Regularization Advantages of Multilingual Neural Language Models for Low Resource Domains

Anonymous Author(s)

Abstract

Neural language modeling (LM) has led to significant improvements in several applications, including Automatic Speech Recognition. However, they typically require large amounts of training data, which is not available for many domains and languages. In this study, we propose a multilingual neural language model architecture, trained jointly on the domain-specific data of several low-resource languages. The proposed multilingual LM consists of language specific word embeddings in the encoder and decoder, and one language specific LSTM layer, plus two LSTM layers with shared parameters across the languages. This multilingual LM model facilitates transfer learning across the languages, acting as an extra regularizer in very low-resource scenarios. We integrate our proposed multilingual approach with a state-of-the-art highly-regularized neural LM, and evaluate on the conversational data domain for four languages over a range of training data sizes. Compared to monolingual LMs, the results show significant improvements of our proposed multilingual LM when the amount of available training data is limited, indicating the advantages of cross-lingual parameter sharing in very low-resource language modeling.

1 Introduction

Language modeling (LM) is a fundamental task in natural language processing which has been an essential component of several language and speech applications, most notably in machine translation (Koehn et al., 2003) and speech recognition (ASR) (Deoras et al., 2011). More recently, neural language models have been shown useful for transferring knowledge from large corpora to downstream tasks, including text classification, question answering and natural language inference (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018).

However, training an effective neural LM typically requires large amount of written text in the required language and domain, which may not be readily available for many rare domains and languages. Even when there is a pre-trained language model trained on out-of-domain data available, its fine-tuning on very small validation sets is prone to over-fitting. This issue is especially challenging for domain-specific tasks such as conversational text of low resource languages (Ragni et al., 2016).

A common approach to avoid over-fitting when dealing with limited amount of data is regularization. Recently, Merity et al. (2018) demonstrated the effectiveness of regularization methods for a multi-layer LSTM language model, named AverageSGD Weight-Dropped LSTM (AWD-LSTM). Simultaneously, Melis et al. (2018) explored the effect of extensive parameter tuning in a multi-layer LSTM model, showing the competitive performance of LSTM to other proposed network architectures for language modeling.

Beside regularization, parameter sharing between models in various domain/languages facilitates knowledge transfer across the models, and can be especially helpful in very low-resource scenarios. Multilingual training of neural networks has grown in the last few years for various language processing tasks such as machine translation (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017) and document classification (Ferreira et al., 2016; Pappas and Popescu-Belis, 2017). Parameter sharing has also been studied across different tasks, such as document summarization (Zhou et al., 2018), reading comprehension (Nishida et al., 2018), and question answering (Sachan and Xing, 2018).

In this study, we propose a new approach for multilingual neural language modeling along the direction of Ragni et al. (2016). Our proposed multilingual architecture consists of a stacked

DRAFT

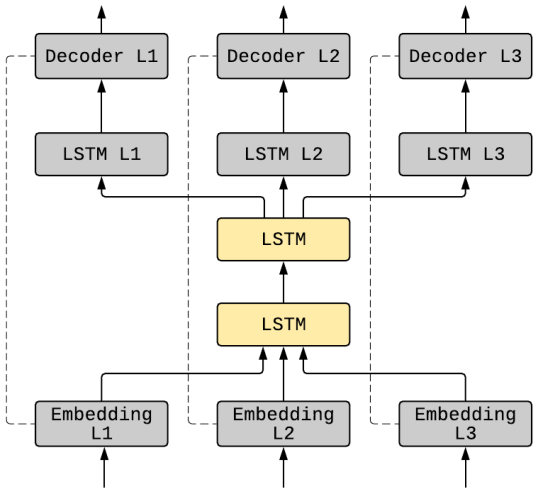


Figure 1: The proposed multilingual architecture.

LSTM model with three layers, where the first two layers are shared across multiple languages, and the last layer is language-specific (Figure 1). The first two LSTM layers capture the common patterns across languages, while the last layer learns language specific subtleties. In contrast to Ragni et al. (2016), our proposed model does not need per language fine-tuning because its language-specific and language-agnostic features are simultaneously trained. Similar to previous work on multilingual training (Pappas and Popescu-Belis, 2017; Firat et al., 2016), every language in our model has separate input and output layers and hence a separate loss function. We integrate the regularization methods proposed by Merity et al. (2018) in our multilingual LM, resulting in a highly regularized multilingual LM, where each language benefits from the shared parameters, and, hence, the training data of other languages.

