Leveraging Multilingual Neural Language Models for On-Device Natural Language Understanding

Huy Tu*
North Carolina State University
hqtu@ncsu.edu

Samridhi Choudhary
Alexa AI, Amazon
samridhc@amazon.com

Hamidreza Saghir
Alexa AI, Amazon
hssaghi@amazon.com

Ross McGowan
Alexa Speech, Amazon
rosmcgow@amazon.com

ABSTRACT

Learning cross-lingual word representations is an effective approach for developing multilingual models. In this work, we lay the groundwork and present preliminary results on learning cross-lingual representations appropriate for deployment to edge devices. Specifically, we learn cross-lingual representations using multilingual language models and use these to seed different parts of a Neural Natural Language Understanding (NLU) model that is designed for an on-device deployment. This is followed by fine-tuning on supervised datasets to perform NLU. We present systematic experiments that show up to 10% relative improvement in NLU error rates across different dataset sizes. We perform experiments for three languages and confirm the effectiveness of using cross-lingual representations for both multi- and monolingual NLU models for low-resource languages that have scarce task-specific annotated data. The effectiveness of cross-lingual transfer indicated by these preliminary experiments will be used in our future work for creating multilingual on-device models.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

1 INTRODUCTION

The recent rise in the application of deep learning approaches to natural language understanding (NLU) [11, 14, 17] has led to an increasing popularity of efficient, feature-rich commercial voice assistants (VAs) like Amazon Alexa, Google Assistant and more. Many VAs are extending their capabilities from cloud to edge-computing [5, 7, 12] in order to improve latency and provide extended availability in case of poor connections. This has lead to a plethora of research on model compression and quantization approaches with the goal to find the most accurate and efficient model that can be directly deployed on devices [1, 9, 10, 13, 15]. Here we investigate multilingual representations as a complementary approach to increase the accuracy of on-device multilingual models without increasing their footprint relative to monolingual models.

NLU is the task of extracting the semantics from user queries. NLU in VAs typically consists of the following sub-tasks: domain classification (DC), intent classification (IC) and named entity recognition (NER). Despite state of the art performance of transformer-based architectures for most NLU tasks [6], their sheer size and intensive computational requirements precludes their deployment to edge devices.

Additionally, in real-world scenarios, VA customers would often like to be able to interact with their devices in multiple languages which often results in multiple on-device model artifacts and multilingual systems that ultimately may increase on-device footprint.

The authors in [9] present a carefully selected model design for a multi-domain, multi-task neural NLU (NNLU) model (also shown in Figure 1) that consists of a Bi-LSTM based shared encoder followed by task-specific classification heads. They also describe a series of compression and quantization approaches that can be applied to this model architecture to produce a low-footprint model compatible for on-device use cases. In this work, we seek to extend this model architecture to multi-language scenarios in order to improve performance in a multilingual scenario without increasing model size.

We systematically experiment with leveraging cross-lingual representations [2–4, 8, 16], learned from training multilingual language models [16], to improve both mono- and multilingual NNLU performance. In particular, we apply the approach proposed in [16] to train Bi-LSTM based mono- and multilingual LMs using data from three different languages. These pre-trained LMs are then used to seed different parts of the on-device NNLU model from [9] that is then fine-tuned to perform the component NLU tasks - DC, IC and NER. We further study the efficacy of this approach for low-resource languages in both mono- and cross-lingual settings. Our contribution in this work is to answer the following three research questions: Given a carefully designed model architecture for on-device use,

(1) Can we improve the performance using monolingual pretraining?
(2) Can we employ crosslingual embeddings to improve the performance of monolingual NLU for low-resource languages?
(3) Can we employ crosslingual embeddings to improve the performance of multilingual NLU across languages?

We present preliminary results of our investigation in different data scenarios and settings. We emphasize that, while promising, our work is still in progress, so we limit our experiments to the backbone model in [9] prior to applying model compression.

