











for  $t = 1, 2, \dots, T$ , 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 soft-max function:

$$\alpha_t = \frac{\exp(W_\alpha u_t)}{\sum_{i=1}^T \exp(W_\alpha u_i)} \quad (3)$$

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

$$c = \sum_{i=1}^T \alpha_i u_i \quad (4)$$

### 3.2 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  $C = c_1, c_2, \dots, c_M = XxY$  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:

$$L_2(C) = \sum_{(x,y) \in C} \log p(y|x) \quad (5)$$

where

$$p(y|x) = p(y|x_1, x_2, \dots, x_m) := \text{softmax}(W_{s_o}) \quad (6)$$

## 4 Datasets

### 4.1 Data Collection

We construct nine low-resource agglutinative language datasets including Uyghur, Kazakh, and Kyrgyz, these 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.<sup>1</sup>

### 4.2 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.

**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.

**Sentiment Analysis** We constructed 3 sentiment analysis datasets for three-category classification, namely positive, negative, and neutral. Each language is related to 900 texts and each category contains 300 texts. We name these datasets as `ug-sen`, `kz-sen` and `ky-sen` as shown in Table-3.

<sup>1</sup>[www.uyghur.people.com.cn](http://www.uyghur.people.com.cn), [uy.ts.cn](http://uy.ts.cn), [Kazakh.ts.cn](http://Kazakh.ts.cn), [www.hawar.cn](http://www.hawar.cn), Sina Weibo, Baidu Tieba and WeChat.

**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.

| 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 2: Statistics of the topic classification dataset.

| 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%                              |

Table 3: Statistics of the sentiment analysis 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%                              |

Table 4: Statistics of the intention classification datasets.

### 4.3 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.

## 5 Experiment

### 5.1 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.

### 5.2 Baselines

We compare our method with the cross-lingual classification model *ULMFiT* Howard and Ruder (2018), which introduces key techniques for fine-tuning language models, and *SemBERT* Zhang et al. (2019b), 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 (2019), which uses a large parallel corpus. We also compare against *BWEs* Hangya et al. (2018), a cross-lingual domain adaptation method for classification text. For cross-lingual pre-training language models, the *XLM-R* model used in this paper is loaded from the torch.Hub. *XLM-R* shows the possibility of training one model for many languages while not sacrificing per-language performance. It is trained on 2.5TB of CommonCrawl data, in 100 languages and uses a large vocabulary size of

|           |          |         |   |
|-----------|----------|---------|---|
| Topic     | Law      | Uyghur  | دۆلەتنى قانۇن بويىچە ئىدارە قىلىشتا چىڭ تۇرۇش                         |
|           |          | Kazakh  | مەملەكەتنى زاڭمەن باسقارۇغا تاياندى بولۇ                              |
|           |          | Kyrgyz  | ماملەكەتنى زاكون بويۇنچا جونگو سالۇۇ                                  |
|           |          | English | Ensuring every dimension of governance is <b>law-based</b>            |
|           | Finance  | Uyghur  | ئامېرىكا نىقتىسادىغا تەسىر كۆرسىتىدىمۇ؟ COVID-19                      |
|           |          | Kazakh  | جاڭا ەتتېپتى وكپە ايدارشا ۆيروسى امەرىكا مەكونومىكاسىنا نەقال مەتەمە؟ |
|           |          | Kyrgyz  | جاڭگچا تاجىسامان ۆبرۇس امەرىكا نىقتىسادىناتاسىر كوتسوتوبۇ             |
|           |          | English | Will the COVID-19 pandemic affect the US <b>economy</b> ?             |
|           | Sports   | Uyghur  | كوبى بىر ئۇلۇغ ۋاسكىتبول تەنھەرىكەتچىسى.                              |
|           |          | Kazakh  | كوبە ۇلى باسكەتبول سپورتشىسى  |
|           |          | Kyrgyz  | كوبى دەمگەن بىر ۇلۇۇ ۋاسكىتبول چەبەرى                                 |
|           |          | English | Kobe is a great <b>basketball player</b> .                            |
| Sentiment | Positive | Uyghur  | شىنجاڭنىڭ مەنزىرىسى سۈرەتتەك گۈزەل                                    |
|           |          | Kazakh  | شىنجاڭنىڭ كورننىسى سۈرەتتەي كوركەم                                    |
|           |          | Kyrgyz  | شىنجاڭدىن كورۇنۇشورۇ سۈرۈتتوي كوركوم                                  |
|           |          | English | Xinjiang is a <b>picturesque landscape</b>                            |
|           | Neutral  | Uyghur  | بىز نىلمى ماقالە يېزىۋاتىمىز.   |
|           |          | Kazakh  | ەبىز علمى ماقالا جازىپ جاتىرمىز                                       |
|           |          | Kyrgyz  | بىز ماقالا جازىپ جاتابىز  |
|           |          | English | We are writing a paper  |
|           | Negative | Uyghur  | سىز نېمىشقا بويسۇنمايسىز؟   |
|           |          | Kazakh  | سەن نەگە بويسىنبايسىڭ؟  |
|           |          | Kyrgyz  | سىز نەگە موپۇن سۇنبايسىز  |
|           |          | English | Why are you <b>disobedient</b> ?                                      |

Table 5: Example from the UKK datasets

250K. For the *ULMFiT* and *BWEs* model, we use English as the source language. *XLM-R* and *ULMFiT* are fine-tuned on target task datasets rather than the fine-tuning datasets that we built.

