KNOWLEDGE-ENRICHED MORAL UNDERSTANDING UPON CONTINUAL 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, China
[2]Department of Computer Science, University of Liverpool, Liverpool, UK
[3]School of Computer Science Technology, University of Bedfordshire, Luton, UK

ABSTRACT

The aim of moral understanding is to comprehend the abstract concepts that hide in a story by seeing through concrete events and vivid characters. To be specific, the story is highly summarized in one sentence without covering any characters in the original story, which requires the machine to behave more intelligently with the abilities of moral perception and commonsense reasoning. The paradigm of "pre-training + fine-tuning" is generally accepted for applying neural language models. In this paper, we suggest adding an intermediate stage to build the flow of "pre-training + continual pre-training + fine-tuning". Continual pre-training refers to further training on task-relevant or domain-specific corpora with the aim of bridging the data distribution gap between pre-training and fine-tuning. Experiments are basing on a new moral story dataset, STORAL-ZH, that composes of 4,209 Chinese story-moral pairs. We collect a moral corpus about Confucius theory to enrich the T5 model with moral knowledge. Furthermore, we leverage a Chinese commonsense knowledge graph to enhance the model with commonsense knowledge. Experimental results demonstrate the effectiveness of our method, compared with several state-of-the-art models including BERT-base, RoBERTa-base and T5-base.

KEYWORDS

Moral Understanding, Continual Pre-training, Knowledge Graph, Commonsense

1. INTRODUCTION

Morality is one of the most complicated topic about humanity [1]. It is tied with commonsense, formed upon ethnic culture, and regulated by rules and laws. Fable stories are must-read books for children, from which they learn morals and ethics to distinguish right from wrong in everyday world. Moral understanding aims to comprehend the abstract concepts that hide in a story by seeing through concrete events and vivid characters, which has become a new challenging task for Natural Language Processing (NLP). Previous works related to story understanding are mainly story ending prediction [2], story completion given constraints (e.g., storylines [3], emotions [4], styles [5], morals [6]). Most of them surround concrete concepts in story itself, whereas our work concentrates on digging out the moral lesson behind it. Table 1 shows one example of moral-story pair from the new dataset STORAL [6].

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 175-185, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130414

Benefit from big data, self-supervised pre-training on an enormous amount of unlabeled corpora from a general domain equips large language models with contextual knowledge and the ability of recognizing n-grams. Originated from Transformer [7], plenty of Pre-trained Language Models (PLMs) have sprung up in succession. They can be roughly categorized in three groups in terms of their architecture, including 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 dissimilar pre-training strategies, PLMs are suitable to different downstream tasks. For instance, the contextual word representations learned via masked language modeling by RoBERTa are beneficial for natural language understanding, while the strategy of auto-regressive language modeling exploited by GPT2 lays the foundation for natural language generation.

Table 1. An English moral-story pair from STORAL.

Moral	What is evil won is evil lost.
Story	A wolf had stolen a lamb and was carrying it off to his lair to eat it. But his plans were very much changed when he met a lion, who, without making any excuses, took the lamb away from him. The wolf made off to a safe distance, and then said in a much injured tone: "You have no right to take my property like that!" The lion looked back, but as the wolf was too far away to be taught a lesson without too much inconvenience, he said: "Your property? Did you buy it, or did the shepherd make you a gift of it? Pray tell me, how did you get it? "

The stage of pre-training equips language models with great potential, especially as the number of model parameters and the scale of unlabeled corpora keep growing, which can be proved by the superior performance achieved by prompt learning on a range of 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, continual pre-training on in-domain data (Domain-Adaptive Pre-Training, DAPT) or task-relevant data (Task-Adaptive Pre-Training, TAPT) is a recommended option when the downstream scenarios are of specific domains and no relevant data shows up in the unlabeled corpora.

Other than domain-specific knowledge and task-dependent information that are gained from incremental unlabeled unstructured text, sometimes it is necessary for PLMs to be equipped with the capability of commonsense reasoning. Furthermore, pre-training can be extended to other data of a different structure, such as Knowledge Graph (KG). A typical KG is composed of RDF triples (*h*, *r*, *t*), where *h* and *r* 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]. In one popular commonsense KG, ATOMIC [20], triples like (*PersonX applies to jobs*, *xEffect*, *gets hired*), (*PersonX asks PersonY for money*, *xWant*, *to go pay bills*) are telling inferential knowledge about everyday life.