In the following sections, we define our methodology, experimental setup, datasets, and our experiments. We then evaluate, compare, and discuss our results, and propose next steps for creating multilingual NNLU models on the edge.

*Work done as an intern at Amazon.

2 METHODOLOGY

The on-device compatible model architecture presented in [9] is shown in figure 1. It is a multi-task model with a Bi-LSTM based shared encoder that receives an embedding for each word in the sentence. There are three task-specific heads that take in the shared layer representations - two sentence classification heads for DC and IC and a sequence classification layer for NER. We use Conditional Random Field (CRF) for sequence classification. The model is trained jointly on the combined task-specific losses. We use the standard cross-entropy loss for DC and IC, and CRF loss for NER.

In order to learn cross-lingual representations, we train a neural language model (LM) (Figure 2) that has a similar architecture as the shared Bi-LSTM encoder of the NNLU model. The LM is trained on next-word prediction. We perform experiments in two settings: (i) *Monolingual* and (ii) *Cross-Lingual*. In the *monolingual* setting, the LM is trained using the utterances of one language. Once trained, different components of the monolingual NNLU model are initialized by the learned LM weights. This boot-strapped NNLU is then fine-tuned on the task-specific dataset to perform the component NLU tasks. The results from this stage help us identify whether the self-supervised representations learned by the LM help the supervised NLU model. In the *cross-lingual* setting, we train a multilingual LM by using utterances from three different languages. The weights learned by this multilingual LM are used to bootstrap both multilingual NNLU and the low-resource mono-lingual NNLU. The results from this stage help us demonstrate the effectiveness of cross-lingual representations for boosting the performance of the NNLU model for languages where training data is scarce.

2.1 Dataset and Evaluation Metrics

We use a random snapshot of a curated set of utterances taken from the live production traffic of a commercial VA system. The data is first processed so that users are not identifiable ("de-identified"). These are annotated with utterance text (audio transcription), domain, intent, and the slot tags that are supported on-device. We prepare two datasets for our experiments: (i) Monolingual Dataset; and (ii) Multilingual Dataset. The monolingual set is a random subsample of the English production traffic (10M training, 1M testing and 50K validation). The multilingual dataset contains subsamples for three languages - English (EN), French (FR) and Spanish (ES).We create smaller subsamples of datasets from each language as proxies for low-resource languages. Therefore, for each language, we create three datasets of varying sizes - 50K, 500K, & 3.5M utterances for training; 5K, 50K, & 350K for testing and 2.5K, 10K, & 50K for validation, respectively. Each subset testset is used to check for general model fit, but all the models are finally tested on a common evaluation set consisting of a set of live utterances, per language. For multi-lingual models, average performance across the component language evaluation sets is reported.

Evaluation Metrics - We use the following metrics to evaluate or models:

(1) *Interpretation Error Rate (IRER)* : Ratio of the number of incorrect interpretations to the total number of utterances. An incorrect interpretation is the one where either the domain, intent or the slots are wrong. This can also be seen as an exact match error rate and is the strictest metric.
(2) *Intent Classification Error Rate (ICER)*: Ratio of number of incorrect intent predictions to the total number of utterances.
(3) *Domain Classification Error Rate (DCER)*: Ratio of number of incorrect domain predictions to the total number of utterances.
(4) *Slot Error Rate (SER)*: Ratio of number of incorrect slot predictions to the total slots.

2.2 Experimental Setup

Monolingual Experimental Setup - Both the LM and the NNLU models are trained using the monolingual dataset. LM training just uses the utterance text, while NNLU training uses the NLU task annotations.

