Entailment Method Based on Template Selection for Chinese Text Few-shot Learning

Zeyuan Wang†, Zhiyu Wei†, Lihui Zhang, Ruifan Li‡ and Zhanyu Ma
† Both authors contributed equally to this research. ‡ Corresponding author.
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
E-mail: {wangzeyuan, zydotwei, elliot_zlh, rfli, mazhanyu}@bupt.edu.cn

***Abstract*—The limit of labeled data has become the bottleneck in numerous text-related tasks. Recently, few-shot learning based on pre-trained language model has become an attractive topic. Entailment-based Few-shot Learning (i.e., EFL) is an effective way through transforming a text classification task into a textual entailment task, which bridges the gap between downstream tasks and pre-trained tasks. However, the performance of the downstream task is sensitive to the manually selected templates in this type of approaches. To alleviate this problem, we improve the EFL method by applying a naïve template selection mechanism, leveraging masked language model to assess the quality of candidate templates. Moreover, we evaluate our method on FewCLUE shared tasks. Extensive experiments demonstrate the effectiveness of our proposed method.**

I. Introduction

There is a prevailing trend to leverage pre-training and fine-tuning paradigm to solve various natural language processing (i.e., NLP) tasks. Specifically, this strategy first train language models on large-scale unlabeled samples, and then perform downstream tasks with the fine-tuning strategy. However, there exist various domains tackling different tasks and languages in real-world applications, which require an enormous cost in crowd-sourcing high quality annotations. Compared with machines, human can understand a specific or conceptual object with only a few samples. Therefore, learning to solve problems from only a few examples is still a challenging task.

It is noteworthy that some pre-trained models with large-scale parameters such as GPT-3 [1] can achieve state-of-the-art performance after learning limited samples. However, the methodology makes it difficult to fine-tune and to deploy services. Therefore, some works such as [2], [3] reformulate the downstream tasks into cloze problems, allowing the pre-trained language model to predict the answer by reusing the Masked Language Model (i.e., MLM) head [2]. However, Gao et al. [4] shows that when the data distributions of the downstream tasks are different from the pre-trained text corpus, such as natural language inference tasks, the convergence performance will be severely limited.

In our solution, MacBERT [5], a pre-training model that optimizes traditional MLM tasks, is chosen as the backbone of the system. Entailment-based Few-shot Learning (i.e., EFL) [6] proposed by Facebook, is applied to fine-tune specific tasks. This method differs from the prompt-based strategy of cloze question such as PET [2] and LM-BFF [4]. The basic idea of EFL behind is to transform the original task into a textual entailment task through a predefined template filled with fine-grained label description, which is selected by the auto-encoder network [7]. After turning into textual entailment tasks, the model is supposed to take a pair of sentences which comprise the original sample and its label description. And then the model predicts whether the fact in the first sentence can necessarily imply the fact in the gpsecond one. Furthermore, we use the pre-trained model as a template screening tool considering the adaptability of the language model itself. The experimental results verify that through transforming to entailment style tasks, our method can achieve competitive performance on multiple few-shot datasets.

We summarize our major contributions as follows:

- We design an automatic template selection method for the problem of templates' sensitivity and evaluate its effectiveness.
- We verify the performance of entailment-based method in multiple Chinese text datasets.
- We share the code of all our experiments for advancing the knowledge. [1]

II. Related Work

A. Meta-learning

Meta-learning achieves great progress in several scenarios of few-shot learning, such as text classification [8], machine translation [9] and text generation [10]. In general, meta-learning methods can be categorized into two groups regarding to optimization-based and metric-based directions, respectively. The former methods work in the perspective of optimization by utilizing better initial parameters [11] or task-specific adjustments [12]. While the latter introduces several effective metrics as similarity evaluation, and performs inference depending on the similarity between the testing data and labeled data [13].

