UNDERSTANDING CHINESE MORAL STORIES WITH FURTHER PRE-TRAINING

Jing Qian[1], Yong Yue[1], Katie Atkinson[2] and Gangmin Li[3]

[1]School of Advanced Technology, Xi'an Jiaotong Liverpool University,
Suzhou, China
[2]Department of Computer Science, University of Liverpool, Liverpool, UK
[3]School of Computer Science Technology, University of Bedfordshire, Luton, UK

ABSTRACT

The goal of moral understanding is to grasp the theoretical concepts embedded in a narrative by delving beyond the concrete occurrences and dynamic personas. Specifically, the narrative is compacted into a single statement without involving any characters within the original text, necessitating a more astute language model that can comprehend connotative morality and exhibit commonsense reasoning. The "pre-training + fine-tuning" paradigm is widely embraced in neural language models. In this paper, we propose an intermediary phase to establish an improved paradigm of "pre-training + further pre-training + fine-tuning". Further pre-training generally refers to continual learning on task-specific or domain-relevant corpora before being applied to target tasks, which aims at bridging the gap in data distribution between the phases of pre-training and fine-tuning. Our work is based on a Chinese dataset named STORAL-ZH that composes of 4k human-written story-moral pairs. Furthermore, we design a two-step process of domain-adaptive pre-training in the intermediary phase. The first step depends on a newly-collected Chinese dataset of Confucian moral culture. And the second step bases on the Chinese version of a frequently-used commonsense knowledge graph (i.e. ATOMIC) to enrich the backbone model with inferential knowledge besides morality. By comparison with several advanced models including BERT-base, RoBERTa-base and T5-base, experimental results on two understanding tasks demonstrate the effectiveness of our proposed three-phase paradigm.

KEYWORDS

Moral Understanding, Further Pre-training, Knowledge Graph, Pre-trained Language Model

1. INTRODUCTION

Morality is one of the most intricate topics related to human behavior [1]. It overlaps with commonsense, molded by cultural norms, and regulated by laws and rules. Children are often introduced to ethical values and morals through fable stories, which teach them to differentiate between right and wrong in their daily lives. Moral understanding is a new challenging task for natural language processing, which aims to comprehend the abstract concepts that are hidden within a narrative by observing the concrete events and vivid characters portrayed. Earlier works about story understanding are mainly centered on story ending prediction [2], or controllable story generation given specific constraints, such as storylines [3], emotions [4], and styles [5]. Most of them attach importance to the surface realization based on the concrete content, whereas our work attempts to uncover the implied morals that are conveyed by the stories. Table 1 shows one moral-story pair from STORAL-ZH [6], a recently released dataset of 4,209 Chinese moral stories.

