CAM-BERT: Chinese Aerospace Manufacturing Pre-trained Language Model

Jinchi Dai
dept. School of Computer Science
Shenyang Aerospace University
Shenyang, China
daijinchi@stu.sau.edu.cn.com

Shengren Wang
dept. Network Information Center
AVIC Shenyang Aircraft Company Limited
Shenyang, China
wangsr002@avic.com

Peiyan Wang*
dept. School of Computer Science
Shenyang Aerospace University
Shenyang, China
wangpy@sau.edu.cn
*Corresponding author

Ruiting Li
dept. School of Computer Science
Shenyang Aerospace University
Shenyang, China
liruiting@stu.sau.edu.cn.com

Jiaxin Chen
dept. School of Computer Science
Shenyang Aerospace University
Shenyang, China
1261047104@qq.com

Xinrong Li
dept. School of Computer Science
Shenyang Aerospace University
Shenyang, China
lixinrong54@163.com

2024 6th International Conference on Natural Language Processing (ICNLP) | 979-8-3503-4911-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICNLP60986.2024.10692464

***Abstract*—In the era of intelligent manufacturing and Industry 4.0, there is a growing demand for specialized Chinese pre-trained language models designed for the aerospace manufacturing. This is essential to overcome the limitations of general-purpose models when processing aerospace manufacturing texts with complex domain-specific terms and frequent proper nouns. This paper introduces the Chinese Aerospace Manufacturing Pre-trained Language Model (CAM-BERT), developed by compiling an aerospace-specific vocabulary and implementing a continual pretraining strategy. CAM-BERT notably enhances the feature extraction capabilities within aerospace manufacturing texts. We conducted experiments in named entity recognition (NER) and relation extraction (RE), under the pretraining fine-tuning and few-shot prompt learning paradigms, to validate CAM-BERT's effectiveness in information extraction tasks within the Chinese aerospace manufacturing.**

***Keywords*—pre-trained language model, aerospace manufacturing, information extraction**

I. Introduction

In Natural Language Processing (NLP), the pre-trained language model is a language model that has been trained on large-scale datasets [1]. It obtains a generalized linguistic representation of language and the extraction of complex semantic information from vast text corpora [2]. Consequently, pre-trained language models have found broad applications across a variety of downstream tasks [3].

In recent years, pre-trained language models have demonstrated considerable success in general-purpose Chinese natural language processing tasks. However, its outperforms in specialized domains like healthcare, finance, and aerospace manufacturing remains less than optimal. Particularly in the aerospace manufacturing, the corpus of the aerospace manufacturing, characterized by an abundance of proper names, domain-specific terms, and intricate features, poses a challenge for general domain pre-trained language models. Presently, general domain pre-trained language models are limited by the characteristics of their training corpora, which frequently lack adequate semantic comprehension and expressive power. This deficiency adversely affects the models' performance in downstream domain-specific tasks. This deficiency significantly impairs their performance in downstream tasks within this specialized domain. Addressing the aforementioned challenges, this paper introduces a specialized Chinese pre-trained language model for the aerospace manufacturing, termed CAM-BERT (Chinese Aerospace Manufacturing BERT). This model is an adaptation of the BERT-base-Chinese[1], the Chinese variant of the Bidirectional Encoder Representations from Transformers [4]. CAM-BERT employs Whole Word Masking [5] (WWM) in its Continual Pretraining strategy, integrated with a tailored aerospace-specific vocabulary. This approach allows for a more profound mining of semantic information pertinent to this domain. CAM-BERT demonstrates enhanced performance in information extraction tasks, underscoring its effectiveness in processing aerospace manufacturing texts. The main contributions of this paper are outlined below:

- We construct a Chinese vocabulary for the aerospace manufacturing, and the text semantic information is deeply mined through the Chinese WWM mechanism, which further improves the feature extraction ability of the text in the aerospace manufacturing.
- We introduce the continual pretraining strategy, combined with the WWM pretraining task, to effectively learn the lexical semantic information in the aerospace manufacturing, which significantly improves the model's ability to extract information in the aerospace manufacturing.