B. Pre-trained Language Model

The idea of using the pre-trained models for few-shot learning is to transform the form of the downstream tasks,

[1] https://github.com/thunderboom/EntailmentTemplateSel

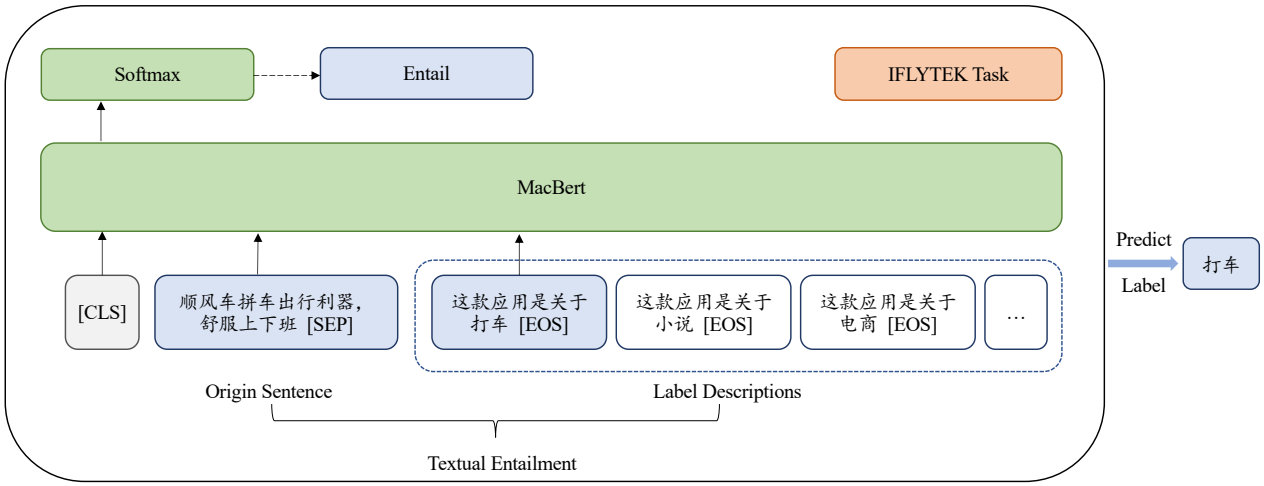

Fig. 1. Entailment-based reformulation of IFLYTEK task. Each class has a label description, and we choose the class with maximum probability of entailment (shown in light blue) between the original sentence and the label descriptions.

so as to bridge the gap between the downstream tasks and the pre-training tasks. In this way, the model can achieve better performance with limited samples. According to the types of pre-training tasks, these methods can be divided into cloze tasks and sentence-pair tasks. LM-BFF [4] is a typical technique in cloze tasks, which converts text classification tasks into word-selection tasks by setting a template. In addition, PET [2] combines unsupervised data and ensemble model to improve the performance of the cloze test. EFL [6] concentrates on transforming to sentence pairs, which greatly improves the effectiveness of few-shot learning by converting classification tasks into entailment tasks.

III. Methodology

A. Pre-trained Backbone

Mainstream pre-trained models in NLP tasks include unidirectional language models, such as ELMo [14], GPT [15] and GPT-2 [16], and bidirectional language models like BERT [17] and its variants [18]–[21]. Starting from BERT, researchers have made great and rapid progress on optimizing the model. Recently, Cui et al. [5] proposed a simple but effective network called MacBERT, aiming to build Chinese pre-trained models. This method replaces the original MLM task of BERT with the correction task. Specifically, certain tokens from the input are randomly masked with similar words. This improvement can mitigate the discrepancy between the pre-trained phase and the fine-tuning phase. For the sake of better evaluation, we therefore employ the MacBERT as our pre-trained model.

B. Entailment Framework

EFL [6] can transform a small-scale language model into a better few-shot learner. The method reformulates the potential NLP task as a textual entailment task, and converts the label of the data sample into a corresponding label description through a predefined template. Few-shot learners based on EFL can even reach competitive performance compared to GPT-3. By converting text classification tasks into entailment tasks, EFL will make the downstream classification tasks better match with the original pre-trained tasks. In addition, since all downstream tasks are in the unified form of entailment, intermediate training can be carried out with data related to textual entailment, such as CMNLI dataset, so as to obtain a language model more suitable for downstream tasks.

1) Entailment-based Reformulation: The pre-trained MacBERT is deployed as the basic framework for few-shot learning. Therefore, our essential consideration is to transform text classifications into textual entailment. During the reformulation process, the input consists of two distinct sentences. The first sentence is the original text to be classified, and the second sentence is the template for each label. The output is the entailment relationship of these two sentences. Take the IFLYTEK task as an example, the reformulation process is illustrated in Fig. 1.

