IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 20, 2020, accepted August 1, 2020, date of publication August 11, 2020, date of current version August 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015854

AgglutiFiT: Efficient Low-Resource Agglutinative Language Model Fine-Tuning

ZHE LI[1], XIUHONG LI[2], JIABAO SHENG[1], AND WUSHOUR SLAMU[3]

[1]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang Multilingual Information Technology Research Center, College of Software, Xinjiang University, Urumqi 830046, China
[2]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
[3]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang Multilingual Information Technology Research Center, College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Corresponding author: Xiuhong Li (xjulxh@xju.edu.cn)

This work was supported in part by the Xinjiang University Ph.D. Foundation Initiated Project under Grant 620312343, in part by the Xinjiang Uygur Autonomous Region Graduate Research and Innovation Project Grant, in part by the Dark Web Intelligence Analysis and User Identification Technology under Grant 2017YFC0820702-3, in part by the National Language Commission Research Project under Grant ZDI135-96, and in part by the National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

ABSTRACT Text classification tends to be difficult when data are inadequate considering the amount of manually labeled text corpora. For low-resource agglutinative languages including Uyghur, Kazakh, and Kyrgyz (UKK languages), in which words are manufactured via stems concatenated with several suffixes and stems are used as the representation of text content, this feature allows infinite derivatives vocabulary that leads to high uncertainty of writing forms and huge redundant features. There are major challenges of low-resource agglutinative text classification the lack of labeled data in a target domain and morphologic diversity of derivations in language structures. It is an effective solution which fine-tuning a pre-trained language model to provide meaningful and favorable-to-use feature extractors for downstream text classification tasks. To this end, we propose a low-resource agglutinative language model fine-tuning *AgglutiFiT*, specifically, we build a low-noise fine-tuning dataset by morphological analysis and stem extraction, then fine-tune the cross-lingual pre-training model on this dataset. Moreover, we propose an attention-based fine-tuning strategy that better selects relevant semantic and syntactic information from the pre-trained language model and uses those features on downstream text classification tasks. We evaluate our methods on nine Uyghur, Kazakh, and Kyrgyz classification datasets, where they have significantly better performance compared with several strong baselines.

INDEX TERMS Transfer learning, pre-training, low-resources text classification, fine-tuning.

I. INTRODUCTION

Text classification is the backbone of most natural language processing tasks such as sentiment analysis, classification of news topics, and intent recognition. Although deep learning models have reached the most advanced level on many Natural Language Processing(NLP) tasks, these models are trained from scratch, which makes them require larger datasets. Still, many low-resource languages lack rich annotated resources that support various tasks in text classification. For UKK languages, as show in Table-1, words are derived from stem affixes, so there is a huge vocabulary. Stems represent of text content and affixes provide semantic and grammatical functions. Diversity of morphological structure leads to transcribe speech as they pronounce while writing and suffer from high uncertainty of writing forms on these languages which causes the personalized spelling of words especially less frequent words and terms Ablimit *et al.* [2]. Data collected from the Internet are noisy and uncertain in terms of coding and spelling Ablimit *et al.* [1]. The main problems in NLP tasks for UKK languages are uncertainty in terms of spelling and coding and annotated datasets inadequate poses a big challenge for classifying short and noisy text data.

Data augmentation can effectively solve the problem of insufficient marker corpus in low-resource language datasets. Şahin and Steedman [17] present two simple text augmentation techniques using ‘‘crops’’ sentences by removing dependency links, and ‘‘rotates’’ sentences by moving the tree fragments around the root. However, this may not be

The associate editor coordinating the review of this manuscript and approving it for publication was Yucong Duan.

Stem	Words	Affixes
ئىش work	worker ئىش+چى = ئىشچى	چى
	office ئىش+خانا = ئىشخانا	خانا
	position ئىش+ تات = ئىشتات	تات
ئوقۇ read	go to school ئوقۇ+ ش = ئوقۇش	ش
	student ئوقۇ+غۇچى = ئوقۇغۇچى	غۇچى
	teach ئوقۇ+ ت = ئوقۇت	ت

TABLE 1. **Examples of Uyghur word variants.**

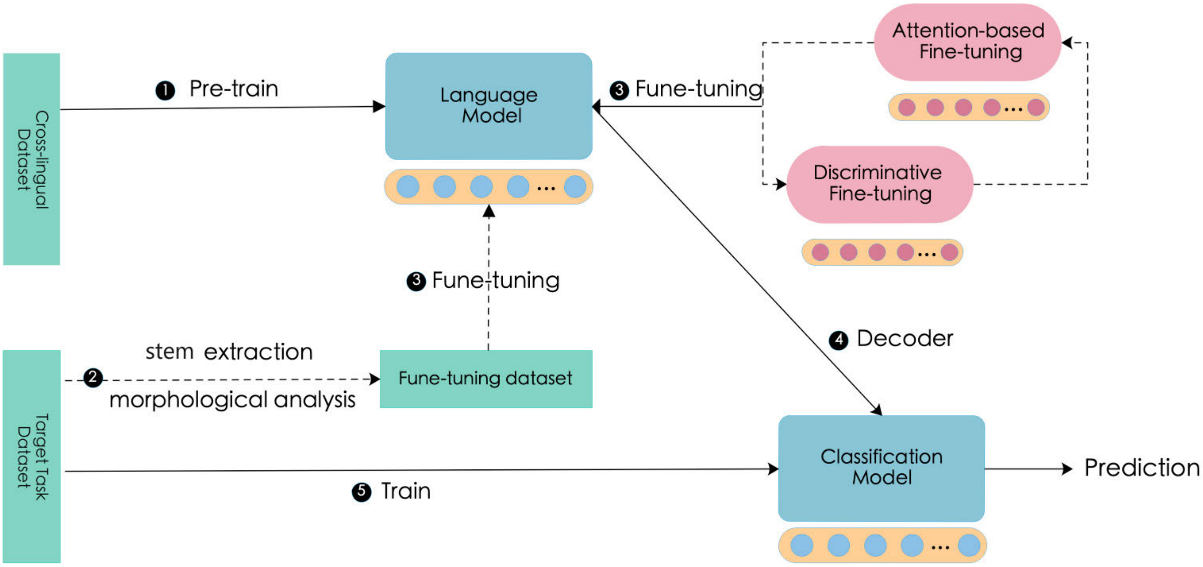

FIGURE 1. **High-level illustration of AgglutiFiT.**

sufficient for several other tasks such as cross-language text classification due to irregularities across UKK languages in these kinds of scenarios. Pre-trained language models such as *BERT* Devlin *et al.* [7] or *XLM* Devlin *et al.* [5] have become an effective way in NLP and yields state-of-the-art results on many downstream tasks. These models require only unmarked data for training, so they are especially useful when there is very little market data. Fully exploring fine-tuning can go a long way toward solving this problem Xu *et al.* [25]. Sun *et al.* [21] conduct an empirical study on fine-tuning, although these methods achieve better performance, they did not perform well on UKK low-resource agglutinative languages due to the morphologic diversity of derivations.

