A Multi-tasking and Multi-stage Chinese Minority Pre-Trained Language Model*

Bin Li[1]**, Yixuan Weng[2]**, Bin Sun[1], Shutao Li[1]***

[1] College of Electrical and Information Engineering, Hunan University
{libincn, shutao_li, sunbin611}@hnu.edu.cn
[2] National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy Sciences, Beijing
wengsyx@gmail.com

Abstract. The existing multi-language generative model is considered as an important part of the multilingual field, which has received extensive attention in recent years. However, due to the scarcity of Chinese Minority corpus, developing a well-designed translation system is still a great challenge. To leverage the current corpus better, we design a pre-training method for the low resource domain, which can help the model better understand low resource text. The motivation is that the Chinese Minority languages have the characteristics of similarity and the adjacency of cultural transmission, and different multilingual translation pairs can provide the pre-trained model with sufficient semantic information. Therefore, we propose the Chinese Minority Pre-Trained (CMPT) language model with multi-tasking and multi-stage strategies to further leverage these low-resource corpora. Specifically, four pre-training tasks and two-stage strategies are adopted during pre-training for better results. Experiments show that our model outperforms the baseline method in Chinese Minority language translation. At the same time, we released the first generative pre-trained language model for the Chinese Minority to support the development of relevant research[3].

Keywords: Multi-task · Multi-stage · Chinese Minority · Generative pre-trained language model.

1 Introduction

With the emergence of the pre-training language model, great progress has been made in the field of natural language processing [1]. The self-supervised method has achieved remarkable success in many tasks [2], which is designed to reconstruct the input text by using the AutoEncoder [3,4]. In the previous works, the generative sequence-to-sequence (seq2seq) model can be applied to a wide range

* Supported by the National Key R&D Program of China (2018YFB1305200), the National Natural Science Fund of China (62171183).
** These authors contribute this work equally.
*** Corresponding author.
[3] All the experimental codes and the pre-trained language model are open-sourced on the website https://github.com/WENGSYX/CMPT.

of downstream tasks. Firstly, the text is destroyed by noise manipulation, and then the original text is reconstructed with the language model. The downstream task performance can be effectively improved by further fine-tuning [5].

For some low-resource languages, the self-supervised method is difficult to adapt to the downstream task directly because the corpus is relatively small. At the same time, the model will have a better understanding of high resource languages but ignore the learning between similar low resource languages [6].

The Chinese Minority languages have similarity and adjacency in cultural transmission [7]. This is because ethnic integration in East Asia has been going on continuously since ancient times [8]. The frequent iterations of the regime have promoted the social development of the Han nationality and the cultural exchanges among all ethnic groups. Chinese and minority languages have long been in contact, influenced, and integrated with each other [9].

Therefore, we propose the Chinese Minority Pre-Trained (CMPT) language model with multi-tasking and multi-stage strategies. The CMPT model improves the ability of cross-language understanding through pre-training, which is designed with denoising and contrastive learning between texts in different low-resourced languages. Specifically, we refer to the settings of the BART [10] to randomly mask the text and require the model to be restored. In order to improve the understanding ability of the encoder model, we refer to the setting of CPT [11] and add a single-layer masked language model (MLM) [12] decoder to the encoder output layer for joint training of generation and understanding. Due to the small number of minority languages, we learn close to the dense vector of language pairs with the same semantic meaning based on the cross-lingual contrastive learning between text pairs. This can pull the language pairs with the same semantic meaning to similar positions in the vector space and can help the model to better realize the migration and understanding of low-resource languages.

In order to further study the feasibility of a large-scale pre-training language model, we use DeepNorm [13] to implement a 256 layer into the CMPT model, which has 128 layers of the encoder and 128 layers of decoder, respectively. We believe that the model with depth can better extract the understanding ability between languages.

In conclusion, we have the following three contributions to this work:

1. We have proposed a CMPT model for Chinese Minority languages. Through the use of denoising tasks and contrastive learning, it has the ability to understand and generate meanwhile.

2. We have trained a 256-layer CMPT model and open-sourced it online, which greatly promotes the research of Chinese Minority language translation.

