Sequence-based Multi-lingual Low Resource Speech Recognition

Siddharth Dalmia

Ramon Sanabria

Florian Metze

Alan W Black

sdalmia@cs.cmu.edu

Carnegie Mellon University
Language Technologies Institute

Language Universal
Multi-lingual Speech Recognition

- Many speech sounds are shared across languages.

- These sounds can be mapped to a set of language independent target units called International Phonetic Alphabet (IPA).

- However, these units are not always language agnostic.

Models trained with these units are prone to such errors!

Multiple target
Multi-lingual Speech Recognition

- What is an alternative?
  - We can train a shared acoustic model with multiple targets, one for each language.
  - The model learns to implicitly share the hidden space without the need of grounding them to language universal phonemes!

Multiple target
Multi-lingual Speech Recognition

- Lots of advantages
  - Removes the need of having language universal phoneme set. They can even be characters of a language!
  - We can use any of the existing datasets without preparing new labels or creating mappings of phonemes!

Previous Explorations

- Shared phone set with target language adaptation (T. Schultz et al, 2001; N. T. Vu et al, 2014)

- Language independent features like articulatory features (S. Stuker et al, 2003)

- Multilingual training of DNNs (A. Ghoshal et al, 2013; G. Heigold et al, 2013)

- Language-independent bottleneck features (K. Vesely et al, 2012; F. Grézl et al, 2014)

- Shared Phone Multilingual CTC Model (M. Müller, 2017)

- and many more…

CTC Based Multi-lingual ASR

- In this paper, we demonstrate that it is possible to train multi-lingual ASR directly on phone sequences and without explicitly using a shared phoneme set.

- We try to understand the effect of adding more languages (related or unrelated) in both multi-lingual and cross-lingual setting.

- We look into learning “bottleneck” like shared hidden acoustic representations and use it for cross-lingual adaptation.

Data - Babel Dataset

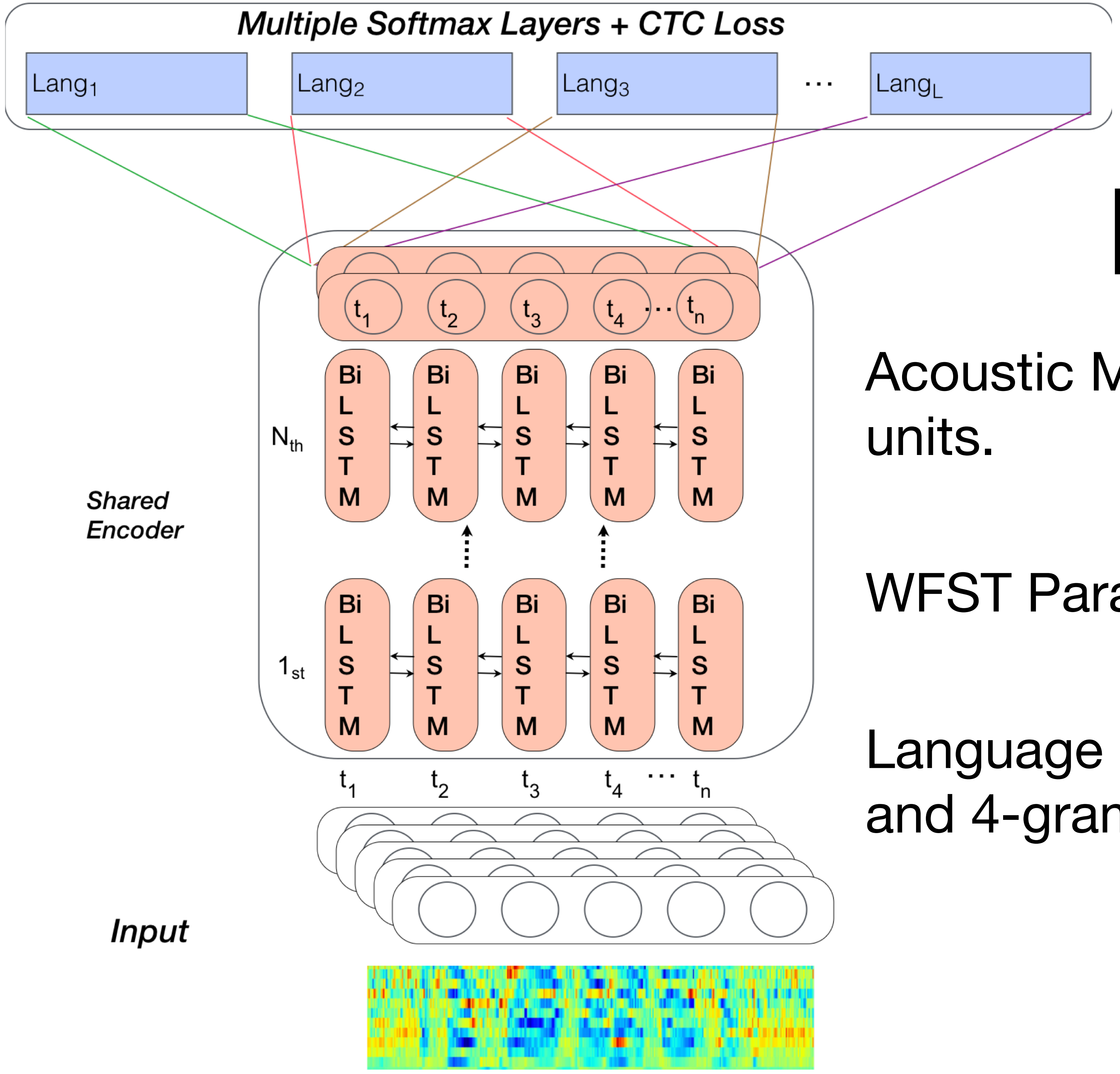
- We chose to perform experiments on a set of four languages which are the closest/have maximum phone overlap with Kurmanji.