The significant challenge of using language model fine-tuning on low-resource agglutinative languages is how to capture feature information. To apprehend rich semantic patterns from plain text, *Zhang et al.* [27] incorporating knowledge graphs (KGs), which provide rich structured knowledge facts for better language understanding. Zhang *et al.* [28] propose to incorporate explicit contextual semantics from pre-trained semantic role labeling (SemBERT) which can provide rich semantics for language representation to promote natural language understanding. UKK languages are a kind of morphologically rich agglutinative languages, in which words are formed by a root (stem) followed by suffixes. These methods are difficult to capture the semantic information of UKK languages. As the stems are the notionally independent word particles with a practical meaning, and affixes provide grammatical functions in UKK languages, morpheme segmentation can enable us to separate stems and remove syntactic suffixes as stop words, and reduce noise and capture rich feature in UKK languages texts in the classification task.

In this paper, as depict in Figure-1, we propose a low-resource agglutinative language model fine-tuning model: *AgglutiFiT* that is capable of addressing these issues. First, we use $XLM - R$ to pre-train a language model on a large cross-lingual corpus. Then we build a fine-tuning dataset by stem extraction and morphological analysis as the target task dataset to fine-tune the cross-lingual pre-training model. Moreover, we introduce an attention-based fine-tuning strategy that selects relevant semantic and syntactic information from the pre-trained language model and uses discriminative fine-tuning to capture different types of information on different layers. To evaluate our model, we collect and annotate nine corpora for text classification of UKK low-resource agglutinative language, including topic classification, sentiment analysis, intention classification. The experimental results show *AgglutiFiT* can significantly improve the performance with a small number of labeled examples.

The contributions of this paper are summarized as follows:

- We construct three low-resource agglutinative languages including Uyghur, Kazakh, and Kyrgyz nine datasets, each of languages datasets contains topic classification, sentiment analysis, and intention classification three common text classification tasks.
- We propose a fine-tuning strategy on low-resource agglutinative language that builds a low-noise fine-tuning dataset by stem extraction and morphological analysis to fine-tune the cross-lingual pre-training model.
- We propose an attention-based fine-tuning method that better select relevant semantic and syntactic information from the pre-trained language model and uses discriminative fine-tuning to capture across different types of information different layers.

II. RELATED WORK

In the field of NLP, low-resource text processing tasks receive increasing attention such as Hangya *et al.* [10] utilizes a delightfully simple method for domain adaptation of bilingual word embeddings overcoming data sparsity in the target language. And Zhang *et al.* [28] proposes Semantics-aware BERT (SemBERT), which is capable of explicitly absorbing contextual semantics over a BERT backbone and it obtains substantially improves results on typical NLP tasks. We briefly review three related directions: data augmentation, language model pre-training, and fine-tuning.

A. DATA AUGMENTATION

Data Augmentation is that addresses the challenge of insufficient data by creating composite examples that are generated from but not identical to the original document. Wei and Zou [24] present EDA, easy data augmentation techniques to improve the performance of text classification task. For a given sentence in the training set, EDA randomly chooses and performs one of the following operations: synonym replacement, random insertion, random swap, random deletion. UKK languages has few synonyms for a certain word, so the substitution of synonyms cannot add much data. Its words are formed by a root (stem) followed by suffixes, and as the powerful suffixes can reflect semantically and syntactically, random insertion, random swap, random deletion may change the meaning of a sentence and cause the original tags to become invalid. In the text classification, training documents are translated into another language by using an external system and then converted back to the original language to generate composite training examples, this technology known as *backtranslation*. Shleifer [19] work experiments with *backtranslation* as data augmentation strategies for text classification. The translation service quality of Uyghur is not good, and Kazakh and Kyrgyz do not have mature and robust translation service, so it is difficult to use the three languages in *backtranslation*. Şahin and Steedman [17] proposes an easily adaptable, multilingual text augmentation technique based on dependency trees. It augments the training sets of these low-resource languages which are known to have extensive morphological case-marking systems and relatively free word order including Uralic, Turkic, Slavic, and Baltic language families.

B. CROSS-LINGUAL PRE-TRAINED LANGUAGE MODEL

Recently, Pre-training language models such as BERT Devlin *et al.* [8] and GPT-2 Radford *et al.* [15] have achieved enormous success in various tasks of natural language processing such as text classification, machine translation, question answering, summarization, etc. The early work in the field of cross-language understanding has proven the effectiveness of cross-language pre-trained models on cross-language understanding. The multilingual *BERT* model is pre-trained on Wikipedia in 104 languages using a shared vocabulary of word blocks. LASER Artetxe and Schwenk [3] is trained on parallel data of 93 languages and those languages share BPE vocabulary. Conneau and Lample [5] also use parallel data to pre-train *BERT*. These models can achieve zero distance migration, but the effect is poor compared with the monolingual model. The $XLM - R$ Conneau *et al.* [6] uses filtered common-crawled data over 2TB to demonstrate that using a large-scale multilingual pre-training model can significantly improve the performance of cross-language migration tasks.

