CLOWER: A Pre-trained Language Model with Contrastive Learning over Word and Character Representations

Borun Chen[1]*, Hongyin Tang[2]*, Jingang Wang[2]†, Qifan Wang[3],
Hai-Tao Zheng[1, 4]†, Wei Wu[2] and Liqian Yu[2]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Meituan [3]Meta AI [4]Peng Cheng Laboratory
cbr20@mails.tsinghua.edu.cn zheng.haitao@sz.tsinghua.edu.cn
{tanghongyin,wangjingang02,wuwei30,yuliqian}@meituan.com
wqfcr@fb.com

Abstract

Pre-trained Language Models (PLMs) have achieved remarkable performance gains across numerous downstream tasks in natural language understanding. Various Chinese PLMs have been successively proposed for learning better Chinese language representation. However, most current models use Chinese characters as inputs and are not able to encode semantic information contained in Chinese words. While recent pre-trained models incorporate both words and characters simultaneously, they usually suffer from deficient semantic interactions and fail to capture the semantic relation between words and characters. To address the above issues, we propose a simple yet effective PLM CLOWER, which adopts the Contrastive Learning Over Word and charactER representations. In particular, CLOWER implicitly encodes the coarse-grained information (i.e., words) into the fine-grained representations (i.e., characters) through contrastive learning on multi-grained information. CLOWER is of great value in realistic scenarios since it can be easily incorporated into any existing fine-grained based PLMs without modifying the production pipelines. Extensive experiments conducted on a range of downstream tasks demonstrate the superior performance of CLOWER over several state-of-the-art baselines.

1 Introduction

Pre-trained language models (PLMs) have gained tremendous success in the field of natural language processing recently. As a major milestone of PLMs, BERT (Devlin et al., 2019) and its variants (Yang et al., 2019; Liu et al., 2019; Clark et al., 2019) have demonstrated outstanding performance on a wide variety of natural language understanding (NLU) tasks, such as sentiment analysis and machine reading comprehension tasks. The architecture of Transformer (Vaswani et al., 2017) is typically the foundation for these models, which models the semantic and syntactic relationships between the tokens of the entire input text and learns the contextual representations for each token.

Early Chinese PLMs (Sun et al., 2019) often take the sequences of Chinese characters as the input. These models require relatively small vocabulary and learn the representations of each character from the corpus, which avoids the Out-Of-Vocabulary problem (Li et al., 2019). However, the meanings of a Chinese word can be totally different from the meanings of each Chinese character in the word. For example, the meaning of "小心" (careful) can not be derived from summing the meaning of "小" (small) and "心" (heart). In general, the phenomenon of semantic gaps between coarse-grained language units and fine-grained language units (e.g., words & characters, phrases & words) exists not only in Chinese but also in many other languages.

To alleviate the gap, prior studies improve the pre-trained models in two directions. One direction is to enrich the masking strategies in the masked language model (MLM) objective to mask coarse-grained units, such as the whole word masking (WWM) (Cui et al., 2021) and phrase masking (Sun et al., 2019). These methods encourage the pre-trained model to recover the coarse-grained masks with fine-grained tokens. However, the relation between the coarse-grained and fine-grained representations is modeled in an implicit manner, leading to less effective representations. The other direction is to leverage the multi-grained tokenizations as input. AMBERT (Zhang et al., 2021) encodes both the fine-grained and coarse-grained token sequences and performs the masked language modeling tasks correspondingly, while LICHEE (Guo et al., 2021) merges the multi-grained token embeddings explicitly to integrate the information. Lattice-BERT (Lai et al., 2021) adopts the lattice graph to construct the multi-grained input. Nevertheless, these mod-

*Equal contribution.

†Corresponding authors.

els require additional computational costs (e.g., tokenization, graph construction, multi-grained encoding) and the multi-grained information is only integrated in the embedding layer other than the full encoder, leading to limited usability with low effectiveness.

To fully leverage the semantic information of multi-granularity and preserve the flexibility of single-grained models in the fine-tuning stage, we propose a novel PLM named CLOWER to efficiently model the multi-grained semantic information in pre-training to improve the representation capability. CLOWER adopts the contrastive learning framework to carry out the semantic interaction between multi-grained representations. Specifically, in the pre-training stage, we perform both character and word level tokenizations separately for each input sequence and feed them into the encoder to obtain the contextual representations. Then we conduct the contrastive learning over character and word representations on both token-level and sentence-level. In this way, the word-level semantic information is encoded into the character tokens by bringing their representations closer. Different from AMBERT or LICHEE, in fine-tuning, CLOWER requires no additional computation and can be directly used in any fine-grained PLMs. The merit makes CLOWER production-friendly since it could be deployed easily without modifying the established production pipeline.