DOI: 10.5121/ijnlc.2023.12201

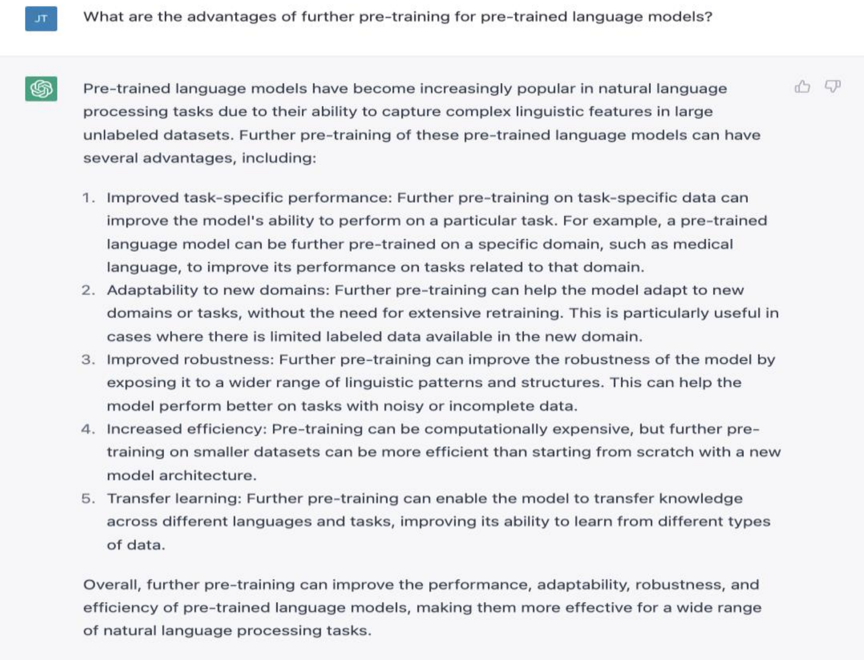
Self-supervised pre-training on massive amounts of unlabeled corpora from a general domain endows large language models with contextual knowledge and the capacity to recognize n-grams. Originated from Transformer [7], plenty of Pre-trained Language Models (PLMs) have sprung up in succession. These models can be broadly classified into three groups based on their architecture: Transformer encoder (e.g., BERT [8], RoBERTa [9]), Transformer decoder (e.g., GPT2 [10], GPT3 [11]), and the full Transformer encoder-decoder network (e.g., T5 [12], MASS [13]). With distinct pre-training strategies, PLMs are capable of handling diverse downstream tasks. For example, RoBERTa's masked language modeling produces advantageous contextual word representations that enhance natural language understanding, while the strategy of auto-regressive language modeling exploited by GPT2 establishes the groundwork for natural language generation.

Table 1. A moral-story pair from STORAL-ZH.

STORY	**晚饭后，母亲和女儿一块儿洗碗盘，父亲和儿子在客厅看电视。** One day after dinner, mother and daughter were washing dishes together, father and son were watching TV in the living room. **突然，厨房里传来打破盘子的响声，然后一片沉寂。** Suddenly there was a sound of breaking plates in the kitchen, and then there was silence. **这时儿子望着父亲说道：“一定是妈妈打破的。”** The son looked at his father and said, “Mom must have broken it.” **“你怎么知道?” “她没有骂人。”** “How do you know?” “She didn’t swear.”
MORAL	**我们习惯以不同的标准来看人看己，以致往往是责人以严，待己宽。** We are used to looking at others and ourselves by different standards, so that we tend to be strict with others and lenient with ourselves.

The pre-training phase enhances the capability of language models, particularly as the number of model parameters and the magnitude of unlabeled corpora continue to expand. This assertion is supported by the impressive performance achieved by prompt learning on various benchmark tasks [14]. Prompt learning [15] wraps the input sequence with a template containing masked tokens to handle downstream tasks by imitating the pre-training objectives. By which, the great potential of PLMs is better stimulated. Therefore, further pre-training on in-domain data (Domain-Adaptive Pre-Training, DAPT) or task-relevant data (Task-Adaptive Pre-Training, TAPT) is a highly recommended option when the downstream scenarios involve specific domains and no relevant data is available in the unlabeled corpora. Figure 1 exhibits the advantages of further pre-training for pre-trained language models as listed by ChatGPT, the most remarkable PLM in recent days.

Apart from domain-specific knowledge and task-relevant information that are attained from unlabeled unstructured text, sometimes equipping PLMs with the capability of commonsense reasoning can further enhance their efficacy. In this regard, the phase of further pre-training can be extended to heterogeneous data, such as Knowledge Graphs (KGs), A typical KG is composed of RDF triples (*h*, *r*, *t*), where *h* and *t* represent head entity and tail entity respectively, *r* represents their relationship. There have been various kinds of KGs, including linguistic [16], encyclopedia [17], commonsense [18], domain-specific [19]. For instance, ATOMIC [20], a frequently-used commonsense KG, contains triples that encode inferential knowledge about everyday life, such as (*PersonX applies to jobs*, *xEffect*, *gets hired*) and (*PersonX asks PersonY for money*, *xWant*, *to go pay bills*).



What are the advantages of further pre-training for pre-trained language models?