3. The CMPT model has achieved better performance in the shared task of CCMT2022[4] compared with the baseline method.

[4] `http://mteval.cipsc.org.cn/`

2 Related Work

2.1 Pre-trained Language Model

In recent years, increasing pre-training methods have been used in the field of natural language processing [14,15]. These methods can learn common knowledge from a large number of unlabeled texts. GPT uses a one-way decoder to perform generation tasks. Bert [4] introduces a mask language modeling (MLM) task, which can significantly improve the performance of the pre-trained language model through pre-training to learn the interaction between context tokens with longer training time and larger model parameter size. In order to realize the conversion of seq2seq, the BART [10] and the T5 [16] use the denoising task and mask restoration task for pre-training respectively. The BART has achieved SOTA in generation tasks like translation, while T5 has SOTA performance in understanding and summarization.

2.2 Multilingual Model

Large-scale multilingual pre-training can significantly improve the performance of cross-lingual migration tasks. The XLM-R [17] uses more than 2TB of multilingual data sets and is pre-trained in 100 languages [18]. This model can model multiple languages without sacrificing language performance through denoising pre-training. The mBART [19] has significantly improved in a variety of machine translation tasks. At the same time, it can also migrate to language pairs without bidirectional corresponding text.

The M2M [20] is the translation model that not only focuses on English but realizes the first real multi- to multi-lingual translation model by collecting supervised data of thousands of language pairs. The M2M model can achieve an improvement of more than 10 BLEU [21] score when focusing on non-English translation. In order to align the context representations between different languages, the VECO [22] adds a cross-attention module [23] to explicitly construct the interdependencies between languages. It can effectively avoid the generation of predicting masked words only conditioned on the context in its own language. In order to improve the efficiency of translation. Switch-GLAT [24] proposes a non-autoregressive translation method, which improves the translation performance by shortening the spatial distance between the replaced words and the original target language.

2.3 Chinese Minority Languages

The existing work for the Chinese Minority languages is relatively small, as the multilingual model is difficult to model indigenous languages and minority languages. Nevertheless, the CINO [25] has developed the first pre-trained language model for Chinese Minority languages, covering Chinese, Cantonese, and other six low-resourced languages. It is believed that for languages with scarce resources, multilingual pre-training can perform better than language pre-training.

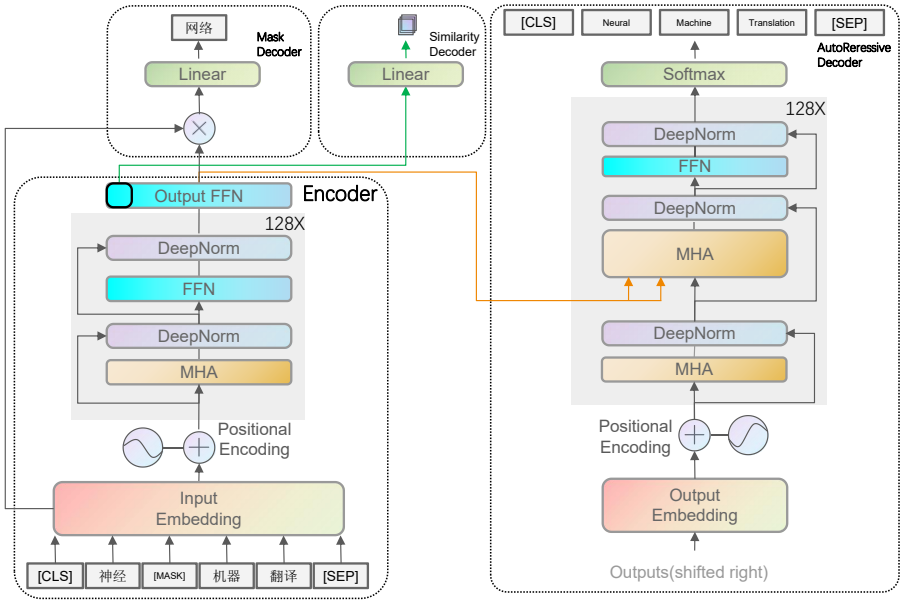

Fig. 1. Overview of the proposed Chinese Minority Pre-Trained (CMPT) language model.