We perform comprehensive experiments on different downstream NLU tasks. The experimental results show that CLOWER achieves considerable improvements over several baselines. Ablation studies demonstrate the effectiveness of contrastive learning in our pre-training framework. Our contributions are summarized as follows:

- We present a novel approach that adopts contrastive learning over both word and character representations, which effectively captures their semantic relations.
- With the help of the aforementioned contrastive learning approach, we introduce a Chinese pre-trained language model that connects multi-grained semantic information for learning high quality word and character encoders.
- We conduct an extensive set of experiments on several benchmarks and demonstrate the effectiveness of the proposed model.

2 Related Work

Multi-grained Pre-trained Language Models There have been some efforts to explore the multi-granularity information on the pre-trained language models (Tay et al., 2021; Xue et al., 2022). Cui et al. (2021) adopts the whole word masking strategy to select the masking tokens for pre-training. Similarly, ERNIE 1.0 and 2.0 (Sun et al., 2019, 2020), utilize named entity masking and phrase masking to encode the coarse-grained information into the models, while ERNIE-Gram (Xiao et al., 2021) uses explicit n-gram identities as predicted targets for the enhancement with coarse-grained information. Besides, Joshi et al. (2020) propose the SpanBERT to mask text spans and train the span boundary objective. However, these methods mainly concentrate on fine-grained tokens. The coarse-grained information is only implicitly explored in the masked language modeling by designing the masking strategies and the coarse-grained representations are absent.

Instead of designing the coarse-grained masking strategy on the fine-grained token sequences, several methods focus on improving the pre-training models with multi-grained tokenization. AMBERT (Zhang et al., 2021) utilizes two encoders with shared parameters to process the fine-grained and coarse-grained token sequences. LICHEE (Guo et al., 2021) proposes to merge the multi-grained tokenizations at the embedding level to incorporate multi-grained information of input. Recently, Lai et al. (2021) propose the Lattice-BERT, which introduces the lattice graph constructed from characters and words to explicitly explore the word representations in a multi-granularity way. However, these models are either computationally intensive or lack the integration of multi-grained information in the deep encoder layers, resulting in the limitations of usability and effectiveness.

Contrastive Learning in Pre-trained Language Models As contrastive learning become popular in visual representation learning (Chen et al., 2020; He et al., 2020; Khosla et al., 2020) and NLP tasks (Wu et al., 2020; Meng et al., 2021; Wang et al., 2021), there have been several works exploring the effects of contrastive learning for pre-trained language models. CERT (Fang et al., 2020) adopts the framework of MOCO (He et al., 2020) and performs the sentence augmentations by back-translation. Zhang et al. (2020) propose

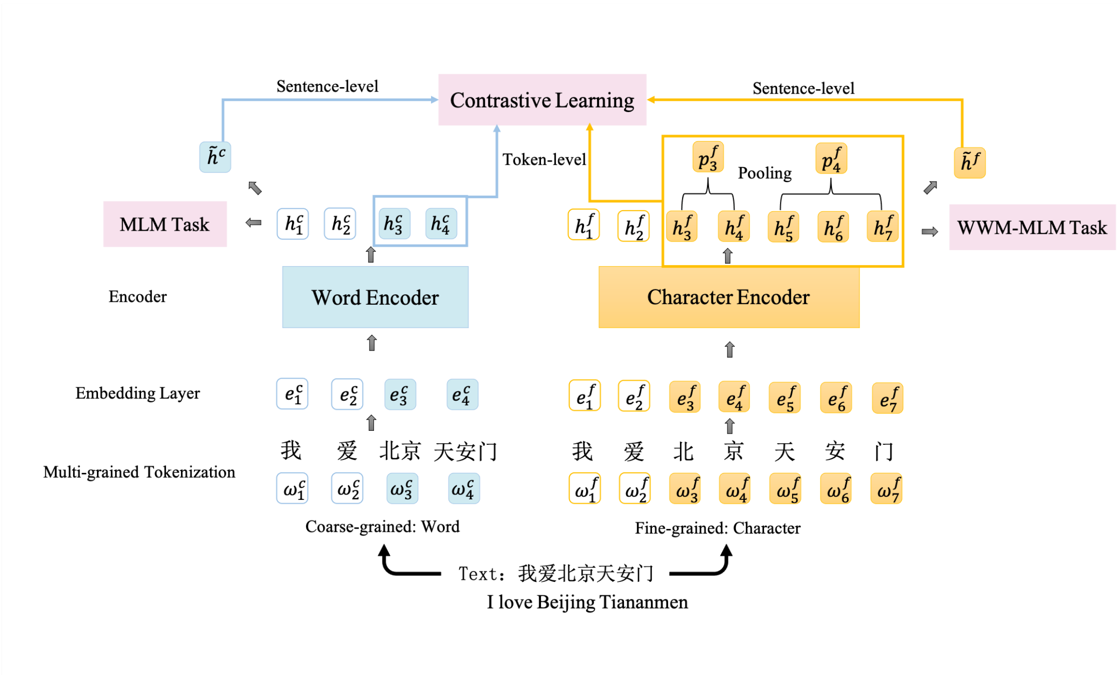
Sentence-level
Contrastive Learning
Sentence-level
Token-level
$\tilde{h}^c$
$\tilde{h}^f$
Pooling
p_3^f
p_4^f
MLM Task
h_1^c h_2^c h_3^c h_4^c
h_1^f h_2^f h_3^f h_4^f h_5^f h_6^f h_7^f
WWM-MLM Task
Encoder
Word Encoder
Character Encoder
Embedding Layer
e_1^c e_2^c e_3^c e_4^c
e_1^f e_2^f e_3^f e_4^f e_5^f e_6^f e_7^f
Multi-grained Tokenization
我 爱 北京 天安门
我 爱 北 京 天 安 门
ω_1^c ω_2^c ω_3^c ω_4^c
ω_1^f ω_2^f ω_3^f ω_4^f ω_5^f ω_6^f ω_7^f
Coarse-grained: Word
Fine-grained: Character
Text：我爱北京天安门
I love Beijing Tiananmen