Pre-trained language models have become increasingly popular in natural language processing tasks due to their ability to capture complex linguistic features in large unlabeled datasets. Further pre-training of these pre-trained language models can have several advantages, including:

1. Improved task-specific performance: Further pre-training on task-specific data can improve the model's ability to perform on a particular task. For example, a pre-trained language model can be further pre-trained on a specific domain, such as medical language, to improve its performance on tasks related to that domain.
2. Adaptability to new domains: Further pre-training can help the model adapt to new domains or tasks, without the need for extensive retraining. This is particularly useful in cases where there is limited labeled data available in the new domain.
3. Improved robustness: Further pre-training can improve the robustness of the model by exposing it to a wider range of linguistic patterns and structures. This can help the model perform better on tasks with noisy or incomplete data.
4. Increased efficiency: Pre-training can be computationally expensive, but further pre-training on smaller datasets can be more efficient than starting from scratch with a new model architecture.
5. Transfer learning: Further pre-training can enable the model to transfer knowledge across different languages and tasks, improving its ability to learn from different types of data.

Overall, further pre-training can improve the performance, adaptability, robustness, and efficiency of pre-trained language models, making them more effective for a wide range of natural language processing tasks.

Figure 1. The advantages of further pre-training for PLMs, according to ChatGPT.

This work proposes a three-phase paradigm that builds upon the conventional two-phase paradigm of “pre-training + fine-tuning”. The proposed paradigm involves an intermediary phase of further pre-training, which will be evaluated using two downstream tasks related to moral understanding. STORAL-ZH [6], a dataset of Chinese human-written moral stories, is employed for the target tasks, and LongLM-base [21], a T5-architectured language model pre-trained on 120G Chinese long novels, is selected as our backbone model. TAPT and DAPT constitute the newly added phase of further pre-training. For TAPT, the language model is further trained on unlabeled STORAL-ZH to possess task-awareness knowledge. For DAPT, we design a two-step process to involve two domains, namely moral culture and commonsense knowledge. Inspired by [22], we utilize triples of a commonsense KG and transform them into readable textual sequences for domain-adaptive further pre-training. To summarize, our contributions are reflected in the following three aspects: (1) different from the conventional two-phase paradigm, we add an intermediary phase of further pre-training before fine-tuning to boost the model’s performance on moral understanding, (2) our model is equipped with commonsense knowledge through continual pre-training on the template-based transformation of triples in ATOMIC-zh [18] to facilitate moral perception beyond concrete characters and events, and (3) a new corpus centered on Chinese traditional moral culture (i.e. the Four Books and Five Classics) is collected to support domain-adaptive pre-training about moral culture.

2. RELATED WORK

2.1. Story Understanding

A wide range of tasks related to story understanding and generation have been proposed in the literature, which includes story ending prediction [2], commonsense story continuation [22] and story ending generation with fine-grained sentiment [23]. To improve story understanding, various factors have been taken into account, such as storylines [3], emotions [4] and styles [5]. Different from that storylines guide the writing of a story, emotions portray the characters' mental and emotional states, and styles determine the overall tone of the narrative, moral understanding emphasizes the ability to bridge story content and implied morals, which is more complicated. To address this challenge, [6] introduces a new dataset of 4k Chinese and 2k English moral stories, and proposes two moral understanding tasks including concept understanding and preference alignment to facilitate the modeling of commonsense and discourse relations among abstract concepts and concrete events.

2.2. Further Pre-training

Pre-training is a crucial initial phase for language model implementation, serving to initialize the model and expedite parameter convergence in downstream tasks. With the rapid expansion in model size, larger unlabeled corpora are necessary to fully pre-train the model and prevent over-fitting. To bridge the gap between pre-training and fine-tuning data distributions, further pre-training exhibits positive effects and brings improved model performance [24, 25]. Additionally, [26] introduces two forms of further pre-training: task-adaptive pre-training (TAPT) and domain-adaptive pre-training (DAPT). To be specific, TAPT refers to continual pre-training on task-specific unlabeled data prior to the phase of fine-tuning, which helps to incorporate task-awareness knowledge with the language model in advance and brings consistent performance improvements [27]. On the other hand, DAPT supports better adaptability to relevant domains and acquires domain-awareness knowledge, which is particularly useful in low-resource cases [28].