Subset	Language	#Phones + Φ	Hours
MLing	Turkish	50	79 hrs
	Haitian	40	67 hrs
	Kazakh	70	39 hrs
	Mongolian	61	46 hrs

- We test the effect of adding more languages by using SWBD (a large well prepared unrelated language) and BAB300 (a set of 4 unrelated languages in babel totaling to 300h).

- We do cross-lingual tests on Kurmanji (related) and Swahili (unrelated).

CTC Based Multi-lingual ASR



Model Parameters

Acoustic Model Params - 6 Layer BiLSTM with 360 hidden units.

WFST Params - Beam size of 9.0 and Lattice Beam of 4.0

Language Model - Lowest dev perplexity between 3-gram and 4-gram models.

CTC Based Multi-lingual ASR

Table 2: Word error rate (% WER) for each language in the MLing subset

Model	Kazakh		Turkish		Haitian		Mongolian	
	WER	PER	WER	PER	WER	PER	WER	PER
Monolingual	55.9	40.9	53.1	36.2	49.0	36.9	58.2	45.2

BASELINE

CTC Based Multi-lingual ASR

It works!

Table 2: Word error rate (% WER) for each language in the MLing subset

Model	Kazakh		Turkish		Haitian		Mongolian	
	WER	PER	WER	PER	WER	PER	WER	PER
Monolingual	55.9	40.9	53.1	36.2	49.0	36.9	58.2	45.2
Multilingual	53.2	36.5	52.8	34.4	47.8	34.9	55.9	41.1

~1.5 % WER⇩

Multi-lingual Training

CTC Based Multi-lingual ASR

Improves further!

Table 2: Word error rate (% WER) for each language in the MLing subset

Model	Kazakh		Turkish		Haitian		Mongolian	
	WER	PER	WER	PER	WER	PER	WER	PER
Monolingual	55.9	40.9	53.1	36.2	49.0	36.9	58.2	45.2
Multilingual	53.2	36.5	52.8	34.4	47.8	34.9	55.9	41.1
+ FineTuning	**50.6**	**35.1**	**49.0**	**32.2**	**46.6**	**33.2**	**53.4**	**39.6**

~4 % WER⇩

Fine-tuning for each language

Note : Improvements are higher for lower resourced languages!

What if you add English Switchboard (300h) !?

Table 2: Word error rate (% WER) for each language in the MLing subset

Model	Kazakh		Turkish		Haitian		Mongolian	
	WER	PER	WER	PER	WER	PER	WER	PER
Monolingual	55.9	40.9	53.1	36.2	49.0	36.9	58.2	45.2
Multilingual	53.2	36.5	52.8	34.4	47.8	34.9	55.9	41.1
+ FineTuning	50.6	35.1	49.0	32.2	46.6	33.2	53.4	39.6
Multilingual + SWBD	**52.3**	**36.6**	**51.3**	**33.0**	**45.8**	**33.9**	**54.5**	**40.2**
+ FineTuning	**48.2**	**33.5**	**48.7**	**31.9**	**44.3**	**31.9**	**51.5**	**37.8**

~6 % WER⇩

Multi-lingual Training with SWBD and Fine-Tuning

SWBD vs Bab300

- Using 300 hours of various Babel languages performs worse than just adding SWBD.

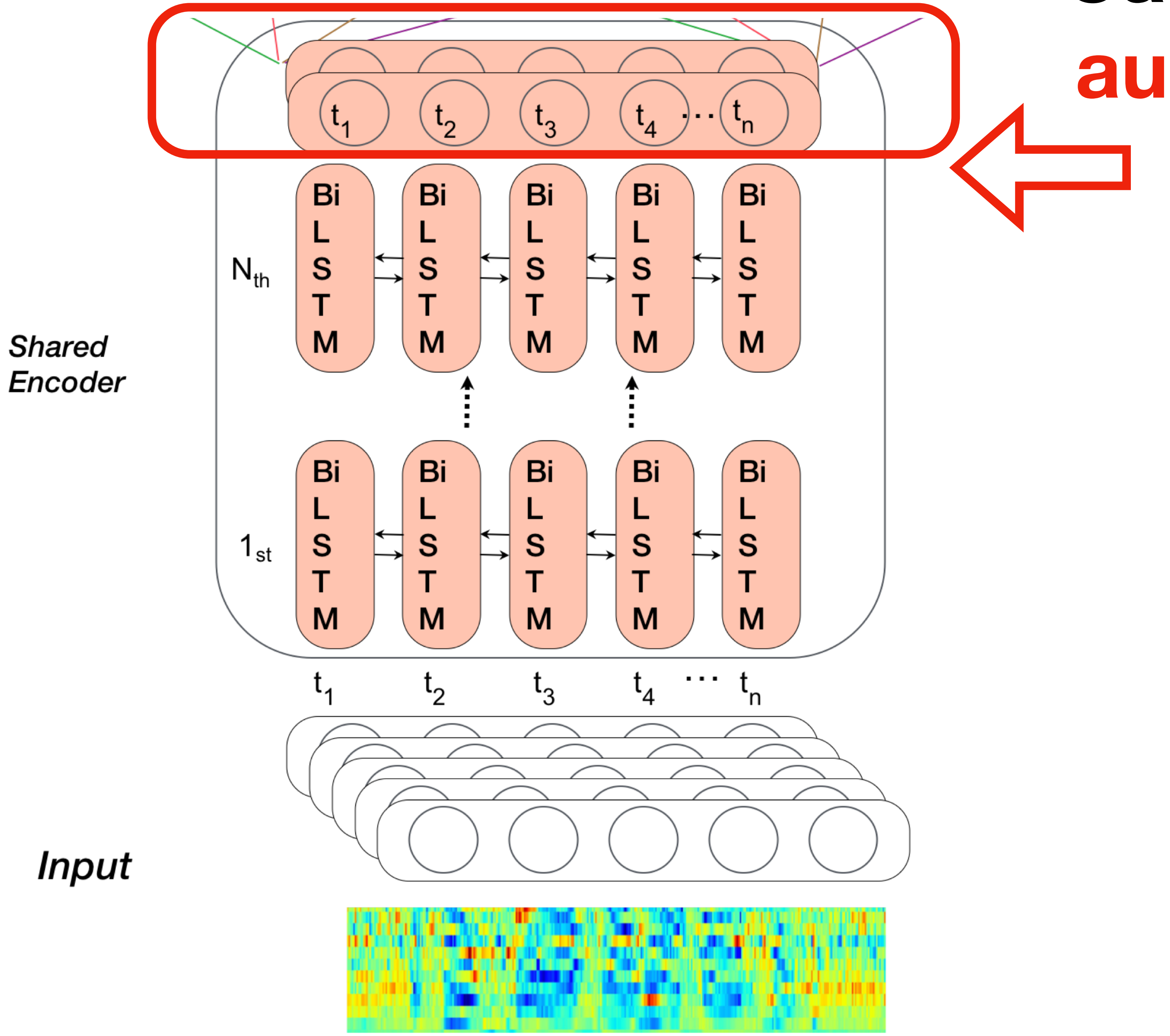
Table 2: Word error rate (% WER) on the test languages.