Figure 1: An overview of CLOWER. Fine-grained and coarse-grained representations are encoded by two encoders.Token-level and sentence-level contrastive learning are conducted together with the MLM and WWM-MLM tasks.

the unsupervised sentence embedding model IS-BERT, increasing the mutual information between the global representations and the local context when training the model. ConSERT (Yan et al., 2021) applies a variety of data augmentation techniques to generate various input views at the embedding level for contrastive learning. Similarly, SimCSE (Gao et al., 2021) utilizes dropout acts as data augmentation in sentence-level. The above methods conduct the contrastive learning to fine-tune the pre-trained language encoder. As for pre-training the language model, DeCLUTR (Giorgi et al., 2021) and CLEAR (Wu et al., 2020) utilize the architecture of SimCLR (Chen et al., 2020) to combine the contrastive learning objective with the masked language modeling. Compared to the above models, our CLOWER conducts the contrastive learning over word and character representations in pre-training and we have the flexibility to fine-tune it in specific downstream tasks.

3 Methodology

In this section, we present CLOWER, the pre-trained language model based on contrastive learning over word and character representations. We first present the overall model architecture of CLOWER, and then we introduce its details in the pre-training stage. Finally, we discuss the strategy of fine-tuning the model efficiently using only the fine-grained input.

3.1 Model Architecture

Figure 1 illustrates an overview of CLOWER pre-training, where the contrastive learning framework is leveraged across multiple granularity information to enhance the representation ability of the model.

CLOWER takes the text sequences as input and performs multi-grained tokenization on the input to obtain the fine-grained and coarse-grained token sequences. It should be noted that the fine-grained and coarse-grained tokens share the same vocabulary, which aims at improving the alignment of embedding spaces between multi-grained tokens. In this paper, we treat the characters and words as fine-grained and coarse-grained tokens respectively. Formally, given the input text sequence s, we denote the fine-grained and coarse-grained token sequences by $\boldsymbol{s_f} = \{\omega_1^f, \cdots, \omega_i^f, \cdots, \omega_m^f\}$ and $\boldsymbol{s_c} = \{\omega_1^c, \cdots, \omega_j^c, \cdots, \omega_n^c\}$, where m and n are the lengths of two tokenized sequences.

Consistent with the shared vocabulary, CLOWER adopts the shared embedding layers to map the tokens ω_i^f and ω_j^c to the embedding representations $\boldsymbol{e}_i^f$ and $\boldsymbol{e}_j^c \in \mathcal{R}^d$ respectively, where d is the dimension of the embedding. The fine-grained and coarse-grained embeddings are then passed to the two encoders to obtain the contextualized representations respectively. The encoders utilized in CLOWER can be any pre-trained language model and two encoders of fine-grained and coarse-grained have independent parameters. In this paper, we adopt Chinese BERT(Devlin et al., 2019) as the encoders.

Token-level and sentence-level contrastive learning are conducted over the fine-grained and coarse-grained contextualized representations from the above encoders, together with the traditional MLM task and WWM-MLM task.

3.2 Pre-Training

Masked Language Model In the pre-training stage, CLOWER adopts the MLM task at multi-grained levels. Specifically, we denote the masked fine-grained and coarse-grained token sequences as $\tilde{\boldsymbol{s_f}}$ and $\tilde{\boldsymbol{s_c}}$. The masked fine-grained and coarse-grained tokens are represented as $\boldsymbol{s_f^m}$ and $\boldsymbol{s_c^m}$ respectively. Then, the object of our MLM task at multi-grained levels is to optimize the following loss function:

$$\begin{aligned}\mathcal{L}_{mlm} = &- \sum_{\omega_f^m \in \boldsymbol{s_f^m}} \log P_\theta(\omega_f^m | \tilde{\boldsymbol{s_f}}) \\ &- \sum_{\omega_c^m \in \boldsymbol{s_c^m}} \log P_\theta(\omega_c^m | \tilde{\boldsymbol{s_c}}),\end{aligned} \tag{1}$$

where θ denotes the model parameters.

We adopt the WWM strategy (Cui et al., 2021) as the strategy of fine-grained token sequences and the conventional masking strategy introduced by BERT(Devlin et al., 2019) for the coarse-grained token sequences.

