Structural-Semantic Constraints for Enhanced Chinese Language Modeling

Zhongyi Deng*

[1] the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
dengzhongyi2024@gmail.com

Abstract

Most Chinese pre-training studies follow word-level strategies of English pre-training. These studies do not consider the exposure imbalance of Chinese characters, resulting in imbalanced performance on different downstream tasks. To address above issues, we propose a structure semantic constraints of Chinese Characters for enhanced language modeling. Since a Chinese character is composed of several structure units (components, strokes and composite types), the structure semantic constraints explore Chinese structure-level semantics via deconstructing and reconstructing between Chinese characters and structure units. In contrast to MLM, which learns the global semantics of characters, this task is designed to focus on their local semantic representations. Different level representation tasks help model performs well in fine-grained Chinese representation, alleviating imbalanced word-level pre-training by balanced structure-level pre-training. In terms of experiments, we implement structure semantic constraints on the BERT architecture. The proposed model achieves overall performance improvement on multiple Chinese NLP tasks. Experimental results and analysis demonstrate the effectiveness of proposed scheme in Chinese pre-training.

Introduction

Chinese pre-training aims to establish general representation for each character in Chinese context (Clark et al. 2019; Qiu et al. 2020). The issue of imbalanced Chinese characters is emerging with larger training corpus and broader domain coverage. For instance, prepositions outnumber nouns in an open domain and professional nouns outnumber common nouns in a specific domain. Research (Li et al. 2020a) indicates word imbalance led to anisotropy, which is a conical distribution between word embeddings of language model. High-frequency words near the cone's origin, while low-frequency words are dispersed at greater distances. Due to the polarization, it's unfair to measure the similarity between words of varying frequencies by the same standard. Furthermore, the sparse distribution of low-frequency word result in incomplete semantic spaces. Therefore, words imbalance prevalent in corpora poses challenges for the pre-training of language models.

Strategies of mitigating word imbalance for English pre-training generally include sampling and weighting. Sampling strategy directly affects word frequency. The subsampling (Mikolov et al. 2013) discard frequently occurring but not significant words with certain probability. Furthermore, statistics of word occurrences in corpora (Pennington, Socher, and Manning 2014) serve as essential data for representing words. As a particular sampling, whole word masking (WWM) is employed in the masked language modeling (MLM) (Devlin et al. 2019), which is primary pre-training task in language models. It treats word pieces as independent tokens of vocabulary. Words containing masked token have a masking probability. Low-frequency words pre-training is improved because it may contain high-frequency tokens. Weighting strategy emphasizes the enhancement of low frequency words (Yu et al. 2022). Low-frequency words are marked as positive samples challenging to classify. Positive samples are assigned higher weights than negative samples to boost impact in training. It is often employed in classification tasks to deal with imbalanced data (Li et al. 2020b; Menon et al. 2020).

Howere, the above methods exhibit limitations in Chinese pre-training by not taking the imbalance of Chinese characters into account. The national standard Chinese character table currently lists 8105 standard characters, of which 3500 are commonly used. It means Chinese has more abundant semantic patterns than English. Sampling may disrupt the natural distribution of Chinese language. As training data expands, the impact intensifies. Weighting focus on fine-tuning rather than pre-training. MLM is a multi-class classification with categories equivalent to tokens in vocabulary. It's not suitable for weighting given that categories in MLM greatly exceeds that of general classification.

We propose the optimization scheme of Chinese character structural-semantic constraints to obtain a robust representation for imbalanced Chinese characters. The scheme consists of MLM task and Chinese structural recombination (CSR) task, as illustrated in Fig. 1. The MLM task extracts global feature of Chinese character to represent word-level semantics. The CSR task extracts local feature of Chinese character and refines semantics into structure-level. Dual tasks present

*This work was funded by the Guangdong Science and Technology Innovation Strategy Special Fund(Student Science and Technology Innovation Cultivation Program) grant under number pdjh2025bk026.