C. FINE-TUNING

When we adapt the pre-training model to NLP tasks in a target domain, a proper fine-tuning strategy is desired. Howard and Ruder [11] proposes the universal language model fine-tuning (*ULMFiT*) with several novel fine-tuning techniques. ULMFiT consists of three steps, namely general-domain LM pre-training, target task LM fine-tuning, and target task classifier fine-tuning. Eisenschlos *et al.* [9] combines the *ULMFiT* with the quasi-recurrent neural network (*QRNN*) Bradbury *et al.* [4] and subword tokenization Kudo [12] to propose multi-lingual language model fine-tuning (*MultiFit*) to enable practitioners to train and fine-tune language models efficiently. The *MultiFiT* language model consists of one subword embedding layer, four *QRNN* layers, one aggregation layer, and two linear layers. Moreover, a bootstrapping method Ruder and Plank [16] is applied to reduce the complexity of training. Although those approaches are general enough and have achieved state-of-the-art results on various classification datasets, the method is considered can not solve the problem of morphologic diversity of derivations in language structures on low-resource agglutinative language. Tao *et al.* [22] proposes an attention-based

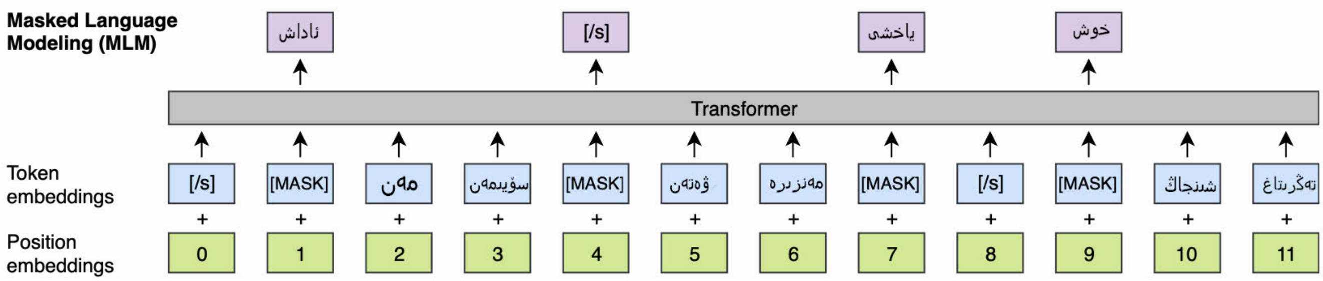

FIGURE 2. **Cross-lingual language model pre-training. The MLM objective is similar to the one in BERT Devlin *et al.* [7], but with continuous streams of text as opposed to sentence pairs.**

fine-tuning algorithm. With this algorithm, the customers can use the given language model and fine-tune the target model by their own data, but that does not capture different levels of syntactic and semantic information on different layers of a neural network. In this paper, we use a new fine-tuning strategy that provides a feature extractor to extract features and use these features for downstream text classification tasks.

III. METHODOLOGY

In this section, we will explain our methodology, which is also shown in Figure-1. Our training consists of four stages. We first pre-train a language model on a large scale cross-lingual text corpus. Then the pre-trained model is fine-tuned by the fine-tuning dataset on unsupervised language modeling tasks. The fine-tuning dataset is constructed by means of stem extraction and morpheme analysis on the downstream classification datasets. Moreover, we use an attention-based fine-tuning to build our classification model and uses discriminative fine-tuning to capture different types of information on different layers. Finally, train the classifier using target task datasets.

A. CROSS-LINGUAL MODEL PRE-TRAINING

Given a text sequence $X = (x_1, x_2, \ldots, x_T)$ and a sequence $Y = (y_1, y_2, \ldots, y_{T'})$ to denote the sequence of input context and target response respectively. The conditional probability $p(x_t|x_{0:t-1})$ can be modeled by a probability distribution over the vocabulary given linguistic context $x_{0:t-1}$. The context $x_{0:t-1}$ is modeled by neural encoder $f_{enc}(\cdot)$, and the conditional probability:

$$p(x_t|x_{0:t-1}) = g_{LM}\left(f_{enc}(x_{0:t-1})\right) \tag{1}$$

where $g_{LM}(\cdot)$ is prediction layer.

Given a huge cross-lingual corpus, we can train the entire network with maximum likelihood estimation (MLE). We have concatenated X and Y, then we can obtain prediction loss over the whole target response sequence as the loss function, the loss term for predicting the dialogue context X. The loss function:

$$L_{LM} = -\sum_{t=1}^{T} \log p(x_t|x < t) \tag{2}$$

In this paper, we aim to utilize $XLM - R$ to model the conditional probability. $XLM - R$ uses the same shared vocabulary to process all languages through Byte Pair Encoding (BPE) Sennrich *et al.* [18]. As shown in Lample *et al.* [14], this method greatly improves the alignment of embedding spaces across languages that share either the same alphabet or anchor tokens such as digits or proper nouns Smith *et al.* [20]. We learn the BPE splits on the concatenation of sentences sampled randomly from the monolingual corpora. Sentences are sampled according to a multinomial distribution with probabilities. And sentences are sampled according to a probable multinomial distribution $\{q_i\}_{i=1,2,3\ldots n}$, where:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha}, \tag{3}$$

where $p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$ and $\alpha = 0.3$. This distributed sampling method increases the number of tokens associated with low-resource languages and alleviates the bias to high-resource languages. In particular, this method prevents words in low-resource languages from being split at the character level.

As shown in Figure-2, $XLM - R$ utilizes a transformer model Vaswani *et al.* [23] to train with the multilingual MLM objective Devlin *et al.* ,Lample *et al.* [7], [14] using only monolingual data. $XLM - R$ samples streams of text from each language, is trained to predict the masked tokens in the input and apply subword tokenization directly on raw text data using sentence piece Kudo and Richardson [13] with a unigram language model Kudo [12]. Masking multi-head self-attention is utilized as the core technical operation to conduct representation learning:

$$Attention(Q, K, V) = Softmax\left(\frac{Q\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{4}$$

To extend the ability of the model to focus on different locations and to increase the representation learning capacity of subspaces for attention units, Transformer adopts the ''multi-head'' mode that can be expressed as:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{5}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^K) \tag{6}$$

B. LM FINE-TUNING BASED ON UKK CHARACTERISTICS

When we apply the pre-training model to text classification tasks in a target domain, a proper fine-tuning strategy is