Contrastive Learning To fully learn from the multi-grained information, we conduct contrastive learning between the fine-grained representations and their corresponding coarse-grained representations at both token-level and sentence-level. Formally, for each pair of multi-grained token sequences $\boldsymbol{s_f}$ and $\boldsymbol{s_c}$, we randomly choose some of the coarse-grained tokens $\boldsymbol{s_a} = \{\omega_1^c, \cdots, \omega_i^c, \cdots, \omega_k^c\} \subset \boldsymbol{s_c}$ as anchors, where k is the maximum number of anchors for each sequence. The strategy of selecting the anchors will be detailed in Section 4.1.

Given the anchor ω_i^c, which is composed of the fine-grained tokens $\omega_{b(i)}^f, \cdots, \omega_{e(i)}^f$ where $b(i)$ denotes the begin index of the anchor ω_i^c and $e(i)$ denotes the end index of the anchor ω_i^c, we can obtain its coarse-grained representation $\boldsymbol{h_i^c}$ generated by the word encoder and its fine-grained representation $\boldsymbol{p_i^f} = \mathrm{AVG}\left(\boldsymbol{h_{b(i)}^f} \cdots \boldsymbol{h_{e(i)}^f}\right)$ generated by the character encoder, where $\mathrm{AVG}(\cdot)$ means the average pooling.

Our motivation is to close the gap between the fine-grained representations and their corresponding coarse-grained representations while enlarge the gap between unrelated representations. Following the contrastive learning paradigm, it can be implemented by constructing positive and negative instance pairs. For the coarse-grained representation $\boldsymbol{h_i^c}$, we mark the fine-grained representations of the same anchor $\boldsymbol{p_i^f}$ as its positive instance and the fine-grained representations of the other anchors in the same mini-batch $\boldsymbol{p_j^f}$ as the negative instances. We further introduce the "`[CLS]`" embeddings of each sentence as the sentence-level representations, namely $\tilde{\boldsymbol{h}}^{\boldsymbol{c}}$ for the coarse-grained representation and $\tilde{\boldsymbol{h}}^{\boldsymbol{f}}$ for the fine-grained representation. Similar to the token-level, we treat the multi-grained representations $(\tilde{\boldsymbol{h}}^{\boldsymbol{c}}, \tilde{\boldsymbol{h}}^{\boldsymbol{f}})$ of the same sentence as the positive instance pair and the multi-grained representations of different sentences in a mini-batch as the negative instance pairs.

Following the contrastive objective in Chen et al. (2020), we utilize the normalized temperature-scaled cross-entropy loss (NT-Xent) for both the token-level and sentence-level representations. We optimize the symmetric cross-entropy loss in the pre-training. Specifically, the objective of contrastive learning in multi-grained token-level representations $\mathcal{L}_{tcl}$ is as follows:

$$\mathcal{L}_{tcl}^c = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathrm{sim}(\boldsymbol{h_i^c}, \boldsymbol{p_i^f})/\tau}}{\sum_j e^{\mathrm{sim}(\boldsymbol{h_i^c}, \boldsymbol{p_j^f})/\tau}}, \tag{2}$$

$$\mathcal{L}_{tcl}^f = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathrm{sim}(\boldsymbol{p_i^f}, \boldsymbol{h_i^c})/\tau}}{\sum_j e^{\mathrm{sim}(\boldsymbol{p_i^f}, \boldsymbol{h_j^c})/\tau}}, \tag{3}$$

$$\mathcal{L}_{tcl} = \frac{1}{2}(\mathcal{L}_{tcl}^c + \mathcal{L}_{tcl}^f), \tag{4}$$

where N indicates the number of in-batch anchors, $\mathrm{sim}(\cdot)$ denotes the similarity function as we use the cosine similarity, and τ is a temperature hyper-parameter. Similarly, we define the symmetric sentence-level contrastive loss $\mathcal{L}_{scl}$ with a mini-batch size M as:

$$\mathcal{L}^c_{scl} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\mathrm{sim}(\tilde{\boldsymbol{h}}^c_i,\tilde{\boldsymbol{h}}^f_i)/\tau}}{\sum_j e^{\mathrm{sim}(\tilde{\boldsymbol{h}}^c_i,\tilde{\boldsymbol{h}}^f_j)/\tau}}, \tag{5}$$

$$\mathcal{L}^f_{scl} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\mathrm{sim}(\tilde{\boldsymbol{h}}^f_i,\tilde{\boldsymbol{h}}^c_i)/\tau}}{\sum_j e^{\mathrm{sim}(\tilde{\boldsymbol{h}}^f_i,\tilde{\boldsymbol{h}}^c_j)/\tau}}, \tag{6}$$

$$\mathcal{L}_{scl} = \frac{1}{2}(\mathcal{L}^c_{scl} + \mathcal{L}^f_{scl}), \tag{7}$$

Therefore, the final object of contrastive learning $\mathcal{L}_{con}$ is the sum of $\mathcal{L}_{tcl}$ and $\mathcal{L}_{scl}$.

