School of Informatics

Informatics Research Review
Pre-trained Models in Multilingual Learning

B202412
January 2022

Abstract

Multilingual learning is an important research direction in natural language processing. With the development of pre-trained models, more and more multilanguage pre-trained models appear in people's sight. In this IRR, we will explain some multilingual pre-trained models and summarize some of their limitations and achievements.

Date: Thursday 27th January, 2022

Supervisor: Pavlos Andreadis

1 Introduction

In recent years, the vigorous development of natural language processing has provided more and more convenience for people. Among them, the development of pre-trained models is getting more and more attention, and the high performance and efficiency of these models can benefit people in different downstream tasks of natural language processing, including text classification, named entity recognition and machine translation.

The earliest pre-training models mainly focus on English tasks or monolanguage tasks, and their performance is poor for multi-language tasks, especially for some low-resource languages. Due to the scarcity of data resources in these languages, it is difficult to complete the natural language processing downstream tasks of these languages. Breaking the gap between machine understanding between different languages is a key issue in natural language processing. Therefore, in order to improve the models in multilingual environments, people have turned their attention to multilingual learning, and proposed different multilingual pre-trained models, these models have achieved excellent results.

In terms of data collection and preprocessing, there are Byte-Pair Encoding (BPE) [1], WordPiece [2], SentencePiece [3] and other methods. Due to the large gap in dataset size between different languages, XLM [4] proposes a sampling method for high-resource and low-resource language dataset to solve this data imbalance problem. XLM-R [5] also proves that increasing the data size can improve model performance.

The methods used in multilingual pre-training includes: 1.Methods based on different objectives: The most classic objectives are Casual Language Modeling(CLM), multilingual Masked Language Modeling(MLM) and Translation Language Modeling(TLM) in XLM. They are improved from the MLM in BERT [6] and provide inspiration for the subsequent multilingual pre-training model. After this Other MLM variants have emerged that continuously improve the model's multilingual learning capabilities. Unicoder [7], ERNIE-M [8], MT5 [9] , and MT6 [10] all propose their own objectives to obtain stronger cross-language transfer capabilities. 2.Denoising Autoencoder [11] based method: mBART [12] and XNLG [13] borrow the idea of Denoising Autoencoder and add noise to the data in different ways and then restoring the data. 3.Constrastive Learning [14] based method: A classic example is HICTL [15]. HICTL uses constrastive learning in sentence-level and word-level. 4.Other method: LaBSE [16] uses new training methods to improve multilingual learning ability.

The problems that these multi-language pre-training models need to overcome include the difficulty of obtaining parallel corpora, their performance on monolanguage tasks is not as good as that of monolanguage pre-training models, the lack of performance in low-resource language dataset, the high training costs, and difficulty in completely unsupervised, etc. For example, the TLM task and the introduction of translation pairs in MT6 require parallel corpora. Although XLM-R outperforms BERT on monolingual tasks, it may not outperform all monolingual pre-trained models.

This Research Review explores some of the most relevant methods for multilingual learning, focusing on some multilingual pre-training models, and summarizes their methods, results, and some limitations of their applications. Our research shows that by proposing new training objectives and borrowing some other methods, the cross-language learning ability of the model can be improved, and the use of larger dataset size can also improve the model performance. However, the unsupervised learning ability of the model needs to be improved, the application of low-resource languages and how to better improve the performance on monolanguage is still a difficult problem. At the same time, we believe that other fields of natural language processing

will benefit from multilingual pre-trained model, such as multimodal tasks, therefore, how to improve the ability of multilingual pre-training models is a question worth exploring.

2 Literature Review

2.1 From Monolingual Learning to Multilingual Learning

In monolingual pre-training, BERT [6] demonstrates its strong performance, achieving the best results on 11 natural language processing benchmarks. One of its key tasks is the Masked Language Model (MLM), where some input tokens are marked as [mask] with a certain probability, and then these masked tokens are predicted. MLM is illustrated in Figure 1. Inspired by it, a multilingual variant of BERT (mBERT) [17] is proposed. The core idea of mBERT is to use the same model and weights to process all target languages, so as to obtain a certain ability of cross-language transfer and realize multi-language learning.