2) Training and Inference Process: In the few-shot classification of specific tasks, it is of necessity to construct entailment tasks during the training and inference stages. In other words, few-shot data is supposed to be combined with templates during training, and the label of the samples is acquired on the basis of the relationship between texts and templates. During the training phase, let N be the set scale, we utilize the cross-entropy loss to fit the model. Specifically, in the multi-classification task, a text usually needs to be joined with multiple templates.

$$L = -\sum_{i=1}^{N} \left(y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log \left(1 - \hat{y}^{(i)}\right) \right) \quad (1)$$

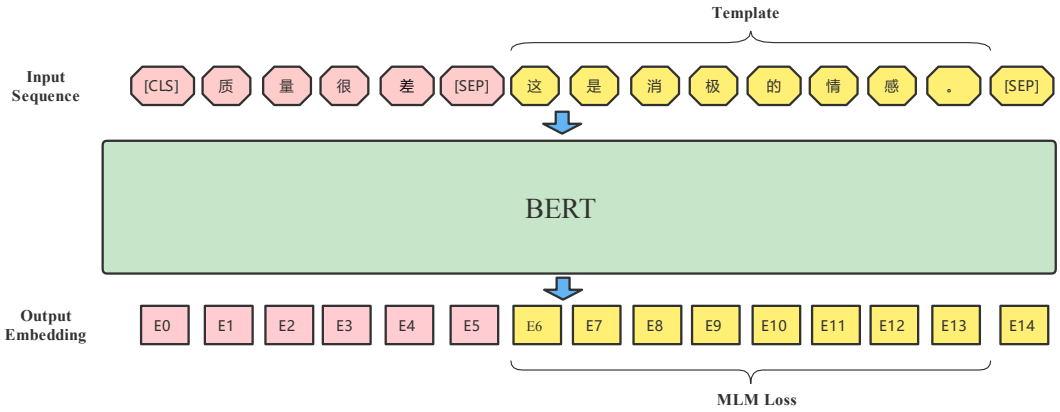

Fig. 2. In the template selection method, each input sentence has a corresponding template. We compute the Masked Language Model (MLM) loss of template token and use the average MLM loss as the template score.

TABLE I
THE BRIEF DESCRIPTIONS OF NINE FEWCLUE TASKS.

Type	Corpus	Train	Dev	Test (Public)	Test	#Labels	Task	Source
Single Sentence	EPRSTMT	32	32	610	753	2	SentimentAnalysis	E-CommrceReview
	CSLDCP	536	536	1780	2999	67	LongTextClassify	AcademicCNKI
	TNEWS	240	240	2010	1500	15	ShortTextClassify	NewsTitle
	IFLYTEK	928	690	1749	2279	119	LongTextClassify	AppDesc
Pair Sentence	OCNLI	32	32	2520	3000	3	NLI	5Genres
	BUSTM	32	32	1772	2000	2	SemanticSimilarity	AIVirtualAssistant
Reading Comprehension	CHID	42	42	2002	2000	7	MultipleChoice,idiom	Novel, EssayNews
	CSL	32	32	2828	3000	2	KeywordRecogntn	AcademicCNKI
	CLUEWSC	32	32	976	290	2	CorefResolution	ChineseFictionBooks

In this equation, $y^{(i)}$ stands for the ground truth of the i-th sample, and $\hat{y}^{(i)}$ correspondingly denotes the model's prediction. During the inference, each sample in the dataset needs to join the templates of all classes to compute the probability of prediction. Consequently, the class with the highest probability is selected as the final result.

C. Intermediate Training

Utilizing the similar dataset as fine-tuning data for downstream tasks, zero-shot downstream task can also achieve preferable results. The entailment method converts all of the text classification tasks into the entailment tasks. Therefore, intermediate training with entailment dataset can bridge the semantic gap between the pre-trained models and the downstream tasks. Specifically, we use CMNLI dataset to conduct intermediate training for MacBERT pre-trained language model. The datasets consists of XNLI [22] (Cross-lingual Natural Language Inference) and MNLI [23] (Multi-genre Natural Language Inference). Each sample contains two sentences and the mutual relationship drawn from three categories, including entailment, neutrality and contradiction, respectively. All downstream tasks will reuse the parameters of the encoder during intermediate training, so as to better integrate the reformulated textual entailment data in the succeeding procedures.