Sentence Order Prediction Apart from the MLM and contrastive learning tasks, we adopt the sentence order prediction (SOP) task (Lan et al., 2019) to effectively model the relationship of sentence pairs and denote the training loss as $\mathcal{L}_{sop}$. Hence, the overall training loss of CLOWER in pre-training is the combination of three tasks:

$$\mathcal{L} = \mathcal{L}_{mlm} + \lambda\mathcal{L}_{sop} + \mu\mathcal{L}_{con} \tag{8}$$

where λ and μ are the hyper-parameters of balancing three task objectives.

3.3 Fine-Tuning

Note that the usage of the character encoder of CLOWER is virtually the same as the fine-grained Chinese PLMs like BERT, thus we can directly substitute them with our character encoder without any modification while having the benefit of the coarse-grained information encoded in the fine-grained representations.

For the sentence-level downstream tasks, like single sentence classification and sentence pair classification, we conduct classification base on the contextualized sentence-level representation $\tilde{\boldsymbol{h}}^f$. As for the token-level tasks, such as question answering, fine-grained contextualized representations of each token are extracted and used for predictions.

4 Experiments

We conducted comprehensive experiments on various Chinese NLU tasks to examine the effectiveness of CLOWER. In this section, we first introduce the details of pre-training and fine-tuning, including the datasets and experimental settings. Then, we present the overall results on different tasks and conduct an in-depth analysis. Ablation studies are also conducted to evaluate the impact of multi-level contrastive learning in our model.

Dataset	MSL	BS	LR	Epoch
ChnSentiCorp	256	32	3e-5	10
THUCNews	512	16	3e-5	10
Tnews	128	32	3e-5	10
Bq Corpus	128	64	3e-5	10
Lcqmc	128	64	3e-5	10
Ocnli	128	32	3e-5	10
Xnli	128	64	3e-5	10
CMRC2018	512	8	3e-5	5
DRCD	512	8	3e-5	5

Table 1: Hyper-parameters settings for 9 fine-tuning tasks. MSL: Maximum Sequence Length; BS: Batch Size; LR: Learning Rate.

4.1 Pre-training Datasets

To the best of our knowledge, WuDaoCorpora (Yuan et al., 2021) is the largest open-source Chinese corpora for pre-training. We utilize the base version of WuDaoCorpora[1], consisting of about 200GB training data and 72 billion Chinese characters in total. Following the settings of most Chinese PLMs, we consider the characters as the fine-grained tokens. We utilize Jieba[2] to perform the word segmentation on texts and the segmented words are treated as the coarse-grained tokens. There are $5,466$ Chinese characters and $40,014$ words in our vocabulary, together with other tokens like digits and some basic English tokens. We conduct the fine-grained and coarse-grained tokenizations based on the vocabulary and the words will be split to characters if they are not in the vocabulary. For contrastive learning, we select up to k anchors whose lengths are between 2 and 4 from each sequence. Note that for semantic integrity, the words that have been masked either on coarse-grained sequences or their fine-grained characters will not be selected as anchors.

4.2 Fine-tuning tasks

To thoroughly examine the effectiveness of CLOWER, an extensive set of experiments are performed on various Chinese NLU tasks, including three single sentence classification (SSC) tasks, four sentence pair classification (SPC) tasks and two machine reading comprehension(MRC) tasks. Specifically, three SSC tasks are ChnSentiCorp (Tan and Zhang, 2008), THUCNews (Li

[1] https://resource.wudaoai.cn/home

[2] https://github.com/fxsjy/jieba

Model	Tnews	THUCNews		ChnSentiCorp		Average
	Dev	Dev	Test	Dev	Test	
BERT-wwm	66.59	98.16	97.41	94.97	95.55	90.53
BERT-wwm-sop	66.42	98.31	97.49	94.87	95.32	90.48
MM-BERT	66.39	98.18	97.53	94.92	94.80	90.36
MM-BERT-sop	66.27	98.16	97.45	94.62	95.65	90.43
MacBERT	67.07	98.29	97.34	95.16	95.18	90.61
CLOWER	**67.15**	**98.39**	**97.74**	**95.18**	**95.84**	**90.86**

Table 2: Experimental results on single sentence classification tasks.

Model	Ocnli	Lcqmc		Xnli		Bq		Average
	Dev	Dev	Test	Dev	Test	Dev	Test	
BERT-wwm	74.87	89.37	86.93	79.50	78.89	85.39	84.37	82.76
BERT-wwm-sop	75.73	89.75	87.30	79.74	78.45	85.73	84.81	83.07
MM-BERT	75.34	89.55	87.08	79.56	78.66	85.37	84.51	82.87
MM-BERT-sop	75.44	89.85	87.18	79.42	78.62	86.00	84.84	83.05
MacBERT	75.90	89.58	86.59	**80.54**	79.10	85.71	84.95	83.20
CLOWER	**76.25**	**89.92**	**88.10**	80.14	**79.19**	**86.01**	**85.26**	**83.55**

Table 3: Experimental results on sentence pair classification tasks.

