Multilingual Transformer Language Model for Speech Recognition in Low-resource Languages

1st Li Miao  
Microsoft Corporation  
Mountain View, United States  
limia@microsoft.com

2nd Jian Wu  
Microsoft Corporation  
Seattle, United States  
jianwu@microsoft.com

3rd Piyush Behre  
Microsoft Corporation  
Mountain View, United States  
piyush@microsoft.com

4th Shuangyu Chang  
Microsoft Corporation  
Mountain View, United States  
shchang@microsoft.com

5th Sarangarajan Parthasarathy  
Microsoft Corporation  
Mountain View, United States  
sarangp@microsoft.com

Abstract—It is challenging to train and deploy Transformer Language Models (LMs) for hybrid speech recognition second pass re-ranking in low-resource languages due to (1) data scarcity in low-resource languages, (2) expensive computing costs for training and refreshing 100+ monolingual models, and (3) hosting inefficiency considering sparse traffic. In this study, we present a novel way to group multiple low-resource locales together and optimize the performance of Multilingual Transformer LMs in ASR. Our Locale-group Multilingual Transformer LMs outperform traditional multilingual LMs along with reducing maintenance costs and operating expenses. Further, for high-traffic locales where deploying monolingual models is feasible, we show that fine-tuning our locale-group multilingual LMs produces better monolingual LM candidates than baseline monolingual LMs.

Index Terms—multilingual language model, transformer language model, speech recognition

I. INTRODUCTION

Automatic Speech Recognition (ASR) usually involves two passes. The first-pass acoustic models and n-gram language models generate n-best hypotheses from the global search space [1]. In the second pass, Neural Network Language Models (NNLM) are widely used to re-rank the n-best hypotheses [2]. It has been demonstrated that re-ranking using NNLM is effective at reducing WER (Word Error Rate) [3], with Transformer language models producing state-of-the-art results in re-ranking [4].

Today our ASR system supports 100+ locales, but re-ranking is only applied to a few high-resource locales, even though we have proven the higher benefits of re-ranking for low-resource locales like Slovenian. The key challenges today are: (1) the low-resource locales’ training data is scarce, which limits our capacity to train the NNLM, (2) training and regularly refreshing 100+ monolingual re-ranking models, one for each locale, is computationally expensive, (3) hosting these monolingual models in production is prohibitively expensive and inefficient, as traffic can be sparse, but each model takes memory and compute to host across all of our Speech clusters. Multilingual Transformer language models [5]–[7] provide a very effective general solution to support ASR with pre-trained components and data sources that can be shared across multiple languages. However, multilingual models that are applied blindly may not always beat monolingual models, and limited serving resources prevent us from having a massive multilingual model for all 100+ locales. Our research found that combining multiple related locales can optimize performance, especially when dealing with low-resource locales.

As a result of this key insight, we have been able to overcome the challenges listed above. All the data available for a locale group can benefit locales with limited resources (scarce data). By training and maintaining just a few Transformer LMs per locale-group, we would be able to cover 100+ locales for re-ranking, and fewer overall 2nd pass LMs would result in a higher level of cluster scalability and hosting efficiency.

Moreover, our key finding regarding grouping low-resource locales has been found to work in other related domains as well, such as improving capitalization and punctuation in recognition outputs.

II. RELATED WORK

Multilingual/Cross-lingual. The effectiveness of sentence encoders’ generative pre-training was first demonstrated for English natural language processing [8]–[10]. Multiple approaches have since been proposed to extend it to multilingual/cross-lingual pretraining and show the success in transfer learning, such as mBERT [9], XLM [5], XLM-R [6], Unicoder [11], etc. Large amounts of unlabeled data from multiple languages are used to train these models, with the goal that low-resource languages can benefit from high-resource languages from shared vocabularies and underlying linguistic similarities. mBert trains a BERT model using Wikipedia corpora in 104 languages. XLM introduced a translation language model (TLM) in addition to masked language model (MLM), in which bilingual sentences are concatenated as inputs. To further improve the performance, Unicoder presents three new cross-lingual pre-training tasks, including cross-lingual word recovery, cross-lingual paraphrase classification and cross-lingual masked language model. XLM-R trains exclusively
with MLM objective on a huge multilingual dataset at an enormous scale.

