Phoneme Level Language Model for Sequence Based Low Resource ASR

Siddharth Dalmia, Xinjian Li, Alan W Black, Florian Metze
sdalmia@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Language Technologies Institute

Carnegie Mellon University

At a Glance

Problems in low resource languages -

- Collection and cleaning of data can be expensive.
- Often we find data which is either out-of-domain or have very little in-domain data.
- This results in **bad word language models**.

Solution: Train LMs on smaller units; characters, phonemes, etc. In this work, we show that with **phoneme language models** -

- We can do parameter sharing (**Multilinguality**).
- Better adaptation to a new language (**Crosslingual Adaptation**).
- Decode with a targeted lexicon to get unseen words (**Domain Robustness**)

Phoneme Level Language Models (PLMs)

The idea of PLMs is simple -

- Instead of training on characters, convert the words of any language into their corresponding IPA symbol. [1]
- Use the phonemic transcriptions sequence of "characters" to train a standard charLM. [2]

Multilinguality using PLMs

- Making **one model for all languages** could not be imagined with word LMs as the sharing of words across language is quite low.
- PLMs present a unique opportunity to **share parameters and transfer knowledge from other languages**.

We train the model on the phonemic transcription of each language by keeping a shared phoneme space but individual word boundary, `<space>`. We apply masked training approach to train the model -

$$\texttt{ind} = \texttt{where}(\texttt{lang\_mask} = \texttt{True})$$
$$\texttt{logits} = \mathbf{W}_{out}\texttt{LSTM}(\texttt{Emb}(x_1, \ldots, x_{t-1})) + \mathbf{b}_{out}$$
$$\texttt{sparse\_softmax} = \texttt{softmax}(\texttt{gather}_{\texttt{ind}}(logits))$$

We can see that with Multilingual PLMs, **we use 6 times fewer parameters** with almost the same performance.

	PLM Small	PLM Large	Multi-PLM Large
# Params	~0.4M×6	~4.5M×6	**~4.6M**
Javanese	3.91	3.80	3.80
Tagalog	3.62	3.43	3.46
Turkish	3.53	3.36	3.38
Kazakh	3.02	2.89	2.89
Swahili	3.63	3.44	3.50
Zulu	4.18	3.95	4.00

Table: PLM (Small and Large) and Multi-PLM (Large) perplexities for different languages in the training set.

Crosslingual Adaptation of PLMs

- **Multilingual PLMs show better adaptations to a new language** than training a new language model.
- Bigger improvements on smaller amounts of data.