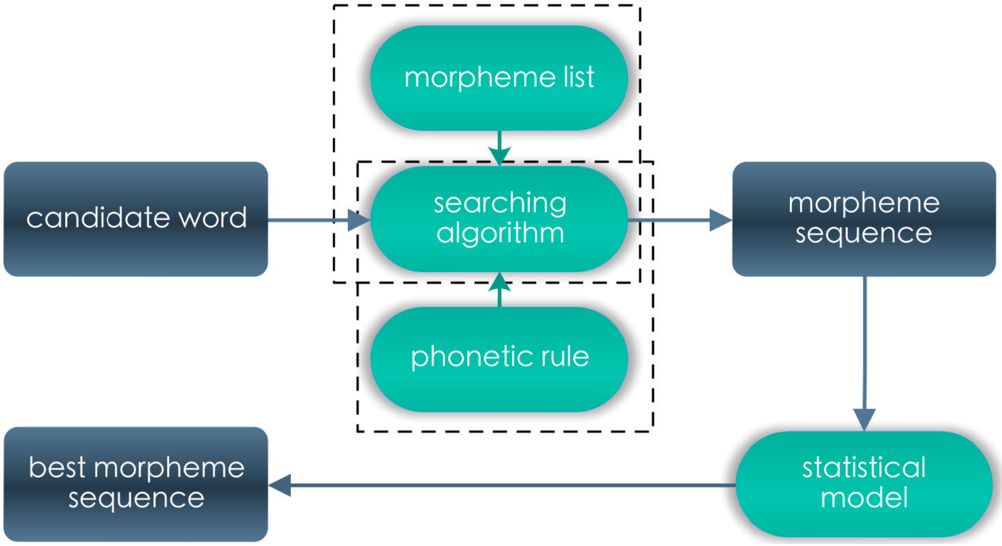

FIGURE 3. Morpheme segmentation flow chart.

desired. In this paper, we employ three fine-tuning methods as below.

1) FINE-TUNING DATASETS BASED ON MORPHEMIC ANALYSIS

UKK languages are agglutinative languages, meaning that words are formed by a stem augmented by an unlimited number of suffixes. The stem is an independent semantic unit while the suffixes are auxiliary functional units. Both stems and suffixes are called morphemes. Morphemes are the smallest functional units in agglutinative languages. Because of this agglutinative nature, the number of words of these languages can be almost infinite, and most of the words appear very rarely in the text corpus. Modeling based on a smaller unit like morpheme can provide stronger statistics hence robust models. The total number of suffixes in each of UKK languages is around 120. New suffixes may be created, but this is the typical case.

As shown in Figure-3, we use a semi-supervised morpheme segmenter based on the suffix set Ablimit *et al.* [2]. For a candidate word, this tool designs an iterative searching algorithm to produce all possible segmentation results by matching the stem-set and the suffix set. The phonemes on the boundaries change their surface forms according to the phonetic harmony rules when the morphemes are merged into a word. Morphemes will harmonize each other, and appeal to the pronunciation of each other. When the pronunciation is precisely represented, the phonetic harmony can be clearly observed in the text. An independent statistical model can be adopted to pick the best result from N-best results in the UKK text classification task.

We adopt this tool to train a statistical model using word-morpheme parallel training corpus, extraction, and greatly improve the UKK text classification task. which include 10,000 Uyghur sentences, 5000 Kazakhh sentences, and 5000 Kyrgyz sentences. We select 75% of them as the training corpus. The remainder is used as the testing corpus to execute morpheme segmentation and stem extraction experiments. We can collect necessary terms compose a less noise fine-tuning datasets by extracting stems in the UKK languages classification task. Then fine-tuning with $XLM - R$ on the fine-tuning datasets for better performance. As the examples given in Table-1 are shown below after morpheme analysis

ئىش+چى ئىش+خانا ئىش+تات ئوقۇ+ش ئوقۇ+غۇچى ئوقۇ+ت

The above sentence morphemes are segmented into the following

ئىشچى ئىشخانا ئىشتات ئوقۇش ئوقۇغۇچى ئوقۇت

There are 6 words in the above sentence divide into two groups, in this way, a stem can grasp the features of other words, and the feature will be greatly reduced.

2) DISCRIMINATIVE FINE-TUNING

Different layers of a neural network can capture different levels of syntactic and semantic information Howard and Ruder [11], Yosinski *et al.* [26]. Naturally, the lower layers of the $XLM - R$ model may contain more general information. Therefore, we can fine-tune them with assorted learning rates. Following Howard and Ruder [11], we use the discriminative fine-tuning method. We separate the parameters θ into $\{\theta^1, \ldots, \theta^L\}$, where θ^l contains the parameters of the l-th layer. Then the parameters are updated as follows:

$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta), \tag{7}$$

where η^l represents the learning rate of the $l - th$ layer and t denotes the update step. Following Sun *et al.* [21], we set the base learning rate to η_L and use $\eta^{k-1} = \xi \cdot \eta_k$, where ξ is a decay factor and less than or equal to 1. When $\xi < 1$, the lower layer has a slower learning rate than the higher layer.

TABLE 2. Statistics of the topic classification dataset.

Corpus	of Class	Average text length	Word Vocabulary	Morpheme Vocabulary	Morpheme-Word Vocabulary Ratio (%)
ug-topic	9	148.3	79,126	23,364	29.5%
kz-topic	8	130.9	68,334	20,600	30.1%
ky-topic	7	145.7	58,137	18,487	31.7%

TABLE 3. Statistics of the sentiment analysis datasets.

Corpus	of Class	Average text length	Word Vocabulary	Morpheme Vocabulary	Morpheme-Word Vocabulary Ratio (%)
ug-sen	3	23.6	8,791	2,794	31.1%
kz-sen	3	20.7	7,933	2,403	30.3%
ky-sen	3	21.3	7,385	2,274	30.8%

When $\xi = 1$, all layers have the same learning rate, which is equivalent to the regular stochastic gradient descent (SGD).