2.3. Joint Knowledge and PLMs

In recent times, there has been a growing interest in integrating knowledge into PLMs. Although self-supervised pre-training over large-scale corpora has proven useful in providing PLMs with ample contextual semantics, it lacks domain-specific [19, 29], factual [30, 31] or commonsense knowledge [22, 32]. To address this limitation, various methods have been proposed. For instance, K-BERT [19] inserts domain-specific knowledge triples into the input sequence explicitly and employs a visible matrix to govern the mutual influence among tokens. BERT-MK [29] integrates graph contextualized knowledge from a medical KG into language models. KEPLER [30] encodes entity descriptions as their embeddings and jointly optimizes knowledge embedding and masked language modeling objectives in the same PLM. ERNIE [31] utilizes the informative entities in KGs to enhance language representation by putting forward a new pre-training objective. KG-BART [32] captures the complex relations between concepts over a commonsense KG for generative commonsense reasoning. Furthermore, [22] conducts incremental pre-training on two commonsense knowledge bases to generate more reasonable stories without spending additional efforts on considering heterogeneous information fusion or sub-graph aggregation, which implicitly and effectively incorporates commonsense knowledge with GPT-2 [10].

3. METHODOLOGY

This section elaborates on the key constituents of our method, encompassing the transformer-based language model, the implementation details of task-adaptive and domain-adaptive pre-training, as well as the phase of fine-tuning.

3.1. Transformer-Based Language Model

This work employs a language model based on the full Transformer architecture [7], in which the encoder accepts an input sequence and uses fully-visible masking, while the decoder produces the target sequence through causal masking and cross-attention. This text-to-text framework is adept at handling both understanding and generation tasks. T5 [12] is a representative encoder-decoder language model trained on the Colossal Clean Crawled Corpus (C4) composed of English, French, Romanian, and German. T5 employs an unsupervised pre-training objective of replacing corrupted spans, which proves to be the best choice after a comprehensive comparison.

LongLM [21], a Chinese version of T5, is pre-trained on 120G Chinese novels with two generative tasks: text infilling [12] and conditional continuation [10]. Text infilling, inspired by Span BERT [33], replaces a few of text spans in the input sequence with special tokens at a corruption rate of 15%, with span lengths following the Poisson distribution with $\lambda = 3$. The pre-training objective is to output the original text spans that are replaced with special tokens using the greedy decoding algorithm. The second task, conditional continuation, aims to generate the back half of a text based on its front half using top-*k* sampling [34] with $k = 40$ and a softmax temperature of 0.7 [35]. In this work, we leverage the pre-trained checkpoint of LongLM-base, which has 223M model parameters, available on Hugging Face [36].

3.2. Further Pre-training

3.2.1. Task-Adaptive Pre-Training (TAPT)

It is worthwhile to conduct further pre-training on the unlabeled data of downstream tasks to enhance the adaptability of the language model before fine-tuning. TAPT offers advantages such as increased efficiency and improved task-specific performance. Pre-training can be computationally expensive, but further pre-training on the far smaller task-specific dataset exhibits higher efficiency and possesses task-awareness knowledge in advance. [6] has already post-trained LongLM [21] on the unlabeled version of STORAL [6], and names it T5-Post that serves as a compared baseline in the original paper. Table 2 provides an example for the pre-training objective of text infilling.

Table 2. An example showing the pre-training objective of text infilling.

Original Story	I was sitting in my room and was busy with my usual things. Knowing through the news of social media the carnage of seven civilians, I was afflicted with a heart trouble and great care.
Inputs	I was sitting in my room and was busy with <X>. Knowing through the news of social media <Y>, I was afflicted with a heart trouble and great care.
Targets	<X> my usual things <Y> the carnage of seven civilians <Z>

3.2.2. Domain-Adaptive Pre-Training (DAPT)

With the objective of boosting the performance of the language model on downstream tasks, it is reasonable to conduct further pre-training on unlabeled corpora that are specific to relevant domains. DAPT enables PLMs to acquire domain-awareness knowledge that can improve their ability to understand the text in the target domain. In order to better grasp the abstract morals conveyed by concrete events and characters in Chinese narratives, background domains can involve traditional moral culture and commonsense knowledge.