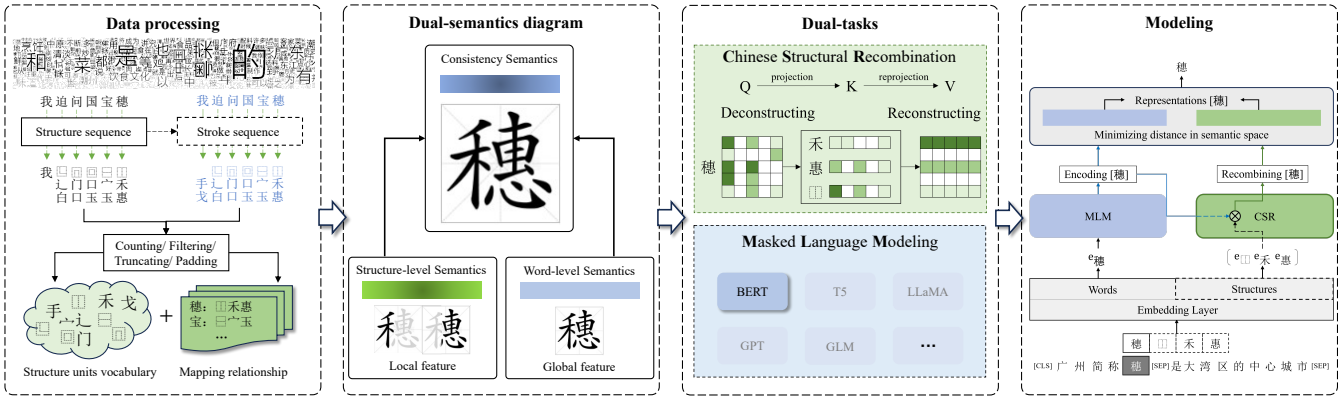

Figure 1: Chinese character structural-semantic constraints optimization scheme. The workflow includes four parts: data processing, dual-semantics diagram, dual-tasks, and modeling.

each Chinese character with two levels of semantics during pre-training. Maintaining semantic consistency between structure-level and word-level aims to enhance the representation of low-frequency Chinese characters. More precisely, low-frequency Chinese characters of semantics benefit from high-frequency Chinese characters when the former share common structures with the latter. It's the core to mitigate the issue of imbalanced Chinese characters. We select BERT as the foundational model for implementing the scheme. Our model shows superior performance than baseline in both multi-domain reading comprehension and fine-grained entity recognition. According to experimental analysis, the proposed scheme aids language models in attaining a balanced semantic recognition ability beyond the limits of word imbalance.

Methodology

Chinese dual-task consistency includes two parts: dual-task execution and representation consistency. Dual task refers to the MLM task of learning word representation and the CSR task of learning structural representation. Representation consistency refers to maintaining semantic consistency at both word-level and structure-level during pre-training.

Masked Language Modeling Task

As crucial pre-training task in language models, MLM task constructs the word-level semantics of Chinese characters and phrases (Che, Li, and Liu 2010).The MLM task of proposed scheme maintains with the version of BERT but introduces new function for dual-task execution. It's not only an extractor of global features but also the source of local features for Chinese characters. The word-level semantics obtained from MLM, the smallest semantic granularity in current Chinese pre-training (Hewitt and Manning 2019; Diao et al. 2020; Zhang, Li, and Li 2021), will be decomposed into structure-level semantics through the CSR task. This means two levels of Chinese semantic coexist in the pre-training and need to be consistent.

Chinese Structure Recombination Task

The CSR task aims to develop a robust Chinese character representation that mitigates the impact of different Chinese character exposures on training.

The task execution process can be divided into deconstructing and reconstructing. The deconstructing stage converts Chinese characters into structural sequences and extracts its local feature. The reconstructing stage recombines a structural sequences into a complete Chinese character, which aims to transfer the semantics of Chinese character representation from local to global.