[1]https://huggingface.co/bert-base-chinese

- This paper presents experiments conducted under two paradigms: pretraining fine-tuning and few-shot prompt learning, applied to two downstream tasks, namely NER and RE. The experiments yield satisfactory results, substantiating the efficacy of the proposed CAM-BERT model.

II. Related Work

The advent of advanced pre-trained language models like ELMO [6], GPT [7], and BERT has marked a significant evolution in NLP. Pre-trained models, epitomized by BERT, have outstripped traditional models in numerous NLP tasks. Initially trained on extensive unlabeled text corpora to develop a generalized language representation, they are subsequently fine-tuned for various downstream tasks. Nevertheless, the performance of these models on domain-specific downstream tasks is often compromised due to disparities between the pretraining and fine-tuning domains. In this context, researchers have proposed the continual pretraining strategy. This approach involves continual pretraining general domain language models within a specific domain, thereby enhancing their applicability to downstream tasks such as NER and RE. For instance, Lee et al. introduce BioBERT [8], which employs this strategy by continuing BERT's training, initially on a general domain corpus, onto a biomedical corpus, thus effectively mining medical text information. Similarly, Araci developes FinBERT [9], also leveraging continual pretraining. This model, initially pre-trained on general domain corpora, is further trained on a vast financial corpus and fine-tuned on specific financial datasets, showing improved performance in financial text analysis. Furthermore, Andrade et al. introduce SafeAeroBERT [10], a pre-trained language model specifically developed for the aviation safety domain. This model employs Continual Pretraining strategy, initially using a corpus of aviation safety documents, and fine-tuned for document categorization task, resulting in satisfactory outcomes.

In summary, the performance of general domain pre-trained BERT models significantly improves in downstream tasks of specific domains through the continual pretraining strategy. Consequently, the adoption of this strategy in our research enables more effective improvement of semantic comprehension and overall performance in specific domains for general-purpose pre-trained language models.

Beltagy et al. introduce the SCIBERT [11], demonstrating that employing a domain-specific vocabulary can significantly enhance performance. SCIBERT is developed with a specific vocabulary for the scientific and technological domains and pre-trained on an unsupervised corpus of 1.14 million papers from these domains. It performs better than BERT in downstream tasks like sequence tagging and sentence classification based on scholarly texts. Similarly, Chandra et al. develop Aviation-BERT [12], a BERT model tailored for the domain of English aviation safety. This model constructe with a specific aviation domain vocabulary, utilized from aviation databases for pretraining. Aviation-BERT is shown to outperform BERT when it comes to text-mining tasks on aviation text datasets. It is also expected to be of tremendous value in numerous downstream tasks in the analysis of aviation text corpora.

In conclusion, domain-specific vocabularies has been shown to significantly enhance performance in downstream tasks. However, the prevalent use of English. Therefore, constructing a Chinese vocabulary for the aerospace manufacturing can enable the model to better learn and adapt to the linguistic features of the Chinese aerospace manufacturing.

III. Method

This study is divided into two phases: model pretraining and model performance validation. In the pretraining phase, the CAM-BERT model is first initialized with the weight parameters of the benchmark model Base-BERT. Then the aerospace-specific vocabulary is added and the WWM language modeling task is used to further pre-train the aerospace manufacturing dataset. The initial evaluation metrics of CAM-BERT focus on the model's perplexity and restored character accuracy. In the model performance validation phase, the CAM-BERT pretraining fine-tuning and few-shot prompt learning paradigms are fine-tuned for the NER and RE tasks, respectively. Fig. 1 provides an overview of the methodology and the following subsections detail the steps taken to develop the CAM-BERT model.