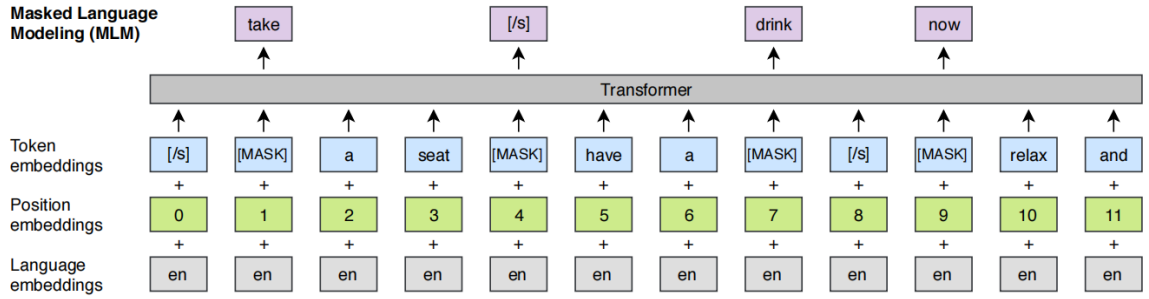

Figure 1: Masked Language Model (MLM)

The data is taken from the corpora of 104 languages in Wikipedia and a shared word piece vocabulary is built, parallel corpora are not used. When trained in the same way as BERT, the experimental results show that the zero-shot task of languages without overlapping vocabulary can be done well in mBERT, and the effect improves with the similarity of languages. However, its limitation is that it is less effective for languages with different language sequences, and cannot learn the correspondence between the two languages, so the ability of cross-language transfer needs to be improved.

2.2 Multilingual pre-training model

2.2.1 Method based on different objectives

A core purpose of multilingual learning is to learn corresponding semantic representations between two languages. To achieve this goal and improve multilingual learning methods, some new multilingual training objectives are proposed, thus bringing multilingual pre-training to a new level.

Facebook AI team first proposed Cross-lingual language model (XLM) [4] based on mBERT. The inclusion of three new training objectives in XLM and the introduction of parallel corpora into pre-training enables the model to outperform mBERT. In XLM, Byte pair encoding (BPE) [1] is used to build a vocabulary shared by different languages. The basic idea of BPE is to break up the corpus to characters, and then learn a merging or splitting rule based on statistical

information to reduce or expand the vocabulary. The advantage of this approach is that it can significantly improve the alignment of different languages in the embedding space. In order to ensure the balance of the corpus, suppose the sentences follow multinomial distribution with probabilities $\{q_i\}_{i=1...N}$, the sampling probability of the sentence is as follows:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \tag{1}$$

Three multilingual training objectives are proposed, including:

- **Casual Language Modeling (CLM)**
  The principle of CLM comes from Transformer [18] model, and the goal is to predict the probability of the current word based on the probability $P(w_t|w1, ..., w_{t-1}, \theta)$ of the previous sentence of the word.

- **Masked Language Modeling (MLM)**
  It can be understood as a multilingual version of the MLM in BERT, however, the difference is that in BERT we only uses text consisting of two sentences, while MLM here uses any number of sentences and there are 256 tokens in one sentence. In addition, in order to balance the difference in the frequency of occurrence of different words, a method similar to the resampling is used when masking tokens.

- **Translation Language Model (TLM)**
  TLM is proposed to improve the effect of cross-lingual language model pretraining, which can be regarded as an extension to the MLM task. TLM concatenates parallel sentences and randomly masks tokens between two sentences.. As shown in the Figure 2 , when predicting a masked token in an English sentence, the model first considers inference from the context of the English sentence. If the English text is not enough to infer the result, the model uses the French sentence to infer the masked English token. This way, the model can align the English and French representations.