3) ATTENTION-BASED FINE-TUNING

For classification tasks, we utilize an attention-based encoder-decoder architecture. Encoder learns the contextualized features from inputs of the dataset. Then the hidden states over time steps denoted as $H = h_1, h_2, \ldots, h_T$, can be seen as the representation of the classified data, which are also the input of the attention layer. We use the self-attention to extract the relevant aspects from the input states since we do not have any additional information from the decoder. The alignment is computed as

$$u_t = \tanh(W_u h_t + bu) \tag{8}$$

for $t = 1, 2, \ldots, T$, where W_u and b_u are the weight matrix and bias term to be learned. Then the alignment scores are given by the following Softmax function:

$$\alpha_t = \frac{\exp(W_\alpha u_t)}{\sum_{i=1}^{T} \exp(W_\alpha u_t)} \tag{9}$$

The final context vector, which is also the input of the classifier, is computed by

$$c = \sum_{i=1}^{T} \alpha_t u_t \tag{10}$$

C. TEXT CLASSIFIER

For the classifier, we add two linear blocks with batch normalization and dropout, and ReLU activations for the intermediate layer and a Softmax activation for the output layer that calculates a probability distribution over target classes. Consider the output of the last linear block is S_o. Further, denote by $C = c_1, c_2, \ldots, c_M = XxY$ the target classification data, where $c_i = (x_i, y_i)$, x_i is the input sequence of tokens and y_i is the corresponding label. The classification loss we use to train the model can be computed by:

$$L_2(C) = \sum_{(x,y)\in C} \log p(y|x) \tag{11}$$

where

$$p(y|x) = p(y|x_1, x_2, \ldots, x_m) := softmax(W_{s_o}) \tag{12}$$

IV. DATASETS

A. DATA COLLECTION

We construct three low-resource agglutinative languages including Uyghur, Kazakh, and Kyrgyz nine datasets, datasets cover common text classification tasks: topic classification, sentiment analysis, and intention classification. We use the web crawler technology to collect our text data, and download from the Uyghur, Kazakh and Kyrgyz's official websites as well as other main websites.[1]

B. CORPUS STATISTICS

In this section, we introduce the detailed information of the corpus. We divided them into morpheme sequences and used morpheme segmentation tools to extract word stems. The method of subword extraction based on stem affix has achieved a good performance on the reduction of feature space. As a result, the vocabulary of morpheme is greatly reduced to about 30%, as shown in Table 2, Table 3 and Table 4. In addition, when the types and numbers of corpora increase, the accumulation of morphemes is only one-third of the accumulation of words.

1) TOPIC CLASSIFICATION

The corpus for the Uyghur language cover 9 topics: law, finance, sports, culture, health, tourism, education, science, and entertainment. Each category has 1,200 texts, resulting in a total of 10,800 texts. We name this corpus as `ug-topic`. The corpus for the Kazakh language cover 8 topics: law, finance, sports, culture, tourism, education, science, and entertainment. Each of them contains 1,200 texts, so there are 9,600 texts totally. We name this corpus as `kz-topic`. The corpus for the Kyrgyz language cover 7 topics: law, finance, sports, culture, tourism, education. Each category contains 1,200 texts (totally 8,400 texts). We name this corpus as `ky-topics`. The details are shown in Table-2.

2) SENTIMENT ANALYSIS

We construct 3 sentiment analysis datasets for three-category classification, namely positive, negative, and neutral. Each language is related to 900 texts and each category contains

[1] `www.uyghur.people.com.cn`, `uy.ts.cn`, `Kazakhh.ts.cn`, `www.hawar.cn`, Sina Weibo, Baidu Tieba and WeChat.

TABLE 4. Statistics of the intention classification datasets.

Corpus	of Class	Average text length	Word Vocabulary	Morpheme Vocabulary	Morpheme-Word Vocabulary Ratio (%)
ug-intent	5	18.9	12,651	3,997	31.6%
kz-intent	5	16.0	10,368	3,182	30.7%
ky-intent	5	15.4	11,343	3,720	32.8%

Topic	Law	Uyghur	دۆلەتنى قانۇن بويىچە ئىدارە قىلىشتا چىڭ تۇرۇش
		Kazakh	مەملەكەتتى زاڭمەن باسقارۋعا تاباندى بولۋ
		Kyrgyz	ماملەكەتتى زاكون بويۇنچا جونگو سالۇۇ
		English	Ensuring every dimension of governance is law-based
	Finance	Uyghur	ئامېرىكا ئىقتىسادىغا تەسىر كۆرسىتەمدۇ؟ COVID-19
		Kazakh	جاڭا ءتيىپتى وكپە ايدارشا ۆيرۋسى امەريكا ەكونوميكاسىنا ىقپال ەتەمە؟
		Kyrgyz	جاڭىچا تاجىسامان ۋىرۇس امەرىكا ىقتىسادىناتاسىر كوتسوتوبۇ
		English	Will the COVID-19 pandemic affect the US economy?
	Sports	Uyghur	.كوبى بىر ئۇلۇغ ۋاسكېتبول تەنھەرىكەتچىسى
		Kazakh	كوبە ۋلى باسكەتبول سپورتشىسى
		Kyrgyz	گوبى دەگەن بىر ۇلۇۇ ۋاسكىتبول چەبەرى
		English	Kobe is a great basketball player.
Sentiment	Positive	Uyghur	شىنجاڭنىڭ مەنزىرىسى سۈرەتتەك گۈزەل
		Kazakh	شينجياڭنىڭ كورىنىسى سۋرەتتەي كوركەم
		Kyrgyz	شىنجاڭدىن كورۇنۇشتورۇ سۇروتتوي كوركوم
		English	Xinjiang is a picturesque landscape
	Neutral	Uyghur	.بىز ئىلمىي ماقالە يېزىۋاتىمىز
		Kazakh	ءبىز علمي ماقالا جازىپ جاتىرمىز
		Kyrgyz	بىز ماقالاجازىپ جاتابىز
		English	We are writing a paper
	Negative	Uyghur	سىز نېمىشقا بويسۇنمايسىز؟
		Kazakh	سەن نەگە بويسىنبايسىڭ؟
		Kyrgyz	سىز نەگە مويۇن سۇنبايسىز
		English	Why are you disobedient?

TABLE 5. Example from the UKK datasets.

300 texts. We name these datasets as `ug-sen`, `kz-sen` and `ky-sen` as shown in Table-3.

3) INTENTION CLASSIFICATION

We construct 3 datasets of five-class user intent identification: news, life, travel, entertainment, and sports. Each language contains 200 texts. We name these datasets as `ug-intent`, `kz-intent` and `ky-intent` as shown in Table-4.

C. CORPUS EXAMPLES

In this section, we present some examples of various language categorization tasks. Different from Kazakhstan and Kyrgyzstan, in China, the Kazakh language used by the Kazakh people and the Kyrgyz language borrowed from the Arabic alphabet. The red keywords indicate the words that have the same meaning. The blue keywords represent their meaning in English. As shown in Table-5 for details.