A. Vocabulary and Tokenizer

In order to obtain an aerospace-specific vocabulary, we use the BERT-BiLSTM-CRF [13] to perform the NER task on the preprocessed dataset from Section IV.A.a, sorting the dataset according to the frequency of recognized occurrences of entities, and a minimum frequency threshold is established to determine the minimum number of occurrences required for a word to be included in the final vocabulary. Furthermore, a cut-off value is applied to restrict the maximum number of word included, effectively managing the vocabulary's size. To ensure the accuracy of entity recognition, the 1200 entities with the highest frequency were manually checked. These filtered domain vocabularies are added to the original Base-BERT model vocabulary, and the new voca-bularies are duplicated and their compatibility with the original vocabulary is ensured. This enhances the tokenizer's familiarity with frequently used out-of-vocabulary words.

B. Continual Pretraining

We use an aerospace manufacturing corpus for continual pretraining task based on the Base-BERT model. During pretraining, only WWM pretraining task is used. This approach presents a greater challenge for MLM pre-training and demonstrates enhanced performance in certain downstream tasks. Therefore, in the current work, we have chosen WWM as the pre-training masking strategy.

IV. Experiments

A. Dataset

We collect manufacturing specification texts and manually annotate entities and relationships to serve as the dataset for

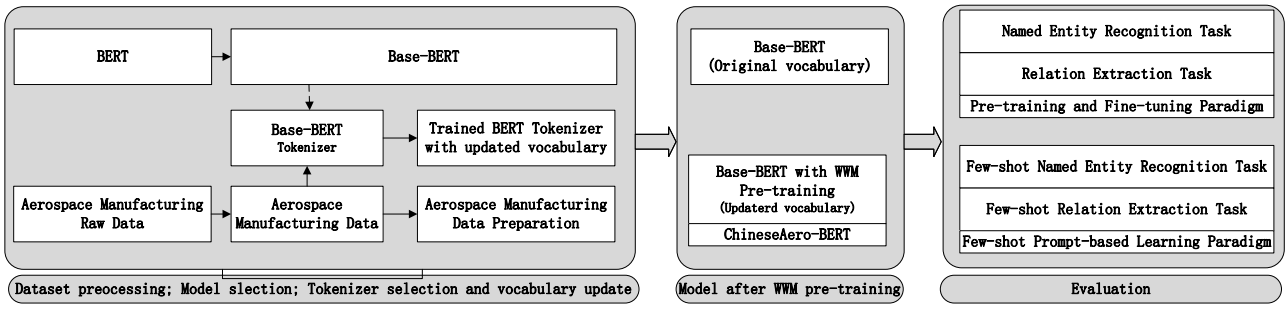

Fig. 1. Overview of the Methodology.

downstream tasks. The dataset is divided in a 7:2:1 ratio to obtain the training set, test set, and validation set. To adequately simulate scenarios with rich-resource and few-shot samples, we further processed the training set.

a) Dataset for Masked Language Model: In this study, texts related to the aerospace manufacturing were obtained from papers in the public manufacturing domain and national product manufacturing standard specifications as a corpus for continual pretraining. To extract valuable corpus data, redundant spaces and blank lines were removed, irrelevant data was eliminated, and some character types were replaced with alternative labels. This preprocessing aimed to shift the model's attention towards the structural and semantic content of the text. After completing the preprocessing, a total of 3,126,301 high-quality corpora were obtained. Subsequently, the optimized corpus was randomly shuffled and divided into training and test sets, with 2,151 pieces of corpus selected for the test set to assess the pre-trained language model's character restoration capability.

b) Dataset for Named Entity Recognition: We establish training sets at sentence-level and entity-level to simulate scenarios with rich-resource and few-shot samples. Sentence-level training sets are extracted from the complete training set according to a certain ratio, denoted as n%-sen, where n% represents the proportion of data extracted. Entity-level training sets are extracted from the complete training set based on the number of entities, denoted as n-shot, where n is the number of entities for each category.

c) Dataset for Relational Extraction: For relation extraction, sentence-level training sets and relation-level datasets are also established. Sentence-level training sets are extracted from the complete training set according to a certain ratio, denoted as n%-sen, where n% represents the proportion of data extracted. Eelation-level training sets are extracted from the complete training set based on the number of relations, denoted as n-shot, where n represents the number of each type of relation.