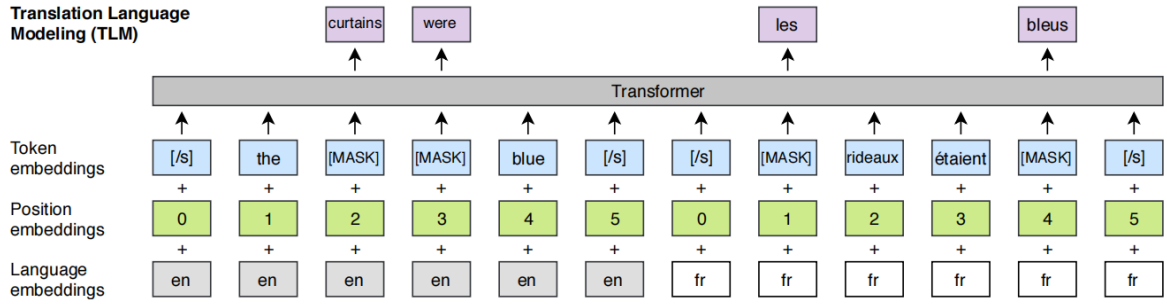

Figure 2: Translation language model(TLM)

CLM and MLM are unsupervised and only require monolingual corpora. TLM is supervised and requires parallel corpora. In CLM and MLM, the data from each batch comes from the same language and contains 64 continuous sentence streams, in which each sentence consists of 256 tokens. The sampling probability of the sentence is given by Eq.1, with $\alpha = 0.7$. In model combining MLM and TLM, MLM and TLM are implemented alternately to complete the training, and the sampling method is similar to the former.

Their experiments demonstrate the powerful performance of XLM. On unsupervised machine translation, MLM pretraining achieves remarkable results and outperforms the previous best method by 9.1 BLEU [19] on WMT'16 German-English. On supervised machine learning, XLM also achieve better results and obtain an improvement of 4.1 BLEU on WMT' Romanian-English. When TLM is used together with MLM, the performance of the model is further improved. In addition, XLM can also improve the perplexity of a Nepali language by using corpora in other languages. Model performance is further improved when TLM and MLM are used in combination. However, although XLM has a significant impact on cross-language pre-training, it still has some limitations, such as poor performance in low-resource languages, and limiting model understanding when using cross-language transfer to scale the model to more languages competence in each language.

To address these issues, in November 2019, Facebook AI team proposed an improved model XLM-R [5]. It draws on the idea of RoBERTa [20] and uses the training method of XLM. Compared with XLM, XLM-R has the following improvements: The number of languages and the number of training datasets has been greatly increased, and uses the filtered CommonCrawl dataset in 100 languages with a vocabulary of 250k generated by SentencePiece [3] . XLM-R uses a similar sampling method as XLM when sampling, but adjusts α to 0.3, sets the vocabulary size to 250k, and increases the overall parameters to 550M. Compared with XLM, the dataset size for the same 88 languages is at least an order of magnitude higher, especially on low-resource language dataset.

XLM-R achieves better results on four cross-lingual understanding benchmarks, including cross-lingual natural language inference(XNLI), named entity recognition, question answering and GLUE [21] benchmark. However, in named entity recognition, while XLM-R has the best average performance across languages, in some languages, such as English and German, XLM-R is less accurate than contextual word embeddings [22] because CRF [23] is not used. It is worth noting that XLM on 7 languages outperforms monolingual model (Bert) on XNLI benchmark. Furthermore, they conducted a data scale study on high-resource language datasets and low-resource language datasets and found that increasing the data scale could improve the performance of cross-language models. It can be seen that despite its large scale, XLM-R still has certain limitations as a multilingual language model, and its effectiveness is higher than some monolingual language models but still needs to be improved. Also, it has certain advantages for low-resource languages and is a big improvement over XLM.