Multilingual is also explored in ASR, primarily from an acoustic perspective. In [12], a massive multilingual acoustic model trained with more than 50 languages and more than 16,000 hours of audio is proven to improve recognition performance especially in low-resource languages. XLSR [13], a cross-lingual speech representation learning method is proposed by pre-training a single model from 56,000 hours raw waveform of speech in 53 languages.

Our work mainly focuses on Multilingual Language Model in ASR 2nd pass re-ranking, where a language model score is interpolated with 1st pass LM and AM to select the 1-best recognition candidate.

III. LOCALE-GROUP TRANSFORMER LM

Our proposed approach involves two steps. The first is to identify the underlying language group of the low-resource locale using our data-driven method, and the second is to process the low-resource locales’ data with shareable Byte Pair Encoding (BPE) tokens [14] and train the large-scale Locale-group Transformer Language Model. A group-based multilingual Transformer Language Model can be deployed whenever we lack the resources and hardware to develop and deploy individual models, providing significant improvements to Speech Recognition accuracy, maintenance, and cost reductions.

A. Language Group Identification

As one of three organizations selected to potentially partner with the European Parliament in 2020, Microsoft developed a real-time AI-based tool for live transcription and translation of debates. To identify the underlying language groups for 26 European languages, we proposed a two-step data-driven mechanism.

Firstly, we computed bi-lingual similarity score, which is a weighted measure of the number of like phonemes, words, phrases, and other similarity that exist in both locales. The Figure 1 shows an example of bi-lingual lexical similarity scores for 26 European languages, where higher score indicates closer linguistic relations between the languages. We observed code-switching/loanwords to be common, especially in English. In contrast, per our experiments Bulgarian (bg-BG) does not appear to be close to any other languages based on our collected data. We suspect that some of the data skew is due to filtered source data.

Secondly, we applied vector-based clustering to categorize similar languages together based on similarity score vectors. As shown in Table I, this mechanism successfully identifies language family like Balto-Slavic, the group 2, which contains Slovenian, Croatian, Slovak and Czech. Group 3 consists of most high-resource Germanic languages such as English and German, and Latin (Romance) languages like Italian, Spanish and French. In Group 4, Greek is supposed to have its own alphabet - the similarity with other languages is mainly due to code-switching or loanwords.

Table I: Language groups of 26 European locales

| Group | Languages |
|-------|-----------|
| 1     | nb-NO, sv-SE, fi-PI, da-DK |
| 2     | sl-SI, hr-HR, cs-CZ, sk-SK |
| 3     | en-all, es-ES, nl-NL, fr-FR, ro-RO, ca-ES, it-IT, pt-PT, pl-PL, de-DE |
| 4     | bg-BG, lv-LV, lt-IT, ga-IE, et-EE, el-GR, mt-MT, tr-TR |

Fig. 1. Example of lexical similarity scores across 26 European languages.
B. Shareable BPE Tokens

In our proposed approach, we process all languages with the same shared vocabulary created through Byte Pair Encoding (BPE) [14]. We provide several BPE format examples in Table II. This approach can greatly improve the token coverage with limited token set size and standardize the sub-word units across languages that share the same alphabet. For example, with 250K BPE tokens, we achieve almost a 100% coverage of 350M unique words across 26 languages.

| Locale   | Word-based Sent | BPE-based Sent |
|----------|-----------------|----------------|
| English  | ask consterna-| ask conster@@ |
|         | tion to my word| nation to my word list |
| Irish    | a naoi scoil | a naoi scoil nach |
|          | nach bhfuil | bhfuil seomra |
|          | seomra | acmhainne acu |
| Estonia  | asukoha nimated | asukoha ni@@ |
|          | on sofia ja | med on sofia ja |
|          | bulgaaria | bulgaaria |

TABLE II: Examples of text in BPE format

C. Transformer Language Model

1) Data Balance: We compile a language group training dataset with a balancing mechanism and train the Locale-group Multilingual Transformer Language Model. To ensure balanced data coverage for multiple regions within the same family, we sample sentences with multinomially distributed probabilities $\{q_i\}_{i=1}^N$, similar as how sentences are sampled in [5].

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

2) Locale-group Model Training: Using balanced language family data, we train the Locale-group Multi-lingual Transformer Language Model, which is similar to the structure used in [8].

As part of the training process, we record the valid loss and perplexity for each locale and family of languages. We found that the average loss minimum is within the range of the individual locale’s loss minimum, which implies that the Locale-group Multilingual Transformer Language Model converges for all locales within the identified language group.