Deconstructing Deconstructing aims to convert a Chinese Character into its structural sequence for structural hard-coding. A structural sequence includes components[1], strokes[2] and composite types[3]. Inputs of the deconstructing stage include a Chinese characters list α of BERT vocab, a component database ϕ containing relationships from Chinese character to its components and a stroke database ζ containing relationships from Chinese character to its strokes. Outputs of the deconstructing stage include a structure vocabulary ν and a mapping table τ. The structure vocabulary contains all structure units of Chinese characters in Unicode. The mapping table contains relationships from Chinese characters to structural sequence, which is used for pre-training data generation. In order to unify pre-training data format, the structural sequence of each Chinese character will be padding to a fixed length ι. The followings describe the settings for key parameters in this paper. The output size of ν is 8521. The output size of τ is 7321. The fixed length ι is set to 7.

Reconstructing Reconstructing aims to balance training benefits between Chinese characters of varying frequencies. Fig. 3 shows an example of reconstructing. Assuming that Chinese character [称] totally occurs 1000 times in corpora, which is 100 times more than [穗]. Fig. 3(a)

[1]https://www.chise.org/ids/
[2]https://github.com/kfcd/chaizi
[3]https://www.unicode.org/charts/PDF/U2FF0.pdf

and 3(b) respectively present sampling and weighting methods, narrowing the disparity proportion from 100 to 1. In Fig. 3(c), component [禾] and composite type [ICD-1] coexist in [称] and [穗]. The improved exposure of [穗] at structure-level narrow the disparity proportion from 100 to 1.49. Intuitively, the recombining can achieve similar effects as sampling and weighting while remaining Chinese characters natural distribution. Common structures of imbalanced Chinese characters serve as information carriers. When low-frequency Chinese characters share common structures with high-frequency Chinese characters, the former benefit from the latter through reconstructing structures into a complete Chinese character during pre-training. Reconstructing stage connects structural semantics to local feature of Chinese characters.

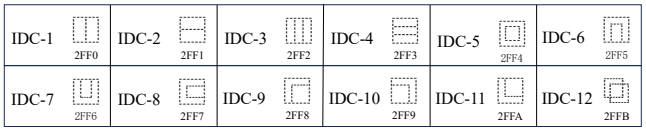

Figure 2: Composite Types of Chinese Characters. There are 12 identifiers in Unicode representing the combination types of Chinese characters

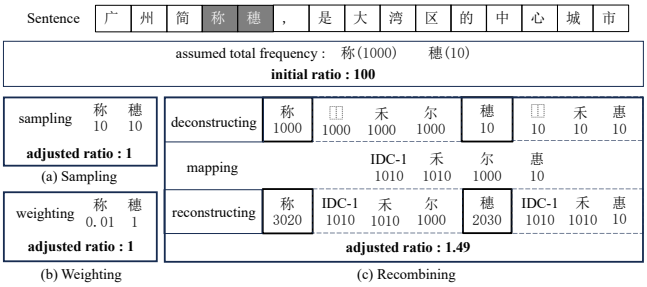

Figure 3: Example of Chinese Structure Recombination. The values in the example are set to illustrate the effect of three methods on the frequency of Chinese characters.

Whole Structures Masking (WSM)

The WSM strategy aims to explore relationships between global and local semantics within a Chinese character. A Chinese character and its structural sequence will be masked simultaneously in the WSM strategy. Chinese characters and their structures are embedded into a unified semantic space interact with each other under attention mechanisms. The WSM is formally similar to the original WWM, which benefit from independent structure representation. Structures are close to word pieces and Chinese character is close to word. Table 1 shows general input-output forms of WSM.

Maintaining consistency of dual-representation is an important step to apply the frequency robustness derived from CSR tasks to final Chinese character representations. For Chinese characters, MLM task extracts global feature to learns word-level semantics, while CSR task extracts local feature to refine semantics into structure-level. The semantics at both levels merge into a complete Chinese character representation during pre-training.

Scheme Implementation

As an implementation of proposed scheme on BERT, the construction of proposed model includes optimizing the architecture to achieve a dual-task training, and designing a loss function to consistently update parameters (Pan et al. 2023). Fig. 4 shows the whole architecture of proposed model.