Also, the cost of data annotation for low-resource languages is reduced significantly. The CINO has the same model architecture as XLM-R [26,17]. In order to adapt to minority languages, additional vocabulary expansion and vocabulary pruning have been carried out and the word embedding matrix is reduced to lower the size of the model. However, the CINO is based on language understanding, which does not have the ability to perform generation downstream tasks. Different from the previous work, our work focuses on the generative pre-trained language model to further advance the development of minority language translation.

3 Main Methods

3.1 Model Architecture

Recently, many works have combined language understanding and generation abilities [27,28,29] into the pre-trained language model. Inspired by the work [11], we incorporate both understanding and generation tasks into our Chinese Minority Pre-Trained (CMPT) language model. In order to better adapt the model to the downstream tasks of minority languages and make full use of the language pre-training tasks in low resource scenarios, we design different decoders into the CMPT with multi-task and multi-stage settings. As shown in

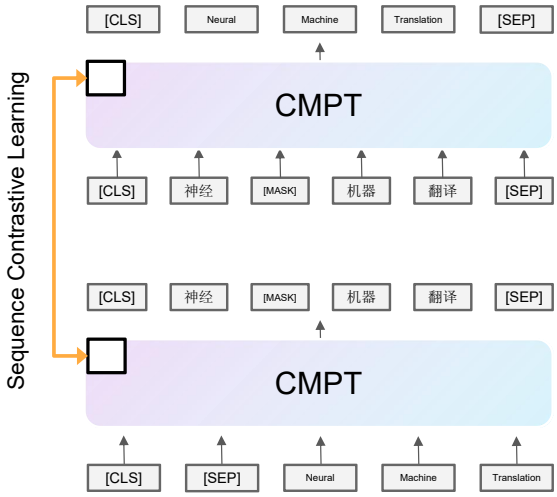

Fig. 2. The illustration of the cross-lingual contrastive learning, where the different translation pairs (Flipped Input-output) are required to learn the same semantics in the vector space between the *CLS*.

Figure 1, we have modified the Transformer [30] structure, which is mainly divided into four parts.

1. Bidirectional Encoder. We use a bidirectional self-attention encoder [30], which can leverage the semantic representation and text meaning.

2. Mask Decoder. We adopt the single linear layer [31] to the output of the Bidirectional Encoder, where the input embedding is multiplied by the output. It is known as the MLM head to support the training of MLM pre-training task.

3. AutoReressive Decoder. We use the original transformer decoder structure, following the settings of BART [10,19] to design our model. The cross-attention is adopted to realize auto-regressive decoding.

4. Similarity Decoder. We input the *CLS* vector of the encoder into the single-layer similarity decoder to extract the semantic vector.

In the downstream tasks of the Chinese Minority language, the encoder can freely choose the decoder accordingly. For example, the comprehension task [32,33] uses the mask decoder, the generation [34] and translation task [35] uses the AutoRegressive Decoder, and the retrieval task [36,37] uses the similarity decoder. This pre-trained language model can meet more diversified requirements, and make efficient use of the parameters with suitable decoders.

3.2 Multi-tasking Multi-stage Pre-training

We designed four pre-training tasks with two-stage strategies to help the model learn language knowledge to make better use of the low resource corpus, which is shown as follows.

1. Mask Language Model (MLM) Task. We randomly mask the input text with a probability of 15%. We require the Mask Decoder to predict the masked token separately so that we can learn deeper semantic information. The input embedding is utilized to multiple with the output from the encoder for this MLM task.

2. Denoising Auto-Encoding (DAE) Task. For the AutoRegressive decoder, we use **two-stage** training to make more efficient use of the provided multilingual translation pairs. We first use DAE for pre-training for the first pre-training stage along with MLM. Specifically, we use the noise function to randomly destroy the input text, and then use the mask to fill in the corresponding position. The motivation is that the AutoRegressive Decoder can learn to reconstruct the original noise input.