3) General & Masked Fine-tuning: In natural language processing (NLP) and machine translation (MT), fine-tuning a pre-trained language model has become the de facto standard for transfer learning [10] [15]. If we fine-tune the multilingual model to a target language, our ability to serve efficiently and reduce computation costs will be compromised. Nevertheless, fine-tuning is also worth exploring, as (1) knowledge transfer will be tested if we are able to achieve better results with fine-tuned monolingual models than those trained from scratch, along with a unified training recipe, and (2) certain high-traffic regions will separate themselves from low-resource groups as more data is collected and traffic increases, allowing us to deploy monolingual models with adequate resources.

In general fine-tuning, we reserve the pre-trained multilingual model parameters, feed the target locale’s data in, and train for several more epochs until it converges to a new minimum.

Masked Fine-tuning is designed to force the model to tune to a specific locale. For BPE tokens that do not exist in the target locale, we freeze the token embedding updates and set the prediction score to a very small negative number, so the token loss will be close to 0, and avoid paying attention to irrelevant tokens. This process is illustrated in Figure 2.

IV. Experiment

A. Dataset

1) Train: Across 26 European locales, we sampled 5B sentences from our in-house text data corpus with an average sentence length of 12. Our own text-normalization pipeline preprocesses all text data into lexical format. Locales with low resources are up-sampled according to Section 2.3.1. Shareable 250K BPE tokens are trained using the same data.

2) Test: Test audio data used for word-error-rate reductions (WERR) measurement of each locale consists primarily of dictations and spontaneous conversations. Minimum coverage per locale is 10K sentences.

B. Model

We experiment with configurations described in Table IV.
| Config          | Description                                                                 |
|-----------------|------------------------------------------------------------------------------|
| Mono:LSTM       | baseline monolingual Long short-term memory (LSTM) LM                        |
| Mono:Trans      | monolingual Transformer LM                                                  |
| Multi:Trans-All | all-languages-together Multilingual Transformer LM                           |
| Multi:Trans-Group| Locale-group Multilingual Transformer LM                                     |
| Multi+FT:Trans-All | fine-tuned all-languages-together Multilingual Transformer LM                |
| Multi+FT:Trans-Group | fine-tuned Locale-group Multilingual Transformer LM                        |
| Multi+MFT:Trans-Group | masked fine-tuned Locale-group Multilingual Transformer LM            |

| TABLE IV: Model Configs |

We train a baseline 1x1024:512 LSTM LM for each language, where 1 is the number of layers, 1024 is the dimensionality of the LSTM state, 512 is the dimensionality of the embedding and also the output dimensionality of the projection layer.

In addition, we trained one multilingual all-languages-together, four multilingual locale-group, and 26 monolingual Transformer language models with the same shareable 250K BPE token set, same data distribution and similar model configuration. These Transformer language models consist of 12 transformer layers, where each transformer layer contains 4096 feedforward dimensions with 16 heads. The warm-up is set to increase the learning rate gradually to improve the convergence of Transformer LMs [16].

We also applied general fine-tuning and masked fine-tuning as described in section III-C3.

C. Results

In this work, we mainly report word-error-rate reductions (WERR) on several low-resource locales: Croatian (hr-HR), Slovenian (sl-SI), Slovak (sk-SK), Lithuanian (lt-LT), Latvian (lv-LV) and Romanian (ro-RO). Language groups are described in section III-A.

As shown in Table III, we observe 3.34% average WERR improvement because of architecture upgrade from LSTM to Transformer, and more parameters. However, it is challenging to deploy those monolingual Transformer LMs as we discussed three limitations in the beginning.

On the other hand, the Locale-group Transformer LM provides a good solution considering the deployment restrictions. Firstly, with one Transformer LM to serve multiple locales in the same group, we can achieve 3.12% average WERR gain compared with the LSTM baseline. Secondly, the locale-group model in general outperforms the all-data-together model by 2.19% when the implicit language similarity information is included, and limited model capacity can spare more attention to the learning of underlying linguistic patterns in similar languages instead of being distracted by irrelevant signals.

Even more, if we can allocate enough resources to train and deploy dedicated Transformer LM for low-resource locales, we can achieve more with masked fine-tuning. Compared with the monolingual LSTM baseline, our multilingual Locale-group model with masked fine-tuning can provide an additional 4.64% average WERR.