Embedding Layer To realize structural hard coding, we assign a unique ID to each structure. This process requires a language model to extend its embedding layer. According to the structure vocabulary created by deconstructing stage described in section , embedding layer adds 8521 structure embedding. Structure embedding dimension is set to 768, aligning with BERT settings. Consequently, the expansion of embedding layer equivalent to adding a matrix with a size of 8521×768.

Structure Activation Union In order to obtain complete representation of Chinese character from its structural sequence, we construct the structure activation union. The Chinese character embedding interacts with each of its structure embeddings in the union. Results serve as the weight of each element in the structural sequence. Details are shown in Fig. 4(b) and Fig. 4(c).

The computation of structural weights, illustrated in the formula 1, is essentially the scaled dot-product attention of the Transformer (Vaswani et al. 2017).

$$\Omega = softmax\left(\frac{a\left(E_{structures}, E_{encoding}\right)}{\sqrt{d}}\right). \quad (1)$$

It's the essence of structures recombination that weighted summation of semantic representations at structure-level. The embedding obtained from reconstructing the structural sequence is represented mathematically as formula 2.

$$\begin{aligned} E_{reconst} &= E_{structures} \times \Omega \\ &= \sum_{i=1}^{H} e_i \omega_i. \end{aligned} \quad (2)$$

- $E_{structures}$ is a structure embedding sequence of Chinese character and show as $(e_1, e_2, \cdots, e_H)$. e_i represents each structure embedding of structure sequence and H is the fix length of structure sequence.
- $E_{encoding}$ is MLM encoder output of Chinese character.
- $a(\cdot)$ is an interaction between Chinese character and its structure sequence, including element-wise subtraction and element-wise product.
- d is a parameter for scaled dot-product. It's set to the embedding size in this paper.
- $softmax(\cdot)$ is a normalization on the weight sequence.
- Ω is a weight sequence and show as $(\omega_1, \omega_2, \cdots, \omega_H)$. ω_i is the weight of e_i, reflecting the importance of each structure in Chinese character.

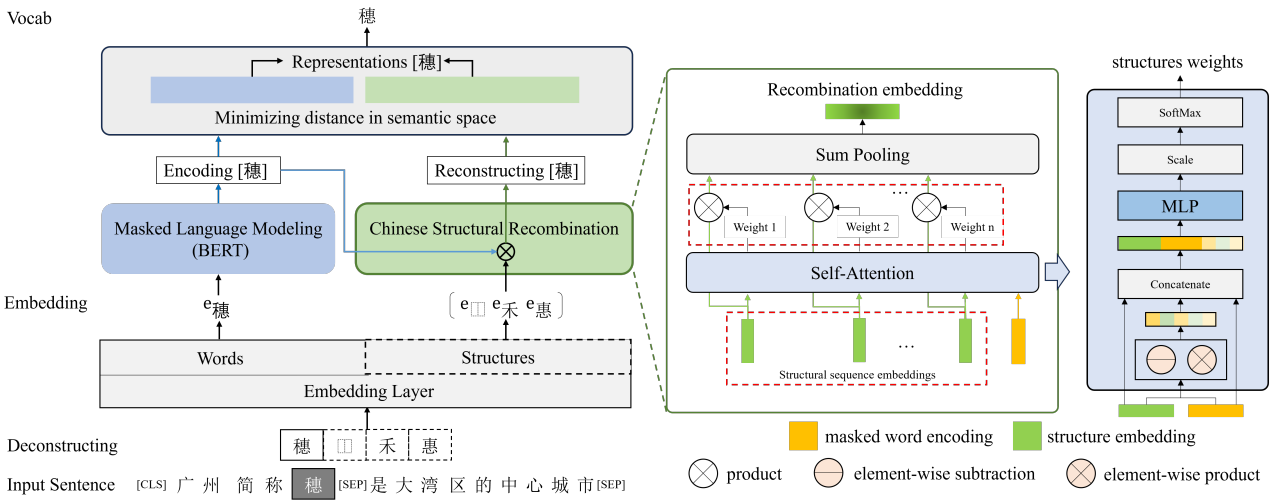

Figure 4: Model Architecture of Chinese character structural-semantic constraints scheme.