Moral Knowledge Confucianism represents the mainstream moral culture of China, and its authoritative books, the Four Books and Five Classics, provide a comprehensive record of the political, economic, diplomatic, cultural, and philosophical developments during the most active period in the history of Chinese ideology. The morals and ethics contained in these books have exerted a significant and lasting influence on Chinese culture. Since the Four Books and Five Classics were originally written in classical Chinese, we have collected translated versions in written vernacular Chinese as the corpus for further pre-training, referred to as **4+5**. Table 3 provides several examples from the Analects, one of the Four Books, to illustrate the type of content included in **4+5**.

Table 3. Three examples from the Analects and translated in Vernacular Chinese and English
.

Example 1	巧言令色，鲜仁矣。
Vernacular Chinese	**花言巧**语，装出和颜悦色的样子，这种人的仁心就很少了。
English Translation	Talking plausibly with feign amiable looks, people of this sort are scarcely benevolent.
Example 2	君子不器。
Vernacular Chinese	**君子不像器具那**样，只有某方面的用途。
English Translation	Gentlemen are not like tools, each of which only has a certain use.
Example 3	学如不及，犹恐失之。
Vernacular Chinese	**学**习如同赛跑，总怕赶不上，可赶上了，又怕被超过。
English Translation	Study is just like race. You’re always afraid of being unable to catch up. Yet when you catch up, you’ll be afraid of being overtaken.

Commonsense Knowledge Integrating commonsense knowledge equips PLMs with the ability of commonsense reasoning in downstream tasks. To this end, we leverage the commonsense knowledge from a frequently-used knowledge graph, ATOMIC [20]. It consists of 877K *if-then* triples (*h*, *r*, *t*) in which the head *h* and the tail *t* are two events and the relation *r* describe their *if-then* relationship. For examples, (*PersonX accomplishes PersonY's work*, *xAttr*, *helpful*) means that if *X* accomplishes *Y*'s work, then *X* is helpful, (*PersonX accomplishes PersonY's work*, *oWant*, *to thank PersonX*) tells that if *X* accomplishes *Y*'s work, then *Y* will thank *X*. *xAttr* represents the persona attribute of *X*, *oWant* states others' event. There are three *if-then* types including *If-Event-Then-Mental-State*, *If-Event-Then-Event* and *If-Event-Then-Persona*. Inferential knowledge brought by ATOMIC [20] facilitates language comprehension, especially commonsense relations among concrete events and abstract concepts for moral stories. Our work is based on Chinese, we utilize the translated ATOMIC dataset, ATOMIC-ZH [18] instead. Inspired by [22] and [37], we linearize KG triples into textual sequences through the template-based transformation, as illustrated in Table 4. Different from previous works that explicitly introduced part commonsense knowledge into PLMs, further pre-training directly on all linearized triples can integrate commonsense knowledge into LongLM [21] implicitly in a more convenient way.

Table 4. Examples of template-based transformation of KG triples.

KG Triples	**Transformed Sentences**
(某人完全放弃某物, **xEffect**, 翻开新的一页) (*PersonX abandons ____ altogether,* ***xEffect****, turns over a new leaf*)	汤姆完全放弃某物，结果他翻开新的一页。 Tom abandons something altogether, as a result, he turns over a new leaf.
(有人接受事实, **xAttr**, 勇敢的) (*PersonX accepts the fact,* ***xAttr****, brave*)	汤姆接受事实，他是勇敢的。 Tom accepts the fact, he is brave.
(有人接受了挑战, **xIntent**, 以证明他能做到) (*PersonX accepts the challenge,* ***xIntent****, to prove he can do it*)	汤姆接受了挑战，因为他想以证明他能做到。 Tom accepts the challenge, because he wanted to prove he can do it.