Model	Kazakh	Turkish	Haitian	Mongolian
MLing + Bab300	57.5	52.0	47.8	56.7
MLing + SWBD	**52.3**	**51.3**	**45.8**	**54.5**

- It is beneficial to add large amounts of well-prepared data from a single language rather than adding many unrelated languages.

- Adding a large number of languages may in fact prevent the model from training well.

Representation Learning



Can this layer be used as a discriminatory audio feature layer that is independent of the input language?

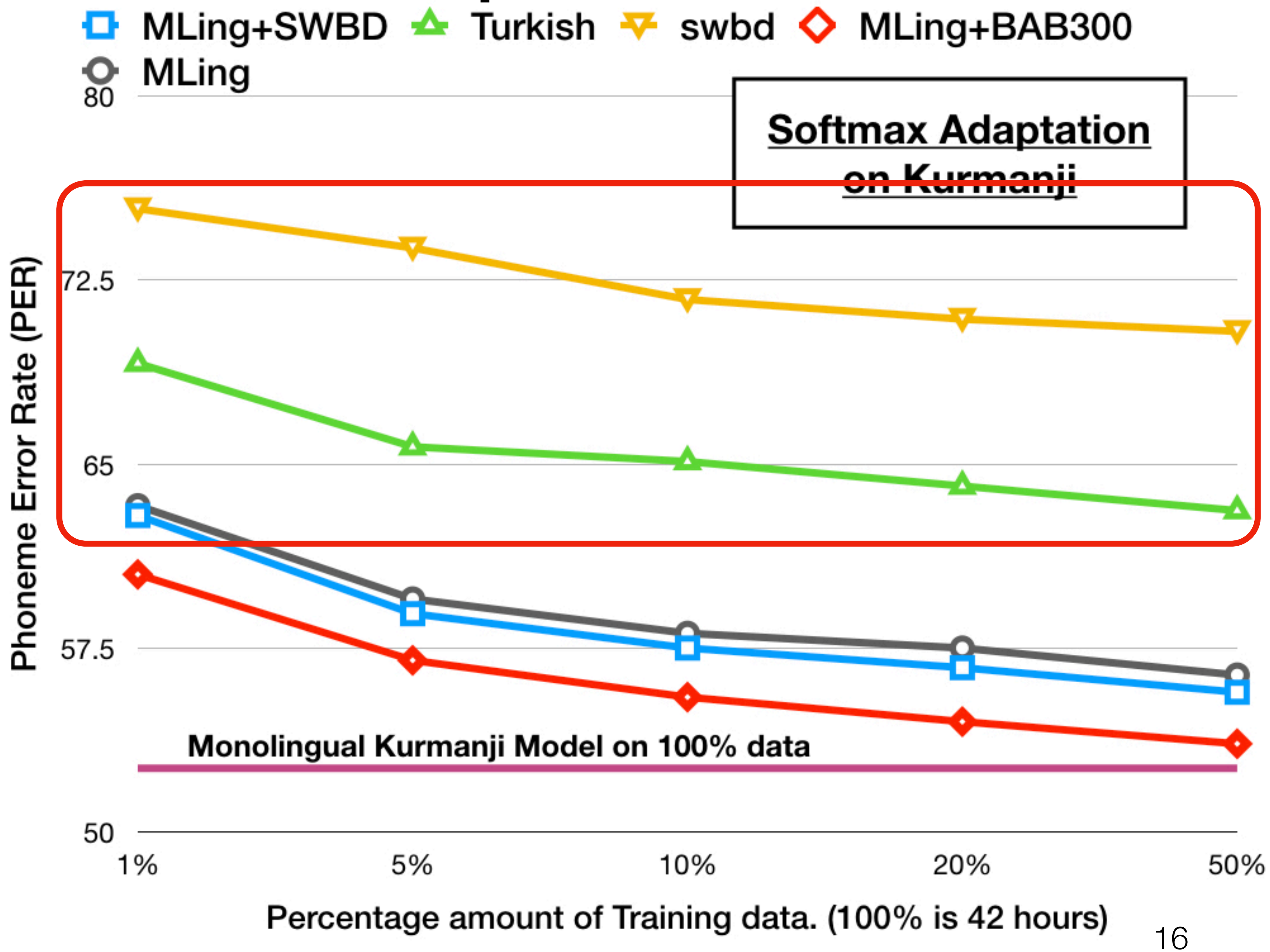
Motivation from bottleneck layers!