Model	CMRC2018		DRCD			
	Dev		Dev		Test	
	EM	F1	EM	F1	EM	F1
BERT-wwm	68.15	86.32	88.20	93.63	87.13	92.55
BERT-wwm-sop	67.47	85.86	87.54	93.15	87.33	92.61
MM-BERT	68.61	86.42	88.45	93.65	87.36	92.85
MM-BERT-sop	67.57	86.18	88.30	93.50	87.18	92.76
MacBERT	68.31	86.38	**88.92**	**94.08**	**88.04**	**93.22**
CLOWER	**68.73**	**86.52**	88.27	93.44	87.68	92.94

Table 4: Experimental results on MRC tasks.

and Sun, 2007) and Tnews (Xu et al., 2020); four SPC tasks include Bq Corpus (Chen et al., 2018), Lcqmc (Liu et al., 2018), Ocnli (Hu et al., 2020) and Xnli (Conneau et al., 2018); two MRC tasks are CMRC2018 (Cui et al., 2019) and DRCD (Shao et al., 2018).

4.3 Experiment Settings

4.3.1 Pre-training

In pre-training of CLOWER, we initiate both the character and word encoder with the Chinese BERT-base released by Google[3] in order to reduce the total convergence time. Given a word not in the vocabulary, we initiate its embedding with the average pooling of the embeddings of the characters that make up the word. For MLM tasks, as with the BERT, 15% of the tokens are masked randomly. For token-level contrastive learning, the maximum number of anchors for each sequence is set as 20 and the temperature is 0.05. The hyper-parameters λ and μ in Equation 8 are both set as 1. We set the maximum sequence length to 512 throughout the pre-training and adopt the ADAM (Kingma and Ba, 2014) optimizer with weight decay whose learning rate is $2e-5$. We train the model with a batch size of 960 (24×40) for $300,000$ steps. The pre-training is carried out on 40 NVIDIA V100 GPUs. To improve efficiency, mixed precision training (Micikevicius et al., 2017) is adopted.

4.3.2 Fine-tuning

To make a fair comparison, we adopt the same hyper-parameters for each fine-tuning task among different models. The detailed parameter settings are shown in Table 1. During fine-tuning, we encode each example using the fine-grained encoder (i.e., character encoder). For three SSP tasks and four SPC tasks, the "[CLS]" embedding is used to represent the sentence and the classification accuracies are reported. For two MRC tasks, the token embeddings are used to extract the answer

[3] https://github.com/google-research/bert

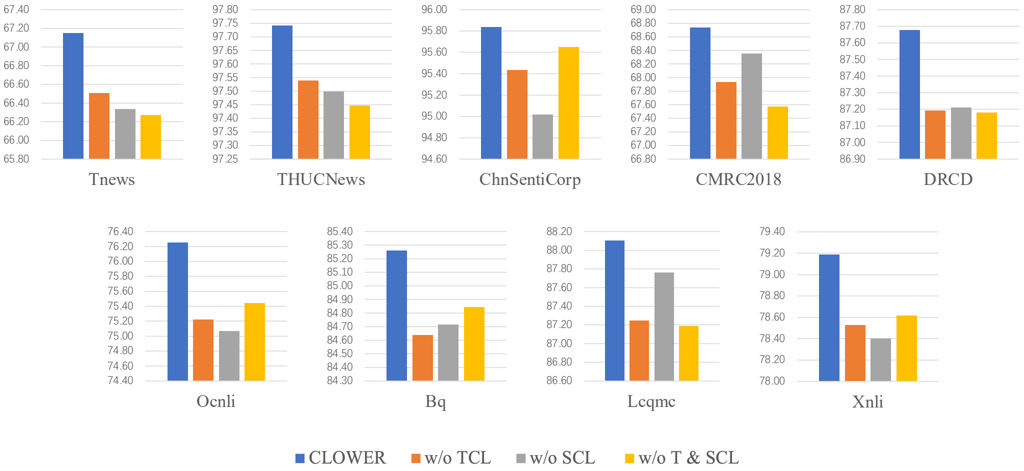

Figure 2: Ablation Results. We report the accuracy for sentence classification tasks and EM for MRC tasks.

span from the sentence, and both exact match (EM) and F1-score are reported. For each task, we perform the experiments five runs with different random seeds and report the average performance to promise the results convincing. We report the results both on the development sets and test sets, except for Tnews, Ocnli and CMRC2018, whose test sets are not publicly available. Since each article in the Tnews dataset consists of a title and several keywords, we associate the titles with keywords as the input sequences to perform the classification task. We fine-tune all the models for each downstream task on one NVIDIA V100 GPU.