3. Text Translation (TT) Task. In the second stage, we will change the DAE task to supervised training, while the MLM task keeps its original setting. Specifically, we input the multilingual translation pairs into the pre-trained language model as the same mini-batch [38]. The model is designed to generate the text of the other language while in this TT task. As for the choice of the loss function, both the Mask Decoder and AutoRegressive Decoder, we choose the Cross-Entropy loss for training [39].

4. Cross-lingual Contrastive Learning (CCL) task. In the second stage, we also add the similarity decoder to compare and learn the *CLS* output of mutual translation pairs, so as to shorten the vector space distance between texts with the same semantics. As shown in the Figure 2, in this similarity decoder, in order to keep the same semantics between flipped translation sequence pairs, we use the sequence contrastive learning loss function, which is presented as follows.

$$L_{CL} = -\sum_{i=0}^{n} \left[\log \frac{\exp\left(f(x)^T f\left(x_i^+\right)\right)}{\exp\left(f(x)^T f\left(x_i^+\right)\right) + \sum_{j=1}^{m} \exp\left(f(x)^T f\left(x_j^-\right)\right)} \right] \tag{1}$$

where the x is the input sample of multilingual, while x_i^+ and x_i^- represent the positive and negative samples of translation pairs.

3.3 Model Parameter Details

Recent studies have shown that a deeper model can have better performance under the same parameter size, as the deeper model can deeply understand the original meaning of the language [40].

We first use the Xavier Norm [41] to initialize model parameters, where E is the number of layers of the encoder and D is the number of layers of the decoder.

$$\alpha^{Encoder} = 0.81(E^4 \cdot D)^{\frac{1}{16}}, \alpha^{Decoder} = (3D)^{\frac{1}{4}} \tag{2}$$

$$\beta^{Encoder} = 0.87(E^4 \cdot D)^{-\frac{1}{16}}, \beta^{Decoder} = (12D)^{-\frac{1}{4}} \tag{3}$$

Referring to DeepNet settings [13], we set the α and β values for the standard parameter normalization

$$std_{Encoder} = \beta^{Encoder} \times \sqrt{\frac{2}{\text{fan\_in} + \text{fan\_out}}} \tag{4}$$

$$std_{Decoder} = \beta^{Decoder} \times \sqrt{\frac{2}{\text{fan\_in} + \text{fan\_out}}} \tag{5}$$

$$W_{Encoder} \sim N\left(0, std_{Encoder}\right), W_{Decoder} \sim N\left(0, std_{Decoder}\right) \tag{6}$$

where the fan_in is the number of incoming network connections, while fan_out is the number of outgoing network connections from that layer.

Then, we use the DeepNorm [13] to implement deep model layers. Specifically, we add residual structure in layernorm for each layer.

$$\mathbf{Layer}_{\mathbf{Encoder}}^{\mathbf{Output}} = LayerNorm(x \times \alpha^{Encoder} + f(x)) \tag{7}$$

$$\mathbf{Layer}_{\mathbf{Decoder}}^{\mathbf{Output}} = LayerNorm(x \times \alpha^{Decoder} + f(x)) \tag{8}$$

where we first use the Encoder to encode sentences into a feature matrix $H \in \mathbb{R}^{x \times d \times t}$, which is then input into three different decoder layers.

For the mask decoder, we first dot product encoded hidden feature H with the weight of the input embedding layer ($CMPT^{Embedding}$), and then adopt the linear layer ($Linear^{Mask}$) to obtain the output vector

$$\mathbf{Output}_{\mathbf{Mask}} = Linear^{Mask}\left(H \cdot CMPT^{Embedding}\right) \tag{9}$$

For the Similarity Decoder, we input the H^{CLS} vector into the linear layer ($Linear^{Sim}$) to obtain the semantic vector of the text.

$$\mathbf{Output}_{\mathbf{Similarity}} = Linear^{Sim}\left(H^{CLS}\right) \tag{10}$$

We adopt the cross-attention mechanism to integrate H into decoder for the AutoRegressive Decoder. The attention function can be described as an output of a Query (Q) and a set of Key-Value (K-V) pairs mapping. The output is the weighting, and calculation between these QKV is presented as follows

$$\dot{H}_D^t = MultiHead\_SelfAtt(H_D^t) \tag{11}$$

$$\ddot{H}_D^t = MultiHeadAtt(\dot{H}_D^t, H, H) \tag{12}$$

$$H_D^{t+1} = LayerNorm(H_D^t \times \alpha^{Decoder} + \ddot{H}_D^t)) \tag{13}$$

where t represents the current time, and the whole calculation is implemented as the recursive process for further auto-regression.