Representation Learning

- We take the encoder representations of various trained model.
- Then train only the softmax layer using various amounts of data from a related unseen language, Kurmanji.

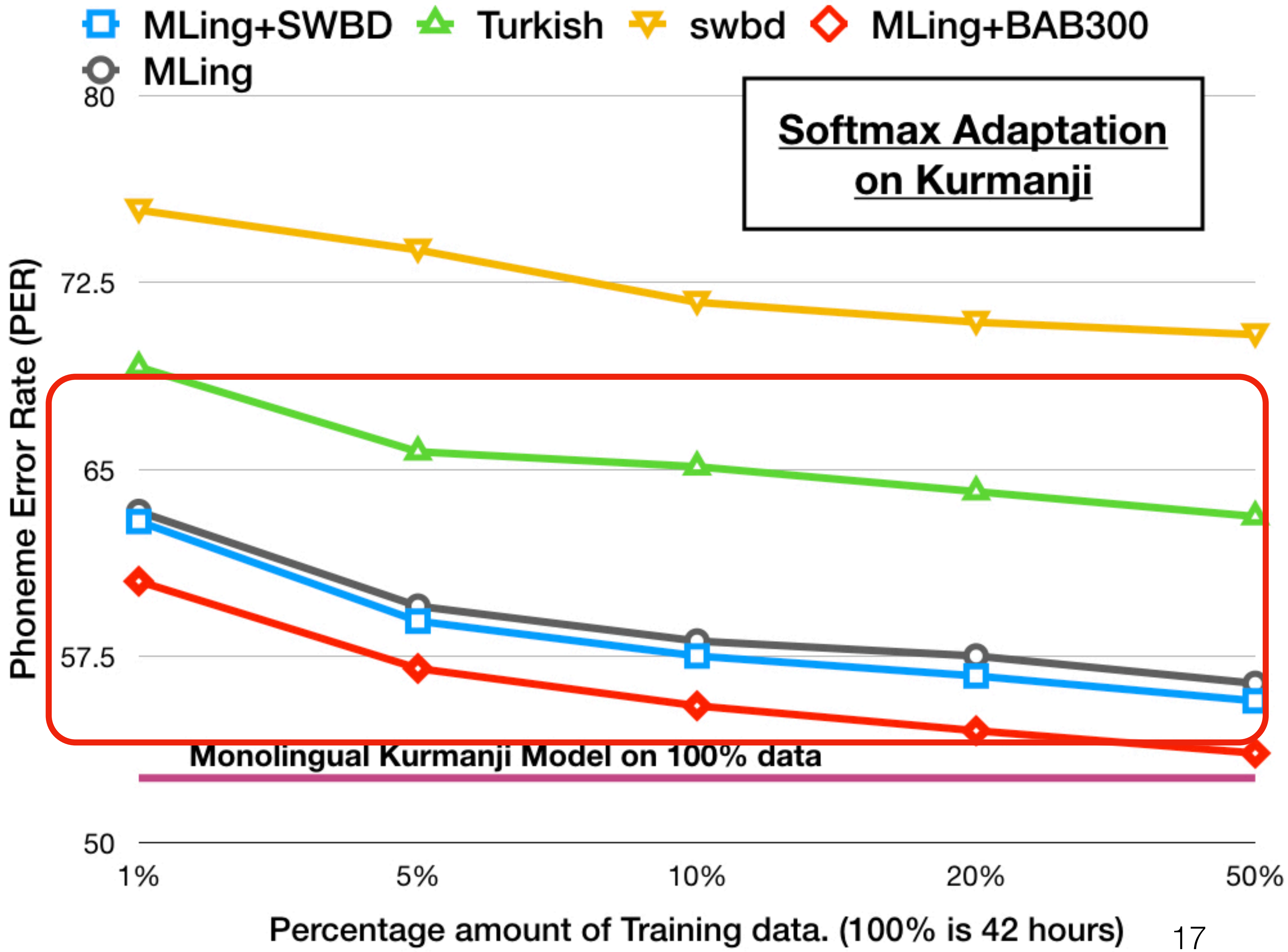
Check for whether the pre-trained hidden representation can linearly separate a new language into it's phoneme sequence.