4.4 Main Results

Since most of the existing Chinese PLMs are trained with different corpus and setups, it is hard to conduct ideally fair comparisons. Therefore, we select the most representative Chinese PLM (i.e., Chinese BERT-base) as the baseline and achieve several pre-training models with different settings on the same corpus. More concretely, we implement the following four baselines: (1) **BERT-wwm** (Cui et al., 2021), a BERT-base model trained with the additional fine-grained WWM task, (2) **BERT-wwm-sop**, a BERT-base model trained with the addtional WWM and SOP tasks, (3) **MM-BERT**, a BERT-base model trained with the multi-grained MLM tasks, including a fine-grained WWM task and a coarse-grained MLM task, (4) **MM-BERT-sop**, Multi-grained MLM on a BERT-base model trained with the multi-grained MLM task and the SOP task. In addition, we also include MacBERT (Cui et al., 2021) as a strong baseline, which is one of the state-of-the-art Chinese PLMs in literature. MacBERT utilize the WWM as well as N-gram masking strategies together during pre-training. In terms of the masking implementation, MacBERT masks the word with a similar word rather than the [Mask] placeholder to improve the performance further. The experimental results of MacBERT are achieved with the released model[4] under the identical settings with the other baselines among all downstream tasks.

For three SSC tasks, the results are shown in Table 2. From the results, we can find that our CLOWER yields consistent improvements over all baselines on all three tasks (both on the development and test sets), which proves the effectiveness and advantages of our model. CLOWER outperforms the 4 baselines pre-trained with the identical data while different settings, which demonstrates the advantages of our multi-level contrastive learning approach. In addition, CLOWER outperforms MacBERT by 0.25 points on average and achieves a new state-of-the art on Chinese SSC tasks.

As for the SPC tasks, fair comparisons are performed and the results are reported in Table 3. From the results, we also observe that CLOWER also achieves consistent improvements over baselines on the four tasks. In comparison to the four baselines pre-trained with the identical data, CLOWER outperforms the best one (i.e., MM-BERT-sop) by 0.33 points on average. In comparison to MacBERT, CLOWER achieves a performance gain of 0.33 points on average. CLOWER performs best on all datasets except Xnli

[4] https://github.com/ymcui/MacBERT

Model	Tnews	THUC	Chn	Ocnli	Lcqmc	Xnli	Bq	Average
CLOWER	**67.15**	**97.74**	**95.84**	**76.25**	**88.10**	**79.19**	**85.26**	**84.22**
w/o tcl	66.51	97.54	95.43	75.22	87.24	78.53	84.64	83.59
w/o scl	66.34	97.50	95.02	75.06	87.76	78.40	84.71	83.54
w/o tcl & scl	66.27	97.45	95.65	75.44	87.18	78.62	84.84	83.64

Table 5: Ablation results on SSC and SPC tasks. For Tnews and Ocnli, the results are on development sets and others are on test sets.

Model	CMRC2018		DRCD			
	Dev		Dev		Test	
	EM	F1	EM	F1	EM	F1
CLOWER	**68.73**	**86.52**	88.27	93.44	**87.68**	**92.94**
w/o tcl	67.93	86.25	88.01	93.34	87.19	92.69
w/o scl	68.35	86.18	88.05	93.38	87.21	92.68
w/o tcl & scl	67.57	86.18	**88.30**	**93.50**	87.18	92.76

Table 6: Ablation results on machine reading comprehension tasks.

Dev set.

The above SSC and SPC tasks are all sequence-level tasks, to further examine the effectiveness of our model, we also perform comparisons on MRC tasks which are document-level span-extraction tasks. The resuls are depicted in Tabel 4. Specifically, for CMRC2018, CLOWER outperforms MacBERT by 0.40 points and 0.14 points on EM and F1 score respectively. As the EM score is a stricter measurement of machine reading comprehension, the improvements over MacBERT are considerable. While for DRCD, we find that the performance of CLOWER is not as competitive as the baselines. We conjecture that the reason may be the original dataset of DRCD is in Traditional Chinese whereas our pre-training corpus is in Simplified Chinese. Although we convert the data to Simplified Chinese literally, there are some differences such as syntax and semantics yet, the performances of the pre-trained models may be affected inevitably.

4.5 Ablation Study

To further investigative the effects of contrastive learning over word and characters in CLOWER, we conduct ablation study on the model variants without token-level or sentence-level contrastive learning tasks. Figure 2 shows the ablation results on sentences classification and machine reading comprehension tasks. The detailed ablation results on 9 downstream NLU tasks are reported in Table 5 and 6 respectively.

When removing the token-level contrastive learning task (w/o TCL) or sentence-level (w/o SCL) from CLOWER, there is a distinct drop in the performance on sentence classification tasks (i.e., SSC and SPC). Furthermore, when removing all the contrastive learning tasks, i.e., actually the MM-BERT-sop model, the performance is almost same as the w/o TCL or w/o SCL models. It indicates that only if the token-level contrastive learning task works jointly with the sentence-level contrastive learning task in pre-training, there will be a positive impact on the sentence-level downstream tasks. We conclude that it is vital for the model to encode the coarse-grained semantic information into the fine-grained sequences at token-level and sentence-level consistently when we apply it on sentence-level downstream tasks.