We evaluate the proposed models in terms of perplexity on conversational data of four low-resource languages, namely Creole, Tagalog, Turkish, and Swahili, against state-of-the-art monolingual AWD-LSTM model. We study the effect of multilingual training on very low resource settings, by limiting the number of words in the training set of each language, and evaluate the performance of models over a range of available training data. Our main contributions are:

- We propose a highly regularized multilingual language model for low-resource domains.
- We demonstrate its superiority and stability against strong monolingual baselines when the amount of training data is very limited.

The benefit of our regularized multilingual model is most pronounced on the Swahili language, the corpus of which is generally much smaller than the rest of the languages (∼230K words). In this case, the multilingual training outperforms monolingual even on the full resource setting.

2 Highly Regularized Multilingual Language Modeling

The proposed architecture is illustrated in Figure 1, for three languages L1,L2,L3. Firstly, a language-specific word embedding maps the input of the given language to its embedding vectors. Secondly, two layers of LSTMs with shared parameters capture the common patterns across the languages, followed by a language-specific LSTM for modeling language-specific characteristics. Thirdly, a language-specific decoder applies a linear transformation followed by the softmax function, and outputs the predicted probability distribution $y_t^{(l)}$ at timestep t over the vocabulary of the given language l. The weights of the input embedding and decoder for each language are tied.

The proposed model is selected based on its superior performance in our preliminary evaluation results among other possible parameter sharing architectures with three LSTM layers, namely sharing all, only first/last, or last two layers.

2.1 Multilingual Training

For training, we use a training objective similar to Firat et al. (2016) and Pappas and Popescu-Belis (2017); we use a joint multilingual objective that facilitates the sharing of a subset of parameters for each language $\theta_1, \ldots, \theta_M$ of our stacked LSTMs:

$$\mathcal{L}(\theta_1, \ldots, \theta_M) = -\frac{1}{Z} \sum_{t}^{N_e} \sum_{l}^{M} \mathcal{H}(y_t^{(l)}, \hat{y}_t^{(l)}) \quad (1)$$

where M is the number of languages, $Z = M \times N_e$, N_e is the epoch size, and $\mathcal{H}$ is the cross-entropy loss between the ground-truth words and the predicted ones. Note that the sentence order in each language is preserved above and that the overall loss is back-propagated through the network, updating both language-specific and language-independent parameters. The sentences are processed in a cyclic fashion for the languages which have lesser number of sentences; once the last sentence of the text corpus is processed for that language, the next sentence that is processed is the beginning one. The joint objective $\mathcal{L}$ is minimized with respect to the parameters $\theta_1, \ldots, \theta_M$.

DRAFT

Table 1: Number of words in the splits obtained from the Babel collection (Gales et al., 2014).

	Creole	Swahili	Tagalog	Turkish
Train	417539	237677	526528	494715
Valid.	87418	6584	5158	67090
Test	84358	6584	5158	72106

2.2 Regularization Techniques

Pursuing the work of Merity et al. (2018) on regularizing neural language models, we apply Weight-Dropped LSTM, Variational Dropout, Embedding Dropout, and Variable Length Backpropagation Sequences. Merity et al. (2018) also use Activation Regularization (AR) and Temporal Activation Regularization (TAR), two weight regularization terms added to the loss function. In our multilingual LM architecture, we add these terms to the loss function of each language while their values are divided by M.

3 Experiments

3.1 Data and Settings

For evaluation we use the conversational data of four low-resource languages, namely Creole, Swahili, Tagalog, and Turkish, taken from language packs released within IARPA Babel program (Gales et al., 2014). Every language pack contains a training and a development set, containing text of audio transcripts. We use the development set for testing and split the given training set into training and validation sets, where the size of the validation set is the same or close to the testing set. The statistics of our training/validation/test sets are reported in Table 1. We apply punctuation removal and set the texts to lower case. Similar to Ragni et al. (2016) for each language, we replace 25% of the vocabulary words with the lowest frequencies with $<unk>$.