Loss Function The loss function of proposed model added constraints on the loss of MLM.

$$\mathcal{L}_{proposed} = \mathcal{L}_{MLM} + \mathcal{L}_{CSR}, \tag{3}$$

where $\mathcal{L}_{CSR}$ is considered as constraints of $\mathcal{L}_{MLM}$.

The standard cross-entropy loss is used to optimize the pre-training task, including MLM and CSR.

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M} y_i \log p_i, \tag{4}$$

where M is the number of masked words in an input, y is the masked word label, p is the activation function output. The activation function of MLM task is shown in formula 5.

$$P_{mlm} = softmax\left(E_{token}, E_{encoding}\right), \tag{5}$$

where E_{token} is the embedding of masked Chinese character and $E_{encoding}$ is the encoder output of MLM. The number of categories in softmax function equals to the size of vocabulary.

Considering the significant memory consumption bring by softmax function, we adopt sigmoid function as the activation function of CSR task.

$$P_{csr} = sigmoid\left(E_{token}, E_{reconst}\right), \tag{6}$$

where $E_{reconst}$ is structure recombination embedding. Sigmoid function minimizes the semantic distance between $E_{reconst}$ and E_{token} through binary classification.

Experiments

Pre-training Settings

We trained a base-level model from scratch, consisting of 12 Transformer layers, with input dimension of 768 and 12 heads per layer.

Masking Strategy On the basis of WWM and CWS, we use the WSM strategy described in section for Chinese characters and their structures masking. 15% of the input tokens are selected for possible replacement. Of the selected tokens, 80% are masked, 10% are left unchanged, and 10% are replaced by a randomly selected vocabulary token. In order to implement dynamic masking in data preprocessing, we use four random number seeds to generate four training samples with different mask patterns for each statement (Liu et al. 2019).

Data The pre-training data contains 5.8B Chinese characters in total, which is collected from multiple sources such as CLUE, Wikipedia, and Sina News. It consists of encyclopedia, community question answering, news articles, etc.

Hyper-parameters The training scheme includes a maximum length of 512, a batch size of 3.2k, a maximum learning rate 1e-4, warm up steps of 20K, and train steps of 500K, which is suggested in previous research (Cui et al. 2020; Sun et al. 2021). For the batch size less than 1024, we adopt the original ADAM (Kingma and Ba 2014) with weight decay optimizer in BERT for optimization and use LAMB optimizer (You et al. 2019) in larger batch.

Fine-tuning Settings

The phenomenon of imbalanced Chinese characters reflects the focus of knowledge in different domain. To assess the effect of structural-semantic constraints in alleviating the imbalance of Chinese characters, we conduct experiments on Chinese tasks that contain multi-domain knowledge. The major tasks include machine reading comprehension (MRC), named entity recognition (NER) and text classification (TC). The baseline of experiments includes multiple BERT-style models, which are shown in Table 2.

The MRC task represents multi-domain knowledge as multiple independent Chinese datasets: simplified Chi-

Table 1: Comparisons of Different Masking Strategies. CWS:Chinese Word Segmentation, WWM:Whole Word Masking, WSM:Whole Structures Masking.

	Input	Output
	广州简称穗，是大湾区的中心城市	
CWS	广州 简称 穗 是 大湾区 的 中心 城市	
WWM	广州 [M][M] 穗 是 大湾区 [M] 中心 城市	Pos_2 : [简] Pos_3 : [称] Pos_9 : [的]
WSM	广州 [M][M] 穗 是 大湾区 [M] 中心 城市	Pos_2 : [简, [IDC-2, 竹, 间]] Pos_3 : [称, [IDC-1, 禾, 尔]] Pos_9 : [的, [IDC-1, 白, 勺]]

Table 2: Training Details of Chinese Pre-Trained Language Models. T: Token, NM: N-gram Masking, CM:Char Masking.