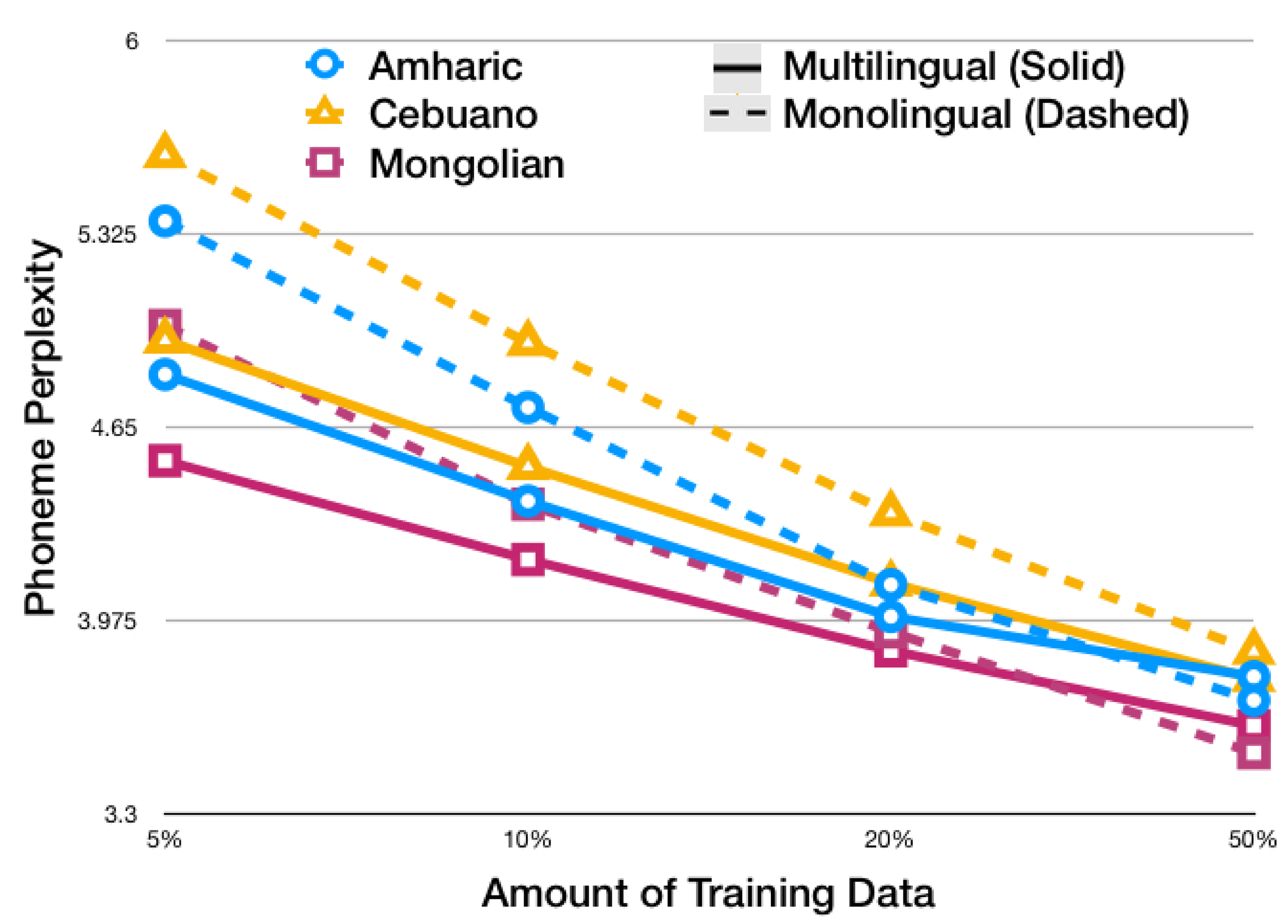

Figure: PPL after adaptation of Multi-PLM to target languages on different amounts of data. Multi-PLM outperforms PLM for small amounts of training data.

Targeted Decoding with PLMs

CTC based acoustic models typically use WFST based decoding [3] or open vocabulary charLM decoding [2].

Open vocabulary decoding is not reliable in low resource languages as it leads to incorrect OOV words. For example, in Zulu, open vocabulary decoding gives **9% incorrect OOV words**.

We propose a modification to better use our PLMs -

- **Targeted decoding** - We decode paths that only produce a valid word.
- This allows us to **control the words produced by the ASR model**.
- **Better than CLM (6% avg)**. Almost as good as WFST.

Babel Languages	WFST	CLM Based Decoding	PLM
Cebuano	57.1	71.1	67.9
Mongolian	60.5	84.3	**59.0**
Amharic	57.2	64.8	**57.6**
Javanese	65.7	68.4	**64.8**
Tagalog	55.7	58.0	**55.8**
Kazakh	57.8	64.2	**61.3**
Turkish	56.9	**58.5**	59.4
Swahili	61.2	**50.7**	**50.8**
Zulu	65.2	75.3	**63.7**

Table: % WER on each languages using different kinds of decoding strategies.

Decoding under Low Resource Conditions

We **study the robustness of our model for typical low resource challenges**–

- **Little training data:** We see PLM based decoding is better than WFST based decoding. This is due to bad word probabilities estimates.
- **Domain Mismatch:** To test domain mismatched conditions we train our model on Bible text and test on the in-domain conversational data from Babel dataset.

We can see that **PLM based decoding outperforms WFST based decoding**, showing its capability of generating words outside language model training data by just using a targeted lexicon.

Babel Languages	WFST Based Decoding	PLM
Cebuano	86.2	**79.8**
Javanese	93.1	**80.8**
Tagalog	83.4	**68.9**
Kazakh	78.3	**72.5**

Table: % WER using different decoding strategies on LMs trained on the Bible text.

Conclusion

In this work we propose a **phoneme level language model** and show

- ✓ With Multilingual PLMs we use **6 times fewer parameters**.
- ✓ Multilingual PLMs **adapt better to a new language** in very low resource settings.
- ✓ Using PLMs with targeted decoding, affords significant **gains over open-vocabulary decoding**
- ✓ **Outperforms WFST in low resource conditions**.

Acknowledgement

This project was sponsored by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114.

References

[1] D. R. Mortensen, *et al.*, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018.

[2] T. Zenkel, *et al.*, "Comparison of decoding strategies for CTC acoustic models," in *Interspeech*. ISCA, 2017.

[3] Y. Miao, *et al.*, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174.