Sentence-level data training sets retain the linguistic characteristics and distribution of entities of domain texts. This aids the model in understanding and learning the specific contexts, grammatical structures, and relationships between entities in the domain, making it suitable for rich-resource scenarios. Entity-level data training sets focus on balancing the distribution of entities within the dataset, ensuring that the model equally learns about each type of entity, making it suitable for few-shot scenarios.

B. Experimental Design

a) Model Validation Experiments: In this paper, the performance of the pre-trained language model is evaluated from three different aspects: accuracy, perplexity and loss function.

Accuracy refers to the proportion of original characters correctly restored by the model, which directly reflects the performance of the model at the character level. N is the length of character sequence, and C is the number of characters correctly predicted.

$$Accuracy = \frac{C}{N} \times 100\%$$

Perplexity is another important metric for evaluating the performance of a language model. It can be used to assess the model's ability to perform in language comprehension and production. Specifically, it is a measure of the model's uncertainty in predicting the next character or word in a text sequence. N is the length of the sequence, x_i is each element in the sequence, and $p(x_i)$ is the logarithm of the model's predicted probability for x_i.

$$Perplexity = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2(p(x_i))}$$

The loss function evaluates the inconsistency between the model's predicted characters and the actual correct characters. It is used to measure the difference between the text generated by the model and the original input text. When the model's generated text matches the original text, a lower value of the loss function indicates a higher accuracy of the model's reduction. N is the length of the generated character sequence, x_i is the character generated at time i , and $p(x_i \mid x_1, x_2, ..., x_{i-1})$ is the conditional probability that the model generates the character at time i.

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\log(p(x_i|x_1, x_2, ..., x_{i-1}))$$

This combined evaluation approach helps to provide an in-depth understanding of the strengths and limitations of the

model and provides an important basis for subsequent model improvement and application.

b) Downstream Tasks: We compare CAM-BERT with Base-BERT model with downstream tasks. we respectively employ pretraining fine-tuning and prompt learning paradigms in rich-resource and few-shot scenarios. Specifically for NER, we use BERT-Softmax, BERT-CRF [14], and BERT-BiLSTM-CRF in rich-resource scenario; whereas EntLM [15] in few-shot scenario, which is specifically designed for few-shot NER. For RE, we use R-BERT [16] in rich-resource scenario and PTR [17] in few-shot scenario.

In NER, TP (True Positive) indicates that the model correctly recognizes the number of entities, FP (False Positive) indicates that the model incorrectly recognizes the number of entities, and FN (False Negative) indicates that the model does not recognize the number of entities. In RE, the relationship class labeled by the to-be-tested instance is the positive class and the relationship classes other than the positive class are the negative classes during the prediction of the current to-be-tested instance by the model.TP denotes the number of positive classes predicted by the model as positive classes, i.e., the model succeeded in predicting the relationship classes labeled by the to-be-tested instance, and FP denotes the number of negative classes predicted as positive classes, i.e., the model incorrectly predicted the relationship classes that are not the to-be-tested instance's labeled as positive classes. predicted as positive classes. FN denotes the number of positive classes predicted as negative classes. In this paper, the accuracy P (Precision), the recall R (Recall), and the $F1$ value are used to evaluate the entity recognition effect of each model.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

C. Analysis of Experimental Results

We employ the masked language task to assess CAM-BERT's ability to understand domain-specific texts. Additionally, through the performance of CAM-BERT in NER and RE, we evaluate the model's application capabilities, including its effectiveness in identifying and understanding aerospace manufacturing specific entities and relationships.