Table 1. Details of the Chinese Minority language corpus.

Language Pair	Dataset	Number
Chinese	Monolingual	11,000,000 words
English Chinese	Train	9,037,417 sentences
	Dev	4003 sentences
Mongolian Chinese	Train	1,262,643 sentences
	Dev	1000 sentences
Tibetan Chinese	Train	1,157,959 sentences
	Dev	1000 sentences
Uyghur Chinese	Train	170,061 sentences
	Dev	1000 sentences

3.4 Model Setting Details

The CMPT is a Transformer-based [30] model that supports multiple languages. It has 256 hidden states, 8 attention heads, 128 encoder layers, and 128 decoder layers. The final model size of CMPT is 390MB, which belongs to the base version for the pre-trained language model. In order to adapt to the minority languages, we adopt the CINO vocabulary [25], which has a number of 135359 in size.

In the pre-training phase, we set the maximum token length to 120 and deleted the excess text. We used a 15% mask rate and a maximum of 3-grams for span masking [42]. We have conducted 200000 steps (about one month) of pre-training in 8 GPUs of RTX6000 on the Pytorch[5] [43] and the hugging-face[6] [44] framework, with a batch size of 256. We implement distributed training with mixed precision based on the DeepSpeed [45]. As a result, we have fine-tuned the corpus officially provided by CCMT datasets, with a total of about 15,000,000 samples. We use an AdamW optimizer [46] with a maximum learning rate of 8e-5, followed by linear attenuation and warm-up optimizing schedules [47].

4 Experiments

We implement the experiments under the minority language corpus shown in Table 1. A variety of evaluation metrics are adopted, which can evaluate the generation quality of sentence level and word level meanwhile and show the detailed performance of the system more comprehensively. Specifically, we adopt "BLEU" [21], "ROUGE" [48], "METEOR" [49] and "CIDER" [50] as the evaluation metrics, which can assess the quality of translate, including fidelity and diversity.

[5] `https://pytorch.org`
[6] `https://github.com/huggingface/transformers`

Type	Source Language	Generation Language	Reference Language
Chinese ↓ English	通过相对集中诊疗和双向转诊，为罕见病患者提供较为高效的诊疗服务，延缓疾病的进展，减轻他们的痛苦。	The clinical services are provided to patients with rare diseases with more effective treatment through relative intensive diagnosis and bi-directional diagnosis, which can alleviate their pains and slow down the progression of diseases.	*Through relatively centralized diagnosis and treatment as well as two-way referral, we can provide more efficient services for patients with rare diseases to delay the progress of the disease and alleviate their pain.*
English ↓ Chinese	In 2008, Armisen, whose mother is Venezuelan and whose father is of German and Korean heritage, told New York Magazine's Intelligencer column that he wore honey colored makeup to portray Obama, who is biracial.	美国杂志《纽约杂志》的英文版《美国人》的英文版《美国人》是《纽约时报》的英文版	*阿米森的母亲是委内瑞拉人，父亲是德韩混血，2008年，他告诉《纽约杂志》的“情报”专栏，他使用蜂蜜色的化妆品来刻画奥巴马形象，因为奥巴马是混血儿*
Mongolian ↓ Chinese	ᠥᠨᠳᠥᠷ ᠤᠯᠤᠰ ᠤᠨ ᠬᠡᠯᠡᠨ ᠳᠡᠭᠡᠷ ᠡᠴᠡ ᠬᠡᠯᠡᠭᠰᠡᠨ ᠦᠭᠡ ᠶᠠᠮᠠᠷ ᠪᠠᠢᠭᠠᠯᠢ ᠪᠠᠢᠭᠠᠭᠰᠠᠨ ᠤᠤ ?	你听见了,也听见了,是照你向那地的妇人所应许的。	*你听说过她和某个当地男子走得很近的闲话吗?*
Tibetan ↓ Chinese	ང་ཚོའི་ཀྱང་གིས་ཧོག་མ་ཉིད་ནས་ཀང་ཡོངས་དང་སྐྱག་པར་དུ་འགྲོ་ཞིང་པས་བྱེད་པའི་སྐོང་སྐྱོང་དམ་འཛིན་ལ་མཚོན་ཏེ་ཆ་ནས་བྱེད་བཞིན་ཡོད།	我们党从来都十分重视对全党特别是领导干部的学习,	*我们党历来重视抓全党特别是领导干部的学习,*
Uyghur ↓ Chinese	ئېكسكۇرسىيە كۆرگەزمىنى 22 كۈنىگە تەخمىنەن قىلىنىغانلار دۆلەت ئېشىپ، قېتىمدىن نادەم مىڭ ئوخشاش ئوتكۈزۈلگەن مۇزېيىدا يېڭى ئىچىدە كۆرگەزمىلەر تۈرىدىكى يارىتىلدى رېكورد	参观者共参观了22000多盏天文天台,参观了来自各国的大型艺术博物馆。	*平均每天有两万两千多人次的参观人数，创下了国博同类型展览的新纪录。*