Following the previous training objectives, three other new multilingual training objectives are proposed to further improve the multilingual learning ability. Unicoder [7] was proposed by Microsoft in September 2019. It proposes three new pre-training tasks that allow the model to better learn the correspondence between different languages:

- **Cross-lingual Word Recovery**
  It is similar to TLM in XLM and requires bilingual sentence pairs. As the Figure 3(a) shows, the task first represents each word in language s by all word embeddungs of language t, then calculates an attention matrix. The obtained representation is used as the input of the Transformer, and finally the original sentence sequence of the predicted language s is output to realize word recovery. Cross-lingual word recovery can learn the correspondence between different languages without using word alignment.

- **Cross-lingual Paraphrase Classification**
  The task is used to determine whether two sentences in different languages have the same meaning. It is similar to the next sentence prediction task in BERT. The structure is

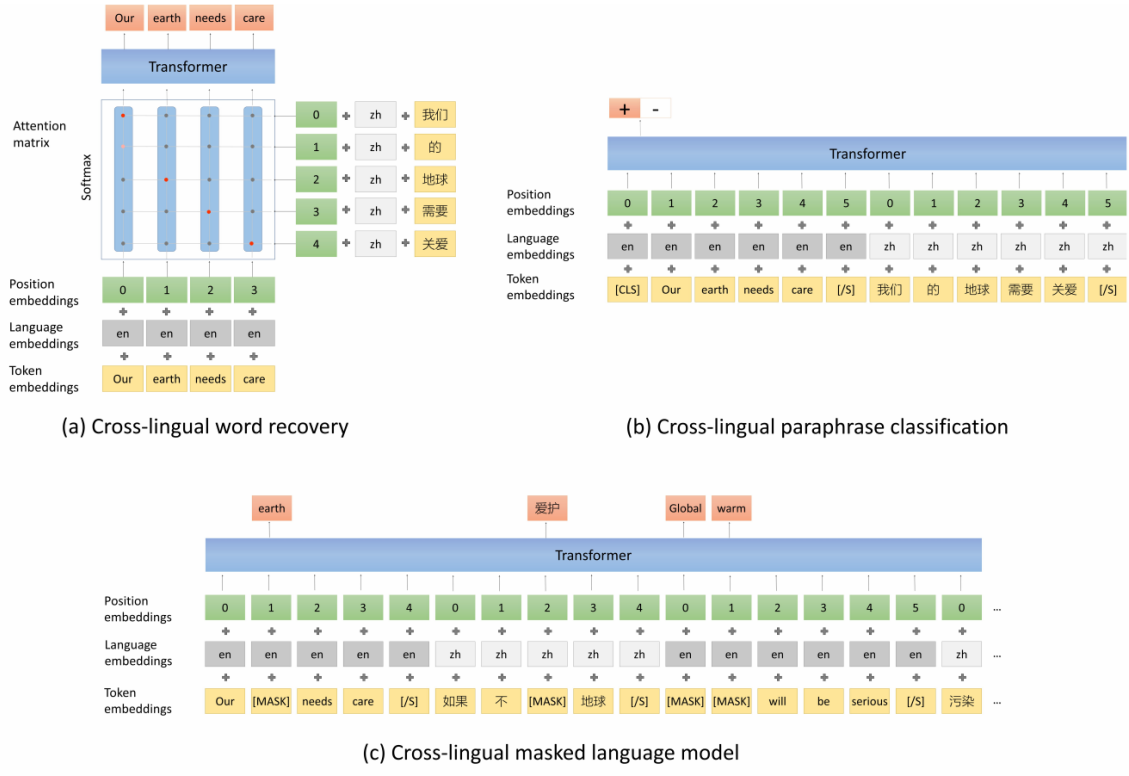

Figure 3: Three new training objectives in Unicoder

shown in Figure 3(b), the input is composed of two sentences in different languages, and the token [CLS] is added at the beginning and end tokens are added at the end. This task can learn sentence-level relations between different languages by training a binary classifier.