3.3. Fine-tuning

Following the standard paradigm “pre-training + fine-tuning”, we fine-tune our model on two moral understanding tasks after task-adaptive pre-training on unlabeled STORAL-ZH [6] and domain-adaptive pre-training on **4+5** and ATOMIC-ZH [18]. Both tasks are designed by [6], they aim to select the correct moral from several choices given a story, but test the abilities of the PLM from two different aspects. One is concept understanding, the other is preference alignment.

ConcePT understanding (CPT) It requires choosing the correct one from the five candidates of morals for each story, that tests the ability of understanding abstract concepts behind concrete events in the story. Apart from the paired moral of the story, the other four candidates are true negative samples that are selected from the morals of stories about irrelevant topics.

PREFerence alignment (PREF) Simpler than CPT, PREF aims to tell the right moral from the other wrong one. There are only two moral candidates for each story in the constructed task dataset [6]. The incorrect candidate is obtained by replacing one random token in the correct moral with its antonym. As some words do not have antonyms, the training data for PREF is a little smaller than CPT.

To handle both tasks of CPT and PREF, we first concatenate the story and its candidate morals, then insert unique special tokens before the story and each candidate, and feed the sequence into the tested language model. Following the default settings of T5 [12], special tokens are <extra_id_i> where i points out the number order. Inspired by [6], we take the hidden states of corresponding special tokens as the representations of the story and each candidate respectively, afterwards we normalize the dot-product scores between the representations of the story and each candidate to predict the probability distribution over all candidates. We optimize the language model by minimizing the cross-entropy loss.

4. EXPERIMENTS

4.1. Datasets

Corpus for Further Pre-training We adopt two kinds of corpora of different domains including moral culture and commonsense knowledge for domain-adaptive pre-training. For moral culture, the corpus is composed of vernacular version for the Four Books and Five Classics. The Four Books are Great Learning, Doctrine of the Mean, Analects and Mencius, while the Five Classics are Classic of Poetry, Book of Documents, Book of Rites, I Ching, and Spring and Autumn Annals. We collect the writings in the vernacular of each work from public web resources and integrate them together to get the unlabeled corpus named "**4+5**".

To enrich our model with commonsense knowledge, we transform the triples in ATOMIC-ZH [18] into readable textual sequences using a template-based method [37] for further pre-training. ATOMIC-ZH [18] is the translated ATOMIC [20] used for Chinese tasks. [18] applies Regular Replacement to alleviate the problems of containing special tokens (i.e., PersonX and PersonY) as well as blank in some triples. To facilitate convenient translation, [18] transform triples into reasonable natural language sentences, then split them into the form of (h, r, t) after being translated via automatic translation system to make up ATOMIC-ZH. The Chinese commonsense knowledge graph provided by [18] is enlarged by other resources, we only select the triples with the nine relations that are mentioned in [20] for our further use.

Corpus for Fine-tuning The corpus for downstream tasks are constructed from STORAL-ZH [6], which composes of 4,209 Chinese story-moral pairs. This new dataset is collected by [6] from multiple web pages of moral stories and is cleansed with de-duplication and decoupling. The average number of words and sentences are 322 and 18 for stories, 25 and 1.5 for morals. When applied in the phase of fine-tuning, the labeled data are randomly splitted by 8:1:1 for training/validation/testing set, respectively.

4.2. Compared Baselines

BERT The BERT-architectured model used in our work is the *bert-base-Chinese* register model [8]. It has been pre-trained for Chinese with the pre-training objective of masked language modeling.

RoBERTa The RoBERTa-architectured model used in our work is the *hfl/chinese-roberta-wwm-ext* register model [38]. It is essentially a Chinese pre-trained BERT model with whole word masking.

T5 The T5-architectured model used in our work is the *thu-coai/LongLM-base* register model [21]. It has been pre-trained on 120G Chinese long novels with two pre-training tasks including text infilling [12] and conditional continuation [10].

4.3. Experiment Settings

Our experiments are basing on Long LM-base [21], a Chinese pre-trained T5 model. All language models are implemented on the codes and pre-trained checkpoints from Hugging Face [36]. The model configurations are following their respective base version. As for the hyper-parameters for all models, we set the batch size to 16, the maximum sequence length to 1,024, and the learning rate to 3e-5. As for tokenization, a sentencepiece vocabulary of 32,000 word pieces [39] is applied. We use accuracy as the metric to evaluate the two understanding tasks.