V. EXPERIMENT

A. DATASETS AND TASKS

We evaluate our method on nine agglutinative language datasets which we construct of three common text classification tasks: topic classification, sentiment analysis, and intention classification. We use 75% of the data as the training set, 10% as the validation set, and 15% as the test set. For cross-lingual pre-training language models, we use the $XLM - R$ model loaded from the torch.Hub that It is trained on 2.5TB of CommonCrawl data, in 17 languages and uses a large vocabulary size of 95K. $XLM - R$ shows the possibility of training one language model for many languages while not sacrificing per-language performance.

B. BASELINES

We compare our method with the cross-lingual classification model *ULMFiT* Howard and Ruder [11], which introduces key techniques for fine-tuning language models, and *SemBERT* Zhang *et al.* [28], which is capable of explicitly absorbing contextual semantics over a BERT backbone. Moreover, we compare against the cross-lingual embedding model, namely *LASER* Artetxe and Schwenk [3], which uses a large parallel corpus. We also compare against *BWEs* Hangya *et al.* [10], a cross-lingual domain adaptation method for classification text.

C. HYPERPARAMETERS

In our experiment, we use the $XLM - R_{Base}$ model, which uses a $BERT_{Base}$ architecture Vaswani *et al.* [23] with a hidden size of 768, 12 Transformer blocks and 12 self-attention heads. We fine-tune the $XLM - R_{Base}$ model on 4 Tesla K80 GPUs and set the batch size to 24 to ensure that the GPU memory is fully utilized. The dropout probability is always 0.1. We use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Following Sun *et al.* [21], we use the discriminative fine-tuning method Howard and Ruder [11], where the base learning rate is $2e-5$, and the warm-up proportion is 0.1. We empirically set the max number of the epoch to 20 and save the best model on the validation set for testing.

D. RESULTS AND ANALYSIS

In this section, we demonstrate the effectiveness of our low-resource agglutinative language fine-tuning model. Our approach significantly outperforms the previous work on cross-lingual classification. Separately, the best results in the metric are bold, respectively.

As given in Table-6, Table-7, and Table-8, we show results for topic classification, sentiment analysis, and intention classification. Our *AgglutiFiT* outperform their cross-lingual and domain adaptation method. Pre-training is most beneficial for tasks with low-resource datasets and enables generalization even with 100 labeled examples when fine-tuning with fine-tuning dataset, our approach has a greater performance boost.

TABLE 6. Results on topic classification accuracy.

Model	ug-topic	kz-topic	ky-topic
ULMFiT	92.99%	92.93%	92.34%
LASER	83.19%	82.32%	82.13%
SemBERT	91.53%	90.12%	90.24%
BWEs	59.24%	59.12%	58.89%
AgglutiFiT	**96.45%**	**95.39%**	**94.89%**

TABLE 7. Results on sentiment analysis accuracy.

Model	ug-sen	kz-sen	ky-sen
ULMFiT	90.49%	90.39%	90.38%
LASER	74.32%	73.99%	72.13%
SemBERT	86.37%	88.47%	86.94%
BWEs	56.59%	56.39%	56.03%
AgglutiFiT	**92.81%**	**92.89%**	**92.23%**

TABLE 8. Results on intention classification accuracy.

Model	ug-intent	kz-intent	ky-intent
ULMFiT	90.97%	91.23%	91.13%
LASER	77.21%	77.89%	77.33%
SemBERT	89.79%	87.28%	89.13%
BWEs	57.50%	57.48%	57.39%
AgglutiFiT	**93.47%**	**93.81%**	**93.28%**

Compared with *ULMFiT*, we perform better on all three tasks, although *ULMFiT* introduces techniques that are key for fine-tuning a language model including discriminative fine-tuning and target task classifier fine-tuning. The reason can be partly explained as we adopt a less noisy datasets in the fine-tuning phase and attention-based fine-tuning which makes it possible to obtain a closer distribution of data in the general domain to the target domain. *LASER* obtains strong results in multilingual similarity search for low-resource languages, but we work better than *LASER* contribute to we use attention-based fine-tuning and different learning rates at a different layer, which allows us to capture more syntactic and semantic information at each layer, moreover, *LASER* has no learn joint multilingual sentence representations for UKK languages. Experimental results on methods *SemBERT* are lower than *AgglutiFiT* on account of lacking the necessary semantic role labels to embedding in the parallel, which leads to does not capture more accurate semantic information. *BWEs* is significantly lower than other models, we conjecture that the source language of method *BWEs* is English, which is quite different from the UKK languages in data distribution, more importantly, the datasets of UKK languages are too inadequacy to create good *BWEs*. Our three task experiments also show that using more high-quality datasets to fine-tune the results would be better.

1) LOW-RESOURCE AGGLUTINATIVE LANGUAGE MODEL

For low-resource agglutinative language, we can use languages with more resources with similar data, especially if their vocabularies are largely the same. The pre-training dataset CommonCrawl contains Uygur, Kazak, and Kyrgyz, and these languages are derived from Arabic which has a large amount of data, and many of these languages are the same in the Shared BPE dictionary. To do so, we train a UKK

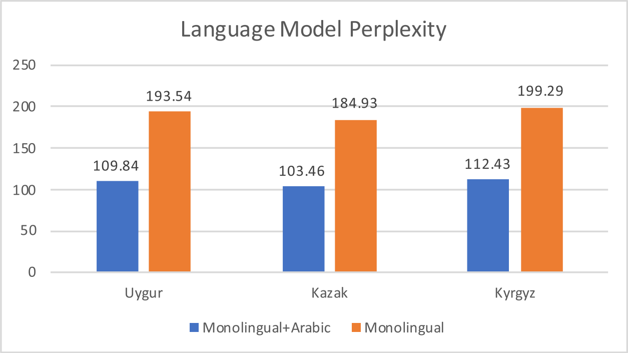

FIGURE 4. Results on language modeling.

language model on Wikipedia, together with additional data from either Arabic. The gains in perplexity from cross-lingual language modeling due to the n-grams anchor points that are shared across languages. Therefore, the cross-lingual language model can transfer the additional context provided by the Arabic monolingual corpora through these anchor points to improve the UKK language model. Figure-4 shows the perplexity of the UKK language model.

E. ABLATION STUDY

To evaluate the contributions of key factors in our method, we perform an ablation study as shown in Figure-5. We run experiments on nine corpora that are representative of different tasks, genres, and sizes.