Options	BERT	RoBERTa	MacBERT	PERT	ChineseBERT	Propossed model
Word	0.4B	5.4B	5.4B	5.4B	5B	5.8B
Batch size	/	384	512	416	3.2K	3.2K
Training Steps	/	1M	1M	2M	500K	500K
Task	MLM	MLM	MAC	PLM	MLM	MLM
Masking	T	WWM	WWM/NM	WWM/NM	WWM/CM	WWM/WSM
Initial Checkpoint	Random	BERT	BERT	Random	Random	Random

nese (Cui et al. 2019), traditional Chinese (Shao et al. 2018), legal (Duan et al. 2019), military, traditional Chinese medicine, wikipedia (Rajpurkar et al. 2016) and internet Q&A (He et al. 2018). For datasets without publicly available test sets, we divide the training set into two independent subsets in a ratio of 8:2. The experiment ensures that the new test set does not participate in fine-tuning and acts as a validation set. The TC task mainly verifies the semantic understanding ability of language models at the sentence level. We conduct experiments on the following text classification tasks. The experiment of short text classification conducts on the TNEWS dataset (Xu et al. 2020), which extracted news from 15 categories, including tourism, education, finance, etc. The experiment of long text classification conducts on the IFLYTEK dataset (Xu et al. 2020), which contains annotation data related to app application descriptions, with a total of 119 categories. The experiment of natural language inference conducts on the OCNLI dataset (Hu et al. 2020). It's the first non-translated native Chinese natural language inference dataset.

Experimental results

Results on TC

Models of bert architecture usually takes [CLS] placeholder output as the main eigenvector in the text classification task. The [CLS] output corresponds to a well-designed sentence representation pre-training task, such as the NSP task of bert and the SOP task of MacBERT. Table 3 shows the experimental results of text classification. It can be found that models with specific sentence representation tasks performs better. Referring to the technical summary in (Liu et al. 2019), our model does not specifically perform the sentence representation task to learn semantics of sentence embeddings. Therefore, the performance of proposed model is broadly in line with the baseline, which meet expectations.

Table 3: Performances of Different Models on TC.

	TNEWS	IFLYTEK	OCNLI	Avg
BERT	56.30	**59.65**	74.60	63.52
RoBERTa	57.40	59.63	76.50	64.51
MacBERT	57.40	59.28	77.00	64.56
PERT	56.70	48.23	74.85	59.93
ChineseBERT	57.49	52.41	76.89	62.26
Ours	**57.81**	58.93	**77.12**	**64.62**

Results on MRC

In order to evaluate the Chinese comprehension ability of proposed model, we conducted experiments on MRC datasets from distinct domains. Results of multi-domain MRC are shown in Table 4 and Table 5. Our model outperforms the baseline model on the majority of datasets, with only a slight drop in the F1 metric on a few datasets. The multi-domain MRC tasks contain knowledge biases caused by word frequency imbalance. The general improvement on baseline originates from mitigating of imbalance. The enhancement of our model is more pronounced in EM metric than in F1 metric, indicating its advantage in detecting exact answer spans. It's an indicator that structure-level semantics can reinforce the model's capacity to identify Chinese semantic boundaries.

Table 4: EM scores of Different Models on Multi-Domain MRC.

	CMRC	DRCD	CJRC	MMRC	QGDTCM	ChineseSquad	DuReader	Avg
BERT	67.10	83.10	57.11	53.30	41.16	60.54	47.12	58.49
RoBERTa	67.40	86.60	58.71	55.20	42.82	52.89	45.95	58.51
MacBERT	68.50	89.40	59.29	55.65	43.82	55.72	46.23	59.80
PERT	68.50	88.90	58.93	55.61	41.94	58.89	47.34	60.02
ChineseBERT	70.43	86.60	56.67	54.71	42.69	62.91	46.46	60.07
Ours	**71.14**	**89.93**	**59.69**	**56.01**	**43.91**	**66.15**	**47.35**	**62.03**

Table 5: F1 scores of Different Models on Multi-Domain MRC.