We train, evaluate and report the results for the NNLU model in the following settings:

(1) **Baseline NNLU**: This model uses pre-trained embeddings [1] for the embedding layer and randomly initialized weights for the Bi-LSTM layers. (Shown as NNLU-Rand-LSTM-PreTrain-*W* in Table 1.)
(2) **LM-initialized NNLU 1**: This model uses either pre-trained or random embeddings for the embedding layer. However, the Bi-LSTM weights are taken from the learned LM Bi-LSTM weights. (Shown as NNLU-LM-LSTM-Rand|PreTrain-*W* in table 1.)
(3) **LM-initialized NNLU 2**: In this model we have both the embedding and the Bi-LSTM weights initialized from the learned LM model. (Shown as NNLU-LM-LSTM-LM-*W* in table 1.)

Cross-Lingual Experimental Setup - We consider two language pairs for these experiments - {ES, FR}-EN. A multilingual LM is trained using these language pairs. We train both mono- and multilingual NNLU (MNNLU) models[2] and test the pre-training effects on both models by initializing NNLU components with multilingual LM weights. We train, evaluate, and report the results for NNLU and MNNLU model in the following settings:

(1) **Cross-Lingual with NNLU**: Both the embedding and LSTM layers of the NNLU model are either initialized randomly, from a monolingual LM or from a multilingual LM.
(2) **Cross-Lingual with MNNLU**: Both the embedding and LSTM layers of the MNNLU model are initialized either randomly or with the multilingual LM.

3 RESULTS

Table 1 reports the results from the monolingual experimental setup (Section 2.2). Owing to the private nature of our data, we report the relative change in each evaluation metric compared to the baseline model for all the experiments. Using pre-trained LM weights helps the NNLU model as indicated by improvements across all of the evaluation metrics for different component initialization settings. Initializing the NNLU LSTM layers with the LM LSTM weights helps the model outperform the baseline regardless of how its word embeddings are initialized. However, initializing both the embedding and LSTM weights with the learned LM weights leads to the best performance in terms of relative improvements across

[1] We use Fast-Text embeddings that have been pre-trained on VA traffic utterances
[2] Bi-lingual in this setting

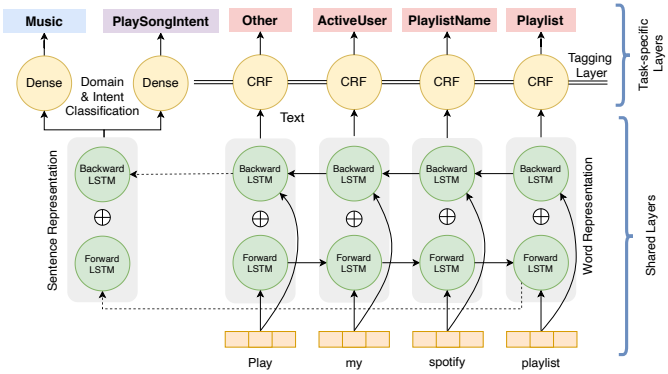

Figure 1: Multi-Domain, Multi-Task on-device Neural NLU Model from [9].

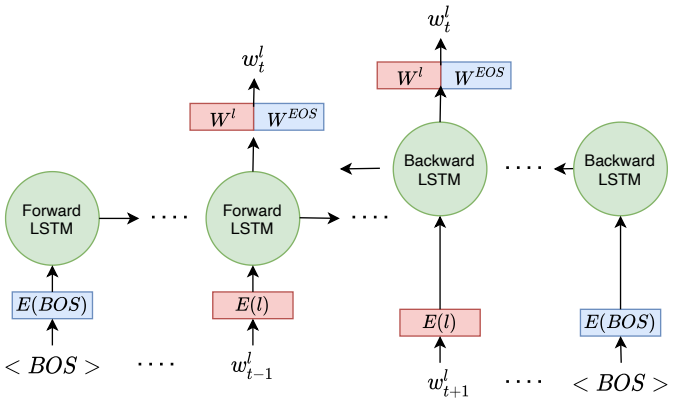

Figure 2: Multilingual Neural Language Model from [16]. The components in blue are shared between multiple languages while those in red are language specific.