- **Cross-lingual Masked Language Model**
  It is based on document-level corpora, and the corpora mixed with multiple languages is treated as one language for MLM. The structure of the model is shown in Figure 3(c). It should be noted that the languages of adjacent sentences in these documents are different, but the meaning of the documents needs to remain smooth.

Unicoder has better performance on both XNLI and cross-lingual question answering (XQA), and has been improved compared to previous methods including BERT, mBERT, and XLM. Among the 15 languages of results in XNLI, Unicoder achieved the best results under different fine-tuning methods. In XQA, unicoder achieves the highest scores in English, French, and German, surpassing the scores of XLM and traditional BERT.

In 2021, Baidu's proposed ERNIE-M [8] introduces two new training objectives, which also enhance the model's multilingual learning capabilities. The objectives include:

- **Cross-attention Masked Language Modeling (CAMLM)**
  The principle of CAMLM is similar to that of MLM, except that it uses the context of another language as the object of Attention, that is, when predicting the masked tokens of the sentence of one language, it uses the context of the sentence in another language and does not depend on the text in original language.

- **Back-translation Masked Language Modeling (BTMLM)**
  First, add several masked tokens to the back of sentences in one language, translate them into corresponding pseudo-tokens in another language, then splicing the pseudo-tokens to the back of the original sentence, mask out some of the preceding tokens, and then proceed to MLM.

Using monolingual and parallel corpora, ERNIE-M outperforms XLM and XLM-R on different datasets and downstream tasks, achieving SoTA results and indicating better transfer ability.

In addition, Multilingual T5 (MT5) [9] and Multilingual T6 (MT6) [10] were proposed by Google and Microsoft in 2020 and 2021. MT5 adds span corruption to the training objective, using the same span to mask tokens, which is an unsupervised MLM task. MT6 introduced translation pairs to improve the cross-lingual transferability and alignment of representations of MT5, and proposed three training targets based on translation pairs, including machine translation, translation pair span corruption and translation span corruption. In addition, they proposed Partially Non-autoregressive Decoding(PNAT) to enable the model to utilize more encoded information for text-to-text learning. Both MT5 and MT6 outperform mBERT, XLM, and XLM-R enough to show their performance.

2.2.2 Denoising Autoencoder based method

Denoising Autoencoder [11] is also used in multilingual learning pretrained models. Facebook AI team proposed mBART [12] in 2020, which can be seen as a multilingual version of BART [24] . it performs sequence-to-sequence pre-training on multilingual dataset using denoising method. Figure 4 shows the framework of mBART.

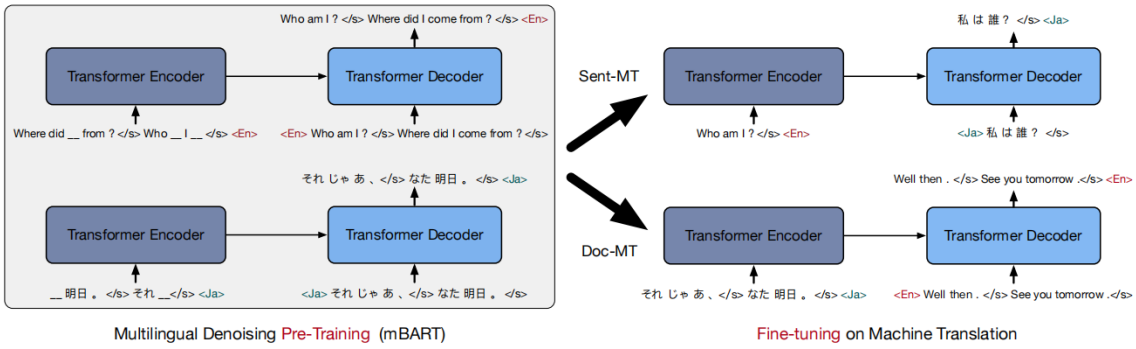

Figure 4: Pre-training and fine-tuning framework of mBART on machine translation tasks.