To measure the effect of data scarcity, in addition to training the models on the full texts, we also train the models on limited parts of training texts. In these scenarios, for every language only a specific number of words (based on a threshold) are used for training, selected from the beginning of the training text of that language. We train the LM models over a range of such threshold values from 20K to 400K as well as on the full training text for the languages. It should be noted that when the text of a language is restricted, the multilingual LM still have access to the full training texts of other languages.

3.2 Model Configuration

We set the hyper-parameters as suggested by Merity et al. (2018) for the Penn Tree Bank language modeling as follows: embedding size of 512, LSTM hidden layer size of 1150, initial learning rate to 30, batch size to 20, maximum number of epochs to 200, and sequence length of 70. The dropout rates for input, output, variational, embedding, and weight dropouts are set to 0.65, 0.4, 0.3, 0.1, 0.5, respectively. The alpha and beta values of the AR and TAR methods are set to 2 and 1. Lastly, we tie the weights between input embeddings and softmax weights for all the models (per language), and use Stochastic Gradient Descent (Bottou, 2010) for optimization. Our code is made available with the submission and will be made publicly available upon publication.

The word embeddings of all the models are initialized randomly and updated during training. We also examined initializing with pre-trained cross-lingual word embeddings using the vectors provided by Lample et al. (2018) as well as creating new ones based on the unsupervised method proposed by Ammar et al. (2016). In both cases, we observe the same LM performance to the models with randomly initialized embeddings, since the embeddings lose their cross-lingual alignment properties as they are being updated.

3.3 Baselines

We compare our multilingual model with two monolingual LSTM models:

- `mono-LSTM`: an out-of-the-box monolingual three-layer LSTM with regularization only through dropouts on the input of output layers of LSTM units.

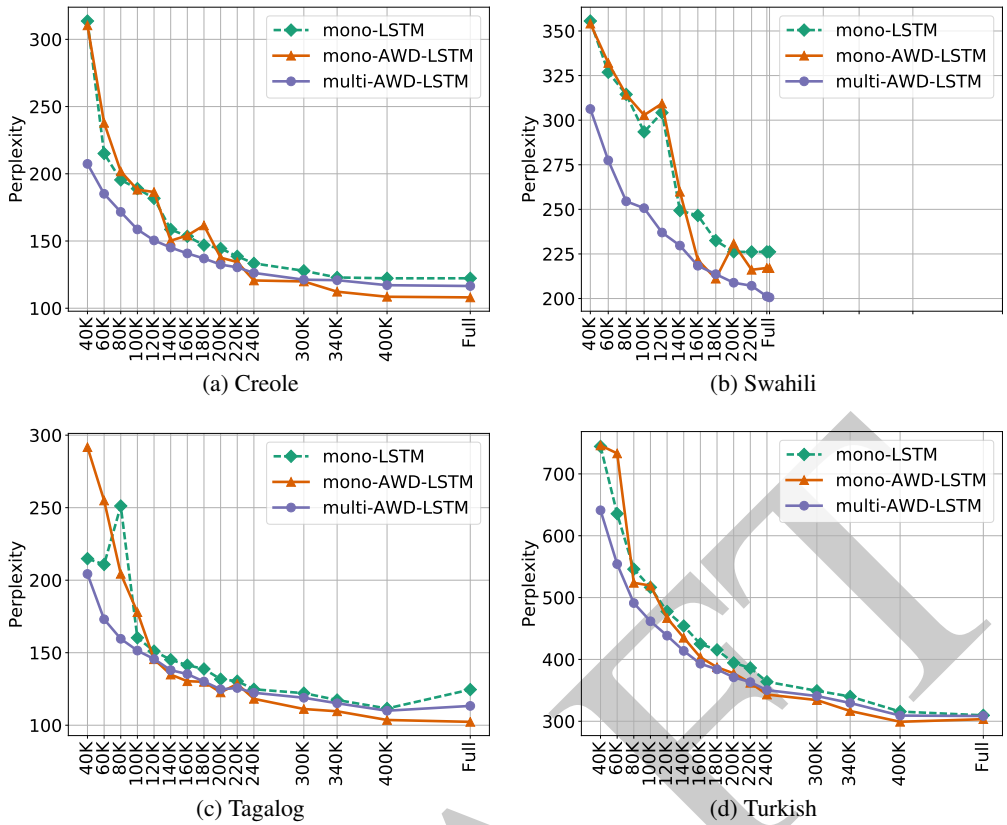
- `mono-AWD-LSTM`: the same model as above but with additional regularization methods, namely Weight-Dropped LSTM, Variational Dropout, Embedding Dropout, Variable Length Backpropagation Sequences, AR, and TAR, as in Merity et al. (2018).