Model	Rel. IRER Change (%)	Rel. ICER Change (%)	Rel. DCER Change (%)	Rel. SER Change (%)
NNLU-Rand-LSTM-PreTrain-W	X	X	X	X
NNLU-LM-LSTM-Rand-W	-6.36	-7.16	-7.76	**-8.67**
NNLU-LM-LSTM-PreTrain-W	-7.65	-9.92	-8.19	-7.61
NNLU-LM-LSTM-LM-W	**-9.74**	**-10.84**	**-9.70**	**-10.44**

Table 1: Results for the mono-lingual (English-EN) experimental setup. The results are reported for the EN evaluation set, as relative change % to the baseline model that has randomly initialized LSTM layers and pre-trained embeddings.

all evaluation metrics. This indicates that maximum knowledge transfer happens when both the LSTM and the embedding weights are used from the pre-trained LM weights for the NNLU model.

Table 2 shows that performance improvements for monolingual NLU models for both ES and FR are higher with the multilingual LM representations as compared to using monolingual LM representations in most cases. This indicates that cross-lingual representations prove to be more effective than monolingual representations in these languages. A similar effect is seen in the multilingual NLU setting, as shown in Table 3. There is a positive transfer from cross-lingual representations toward multilingual NLU model since all the reported errors are relatively smaller than the baseline multilingual NLU model. Moreover, using cross-lingual representations from a multilingual LM trained on the particular language pair leads to higher relative improvements across all evaluation metrics in most cases. Furthermore, for both monolingual and multilingual NLU, as the size of the dataset decreases, the improvements obtained from using cross-lingual embeddings drastically increases.

Overall, the improvements shown in both Table 2 and 3 show that learning cross-lingual representation by multilingual LM in small data regimes **boosts** performance for (1) monolingual NLU tasks (specifically NNLU model) on non-English Languages and (2) multilingual NLU tasks (MNNLU model).

4 CONCLUSION AND FUTURE WORK

In this work, we lay the groundwork for building an on-device multilingual NLU model by experimenting with cross-lingual representations on a previously established on-device NNLU architecture. First, we validated that, consistent with the literature, monolingual pre-training is an effective strategy for improving on-device NNLU model performance. Also, maximum knowledge transfer happens when both the LSTM and the embedding weights are seeded by corresponding components from the pre-trained LM.

We then showed that cross-lingual representations can help improve NLU performance in both monolingual and multilingual settings. In particular, we show that the performance improvements for non-English monolingual NLU models are higher when they are seeded with cross-lingual representations, as compared to seeding with monolingual representations. Our multilingual experiments ({ES, FR} - EN) suggest that in the scarcer the available data-resources, the more beneficial it is to use cross-lingual representations. This is a particularly encouraging finding since in many practical scenarios, data resources for less commonly used languages are very limited. Additionally, for on-device models, memory and computational power restrictions also preclude the use of large pre-trained models. The multilingual neural NLU model that we introduced here, which has the same architecture as a proven on-device model, can address these challenges.

Model	Rel. IRER Change (%)	Rel. ICER Change (%)	Rel. DCER Change (%)	Rel. SER Change (%)
		ES		
NNLU-Random (50K, 500K, 3.5M)	X	X	X	X
NNLU-LM (50K)	-3.92	-4.04	-11.36	-8.68
NNLU-LM (500K)	-2.53	-2.64	-4.05	-5.26
NNLU-LM (3.5M)	-6.36	-8.18	-7.76	-7.61
NNLU-MLM (50K)	-10.82	-10.03	-12	-14.38
NNLU-MLM (500K)	-5.26	-5.28	-3.37	-6.36
NNLU-MLM (3.5M)	-0.19	-1.32	-1.37	-1.7
		FR		
NNLU-Random (50K, 500K, 3.5M)	X	X	X	X
NNLU-LM (50K)	-4.95	+0.47	+6.99	-11.12
NNLU-LM (500K)	-1.97	-4.29	-3.11	-2.93
NNLU-LM (3.5M)	-0.23	-0.41	-0.47	-1.81
NNLU-MLM (50K)	-3.01	-8.76	-6.99	-7.62
NNLU-MLM (500K)	-3.94	-5.72	-5.10	-6.45
NNLU-MLM (3.5M)	-0.23	-0.41	-0.47	-1.81

Table 2: Results for Cross-Lingual Experimental Setup - Using language-specific LM and the {ES, FR}-EN multilingual language model (MLM) representations for mono-lingual NNLU models for non-English languages. The results are reported as relative change %, for a common, language specific, evaluation set across each dataset-size setting.