Representation Learning



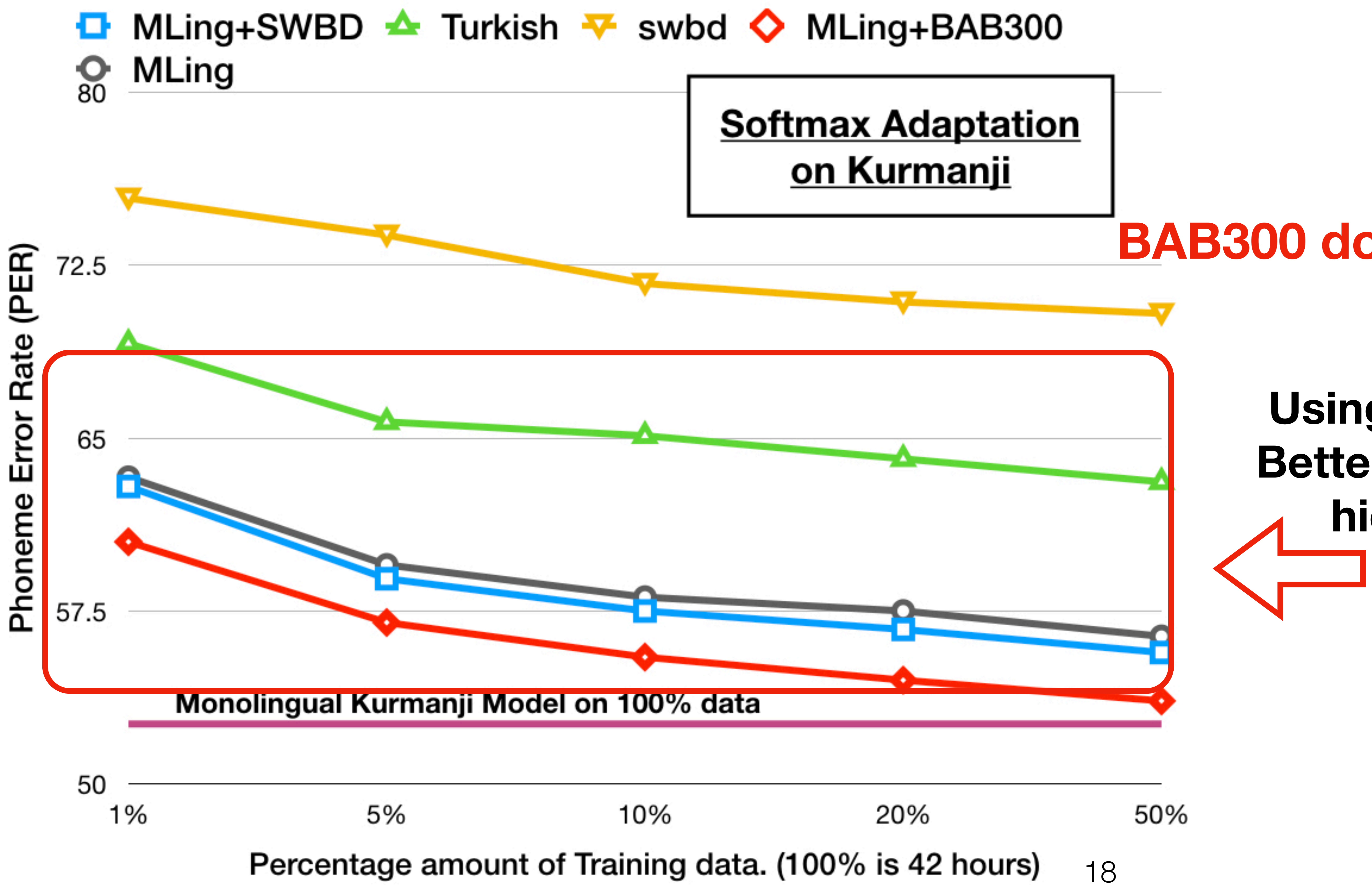
Monolingual Turkish model performs much better than SWBD

Representation Learning



Multilingual Models are considerably better than the Monolingual models!