5. RESULTS AND ANALYSIS

This section is going to specify and analyze the experimental results. Based on previous work done by [6], we conduct further domain-adaptive pre-training focusing on two relevant domains, moral culture and commonsense knowledge. [6] Has post-trained RoBERTa [38] and T5 [21] on the unlabeled data and names them RoBERTa-Post and T5-Post in the original paper. Such post-training is the task-adaptive pre-training that we call in our paper, thus we rename them RoBERTa-T and T5-T in Table 5 for better distinguishment with our methods. The T in their names means Task-adaptive pre-training, TD means both Task- and Domain-adaptive pre-training, but the domain is moral culture. TD+ means further pretraining about the domain of commonsense upon TD. Human means human performance on the two tasks, which has been tested by [6]. #Para is the approximate number of model parameters. For each task, the best performance is highlighted in bold and the second best is underlined, except for human performance.

Table 5. Accuracy (%) for CPT and PREF with different pre-training strategies.

Models	**CPT**	**PREF**	**#Para**
BERT [8]	59.62	82.97	110M
RoBERTa [38]	62.71	**89.54**	110M
RoBERTa-T [6]	64.61	<u>87.59</u>	110M
T5 [21]	69.60	82.00	220M
T5-T [6]	70.07	81.75	220M
T5-TD	<u>70.42</u>	82.68	220M
T5-TD+	**71.86**	82.41	220M
Human [6]	*95.00*	*98.00*	N/A

By analyzing the accuracy results in Table 5, we summarize our findings on two moral understanding tasks as follows: (1) T5 performs better than BERT and RoBERTa on CPT but worse on PREF, that tells that the encoder-only architecture might be good at aligning preferences; (2) We find that further pre-training does not always improve the performance on target tasks after comparing RoBERTa-T with RoBERTa and T5-TD+ with T5-TD on PREF, which advises that a better way is required to make use of these data especially when handling tasks similar with PREF; (3) We observe that different pre-training corpus brings different degrees of effects, which might depend on target tasks. T5-TD makes smaller progress than T5-TD+ on CPT, but the reverse happens on PREF, which indicating that the corpus of commonsense is more needed by CPT to enhance the ability of commonsense reasoning while PREF requires more moral data to capture value preferences; (4) Although a big gap exists between our models and human performance, further pre-training has proved its effectiveness. Zero-shot or few-shot learning has been an important trend, which is supported by PLMs with strong generalization capability.

6. CONCLUSIONS

In this paper, we suggest to leverage a three-phase paradigm ("pre-training + further pre-training + fine-tuning") instead of the traditional two-phase paradigm ("pre-training + fine-tuning"). The effects of the intermediate phase is tested on two downstream tasks of moral understanding. Specifically, the further pre-training is categorized in two types, task-adaptive and domain-adaptive, with the aim of enriching the language model with task- and domain-awareness knowledge. Task-adaptive pre-training refers to further pre-training on unlabeled training corpus for target tasks before fine-tuning on labeled corpus. As for domain-adaptive pre-training, we utilize corpora from two different domains including moral culture and commonsense knowledge. To be specific, the corpus about moral culture is composed of Vernacular Chinese of Confucius theory. Furthermore, we linearize the triples of a Chinese commonsense knowledge graph into readable natural language sentences for incremental domain-adaptive pre-training. Experimental results reveals the effectiveness of our method, and requires paying attention to specific task property and the relevance between the domains and the target task. Further pre-training performs better when the language model is more adaptable to the downstream tasks or when the content of the further pre-training corpus is more supportive for them. Larger-scale pre-training over multitasks and multi-domains is of high computational cost but still necessary, especially in low-resource settings. For future work, we will figure out a better way to make the best of the corpora of further pre-training, such as novel pre-training strategies and preferable data preparation.

