The Language Model Can Have the Personality: Joint Learning for Personality Enhanced Language Model (Student Abstract)

Tianyi Chen[1], **Feiqi Cao**[1], **Yihao Ding**[1], **Caren Han**[1,2,3]

[1]School of Computer Science, University of Sydney
[2]School of Physics, Maths and Computing, Computer Science and Software Engineering, University of Western Australia
[3]School of Computing and Information Systems, University of Melbourne
{tche8310, fcao0492, ydin0771}@uni.sydney.edu.au, caren.han@sydney.edu.au

Abstract

With the introduction of large language models, chatbots are becoming more conversational to communicate effectively and capable of handling increasingly complex tasks. To make a chatbot more relatable and engaging, we propose a new language model idea that maps the human-like personality. In this paper, we propose a systematic Personality-Enhanced Language Model (PELM) approach by using a joint learning mechanism of personality classification and language generation tasks. The proposed PELM leverages a dataset of defined personality typology, Myers-Briggs Type Indicator, and produces a Personality-Enhanced Language Model by using a joint learning and cross-teaching structure consisting of a classification and language modelling to incorporate personalities via both distinctive types and textual information. The results show that PELM can generate better personality-based outputs than baseline models.

Introduction

With the success of large language model, chatbots are becoming more conversational to communicate effectively and capable of handling increasingly complex tasks. However, would it be wonderful if we map the human-like personality and make your chatbot exhibits more relatable and engaging to users (Fernau et al. 2022)? Zheng et al. (2020) incorporate character profiles on Persona-Chat (Zhang et al. 2018), while Majumder et al. (2021) further leverage the background stories at inference time to equip personalised information into the language model. However, injecting simple character profiles or personalised information is insufficient to make language model-based chatbot feel more human (Chaves and Gerosa 2021). We consider adding recognisable human traits as part of a defined personality to create that connection with the user. To do so, we could map out the chatbot's personality using models like the Myers and Briggs Personality Indicator (MBTI), that classifies people into 16 distinct classifications. The MBTI dataset[1] provides a large selection of people with their MBTI type and corresponding textual corpora, which is perfect for teaching language models to intimate different personalities. In this work, we construct a novel framework called the Personality-Enhanced Language Model (PELM) by using joint learning of MBTI personality classification and language modelling.[2] The language model receives explicit guidance on the personalities at the training stage through a classification task and learns the coherent, grammatically correct text with personality information from the language modelling task. Experimental results show the feasibility and case studies of our PELM personality-based text generation output.

[1]https://www.kaggle.com/datasets/datasnaek/mbti-type

Methodology

To incorporate personality information (MBTI type) at the training stage, our proposed PELM model applies a joint learning framework to include an MBTI classifier unit and a language model, trained with a joint loss. The parameter-sharing mechanism is applied between the classifier unit and the language model. Figure 1 shows our PELM framework.

The **Language Model** (LM) is a transformer-based pre-trained decoder to perform the next token prediction on the input sequence $x = (x_1, x_2, \cdots, x_t)$ where t denotes the number of tokens of the given sentence. In parallel, the **classifier** is a transformer-based encoder to predict MBTI type on the same input sequence x. The classifier has an identical structure to the LM module except that it does not apply the masked multi-head attention.

The **parameter sharing** mechanism enables the weights in the language model to be optimised towards the performance of the classification task where it incorporates the personality information to recognise each MBTI type. The parameters in all layers in the language model, W_{lm}, are identical to the parameters at the same position in the classifier, W_{cls}, meaning they share the same nodes in the network computation graph, i.e. $W_{lm} \equiv W_{cls}$. Consequently, all shared weight matrices are updated twice during the backward propagation: once when the gradient is back-propagated through each block in the classifier and once when the gradient is back-propagated through each block in the language model.

The **joint loss** unit aims to aggregate the language modelling loss and the MBTI classification loss for joint learning. The output logits of both the language model and the classifier are used to compute losses for two tasks, L_{lm} and

[2]The model checkpoints and code will be released.