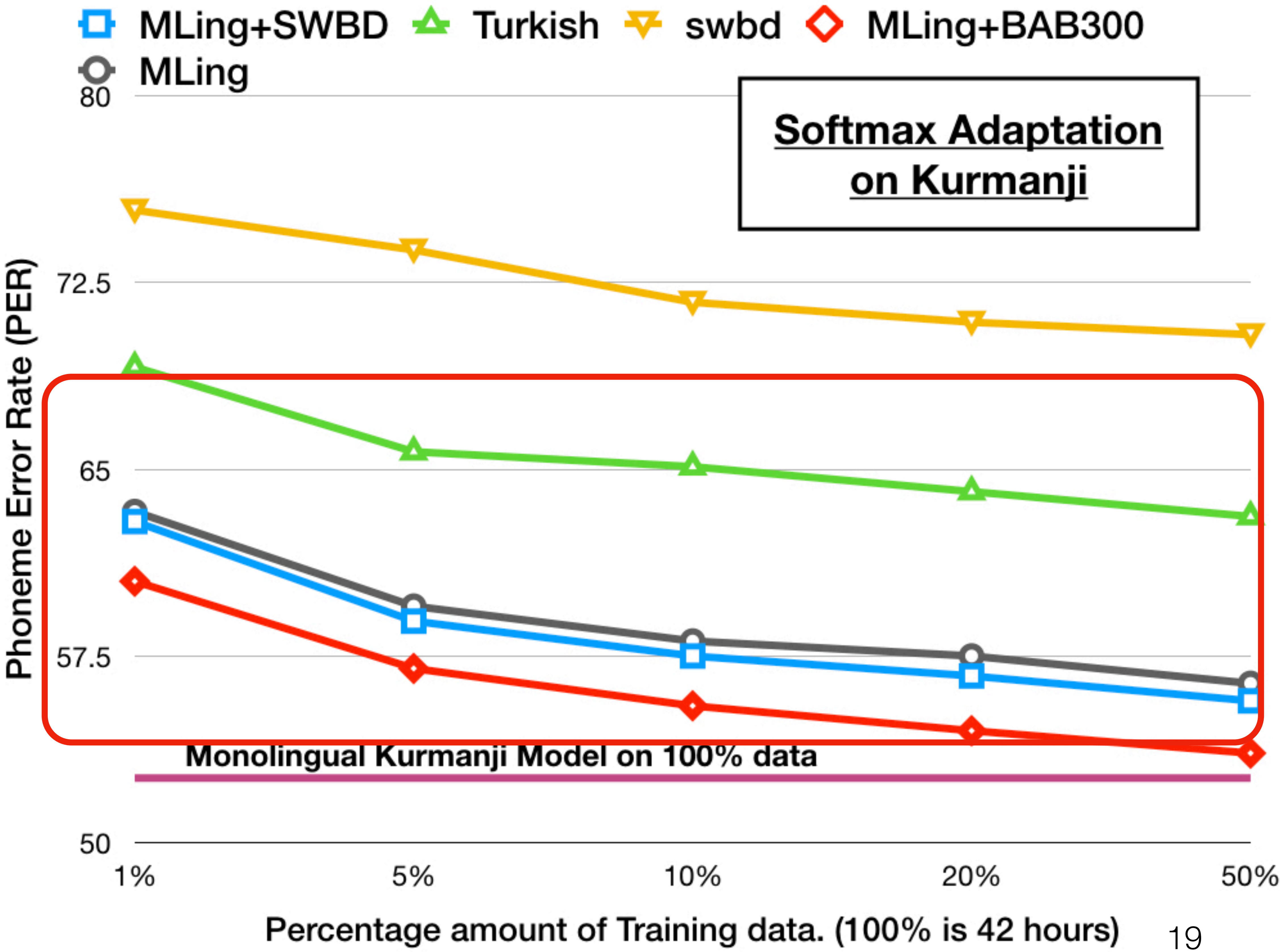
Representation Learning



BAB300 does better than SWBD!

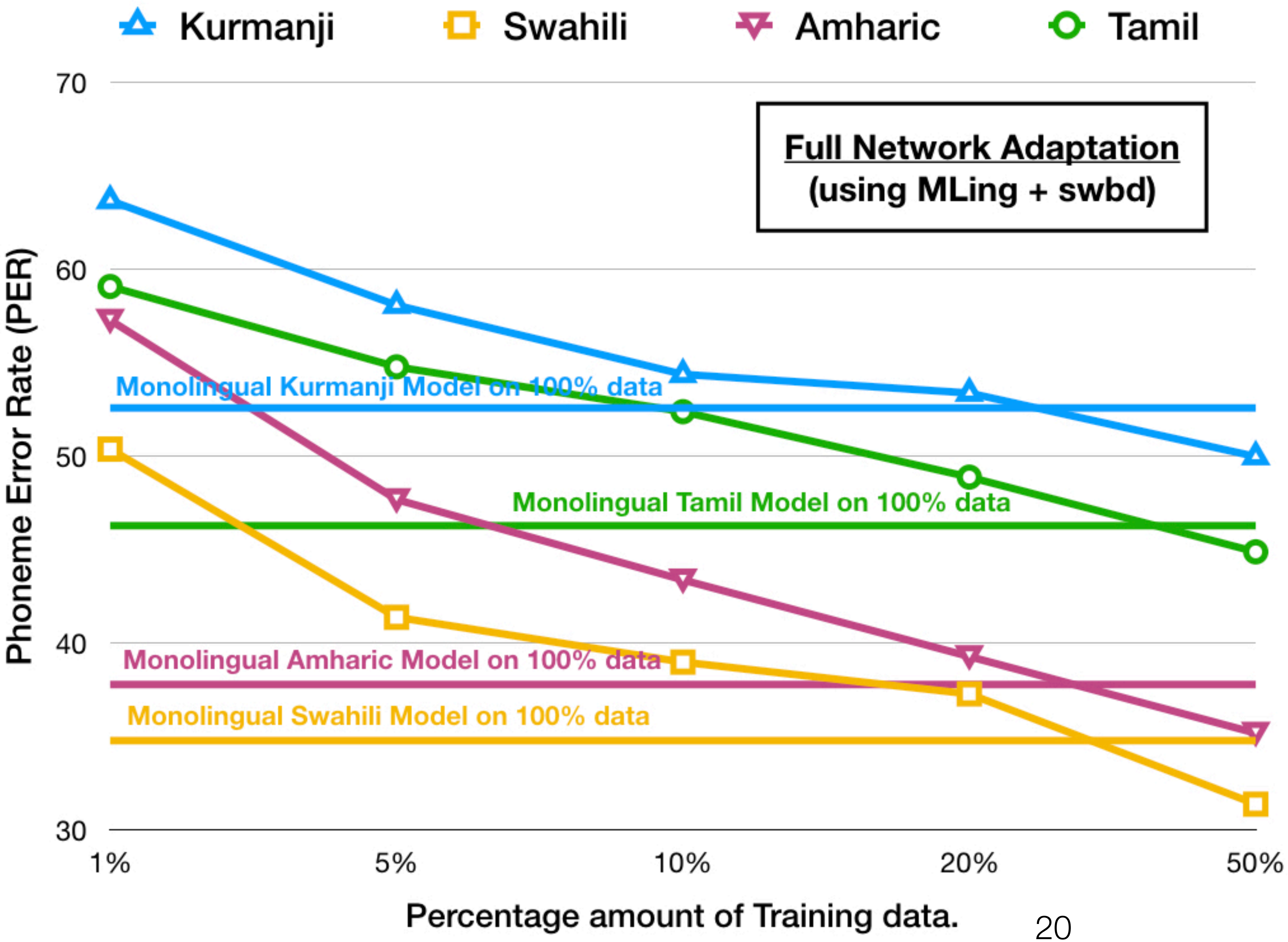
Using more languages help!
Better language independent hidden representation