They use the 25 languages in CommonCrawl as the dataset (CC25), and use the same sampling method as XLM in dealing with the problem of sample imbalance. In pre-training, mBART uses monolingual data. The input of the encoder contains two types of noise including masked tokens and permutated sentences with language id symbols added at the end and a language id symbol added at the head. In this way, mBART implements text recovery by inputting data with noise and learning a single Transformer [18] .

Compared with previous pre-training methods, including XLM, XLM-R and Unicoder, mBART outperforms previous methods with and without back-translation [25]. At the same time, they compared the fine-tuning effects of different languages for English and found that on low-resource data (data volume less than 10M), mBART performed the best on different translation tasks,

and in resources with data volume greater than 25M, it was better to use its own data set alone. Inaddition, mBART is found to perform better after comparison with traditional BART models trained only on the same En and Ro data, which also illustrates the importance of pre-training in a multilingual environment.

XNLG [13] also borrows the idea of Denoising Autoencoder, which was proposed by Microsoft in 2019. XNLG first trains the encoder with monolingual MLM and cross-lingual MLM, and then trains the decoder with denoising auto-encoding (DAE) and Cross-Lingual Auto-Encoding (XAE). Experiments show that XNLG has strong performance on the task of zero-shot cross-lingual natural language generation, which is an improvement over XLM.

2.2.3 Constrastive Learning based method

Another way to improve multilingual learning ability is to use constrastive learning [14] . Figure 5 illustrates Hierarchical Contrastive Learning (HICTL) [15] proposed by Wei et al. Based on XLM-R, they used XLM-R to initialize the parameters of HICTL, and conducted constrastive learning at both sentence-level and word-level.

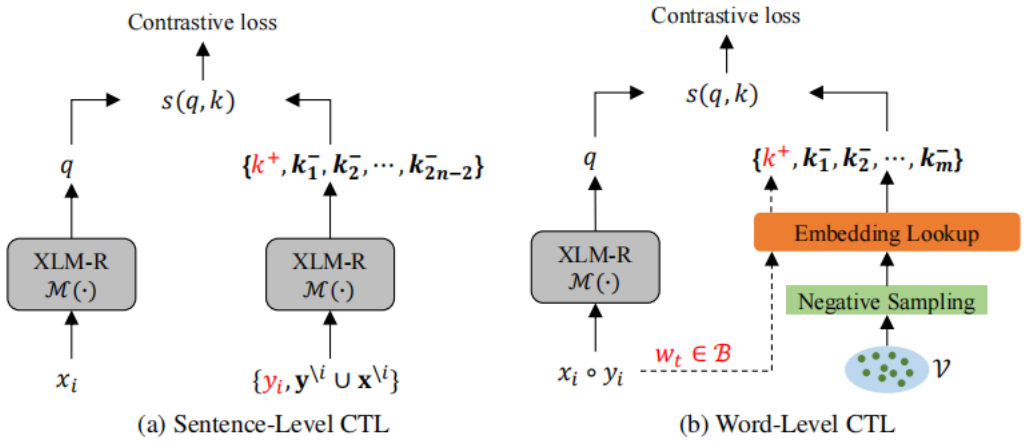

Figure 5: Hierarchical Contrastive Learning (HICTL)

HICTL uses parallel corpora or perturbated corpora as input data. TLM and MLM are included in HICTL by default, and constrastive learning is performed on this basis. At sentence-level, they make the parallel corpus have similar meanings and the meanings expressed by other sentences in the same training batch as far as possible, so as to achieve constrastive learning. When the number of negative samples is insufficient, they employ smoothed linear interpolation [26] to address this issue. At word-level, they regarded words that appeared in the sentence as positive samples and words that did not appear in the sentence as negative samples. Representations across languages are encouraged to learn and reduce semantic differences between different languages.