In this paper, we transform the traditional two-stage paradigm of "pre-training + fine-tuning" into three stage by adding an intermediate step of continual pre-training that will be tested by two downstream tasks about moral understanding. We use STORAL-ZH [18], the Chinese part of STORAL [6], as the dataset for target tasks. Furthermore, LongLM-base [21] is selected as our model, that has been pre-trained on 120G Chinese long novels. For TAPT, the language model is further trained on unlabeled STORAL-ZH to equip itself with task-awareness knowledge. For DAPT, we prepare training corpora for two domains including moral culture and commonsense knowledge. Inspired by [22], we utilize triples of a KG by transforming each triple into one readable textual sequence for continuing to pre-train the language model. To summarize, our contributions are reflected in the following three aspects: (1) Different from the standard paradigm, we choose continual pre-training before fine-tuning to boost model performances on moral understanding. (2) For facilitating the moral perception out of concrete characters and events, we equip our model with the ability of commonsense reasoning by further pre-training on a commonsense KG. (3) We collect a corpus about Chinese traditional moral culture about the Four Books and Five Classics to support domain-adaptive pre-training.

2. Related Work

2.1. Story Understanding

There have been a range of tasks proposed about story understanding and generation, including story ending prediction [2], commonsense story generation [22] and story ending generation with fine-grained sentiment [23]. A variety of attributes are considered for better story understanding, such as storylines [3], emotions [4], styles [5], and morals [6]. Different from that storylines lead the story writting, emotions describe characters’ states, styles decide the story’s tone, moral understanding aims to discover the implied and abstract theme behind concrete events, that is a more challenging task. [6] firstly proposed moral understanding and generation, and published a new moral story dataset, STORAL.

2.2. Continual Pre-training

Pre-training is definitely the most essential stage for employing language models, that facilitates model initialization and accelerates the parameter convergence on downstream tasks. As the model size grows rapidly, larger-scale unlabeled corpora are required to fully pre-train the model to avoid over-fitting. To bridge the data distribution gap between pre-training and fine-tuning, continual pre-training has been applied and shows to be beneficial for model performance [24, 25]. [26] proposes two concepts about continual pre-training, task-adaptive pre-training (TAPT) and domain-adaptive pre-training (DAPT). The TAPT refers to further pre-training on the unlabeled data of the given task before fine-tuning, which brings consistent improvements [27]. The DAPT requires collecting target domain-relevant corpus, which is probably computationally expensive but still helpful [28].

2.3. Knowledge-Enhanced PLMs

Recently, incorporating knowledge into PLMs is experiencing a surge of interest. Thorough self-supervised pre-training over large-scale corpora provides PLMs with abundant contextual semantics but lacks domain-specific [19, 29], factual [30, 31] or commonsense knowledge [22, 32]. K-BERT [19] explicitly injects triples from domain-specific KG into the input sequence and designs a visible matrix to control the mutual effects among tokens. BERT-MK [29] integrates the graph contextualized knowledge of a medical KG into language models. KEPLER [30] encodes entity descriptions as their embeddings and jointly optimize the knowledge embedding and masked language modeling objectives on 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 of concepts over a commonsense KG for generative commonsense reasoning. [22] conducts incremental pre-training on commonsense knowledge bases to generate more reasonable stories without considering heterogeneous information fusion and sub-graph aggregation, which implicitly and efficiently incorporates commonsense knowledge into GPT-2 [10].

3. METHODOLOGY

This section expatiates the main components of our method, including the PLM, the details about task-adaptive and domain-adaptive pre-training, and the stage of fine-tuning.

3.1. Transformer-based Language Model

The language model adopted in this work is based on the full Transformer architecture [7], where the encoder is fed an input sequence and uses fully-visible masking, the decoder generates the target sequence through causal masking and cross-attention. The text-to-text framework is capable of handling both understanding and generation tasks. One representative encoder-decoder model is T5 [12], it is trained on the Colossal Clean Crawled Corpus (C4) of languages of English, French, Romanian, and German with the best-fit unsupervised pre-training objective of replacing corrupted spans.

A Chinese version of T5, LongLM, is released by [21] after being pre-trained on 120G Chinese novels with two generative tasks, i.e., text infilling [12] and conditional continuation [10]. Inspired by SpanBERT [33], text infilling replaces a few of text spans of input sequence by special tokens with a corruption rate of 15%, while the span lengths are following the Poisson distribution with $\lambda = 3$. Then the target is to output the original text spans replaced by special tokens with the greedy decoding algorithm. The second task, conditional continuation, aims to generate the back half of a text given 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 with the number of parameters of 223M on HuggingFace [36].