### 5.3 Hyperparameters

In our experiment, we use the *XLM-R<sub>Base</sub>* model, which uses a *BERT<sub>Base</sub>* architecture Vaswani et al. (2017) with a hidden size of 768, 12 Transformer blocks and 12 self-attention heads. We fine-tune the *XLM-R<sub>Base</sub>* 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. (2019), we use the discriminative fine-tuning method Howard and Ruder (2018), 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.

### 5.4 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.

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-

| 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 | <b>96.45%</b> | <b>95.39%</b> | <b>94.89%</b> |

Table 6: Results on topic classification 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 | <b>92.81%</b> | <b>92.89%</b> | <b>92.23%</b> |

Table 7: Results on sentiment analysis accuracy.

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* obtain 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 lack of the necessary semantic role labels to embedding in the parallel lead to does not capture more accurate semantic information. *BWEs* is significantly lower than other models, we conjecture is 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.

## 5.5 Ablation Study

To evaluate the contributions of key factors in our method, we perform an ablation study as shown in Figure-3. We run experiments on nine corpora that are representative of different tasks, genres, and sizes.

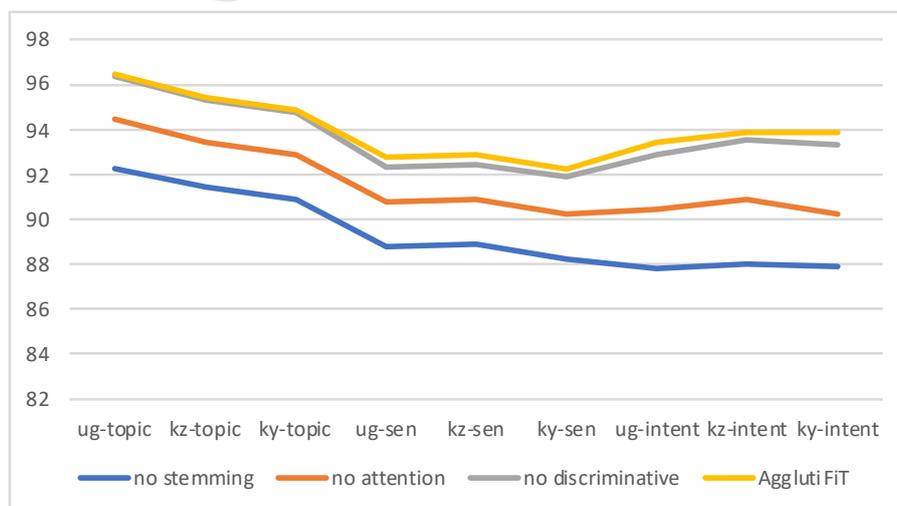


Figure 3: Explore the influence of important factors on accuracy

| 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 | <b>93.47%</b> | <b>93.81%</b> | <b>93.28%</b> |

Table 8: Results on intention classification accuracy.

**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.

**The effect of attention-based fine-tuning** As given in Figure-3, 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-4, where darker color means higher attention scores.



Figure 4: Examples of attention visualization on Uyghur with respect to different classes

**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.

## 6 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.

## 7 Acknowledgments

This paper support by Xinjiang University Ph.D. Foundation Initiated Project Grant Number 620312343, National Language Commission Research Project Grant Number ZDI135-96.

## References

- Mijit Ablimit, Tatsuya Kawahara, Akbar Pattar, and Askar Hamdulla. 2016. Stem-affix based uyghur morphological analyzer. *International Journal of Future Generation Communication and Networking*, 9(2):59–72.
- Mijit Ablimit, Sardar Parhat, Askar Hamdulla, and Thomas Fang Zheng. 2017. A multilingual language processing tool for uyghur, kazak and kirghiz. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 737–740. IEEE.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- James Bradbury, Stephen Joseph Merity, Caiming Xiong, and Richard Socher. 2018. Quasi-recurrent neural network, May 10. US Patent App. 15/420,710.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 0(0).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186.
- Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. Multifit: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5706–5711.
- Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *arXiv:1801.06146*.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.
- Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI Blog* <https://openai.com/blog/better-language-models>.
- Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.
- Gözde Gül Şahin and Mark Steedman. 2019. Data augmentation via dependency tree morphing for low-resource languages. *arXiv preprint arXiv:1903.09460*.
- Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation. *arXiv preprint arXiv:1903.09244*.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.
- Yunzhe Tao, Saurabh Gupta, Satyapriya Krishna, Xiong Zhou, Orchid Majumder, and Vineet Khare. 2019. Fine-text: Text classification via attention-based language model fine-tuning. *arXiv preprint arXiv:1910.11959*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

- Jason W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.
- Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*.
- Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.
- Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.

JCL 2020