Representation Learning



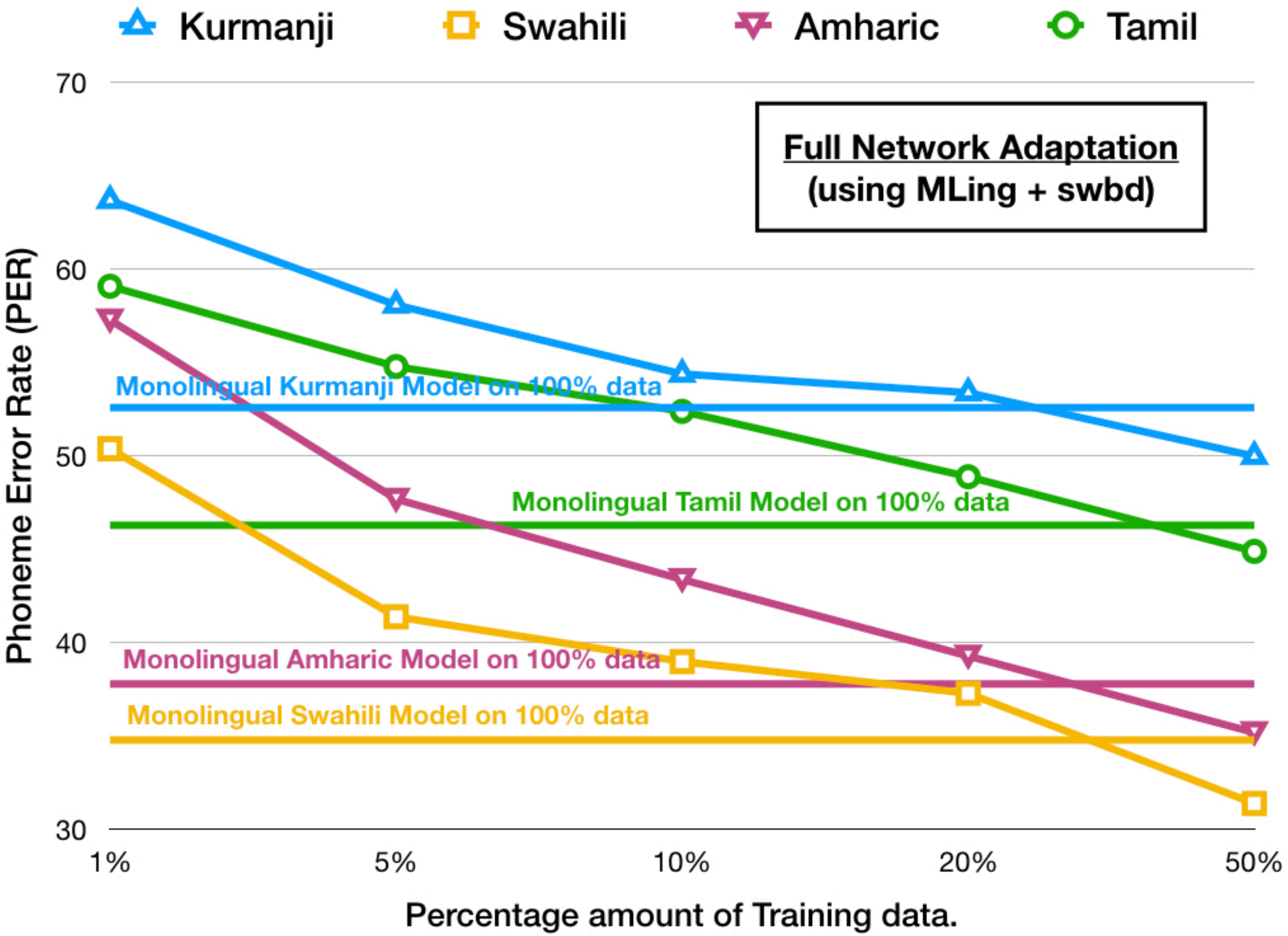
Multilingual models by JUST using 10-20% data to ONLY adapt the softmax layer => almost close to a Monolingual Kurmanji Model on 100% data with ALL layers trained.

Cross-lingual Explorations



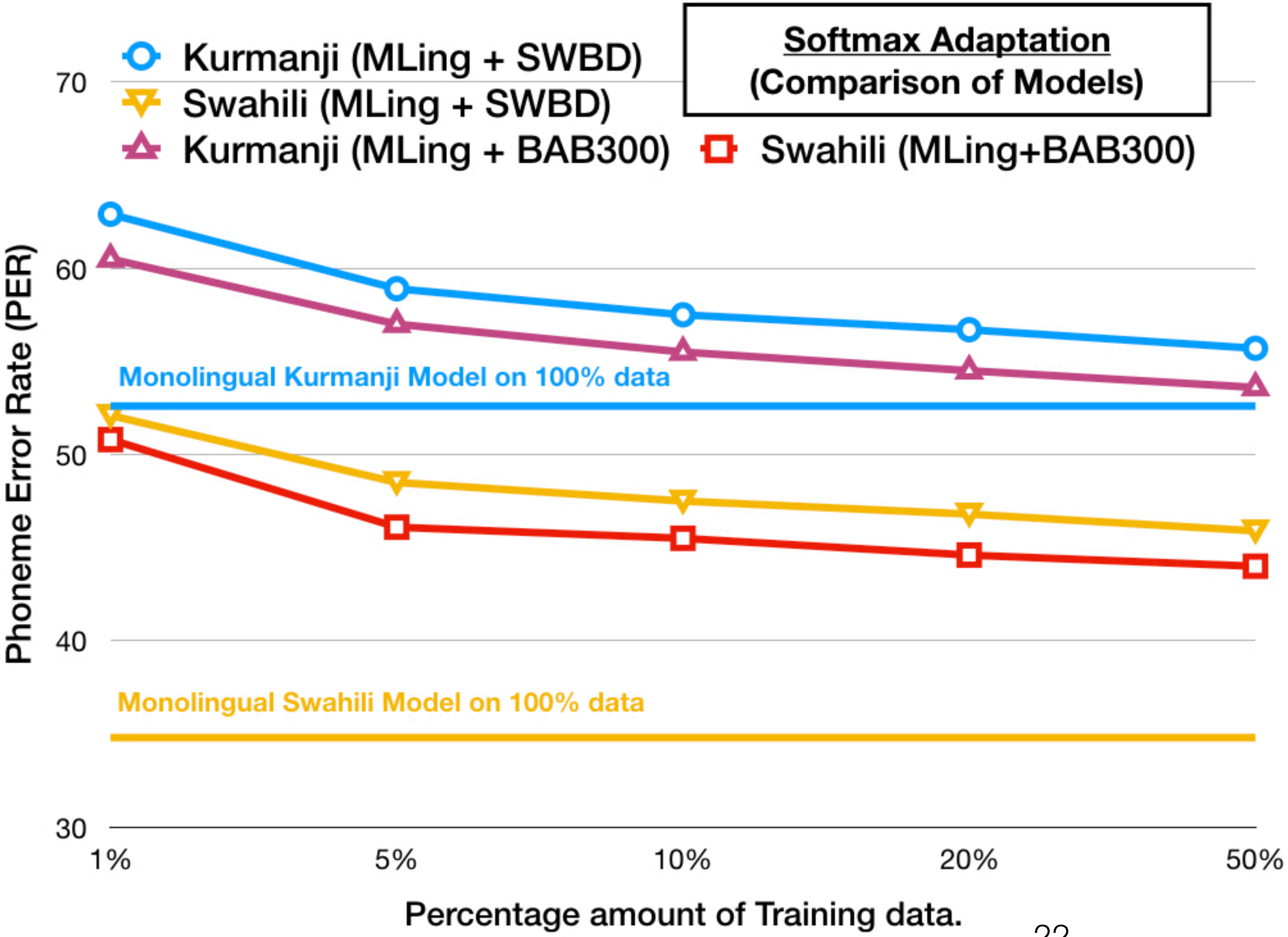
Multilingual system surpasses the mono-lingual baseline when just 25% of the original data has been seen!