As for the MRC tasks, the EM score on CMRC2018 drops a lot when removing the token-level contrastive learning task, which demonstrates the effectiveness of token-level task on the extractive MRC task. While removing the sentence-level contrastive learning task, the EM metric of the model drops less than that without the token-level. Also, the performance of model without both contrastive learning tasks perform worst among these models on CMRC2018. The results on DRCD reveal the similar trend.

5 Discussions

5.1 Flexibility

Compared to other Chinese PLMs which utilize fine-grained and coarse-grained information, one notable advantage of CLOWER is the high flexibility of deployment. In real-world scenarios, fine-grained PLMs are more popular due to its flexibility on processing inputs/outputs and low computational costs. Please recap that CLOWER could be deemed as a fine-grained character encoder during inference, which is enhanced with the coarse-

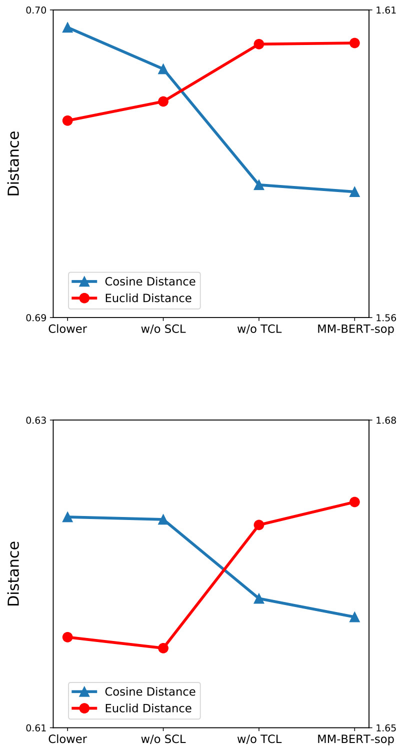

Figure 3: Similarity Analysis of Embeddings. Top: the words with length 2; Bottom: the words with length longer than 2.

grained word encoder during pre-training. In particular, if a production system already deploys a fine-grained Chinese PLM (e.g., the vanilla BERT), the fine-grained encoder of CLOWER can be adopted as a substitute without extra tailor cost seamlessly. CLOWER also provides the coarse-grained encoder (i.e., word encoder) for scenarios where Chinese word sequences are designed as input. The coarse-grained encoder of CLOWER has also been updated and acquired the knowledge from the large corpus during pre-training. We can make flexible choices according to downstream scenarios and conditions when utilizing CLOWER.

5.2 Multi-grained Information Modeling

Through the pre-training, CLOWER implements the multi-grained semantic information modeling by performing the contrastive learning over words to characters and thus implicitly encodes the coarse-grained semantic information into fine-grained tokens and vice versa. To evaluate the character/word representations learned by the interactions, we adopt the measures of cosine similarity and Euclid distance as proxies. We calculate the cosine similarity and Euclid distance between the embeddings of words and the mean embeddings of the characters that compose the words. In our corpus, 72.1% words are composed of two characters. So we conduct the similarity analysis by split the words into two groups, two-character words and the other words composed at least three characters. The similarities produced by four models are shown in Figure 3. We can clearly see that the token-level contrastive learning task play an important role of bringing the word and character embeddings closer, as the similarity of CLOWER and w/o scl are higher than the other two models and so is the Euclid distance. According to the intuitive results, we corroborate that our model indeed achieves our motivation to encode the coarse-grained information into fine-grained tokens.

6 Conclusion

To fully leverage the information of characters and words in Chinese PLMs, we propose a novel PLM CLOWER based on contrastive learning over word and character representations jointly. Through the token-level and sentence-level contrastive learning in the pre-training stage, the model encodes the coarse-grained semantic information into fine-grained tokens. We can not only enhance the model with coarse-grained semantics but also enjoy the flexibility of fine-grained inputs/outputs. The flexibility promises that our model could be deployed conveniently in real scenarios, where certain PLMs like BERT have been established. Comprehensive experiments on a variety of downstream natural language understanding tasks demonstrate the competitive performance of CLOWER. We also conduct a ablation study to evaluate the multi-grained contrastive learning mechanism in CLOWER.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 6201101015), Beijing Academy of Artificial Intelligence (BAAI), Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012640), the Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JCYJ20190813165003837), National Key R&D Program of China (No. 2021ZD0112905) and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (Grant No. HW2021008).

References

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4946–4951.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *Proceedings of International Conference on Learning Representations*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019*, pages 4171–4186.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Weidong Guo, Mingjun Zhao, Lusheng Zhang, Di Niu, Jinwen Luo, Zhenhua Liu, Zhenyang Li, and Jianbo Tang. 2021. Lichee: Improving language model pre-training with multi-grained tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1383–1392.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021. Lattice-bert: Leveraging multi-granularity representations in chinese pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1716–1731.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jingyang Li and Maosong Sun. 2007. Scalable term selection for text categorization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 774–782.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, 34(4):2622–2629.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. Cline: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2332–2342.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1702–1715.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Xinsong Zhang, Pengshuai Li, and Hang Li. 2021. Ambert: A pre-trained language model with multi-grained tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 421–435.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.