Fig. 3. Case study of the proposed method.

In all experiments, we implement the Transformer [30] method as the baseline for fair comparisons since there is no other suitable method for the Chinese Minority language translation. This model is pre-trained in the same Chinese Minority language corpus, which is trained with equal limited training epochs. We repeated the experiment three times by changing different random seeds to ensure the fairness of the experiment. We use the learning rate of 1e-5 to fine-tune the translation of individual language pairs. Each experiment was conducted for 10 rounds. After each round, we conducted experiments in the dev set, and reported in the result table.

4.1 Main Results

In the experiment of the shared task shown in Table 2, we can find that the CMPT has many to many translation abilities and supports translation tasks

Table 2. Experiments for the CCMT 2022 shared task.

Language	Evaluation Metrics	Ours	Baseline
Chinese ↓ English	BLEU_1	**0.531**	0.416
	BLEU_2	**0.378**	0.278
	BLEU_3	**0.277**	0.153
	BLEU_4	**0.207**	0.129
	METEOR	**0.288**	0.172
	Rouge_L	**0.468**	0.324
	CIDEr	**2.016**	1.564
English ↓ Chinese	BLEU_1	**0.143**	0.098
	BLEU_2	**0.031**	0.015
	BLEU_3	**0.011**	0.009
	BLEU_4	0.005	0.005
	METEOR	**0.129**	0.107
	Rouge_L	**0.177**	0.164
	CIDEr	**0.079**	0.067
Mongolian ↓ Chinese	BLEU_1	**0.155**	0.117
	BLEU_2	**0.052**	0.044
	BLEU_3	**0.029**	0.012
	BLEU_4	**0.016**	0.009
	METEOR	**0.136**	0.094
	Rouge_L	**0.149**	0.101
	CIDEr	**0.122**	0.114
Tibetan ↓ Chinese	BLEU_1	**0.363**	0.333
	BLEU_2	**0.253**	0.207
	BLEU_3	**0.202**	0.149
	BLEU_4	**0.168**	0.112
	METEOR	**0.419**	0.374
	Rouge_L	**0.380**	0.365
	CIDEr	**1.295**	0.774
Uyghur ↓ Chinese	BLEU_1	**0.217**	0.124
	BLEU_2	**0.053**	0.041
	BLEU_3	**0.028**	0.015
	BLEU_4	**0.017**	0.010
	METEOR	**0.148**	0.117
	Rouge_L	**0.249**	0.204
	CIDEr	**0.127**	0.108

in different minority languages compared with the baseline method. For the same experimental setting, our method can achieve better performance than the baseline. Specifically, we can also find that CMPT has good translation performance, whether it is Chinese to English or ethnic minorities to Chinese. This further demonstrates the effectiveness of the proposed method. However, our method fails to obtain a high score for the English-Chinese translation, where the reason may be the limited size of the pre-training process.