D. Template Selection

There are nine subtasks in this evaluation, and each task has a different form in its domain. In addition, entailment tasks are sensitive to templates where different templates will have a great impact on the performance. Consequently, how to obtain task-related templates is the central issue which requires elaborate design in the entire structure. Our approach is to enable a trainable classifier for deciding which template is more appropriate for the specified task from a set of candidates. Specifically, each sentence corresponds to a template based on its label. Therefore, as shown in Fig. 2, we take the advantage of the MLM task of the pre-trained language model and then calculate the loss of the template. We choose the candidate with the least loss as the final template for the current task. Intuitively, the template loss is similar to the language confusion of the template based leading-sentence. In addition, pre-trained language models play an important role in downstream tasks. This calculating method with pre-trained language models can properly match downstream tasks. Furthermore, candidate set of templates can be defined artificially or generated automatically by network. In other words, seed templates

TABLE II
MAIN RESULT OF DIFFERENT METHODS ON TEST (PUBLIC) DATASETS

METHOD	SCORE	CS	BUSTM	OCNLI	CSLDCP	TNEWS	WSC	IFLTEK	CSL	CHID
Human	82.49	90.00	88.00	90.30	68.00	71.00	98.00	66.00	84.00	87.10
FineTuningB	39.35	61.90	54.10	33.60	25.60	40.50	50.30	22.60	50.50	15.00
PET [2]	57.36	87.20	64.00	43.90	56.90	53.70	59.20	35.10	55.00	61.30
PtuningB [3]	51.81	88.50	65.40	35.00	44.40	48.20	51.00	32.00	50.00	57.60
PtuningGPT [3]	46.44	75.65	54.90	35.75	33.69	45.30	49.00	24.00	53.50	13.70
Zero-shotG [15]	43.36	57.54	50.00	34.40	26.23	36.96	50.31	19.04	50.14	65.63
Zero-shotR [1]	44.61	85.20	50.60	40.30	12.60	25.30	50.00	27.70	52.20	57.60
EFL(base) [6]	53.40	85.60	67.60	67.50	46.70	53.50	54.20	44.00	61.60	28.20
EFL_our (base)	57.93	82.36	75.18	70.35	46.71	68.39	50.07	38.55	59.53	40.20
EFL_our (large)	59.77	84.46	72.92	66.68	48.86	70.07	56.14	41.92	59.95	48.64

are carefully defined manually and other candidates with similar semantic can be produced by SimBERT [24].

E. Task-dependent Adjustment

In this evaluation, the downstream tasks include single-sentence classification, sentence pair classification, and reading comprehension tasks. This is shown in Table I. We therefore make several adjustments aiming for different tasks. In single-sentence tasks, we use the template selection method described above to select from the candidate set of templates, standing for the second sentence in the sentence-pair task. The reading comprehension task contains three tasks, namely keyword recognition, relational reference and idiom filling in the blanks, which can be redefined by using corresponding templates. Significantly, in the CHID task of idiom filling, the first sentence was the sentence in front of the idiom, and the second sentence was the idiom combined with the sentence after the idiom. By this means, the performance of the model was significantly improved. It is reasonable to speculate that textual entailment pays more attention to the concatenation of two sentences.

IV. EXPERIMENTS AND RESULTS

A. Dataset

We evaluate nine Chinese few shot datasets of FewCLUE, which include sentiment analysis, short text classification, long text classification, natural language inference, sentence similarity, Chinese cloze and co-reference resolution. These are shown in Table I. To be fair, each task provides five different training sets and a training set containing all the data in each task. Moreover, we use the CMNLI dataset for intermediate training, and the detail of CMNLI dataset can be found in [25].

B. Experiment Setup

Our code is implemented based on PyTorch framework using HuggingFace toolkit [26]. We use MacBERT large [5] in Chinese as pre-trained language model for all tasks. During the fine-tuning phase, we use AdamW [27] optimizer with a learning rate of 2×10^{-5}, batch size of 8 (unless specified otherwise), max epochs 10, and dropout rate of 0.1. We also use the same settings during the intermediate training stage.