1) Masked Language Task: The results of the masked character reduction experiment are shown in Table I. It can be seen that the CAM-BERT model achieves 75.10% in terms of accuracy, which is a significant improvement compared to the 71.92% of the Base-BERT model. In addition, CAM-BERT performs better in terms of perplexity, which is only 3.7835, compared to Base-BERT's perplexity of 4.7638. The low perplexity indicates that CAM-BERT is more accurate in predicting the language patterns in the test set. In terms of loss value, CAM-BERT also outperforms Bert-base-Chinese with 1.3306 and 1.561 respectively, indicating that CAM-BERT has a smaller overall error in the prediction process. In summary, the CAM-BERT model shows better performance on the test set, outperforming the Base-BERT model in terms of accuracy, perplexity and loss value. This result indicates that the CAM-BERT model has higher efficiency and accuracy in understanding and processing the corpus in the aerospace manufacturing, and can be better adapted to the application requirements in the aerospace manufacturing.

TABLE I
$Accuracy$, $Perplexity$ AND $loss$ IN MASKED LANGUAGE TASK.

	$Accuracy$	$Perplexity$	$loss$
Base-BERT	71.92	4.76	1.56
CAM-BERT	75.10	3.78	1.33

TABLE II
$F1$ RESULTS (%) OF NER IN RICH-RESOURCE.

		Base-BERT	**CAM-BERT**
BERT-Softmax	10%-sen	58.20	62.11
	50%-sen	74.68	74.83
	100%-sen	76.12	77.08
BERT-CRF	10%-sen	66.99	67.46
	50%-sen	68.80	69.92
	100%-sen	75.84	76.72
BERT-BiLSTM-CRF	10%-sen	67.67	69.09
	50%-sen	69.92	74.38
	100%-sen	76.84	77.89

2) NER Task:

a) Pretraining fine-tuning in rich-resource: Table II shows the results of NER experiment in rich-resource. We observe that CAM-BERT always outperforms Base-BERT. In 100%-sen, our CAM-BERT demonstrated a 1.05% improvement. The performance of CAM-BERT highlights the importance of domain-adaptive pretraining in rich-resource. BERT-BiLSTM-CRF performs the best, followed by BERT-CRF, while BERT-Softmax exhibits the lowest performance. This observation suggests that the BiLSTM may provide additional contextual information to the model, thereby enhancing the ability of entity recognition.

TABLE III
$F1$ RESULTS (%) OF NER IN FEW-SHOT.

	5-shot	**10-shot**	**20-shot**
Base-BERT	17.94	22.33	35.82
	50-shot	100-shot	200-shot
	42.99	50.68	58.79
	5-shot	**10-shot**	**20-shot**
CAM-BERT	23.69	29.94	40.06
	50-shot	100-shot	200-shot
	46.86	5236	60.03

b) Prompt learning in few-shot: The few-shot NER experiment results are listed in Table III. As the training sets increase from 5-shot to 200-shot, the model's performance gradually improves. Additionally, CAM-BERT consistently outperforms Base-BERT in few-shot scenario for prompt

learning. It's worth noting that the performance of CAM-BERT in 5-shot is even better than that of Base-BERT in 10-shot. This highlights the significance of domain-adaptive pretraining in few-shot.

TABLE IV
$F1$ RESULTS (%) OF RE IN RICH-RESOURCE.

	11- types		62-types	
	Base-BERT	CAM-BERT	Base-BERT	CAM-BERT
10%-sen	82.76	82.76	72.18	73.84
50%-sen	90.86	91.88	88.84	90.61
100%-sen	93.85	94.30	92.82	93.26

3) RE Task:

a) Pretraining fine-tuning in rich-resource: The extraction results for 11-types and 62-types relations are presented in Table IV.The conclusions drawn for NER are equally applicable. In both 11-types and 62-types relation extraction tasks, CAM-BERT consistently yields performance improvements. The most significant enhancements are observed on 50%-sen dataset, with an increase of 1.02 for 11-types relations and 1.77 for 62-types relations.

