, as the encoder (Kenton and Toutanova, 2019), and we have used LSTM (Hochreiter and Schmidhuber, 1997) with hidden dimension of 128 and one hidden layer as the decoder. We used a dropout of 0.6 over the input. We implemented the norm classifier with a 2-layers MLP with `tanh` non-linearity on top. We used CrossEntropyLoss (Zhang and Sabuncu, 2018) as loss function over the predictions of our model. We used AdamW (Loshchilov and Hutter, 2018) as the optimiser with the learning rate of 1e-5 for the encoder and 1e-3 for the rest of the network. We trained our model for 50 epochs, on a single machine with NVIDIA one A100 gpu, with an early stop if the validation accuracy is not improved for more than 20 iterations.

For the baselines, we have developed a network with a two-stacked LSTM layers followed by two linear layers. We compared out model with BERT, where uses the `'bert-base-chinese'` pre-trained model. Each of these two models where trained for 100 epochs, using AdmaW optimiser with the learning rates of 1e-3 and 5e-5, respectively. For WS-VAE-BERT (Chen and Yang, 2021), we followed the source code provided in the paper. For replicating the document level labels, when a segment within the sequence of segments contained a socio-cultural norm, we labeled them 1, otherwise 0. We trained SetFit (Hong et al., 2022) by following the online instructions on their GitHub repository [1]. Figure 4 shows the variations of our model, which we used in for the ablation study.

GPT-3 (Brown et al., 2020) was used by incorporating the list of socio-cultural norms into the prompt as well as dialogues, and asking to generate the corresponding label. Our experiments on GPT-3 showed that using random examplars from the training set of LDC2022E20 results in a decrease in the performance. The LDC2022E20 dataset is the copyrighted property of (c) 2022 Trustees of the University of Pennsylvania and has been used for research purposes in CCU program. This dataset was developed to help models to identify socio-cultural norms in courses of dialogues.

In all of our experiments we used a fix random seed, hence all results are reported based on single-run of the models.

[1] https://github.com/huggingface/setfit

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section &

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 5

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 and Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Section 4 and Appendix A

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We reached out to the provider of the dataset to get information about data collection, but we haven't got any responses back yet.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 and Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Appendix A

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.