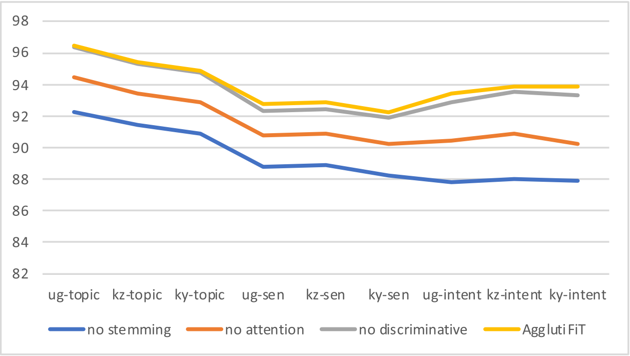

FIGURE 5. Explore the influence of important factors on accuracy.

1) THE EFFECT OF MORPHEMIC ANALYSIS

In order to gauge the impact of fine-tuning datasets quality, we compare the fine-tuning on the constructed fine-tuning datasets with the target task datasets without stem-word extraction. The experimental results show that the performance of all tasks is greatly improved by using our fine-tuning datasets. Stem is a practical unit of vocabulary. Stem extraction enables us to capture effective and meaningful features and greatly reduce the repetition rate of features.

2) THE EFFECT OF ATTENTION-BASED FINE-TUNING

As given in Figure-5, we can observe that by adding an attention fine-tuning, our model advances accuracies. Attention-based fine-tuning relies on a semantic between words that would influence the overall model performance. In order to see the effectiveness of the attention-based fine-tuning more clearly, we visualize the attention scores with respect to the input texts on Uyghur. The randomly chosen examples of visualization with respect to different classes are given in Figure-6, where darker color means higher attention scores.

3) THE EFFECT OF DISCRIMINATIVE FINE-TUNING

We compare with and without discriminative fine-tuning on the model. Discriminative fine-tuning improve performance across all three tasks, however, the role of improvement is limited, we still need a better optimization method to explore how discriminative fine-tuning can be better applied in the model.

4) FINE-TUNING DATASETS SIZE BEHAVIOR

From Figure-5, we also observe the larger the size of fine-tuning datasets, the higher the classification accuracy tends to be, which is more obvious in the comparison between different classification tasks. The reason is that the larger the fine-tuning datasets, the more likely it is to get the same data distribution as the target task.

5) IMPACT OF LM QUALITY

Regarding the language model pre-training, we have tried four different ways: 1) XLM-R on nine low-resource agglutinative datasets of three languages; 2) XLM-R model of the 15 languages. 3) XLM-R model of the 17 languages. and 4) XLM-R model of the 100 languages. As the results in Figure-7, we can have the following observations.

- When we have a large enough pre-trained dataset, the size of source data is not vital. This observation indicates the possibility that when the source dataset is large enough, the performance of language modeling is a significant factor in transfer learning.
- Pre-training on larger source datasets does not always improve downstream task performance. XLM-R with 17 languages is a subset of XLM-R with 100 languages and it is much smaller than XLM-R with 100 languages. But pre-training on XLM-R with 17 languages leads to the best performance among the three. We speculate that the UKK languages borrow the Arabic alphabet and that the XLM-R of 17 languages contains Arabic and has a higher weight. The language model train by it has more similar data distribution to the target task datasets. The vast majority of XLM-R with 100 languages is unnecessary and can affect its data distribution.

F. DISCUSSION AND FUTURE DIRECTIONS

When tasks with scarce amounts of labeled data and provide limited semantics for language representation, language model fine-tuning will be particularly beneficial. Transfer learning and especially fine-tuning for low-resource language

توكيو ئولىمپىك تەشكىللەش كومىتېتىنىڭ رەئىسى: توكيو ئولىمپىك تەنھەرىكەت مۇسابىقىسى بەلكىم بىۋاستە ئەمەلدىن قالدۇرۇلۇشى مۇمكىن

In English: Chairman of the Tokyo Olympic Organizing Committee: The Tokyo Olympics may be canceled directly.

(a) Sports

.شىنجاڭ گۈزەل جاي، بۇ يەرگە كېلىپ كەيپىياتىم خېلى ياخشى بولۇپ قالدى

In English: Xinjiang is a beautiful place and My mood feels very happy when I come here.

(b) Positive

.تۇرمۇش خۇددى نانغا ئوخشايدۇ، سىز نېمىگە ئېرىشىدىغانلىقىڭىزنى مەڭگۈ بىلمەيسىز

In English: Life is like bread, you never know what you will get.

(c) Life

FIGURE 6. Examples of attention visualization on Uyghur with respect to different classes.

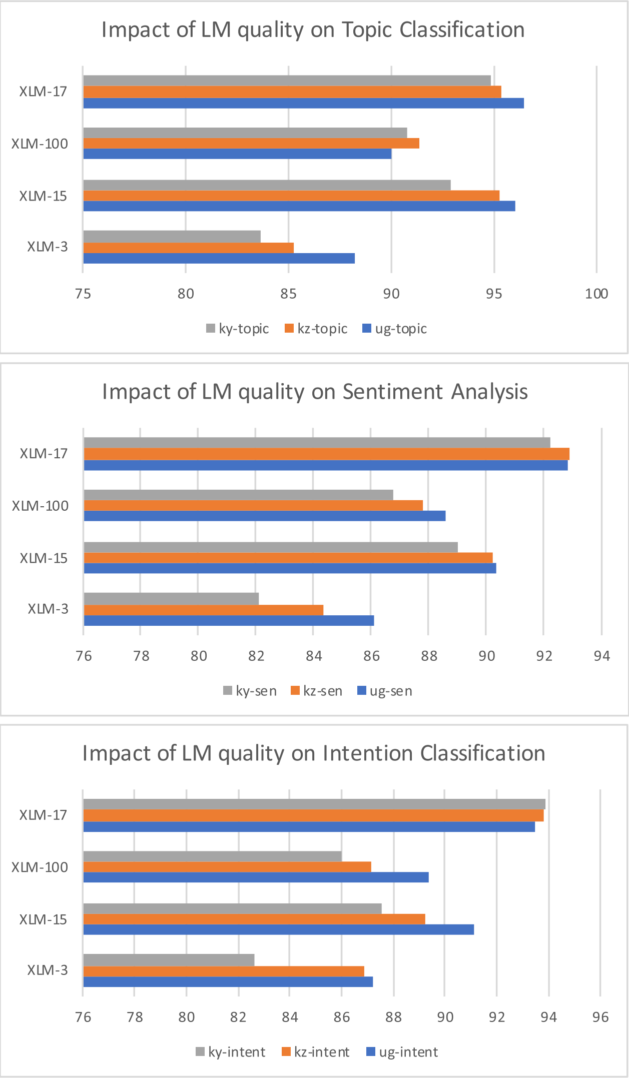

FIGURE 7. Explore the influence of important factors on accuracy.