TABLE V
$F1$ RESULTS (%) OF RE IN FEW-SHOT.

		4-shot	8-shot	16-shot
11-types	Base-BERT	48.73	58.60	79.38
	CAM-BERT	38.84	56.12	75.08
		32-shot	128-shot	
	Base-BERT	82.22	89.15	-
	CAM-BERT	83.00	90.40	
		4-shot	8-shot	16-shot
62-types	Base-BERT	72.75	72.73	81.65
	CAM-BERT	73.83	75.71	82.84
		32-shot	128-shot	
	Base-BERT	85.78	80.16	-
	CAM-BERT	86.06	80.30	

b) Prompt learning in few-shot: The experimental results for few-shot RE are listed in Table V. As the number of samples increases, the performance gap between CAM-BERT and Base-BERT gradually diminishes, but CAM-BERT continues to maintain its advantage. Compared to the 11-types relations, the 62-types relations involve a larger number of relation types and a more complex data distribution, which can lead to relatively lower performance. However, CAM-BERT demonstrates significant potential in handling this diversity and complexity, resulting in improved relation extraction performance.

V. CONCLUSION

In summary, the continual pretraining strategy and the construction of an aerospace manufacturing vocabulary can effectively improve the ability of the pre-trained language model to learn aerospace manufacturing knowledge and achieve better performance in downstream tasks in the aerospace manufacturing. In future research, we will further optimize the model and explore more advanced preprocessing and data enhancement techniques for more accurate NER and RE.

REFERENCES

[1] Q.-N. Nguyen, T. C. Phan, D.-V. Nguyen, and K. V. Nguyen, "Visobert: A pre-trained language model for vietnamese social media text processing," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264172556

[2] K. Kanclerz and M. Piasecki, "Deep neural representations for multiword expressions detection," in *Annual Meeting of the Association for Computational Linguistics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248780008

[3] N. J. Hu, E. Mitchell, C. D. Manning, and C. Finn, "Meta-learning online adaptation of language models," *ArXiv*, vol. abs/2305.15076, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258866057

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[5] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:260471499

[6] "Embedding from language models (elmos)- based dependency parser for indonesian language," *International Journal of Advances in Soft Computing and its Applications*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245581436

[7] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[8] 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, pp. 1234 – 1240, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59291975

[9] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *ArXiv*, vol. abs/1908.10063, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201646244

[10] S. R. Andrade and H. S. Walsh, "Safeaerobert: Towards a safety-informed aerospace-specific language model," *AIAA AVIATION 2023 Forum*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259788141

[11] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202558505

[12] C. Chandra, X. Jing, M. V. Bendarkar, K. Sawant, L. R. Elias, M. R. Kirby, and D. N. Mavris, "Aviation-bert: A preliminary aviation-specific natural language model," *AIAA AVIATION 2023 Forum*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259786699

[13] Y. Gan, R. Yang, C. Zhang, and D. Jia, "Chinese named entity recognition based on bert-transformer-bilstm-crf model," *2021 7th International Symposium on System and Software Reliability (ISSSR)*, pp. 109–118, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244857678

[14] S. Hu, H. Zhang, X. Hu, and J. Du, "Chinese named entity recognition based on bert-crf model," *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pp. 105–108, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252224067

[15] R. Ma, X. Zhou, T. Gui, Y. C. Tan, Q. Zhang, and X. Huang, "Template-free prompt tuning for few-shot ner," *ArXiv*, vol. abs/2109.13532, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238198383

[16] S. Wu and Y. He, "Enriching pre-trained language model with entity information for relation classification," *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160009395

[17] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "Ptr: Prompt tuning with rules for text classification," *ArXiv*, vol. abs/2105.11259, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235166723