In the training phase, we randomly select $k = 8$ negative samples to form a non-entailment relationship with the labels of the positive samples. During the inference stage, for N samples and M label descriptions, we generate $N \times M$ input data. For each sample, we select the label with the highest probability as the prediction.

C. Result and Analysis

We compare our entailment-based method with various baselines using pre-trained network. Table II shows the main results on the testing (public) datasets of the nine NLP tasks on FewCLUE. We compare with several few-shot methods using manual templates. From this table, we have the following findings. EFL method with automatic template selection outperforms other methods. Moreover, compared with PET, EFL has a better effect on sentence-pair task, but a lower effect on single sentence classification. A more plausible explanation would seem to be that EFL can capture the relationship of sentence granularity, while PET can capture the relationship of token level. Moreover, in EFL method, negative sampling is required in the case of a large number of categories. In the single sentence task with a large number of categories, negative sampling may loses part of the label information. In addition, MacBERT-large outperforms MacBERT-base, shown as EFL_our (large) and EFL_our (base), respectively. This indicates that large pre-trained language model is beneficial to better performance in this evaluation.

D. On Templates Effectiveness

In this section, in order to show the effectiveness of our designed template choosing method, we explore the relationship between the masked language loss and accuracy of a template. Table III shows the evaluation scores of different templates on the masked language model in EPRSTMT task. The loss of the language model (i.e., LM) score is chosen as filtering condition. We observe that as the loss of templates decreases, the accuracy of prediction has steady increase accordingly.

TABLE III
THE LM SCORE AND ACCURACY OF CANDIDATE TEMPLATES.

EPRSTMT Task	LM SCORE	ACC
这是 xxx 的情感	2.91	84.6
这表达了 xxx 的情感	2.65	85.05
评论表达了 xxx 的情感	2.30	87.04

V. CONCLUSION

In this paper, we present our entailment method for FewCLUE shared tasks. Our proposed method achieves substantial improvement over several given baselines by integrating the following techniques: (i) take the pre-trained MacBERT in small scale as our infrastructure; (ii) reformulate different NLP tasks into corresponding entailment templates; (iii) select template with high confidence from candidate set generated by SimBERT; (iv) further adjust the template for CHID task with idiom-filling.

In the future, we will consider in the following aspects to enhance the performance of few-shot learning: (i) leverage self-supervised contrastive learning [28] to optimize the model representation; and (ii) utilize graph convolutional networks [29] with dynamic routing to better merge the semantic feature between sentences and multiple templates.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and Subject II under Grant 2019YFF0303302.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. 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. M. 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," *arXiv e-prints*, p. arXiv:2005.14165, May 2020.

[2] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, 2021.

[3] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[4] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[5] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for chinese natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 657–668, 2020.

[6] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as few-shot learner," *arXiv preprint arXiv:2104.14690*, 2021.

[7] J. Li, M.-T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1106–1115, 2015.

[8] R. Geng, B. Li, Y. Li, J. Sun, and X. Zhu, "Dynamic memory induction networks for few-shot text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1087–1094, 2020.

[9] M. Moradshahi, G. Campagna, S. J. Semnani, S. Xu, and M. S. Lam, "Localizing open-ontology qa semantic parsers in a day using machine translation," *arXiv preprint arXiv:2010.05106*, 2020.

[10] Z. Chen, H. Eavani, W. Chen, Y. Liu, and W. Y. Wang, "Few-shot nlg with pre-trained language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 183–190, 2020.

[11] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.

[12] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

[13] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[17] 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)*, pp. 4171–4186, 2019.

[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[19] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8968–8975, 2020.

[20] 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 preprint arXiv:1907.11692*, 2019.

[21] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5753–5763, 2019.

[22] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018.

[23] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1112–1122, 2018.

[24] J. Su, "Simbert: Integrating retrieval and generation into bert," tech. rep., 2020. https://github.com/ZhuiyiTechnology/simbert.

[25] L. Xu, H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu, and et al., "Clue: A chinese language understanding evaluation benchmark," *In Proceedings of the 28th International Conference on Computational Linguistics, pages 4762–4772.*, 2020.

[26] 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[28] Y. Ouali, C. Hudelot, and M. Tami, "Spatial contrastive learning for few-shot classification," *arXiv preprint arXiv:2012.13831*, 2020.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.