	CMRC	DRCD	CJRC	MMRC	QGDTCM	ChineseSquad	DuReader	Avg
BERT	85.70	89.90	73.68	80.40	76.23	64.56	60.38	75.84
RoBERTa	87.20	92.50	75.21	81.93	77.83	58.79	62.30	76.54
MacBERT	87.90	94.30	75.91	81.91	**78.42**	62.01	**62.61**	77.58
PERT	87.20	93.60	75.31	81.91	77.01	64.16	61.63	77.26
ChineseBERT	88.20	92.36	73.09	81.24	77.33	68.50	61.73	77.49
Ours	**88.22**	**94.68**	**76.01**	**82.47**	78.14	**69.40**	61.74	**78.67**

Analysis

Ablation on Masking Strategies

To obtain the optimal dual-task combination strategy, we conduct experiment with different masking strategies. The candidate tasks for learning word-level semantics include MLM, MAC and PLM, adopting RoBERTa, MacBERT and PERT as basic model respectively. Moreover, the model setups of base and large respectively consist of 12/24 Transformer layers, with input dimension of 768/1024 and 12/16 heads per layer. We continue to train 100K steps with a batch size of 512 on basic models under the combination strategy.

The comparison experiments between the training model and the basic model is conducted on simplified Chinese and traditional Chinese MRC tasks. Table 6 shows the results of performance comparison. It's an acceptable semantic boost between structures and phrases, but weak boost between structures and orders. There is no significant improvement in the combination between structures and synonyms. Although the improvement of incremental pre-training is slight, the experimental results show a tendency to have a fusion effect of Chinese language knowledge. The positive and negative tendencies can be used as a basis for the selection of pre-trained combination strategies.

Table 6: Performances on Different Strategies.

	CMRC		DRCD	
	EM	F1	EM	F1
base				
MLM	66.73	86.81	87.67	93.14
MLM+CSR	**68.40**	**87.06**	**88.31**	**93.51**
MAC	67.14	87.68	88.64	94.07
MAC+CSR	**67.88**	87.14	88.67	93.90
PLM	68.88	87.64	88.06	93.10
PLM+CSR	68.58	87.34	88.04	**93.41**
large				
MLM	70.11	88.11	89.77	94.18
MLM+CSR	**70.24**	**88.29**	**90.20**	**95.00**
MAC	69.00	89.15	90.57	95.29
MAC+CSR	**69.79**	88.60	90.07	94.88
PLM	71.34	89.53	90.71	95.12
PLM+CSR	**71.46**	89.14	89.83	94.49

Statistics on Frequency

In order to compare Chinese characters frequency at different levels, we sample and count Chinese characters exposure in Wikipedia dump recombining articles, templates, media, file descriptions (as of May 1, 2024). After cleaning the raw text, about 700M simplified and traditional Chinese characters are obtained for analysis and only those in the BERT vocabulary are considered for statistics. Sampling lengths of 128/256/512/1024/2048 are independently used and the results are averaged.

Fig. 5 shows the improvement for imbalanced Chinese characters exposure. The word-level frequency means Chinese character frequency of occurrence. The structure-level frequency means total frequency of structures in a Chinese character. Compared to the substantial variations at word-level, the frequency of most Chinese characters is close to the same order of magnitude at structure-level, which can ensure balanced learning of structural semantics. This serves as a foundation for addressing imbalances in word-level pre-training through data-driven perspective.