is to be explored, many future directions are possible. Improve language model pre-training and fine-tuning and make them more robust is a potential direction. Language modeling can also be augmented with further pre-train with target domain data, fine-Tuning Strategies, and multi-task fine-tuning. Another direction is that explicit contextual semantics can be effectively integrated with state-of-the-art pre-trained language representation for even better performance improvement. More studies are required to better understand what knowledge a pre-trained language model captures, how these changes during the fine-tuning stage, and what features different tasks require.

VI. CONCLUSION

We propose *AgglutiFiT*, an effective language model fine-tuning method that can be applied to a low-resource agglutinative language classification tasks. This novel fine-tuning technique that via stem extraction and morphological analysis builds a low-noise fine-tuning dataset as the target task dataset to fine-tune the cross-lingual pre-training model. Moreover, we propose an attention-based fine-tuning strategy that better selects relevant semantic and syntactic information from the pre-trained language model to provide meaningful and favorable-to-use feature for downstream text classification tasks. We also use discriminative fine-tuning to capture different types of information on different layers. Our method significantly outperformed existing strong baselines on nine low-resource agglutinative language datasets of three representative low-resource agglutinative text classification tasks. We hope that our results will catalyze new developments in low-resource agglutinative languages task for NLP.

ACKNOWLEDGMENT

(Zhe Li and Xiuhong Li contributed equally to this work.) We thank the anonymous reviewers for their valuable feedback. This article was presented in part at The Nineteenth China National Conference on Computational Linguistics, CCL 2020. They initial conference paper does not address the problem of saturation. This manuscript addresses this issue and provides additional analysis of the trade-off between using more data.

REFERENCES

[1] M. Ablimit, T. Kawahara, A. Pattar, and A. Hamdulla, "Stem-affix based Uyghur morphological analyzer," *Int. J. Future Gener. Commun. Netw.*, vol. 9, no. 2, pp. 59–72, Feb. 2016.

[2] M. Ablimit, S. Parhat, A. Hamdulla, and T. F. Zheng, "A multilingual language processing tool for Uyghur, Kazak and Kirghiz," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 737–740.

[3] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 597–610, Mar. 2019.

[4] J. Bradbury, S. J. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural network," U.S. Patent 15 420 710, May 10, 2018.

[5] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7057–7067.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*. [Online]. Available: http://arxiv.org/abs/1911.02116

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[9] J. Eisenschlos, S. Ruder, P. Czapla, M. Kadras, S. Gugger, and J. Howard, "MultiFiT: Efficient multi-lingual language model fine-tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5706–5711.

[10] V. Hangya, F. Braune, A. Fraser, and H. Schütze, "Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 810–820.

[11] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*. [Online]. Available: http://arxiv.org/abs/1801.06146

[12] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 66–75.

[13] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2018, pp. 66–71.

[14] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," 2017, *arXiv:1711.00043*. [Online]. Available: http://arxiv.org/abs/1711.00043

[15] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever. (2019). Better language models and their implications. OpenAI Blog. [Online]. Available: https://openai.com/blog/better-language-models

[16] S. Ruder and B. Plank, "Strong baselines for neural semi-supervised learning under domain shift," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 1044–1054.

[17] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," 2019, *arXiv:1903.09460*. [Online]. Available: http://arxiv.org/abs/1903.09460

[18] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1715–1725.

[19] S. Shleifer, "Low resource text classification with ULMFit and backtranslation," 2019, *arXiv:1903.09244*. [Online]. Available: http://arxiv.org/abs/1903.09244

[20] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," 2017, *arXiv:1702.03859*. [Online]. Available: http://arxiv.org/abs/1702.03859

[21] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Cham, Switzerland: Springer, 2019, pp. 194–206.

[22] Y. Tao, S. Gupta, S. Krishna, X. Zhou, O. Majumder, and V. Khare, "FineText: Text classification via attention-based language model fine-tuning," 2019, *arXiv:1910.11959*. [Online]. Available: http://arxiv.org/abs/1910.11959

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[24] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 2019, *arXiv:1901.11196*. [Online]. Available: http://arxiv.org/abs/1901.11196

[25] Y. Xu, X. Qiu, L. Zhou, and X. Huang, "Improving BERT fine-tuning via self-ensemble and self-distillation," 2020, *arXiv:2002.10345*. [Online]. Available: http://arxiv.org/abs/2002.10345

[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[27] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451.

[28] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," 2019, *arXiv:1909.02209*. [Online]. Available: http://arxiv.org/abs/1909.02209

ZHE LI is currently pursuing the master's degree in software engineering with Xinjiang University. His research interests include text generation and social computing. He is a member of the Chinese Information Processing Society of China and the Chinese Association for Artificial Intelligence.

XIUHONG LI received the bachelor's degree from Jilin University, and the master's and Ph.D. degrees from Xinjiang University. She is currently an Associate Professor and a Master's Supervisor with Xinjiang University. She has presided over four key projects. She has published ten articles. Her research interests include natural language processing and computer architecture.

JIABAO SHENG is currently pursuing the master's degree in software engineering with Xinjiang University. Her research interest includes knowledge graph. She is a member of the Chinese Information Processing Society of China.

WUSHOUR SLAMU is currently a Professor and a Ph.D. Supervisor with Xinjiang University. He is an also an Academician of the Chinese Academy of Engineering. He is also an Executive Director of the Chinese Association for Artificial Intelligence. He has published more than 200 articles and presided over 65 key projects, including seven national 863 projects and one national 973 project. He has presided over the formulation of five international standards and more than 14 national standards. His research interest includes multilingual natural language processing. He received three National Science and Technology Progress Award.