Model	Rel. IRER Change (%)	Rel. ICER Change (%)	Rel. DCER Change (%)	Rel. SER Change (%)
		EN-ES		
NNLU-Random (50K, 500K, 3.5M)	X	X	X	X
MNNLU-MLM (50K)	**-10.05**	**-13.31**	**-8.86**	**-10.47**
MNNLU-MLM (500K)	-8.44	-5.25	-5.56	-13.17
MNNLU-MLM (3.5M)	-0.77	-1.47	-1.40	-1.77
		EN-FR		
NNLU-Random (50K, 500K, 3.5M)	X	X	X	X
MNNLU-MLM (50K)	**-7.89**	**-12.8**	**-7.18**	**-10.43**
MNNLU-MLM (500K)	-3.66	-2.78	-4.14	-3.62
MNNLU-MLM (3.5M)	-1.87	-2.22	-3.91	-3.85

Table 3: Cross-Lingual Experimental Setup - Using the {ES, FR}-EN multilingual language model (MLM) representations for multilingual NNLU model. The results are reported as average relative change % across common, component language specific, evaluation sets across each dataset-size setting.

Although our current results are promising, we note that we have only experimented with three languages with different data sizes as proxies for low-data-resource scenarios. These are exploratory results that encourage the use of cross-lingual representations in an on-device, multi-task NLU model. Future work will include expanding to more languages, including actual low-data-resource languages, and investigating the effect of compression and quantization approaches on cross-lingual representations.

REFERENCES

[1] Anish Acharya, Rahul Goel, Angeliki Metallinou, and Inderjit Dhillon. 2019. Online embedding compression for text classification using low rank matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*,

Vol. 33. 6196–6203.

[2] Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3874–3884. https://doi.org/10.18653/v1/N19-1388

[3] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges.

[4] Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. 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)*. Association for Computational Linguistics, Hong Kong, China, 1538–1548. https://doi.org/10.18653/v1/D19-1165

[5] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190* (2018).

[6] 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* (2018).

[7] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, et al. 2016. Personalized speech recognition on mobile devices. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5955–5959.

[8] David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of Transfer in Multilingual Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8093–8104. https://doi.org/10.18653/v1/2020.acl-main.720

[9] Kanthashree Mysore Sathyendra, Samridhi Choudhary, and Leah Nicolich-Henkin. 2020. Extreme Model Compression for On-device Natural Language Understanding. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*. International Committee on Computational Linguistics, Online, 160–171. https://doi.org/10.18653/v1/2020.coling-industry.15

[10] Vikas Raunak. 2017. Simple and effective dimensionality reduction for word embeddings. *arXiv preprint arXiv:1708.03629* (2017).

[11] Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[12] Alaa Saade, Alice Coucke, Alexandre Caulier, Joseph Dureau, Adrien Ball, Théodore Bluche, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, et al. 2018. Spoken language understanding on the edge. *arXiv preprint arXiv:1810.12735* (2018).

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[14] Ruhi Sarikaya, Geoffrey E Hinton, and Anoop Deoras. 2014. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 4 (2014), 778–784.

[15] Raphael Shu and Hideki Nakayama. 2017. Compressing word embeddings via deep compositional code learning. *arXiv preprint arXiv:1711.01068* (2017).

[16] Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3113–3124. https://doi.org/10.18653/v1/P19-1300

[17] Puyang Xu and Ruhi Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 136–140.