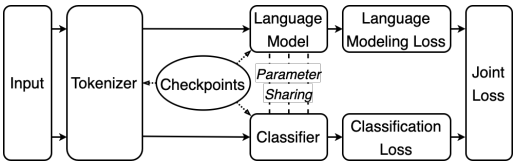

Figure 1: Overview of the PELM framework

L_{cls} respectively. The joint loss is then the sum of two individual losses with an additional parameter γ to control the weight of the aggregation: $L_{joint} = L_{lm} + \gamma \times L_{cls}$.

Preliminary Evaluation and Results

Evaluation Setup

This section briefly describes evaluation setups to demonstrate the reproducibility of the proposed methods.

Baselines BERT and GPT-2 [3] are adopted as the fundamental frameworks. We evaluate the result of each framework by fine-tuning pre-trained models on specific classification (clsBERT and clsGPT-2) or LM tasks (lmBERT and lmGPT-2).

Evaluation Metrics We use weighted *F1-score* and Perplexity to measure the classification language generation performance, respectively. Additionally, we introduce the P-Index to account for the personality information in the generated content. It is defined as $\mu \pm \sigma$ where μ is the average keyword counts in the generated 200-token-long outputs for 20 tests, and σ is the standard deviation. Keywords are selected via TF-IDF for each MBTI type in the corpora.

Implementation Details We use *cross-entropy* loss for classification and LM, setting $\gamma = 0.25$. All models are trained with 2e-05 learning rate for 5 epochs with Adam optimizer.

Result Analysis and Case Study

To demonstrate the effectiveness of the proposed PELM, we compare the performance with mono-task vanilla BERT and GPT-2. Table 1 shows PELM yields the best F1 and P-index scores compared to the baseline models, which may result from effective knowledge sharing via a joint learning setting. Notably, transformer decoder-based GPT-2 improves classification (F1 from 59.56% to 61.06%) and language generation (P-Index from 47.4 to 50.0) more.

As Perplexity displays a slight decreasing trend, we conducted case studies to manually assess content generated by lmGPT-2 and PELM, highlighting the advantages of incorporating personality information. In Table 2, both models produce coherent text from the same prompt. Nevertheless, PELM could generate more personality-aware descriptions, aligning better with ISFJ traits. Phrases like *"Love is a mighty power"* and *"consumes even the strongest of people"* tend to reflect ISFJs' warm and relationship-focused nature.

Conclusion

In this paper, we leveraged MBTI as a defined personality typology. We established a joint learning framework with a language modelling task and classification task for personality-based generation called PELM. Experimental results and case studies indicate that our PELM can generate better personality-based outputs than baseline models.

[3] We use *bert-base-uncased* and *gpt2* as the checkpoints.

Checkpoint	Model	CLS	LM	
		F1(%)	*Perplexity*	*P-Index*
BERT	clsBERT	60.19	-	-
	lmBERT	0.50	**75.63**	17.7±3.9
	PELM	**60.68**	77.31	**19.9±4.0**
GPT-2	clsGPT-2	59.56	-	-
	lmGPT-2	5.31	**52.28**	47.4±2.6
	PELM	**61.05**	52.77	**50.0±3.1**

Table 1: Overall performance of PELM framework and baseline models on MBTI dataset

ISFJ's description: Very dedicated and warm protectors, always ready to defend their loved ones.
Prompt: Love only grows by sharing; one can only have more for oneself by giving it to others.

Models	**Generated Text**
lmGPT-2	I am just curious if anyone else has these and if anyone has them all done. I only took this test a while back.
PELM	Love is a mighty power as it can become so powerful that it consumes even the strongest of people.

Table 2: A case study of type ISFJ showing the generation quality of PELM

References

Chaves, A. P.; and Gerosa, M. A. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8): 729–758.

Fernau, D.; Hillmann, S.; Feldhus, N.; Polzehl, T.; and Möller, S. 2022. Towards Personality-Aware Chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 135–145.

Majumder, B. P.; Berg-Kirkpatrick, T.; McAuley, J.; and Jhamtani, H. 2021. Unsupervised Enrichment of Persona-grounded Dialog with Background Stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 585–592. Online: Association for Computational Linguistics.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 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)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Zheng, Y.; Zhang, R.; Huang, M.; and Mao, X. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9693–9700.