We note as `multi-AWD-LSTM` our multilingual model with the same regularization as `mono-AWD-LSTM`.

4 Results and Discussion

Figure 2 shows the perplexity curves of the three language models on the test sets of four

DRAFT



(a) Creole (b) Swahili

(c) Tagalog (d) Turkish

Figure 2: Perplexity of the three LM models on the test data of four low-resource languages. The X axis shows the size of the training data (number of words), used to learn the models.

Table 2: Perplexity of the three LMs.

	Model	Number of words in training data				
		40K	100K	200K	300K	FULL
Cr	LSTM	313.64	188.96	144.42	127.84	122.27
	AWD-LSTM	310.41	188.18	137.62	**119.92**	**108.06**
	Multilingual	**207.38**	**158.62**	**132.49**	121.23	116.52
Sw	LSTM	355.59	293.43	226.16	-	226.16
	AWD-LSTM	354.14	302.83	230.81	-	217.14
	Multilingual	**306.27**	**250.67**	**208.87**	-	**201.20**
Tl	LSTM	214.85	160.28	131.73	122.12	124.56
	AWD-LSTM	291.74	177.96	**122.63**	**111.18**	**102.32**
	Multilingual	**204.29**	**151.46**	124.77	118.94	113.31
Tr	LSTM	744.41	516.49	394.46	349.26	309.58
	AWD-LSTM	746.09	519.40	377.47	**334.25**	**303.36**
	Multilingual	**641.14**	**461.69**	**371.09**	340.79	308.51

languages when varying the training set size. The results for five training size thresholds are also reported in Table 2, and the full perplexity scores for each threshold are provided in Appendix A. As can be observed, `mono-AWD-LSTM` and `mono-LSTM` perform similarly weak in very low resource settings, while `multi-AWD-LSTM` outperforms both by a large margin in all four languages. When training data is sufficiently large, `mono-AWD-LSTM` achieves the best performance among all models, where `multi-AWD-LSTM` performs on par with or similar to it.

Based on our observations, the threshold below which our multilingual model performs better is between 100K to 250K words, depending on the language. The Swahili language in our collection is such a case, as its training data only consists of ∼240K words. In this case, `multi-AWD-LSTM` outperforms all other models even on the full resource setting.

These results show a consistent improvement for our multilingual language models in transferring knowledge across languages when the training data is limited. We attribute this improvement to the parameter sharing at the lower layers, which allows the model to capture language-independent patterns, facilitating better generalization.

5 Conclusion

We proposed a novel multilingual language model for handling low-resource domains and languages. Compared to a state-of-the-art monolingual LM, AWD-LSTM, on four languages, the proposed multilingual LM achieves significant improvements consistently in very low resource scenarios, namely when the size of training data is between 100K to 250K words. The results highlight the benefits of cross-lingual transfer learning for a more effective generalization of LMs on extreme data scarcity scenarios.

DRAFT

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Physica-Verlag HD.

Anoop Deoras, Tomáš Mikolov, Stefan Kombrink, Martin Karafiát, and Sanjeev Khudanpur. 2011. Variational approximation of long-span language models for lvcsr. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Spoken Language Technologies for Under-Resourced Languages*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Melvin Johnson et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics (TACL)*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Gbor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJNLP)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

A Ragni, E Dakin, X Chen, MJF Gales, and KM Knill. 2016. Multi-language neural network language models. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL)*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

DRAFT