3.2. Continual Pre-training

3.2.1. Task-Adaptive Pre-Training (TAPT)

To make the pre-trained model more adaptive to downstream tasks, further pre-training on the unlabeled data of the tasks before fine-tuning is worth considering. The advantages of TAPT are reflected in much less computational cost and possible performance boost because the training corpus is far smaller and much more task-relevant. [6] has post-trained the Chinese long-text pre-training model named LongLM [21] on the unlabeled version of STORAL [6], and named it as T5-Post as one compared baseline in the original paper. Table 2 shows an example for the pre-training task of text infilling.

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

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)

Apart from continual pre-training on unlabeled data of downstream tasks, further pre-training on much more unlabeled corpora that are collected from relevant domains is more reasonable. By DAPT, the already powerful PLMs are enriched with additional domain-specific knowledge. As for better moral understanding of fable stories, background domains include moral culture and commonsense knowledge.

Moral Knowledge Confucianism is the mainstream moral culture of China, and the Four Books and Five Classics are its authoritative books, which record in detail the politics, economy, diplomacy, culture and other aspects of the most active period in the development of Chinese ideology, as well as the Confucian philosophy which has influenced Chinese culture for thousands of years. Up to now, the morals and ethics conveyed by the Four Books and Five Classics still regulate, correct and improve the ways we think and behave. The Four Books and Five Classics were written in classical Chinese, we collect the translated version in written vernacular Chinese as the corpus for continual pre-training and named it as **4+5**. Table 3 gives several examples from the Analects out of the Four Books.

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

Example 1	不患人之不己知，患不知人也。
Vernacular Chinese	**不要担心别人不了解自己，应该担心的是自己不了解别人。**
English Translation	I am not bothered by the fact that I am unknown. I am bothered when I do not know others.
Example 2	学而不思则罔，思而不学则殆。
Vernacular Chinese	**学习而不思考就会迷惘无所得，思考而不学习就不切于事而疑惑不解。**
English Translation	To study and not think is a waste. To think and not study is dangerous.
Example 3	德不孤，必有邻。
Vernacular Chinese	**品德高尚的人不会孤独，一定有志同道合的人和他做伴。**
English Translation	If you are virtuous, you will not be lonely. You will always have friends.

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

Triples	**Transformed Sentences**
(某人完全放弃某物, **xEffect**, 羞愧地低下头) (*PersonX abandons ____ altogether, **xNeed**, hangs head in shame*)	汤姆完全放弃某物，结果他羞愧地低下头。 Tom abandons something altogether, as a result, he hangs head in shame.
(有人被大学录取了, **xAttr**, 好学的) (*PersonX accepts into college, **xAttr**, studious*)	汤姆被大学录取了，他是好学的。 Tom accepts into college, he is studious.
(某人完成了他的任务, **xIntent**, 赶上最后期限) (*PersonX accomplishes PersonX's task, **xIntent**, to meet a deadline*)	汤姆完成了他的任务，因为他想赶上最后期限。 Tom accomplishes his task, because he wanted to meet a deadline.

Commonsense Knowledge Incorporating commonsense knowledge equips PLMs with the ability of commonsense reasoning in downstream tasks. In this work, we leverage the commonsense knowledge from a structured data, i.e., knowledge graph. ATOMIC [20], one of the most commonly used commonsense knowledge graph, 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, continual pre-training directly on all linearized triples can integrate commonsense knowledge into LongLM [21] implicitly in a more convenient way.

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 Continual 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 continual 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 4209 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 stage 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 LongLM-base [21], a Chinese pre-trained T5 model. All language models are implemented on the codes and pre-trained checkpoints from HuggingFace [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 wordpieces [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 continual 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 **T**ask-adaptive pre-training, **TD** means both **T**ask- and **D**omain-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	87.59	110M
T5 [21]	69.60	82.00	220M
T5-T [6]	70.07	81.75	220M
T5-TD	70.42	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 continual pre-training does not always imrpove 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, continual 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-stage paradigm ("pre-training + continual pre-training + fine-tuning") instead of the traditional two-stage paradigm ("pre-training + fine-tuning"). The effects of the intermediate stage is tested on two downstream tasks of moral understanding. Specifically, the continual 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. Continual pre-training performs better when the language model is more adaptable to the downstream tasks or when the content of the continual 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 continual pre-training, such as novel pre-training strategies and preferable data preparation.

ACKNOWLEDGEMENTS

This research was funded by the Research Development Fund at Xi'an Jiaotong Liverpool University, contract number KSF-A-17.

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, "Wikidata: 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 Goodfellow 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, “SentencePiece: 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.