V. DISCUSSION

A. BPE Token Size

We also trained models with 64k shareable BPE tokens to evaluate the impact of BPE token size. WERR on individual locales varies, but compared with 250K, the 64K BPE based Locale-group models generally regress around 2.48%. When we do masked fine-tuning, the gap is reduced to 0.66%. Our hypothesis is that the average text sequence length are elongated by smaller BPE token size, therefore brings in challenges to learn generic sequence patterns under the same number of model parameters.
B. Hosting Efficiency

A key motivation for exploring this idea was to improve production efficiency. We need to deploy our speech service globally in tens of clusters. When it comes to hosting costs, monolingual and general multilingual are at opposite ends of the spectrum. Deploying over 100 monolingual models everywhere is quite resource-intensive and often wasteful because traffic can be sparse in some locales. However, general multilingual models oversimplify this, resulting in WERR regression for some locales. At this scale, natural data imbalances can also be challenging to overcome. Multilingual models are typically able to defeat monolingual models by increasing their complexity, like adding more Transformer layers. Individual request latency is also affected by this, which is undesirable. By using locale-group multilingual models, we achieve a better WERR than either monolingual or general multilingual models in most cases, requiring us to deploy only a few locale-group models across our speech clusters. Certain high-traffic locations will have the option to deploy monolingual or fine-tuned monolingual models.

C. Parameter Tuning

We haven’t tuned dropout and hyper parameters extensively for all the models. With parameter tuning, we expect to achieve even greater gains, and will explore this in the future.

D. Application in other areas

In our hybrid ASR system, there are many other areas where we use monolingual models or general multi-lingual models. Using this locale-grouping technique, we investigated the applicability of capitalization models. A similar result was achieved - locale-group capitalization models outperformed monolingual and general multilingual models.

VI. CONCLUSION

The neural network language model (NNLM) is an essential module in hybrid automatic speech recognition to achieve optimal recognition accuracy. Our work proposes a general and scalable method of training and deploying large-scale Locale-group Transformer NNLM models for ASR in low-resource languages, with significant accuracy improvements and reduced development and maintenance costs. Moreover, our fine-tuning experiments demonstrate that locale-grouping helps create better monolingual models.

REFERENCES

[1] D. Yu and L. Deng, Automatic Speech Recognition - A Deep Learning Approach. Springer, October 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/automatic-speech-recognition-a-deep-learning-approach/

[2] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. L. Roux, Z. Meng, and S. Watanabe, “Multi-channel speech recognition: Lstms all the way through,” in CHiME 2016 - 4th International Workshop on Speech Processing in Everyday Environments, September 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/multi-channel-speech-recognition-lstms-all-the-way-through/

[3] X. Liu, Y. Wang, X. Chen, M. J. F. Gales, and P. C. Woodland, “Efficient lattice rescoring using recurrent neural network language models,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4908–4912.

[4] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, “Language modeling with deep transformers,” CoRR, vol. abs/1905.04226, 2019. [Online]. Available: http://arxiv.org/abs/1905.04226

[5] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” CoRR, vol. abs/1901.07291, 2019. [Online]. Available: http://arxiv.org/abs/1901.07291

[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,” CoRR, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[7] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?” CoRR, vol. abs/1906.01502, 2019. [Online]. Available: http://arxiv.org/abs/1906.01502

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., “Improving language understanding by generative pre-training,” 2018.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” CoRR, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] J. Howard and S. Ruder, “Fine-tuned language models for text classification,” CoRR, vol. abs/1801.06146, 2018. [Online]. Available: http://arxiv.org/abs/1801.06146

[11] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, and M. Zhou, “Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks,” 2019.

[12] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, “Massively multilingual asr: 50 languages, 1 model, 1 billion parameters,” arXiv preprint arXiv:2007.03001, 2020.

[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” CoRR, vol. abs/2006.13979, 2020. [Online]. Available: https://arxiv.org/abs/2006.13979

[14] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” CoRR, vol. abs/1508.07909, 2015. [Online]. Available: http://arxiv.org/abs/1508.07909

[15] Y. Tang, C. Tran, X. Li, P. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, “Multilingual translation with extensible multilingual pretraining and finetuning,” CoRR, vol. abs/2008.00401, 2020. [Online]. Available: https://arxiv.org/abs/2008.00401

[16] M. Popel and O. Bojar, “Training tips for the transformer model,” CoRR, vol. abs/1804.00247, 2018. [Online]. Available: http://arxiv.org/abs/1804.00247