4.2 Case Study

We randomly selected some translation results for comparison, which is shown in the Figure 3. The further conclusion can be found that in the scenario of English to Chinese translation, the model generates repetition results, which may be because the model has a weak understanding of English due to insufficient training corpus. When using minority languages for translation, we can find that the model can generate relatively complete text. However, due to the insufficient understanding ability of the model, the semantics of the generated text may be biased.

5 Conclusion

In this work, we introduced a multilingual model CMPT that supports downstream generative tasks. It uses Chinese Minority languages for multi-task and multi-stage pre-training to comprehensively improve the ability of understanding and generation. We have conducted an evaluation of the translation task of the CCMT-2022, where the experimental results show that CMPT has achieved better performance both in understanding and generation compared with the baseline method. In the future, we believe that with more pre-training minority language corpus being used for the pre-training, the performance of the CMPT is expected to be further improved.

References

1. Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *AI Open*, 2021.
2. Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
3. Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *north american chapter of the association for computational linguistics*, 2018.

5. Bin Li, Yixuan Weng, Fei Xia, and Hanjun Deng. Towards better chinese-centric neural machine translation for low-resource languages. *arXiv preprint arXiv:2204.04344*, 2022.
6. Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Computing Surveys*, 2019.
7. Xiulan Zuo. China' s policy towards minority languages in a globalising age. *TCI (Transnational Curriculum Inquiry)*, 4(1):80–91, 2007.
8. Hong-Cai Zhou, Jeffrey R Long, and Omar M Yaghi. Introduction to metal–organic frameworks, 2012.
9. Isabelle Attané and Youssef Courbage. Transitional stages and identity boundaries: The case of ethnic minorities in china. *Population and Environment*, 21(3):257–280, 2000.
10. Michael Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *meeting of the association for computational linguistics*, 2019.
11. Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv: Computation and Language*, 2021.
12. Wilson L. Taylor. Cloze procedure: A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 1953.
13. Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. 2022.
14. Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *Learning*, 2020.
15. Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv: Computation and Language*, 2021.
16. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.
17. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *meeting of the association for computational linguistics*, 2020.
18. Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *Language Resources and Evaluation*, 2019.
19. Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Michael Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 2020.
20. Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *arXiv: Computation and Language*, 2020.

21. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
22. Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. Veco: Variable and flexible cross-lingual pre-training for language understanding and generation. *meeting of the association for computational linguistics*, 2021.
23. Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. An attentive survey of attention models. *arXiv: Learning*, 2019.
24. Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li. switch-glat: Multilingual parallel machine translation via code-switch decoder. In *International Conference on Learning Representations*, 2021.
25. Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. Cino: A chinese minority pre-trained language model. 2022.
26. Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *neural information processing systems*, 2019.
27. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *neural information processing systems*, 2019.
28. Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. Unilmv2: Pseudo-masked language models for unified language model pre-training. *international conference on machine learning*, 2020.
29. Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. All nlp tasks are generation tasks: A general pretraining framework. *arXiv: Computation and Language*, 2021.
30. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
31. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
32. Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. End-to-end neural text classification for tibetan. *CCL*, 2017.
33. Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

34. Baotian Hu, Qingcai Chen, and Fangze Zhu. Lcsts: A large scale chinese short text summarization dataset. *empirical methods in natural language processing*, 2015.
35. Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi kiu Lo, Nikola Ljubeśić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (wmt20). *empirical methods in natural language processing*, 2020.
36. Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset. *arXiv: Computation and Language*, 2016.
37. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.
38. Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. Efficient mini-batch training for stochastic optimization. *knowledge discovery and data mining*, 2014.
39. Cong Chen, Qinqin Zong, Qi Luo, Bailian Qiu, et al. Transformer-based unified neural network for quality estimation and transformer-based re-decoding model for machine translation. In *Machine Translation: 16th China Conference, CCMT 2020, Hohhot, China, October 10-12, 2020, Revised Selected Papers*, volume 1328, page 66. Springer Nature, 2020.
40. Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv: Computation and Language*, 2021.
41. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *international conference on artificial intelligence and statistics*, 2010.
42. Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.
43. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
44. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander 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*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

45. Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020.
46. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.
47. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
48. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
49. Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.
50. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.