ACKNOWLEDGEMENTS

This research was supported by the Research Development Fund (RDFSP16) at Xi'an Jiaotong-Liverpool University (XJTLU), XJTLU Key Programme Special Fund (KSF-A-17) and the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004).

REFERENCES

[1] L. Jiang, C. Bhagavatula, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, R. A. Rini, and Y. Choi, "Can machines learn morality? the delphi experiment," 2022.

[2] Z. Li, X. Ding, and T. Liu, "Story ending prediction by transferable bert," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[3] L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, "Plan-and-write: Towards better automatic storytelling," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7378–7385, Jul. 2019.

[4] F. Brahman and S. Chaturvedi, "Modeling protagonist emotions for emotion-aware storytelling," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020.

[5] X. Kong, J. Huang, Z. Tung, J. Guan, and M. Huang, "Stylized story generation with style-guided planning," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Aug. 2021.

[6] J. Guan, Z. Liu, and M. Huang, "A corpus for understanding and generating moral stories," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jul. 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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), Jun. 2019.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," ArXiv, vol. abs/1907.11692, 2019.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[13] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," in Proceedings of the 36th International Conference on Machine Learning, 2019.

[14] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general purpose language understanding systems," in Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., 2019.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," ACM Computing Surveys, vol. 55, pp. 1 – 35, 2021.

[16] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in SIGMOD Conference, 2008.

[17] D. Vrandeˇci´c and M. Kr¨otzsch, "Wiki data: A free collaborative knowledge base," Communications of the ACM, 2014.

[18] D. Li, Y. Li, J. Zhang, K. Li, C. Wei, J. Cui, and B. Wang, "C3KG: A Chinese commonsense conversation knowledge graph," in Findings of the Association for Computational Linguistics: ACL 2022, May 2022.

[19] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," in AAAI Conference on Artificial Intelligence, 2019.

[20] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in AAAI Conference on Artificial Intelligence, 2019.

[21] J. Guan, Z. Feng, Y. Chen, R. He, X. Mao, C. Fan, and M. Huang, "LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation," Transactions of the Association for Computational Linguistics, vol. 10, 2022.

[22] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, "A knowledge-enhanced pretraining model for commonsense story generation," Transactions of the Association for Computational Linguistics, vol. 8, 2020.

[23] F. Luo, D. Dai, P. Yang, T. Liu, B. Chang, Z. Sui, and X. Sun, "Learning to control the fine-grained sentiment for story ending generation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Jul. 2019.

[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 09 2019.

[25] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in The Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, 2020.

[26] S. Gururangan, A. Marasovi´c, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020.

[27] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in Chinese Computational Linguistics, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Springer International Publishing, 2019.

[28] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in Proceedings of the 20th Chinese National Conference on Computational Linguistics, Aug. 2021.

[29] B. He, D. Zhou, J. Xiao, X. Jiang, Q. Liu, N. J. Yuan, and T. Xu, "BERT-MK: Integrating graph contextualized knowledge into pre-trained language models," in Findings of the Association for Computational Linguistics: EMNLP 2020, Nov. 2020.

[30] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "KEPLER: A unified model for knowledge embedding and pre-trained language representation," Transactions of the Association for Computational Linguistics, vol. 9, 2021.

[31] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Jul. 2019.

[32] Y. Liu, Y. Wan, L. He, H. Peng, and P. S. Yu, “Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning,” ArXiv, vol. abs/2009.12677, 2020.

[33] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, “SpanBERT: Improving pre-training by representing and predicting spans,” Transactions of the Association for Computational Linguistics, vol. 8, 2020.

[34] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical neural story generation,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2018.

[35] Y. B. Ian Good fellow and A. Courville, “Deep learning,” MIT Press, 2016.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Oct. 2020.

[37] P. Hosseini, D. A. Broniatowski, and M. Diab, “Knowledge-augmented language models for cause-effect relation classification,” in Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022), May 2022.

[38] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, “Revisiting pre-trained models for Chinese natural language processing,” in Findings of the Association for Computational Linguistics: EMNLP 2020, Nov. 2020.

[39] T. Kudo and J. Richardson, “Sentence Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Nov. 2018.