Conclusion

In this paper, we proposed the dual-task consistency optimization scheme for Chinese pre-training. The scheme designed a balanced structure-level pre-training to improve imbalanced word-level pre-training and enhanced Chinese language model's generalization in multiple semantic represen-

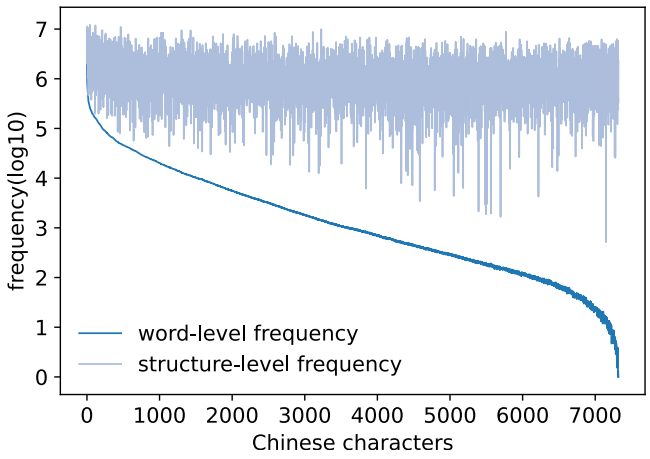

Figure 5: Statistics on Chinese characters frequency. The horizontal axis represents the index of Chinese characters and the vertical axis represents the frequency.

tations. We implemented the scheme on the BERT architecture and verified on various NLP task. The experimental results showed that our model has superior performance over baseline models, demonstrating that introducing structural-semantic constraints into Chinese pre-training is an effective method to achieve optimal performance in Chinese semantic recognition. Future work involves conducting extensive experiments on a variety of Chinese tasks and exploring more levels of Chinese semantics on more large-scale model parameters.

References

Che, W.; Li, Z.; and Liu, T. 2010. LTP: A Chinese Language Technology Platform. In *23rd International Conference on Computational Linguistics*, 13. Citeseer.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; and Hu, G. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 657–668.

Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; and Hu, G. 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)*, 5883–5889.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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, volume 1 (long and short papers)*, 4171–4186.

Diao, S.; Bai, J.; Song, Y.; Zhang, T.; and Wang, Y. 2020. ZEN: Pre-Training Chinese Text Encoder Enhanced by N-gram Representations. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*.

Duan, X.; Wang, B.; Wang, Z.; Ma, W.; Cui, Y.; Wu, D.; Wang, S.; Liu, T.; Huo, T.; Hu, Z.; et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, 439–451. Springer.

He, W.; Liu, K.; Liu, J.; Lyu, Y.; Zhao, S.; Xiao, X.; Liu, Y.; Wang, Y.; Wu, H.; She, Q.; et al. 2018. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, 37–46.

Hewitt, J.; and Manning, C. D. 2019. A structural probe for finding syntax in word representations. 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)*, 4129–4138.

Hu, H.; Richardson, K.; Xu, L.; Li, L.; Kübler, S.; and Moss, L. S. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3512–3526.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; and Li, L. 2020a. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9119–9130.

Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; and Li, J. 2020b. Dice Loss for Data-imbalanced NLP Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 465–476.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Pan, M.-Z.; Liao, X.-L.; Li, Z.; Deng, Y.-W.; Chen, Y.; and Bian, G.-B. 2023. Semi-Supervised Medical Image Segmentation Guided by Bi-Directional Constrained Dual-Task Consistency. *Bioengineering*, 10(2): 225.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Shao, C. C.; Liu, T.; Lai, Y.; Tseng, Y.; and Tsai, S. 2018. DRCD: A Chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; and Li, J. 2021. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. 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)*, 2065–2075.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; et al. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4762–4772.

You, Y.; Li, J.; Hseu, J.; Song, X.; Demmel, J.; and Hsieh, C.-J. 2019. Reducing BERT pre-training time from 3 days to 76 minutes. *arXiv preprint arXiv:1904.00962*, 12.

Yu, W.; Zhu, C.; Fang, Y.; Yu, D.; Wang, S.; Xu, Y.; Zeng, M.; and Jiang, M. 2022. Dict-BERT: Enhancing Language Model Pre-training with Dictionary. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1907–1918.

Zhang, X.; Li, P.; and Li, H. 2021. AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 421–435.