Their results show that HICTL outperforms XLM-R in both cross-language understanding and generation, and achieves 4.2% higher accuracy than XLM-R on XTREME [27], especially in zero-shot sentence retrieval task, in which HICTL achieves 6.0% higher accuracy than XLM-R. Furthermore, HICTL achieves better performance than XLM-R in both low-resource and high-resource language machine translation.

2.2.4 Other method

In addition to the above methods, other methods have been used to improve multilingual pre-trained models. For example, to change the training method of the model, Language-agnostic BERT Sentence Embedding (LaBSE) [16] was proposed by Google in 2020, and its main innovation comes from a new training method for multilingual learning. LaBSE uses MLM and TLM as encoder and is trained with stage progressive stacking algorithm. That is, for an L-layer model, first train the first L/4 layers, then train the L/2 layers, and finally train the L layers, and the weights of the previous session are used to initialize the current session. They also found that softmax is a key factor in training the model, and using MLM pre-training can greatly reduce the amount of data in parallel corpora.

3 Summary & Conclusion

This review provide a overview of multilingual pre-trained model, they are mainly divided into several categories, including XLM, XLM-R, Unicoder, ERNIE-M, MT5 and MT6 that proposed new objectives, mBART, XNLG borrowed from Denoising Autoencoder, HICTL using constrastive learning, and LaBSE using new training method. Among them, XLM is a classic pre-training model in multilingual learning, and the introduction of multilingual MLM and TLM has had a profound impact on the development of multilingual pre-training models.

Most of the multilingual pre-training models are based on Transformer, and different training objectives have created good results. While exploring these models, we realized that increasing the size of the dataset can improve the power of the model, such as XLM-R outperforming XLM. On low-resource language tasks, multilingual pre-training models still need to be improved. In addition, adding supervised learning to unsupervised learning can improve the performance of the model, such as adding TLM to XLM, but it is not easy to obtain parallel corpus, how to achieve completely unsupervised learning is a problem worth thinking about. In monolingual tasks, the multilingual model has some limitations. Although it may surpass some monolingual pre-training models, the performance still needs to be improved. It is worth noting that multilingual models can improve the perplexity of a language by using corpora in other languages, which can also provide inspiration for us to solve some downstream tasks.

The current multilingual pre-trained models still have certain limitations, however, more and more methods are used to try to solve these problems and achieve considerable results. There are many ways to improve the multilingual pre-trained model, including selecting a new monolingual model and changing it into a multilingual variant, proposing new training objectives, using a larger dataset on the model, learning from other machine learning methods, and changing the training order or method, etc. In addition, the development of monolingual pre-trained models and the development of artificial intelligence will also promote the development of multilingual pre-trained models. It is foreseeable that the development of multilingual pre-trained models will benefit all aspects. When multilingual pre-training performance reaches a certain level, other fields in natural language processing can also benefit accordingly, for example, adding multilingual pre-trained models to multimodal tasks. Multilingual pre-trained models can also inspire other fields in machine learning.

References

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

[2] Xinying Song, Alexandru Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization. In *EMNLP*, 2021.

[3] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.

[4] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[5] 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. In *ACL*, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks, 2019.

[8] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora, 2021.

[9] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, 2021.

[10] Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang Xian-Ling Mao, Heyan Huang, and Furu Wei. mt6: Multilingual pretrained text-to-text transformer with translation pairs. In *EMNLP*, 2021.

[11] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

[12] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[13] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training, 2019.

[14] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019.

[15] Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages. *ArXiv*, abs/2007.15960, 2021.

[16] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2020.

[17] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[19] 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.

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

[21] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[22] A. Akbik, Duncan A. J. Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING*, 2018.

[23] John D. Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16, 2016.

[26] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics.

[27] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080, 2020.