Cross-lingual Explorations



This behavior of retraining ("full network adaptation") seems independent of the target language.

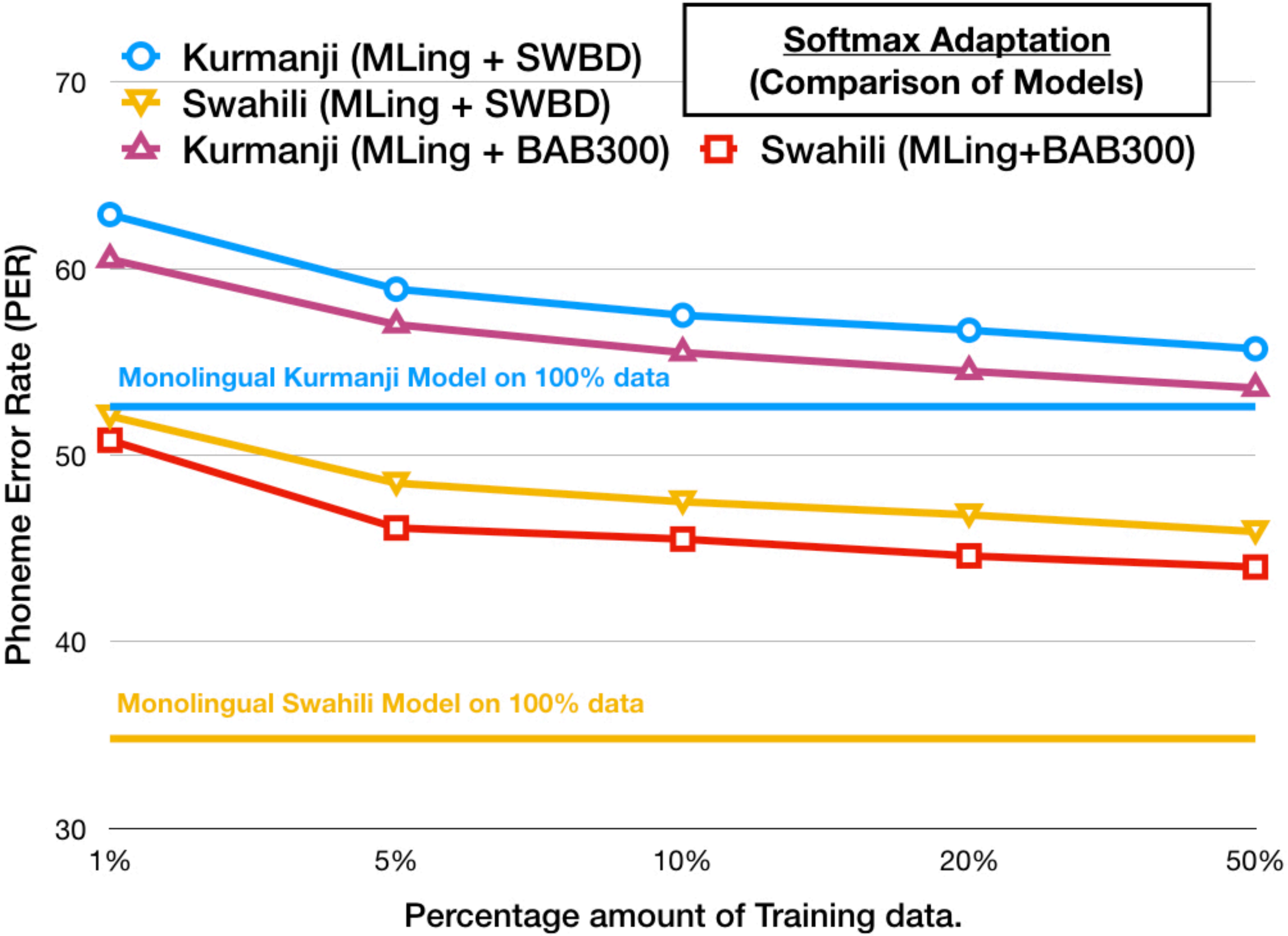
Cross-lingual Explorations



Kurmanji performs well, because the language is similar to the training languages.

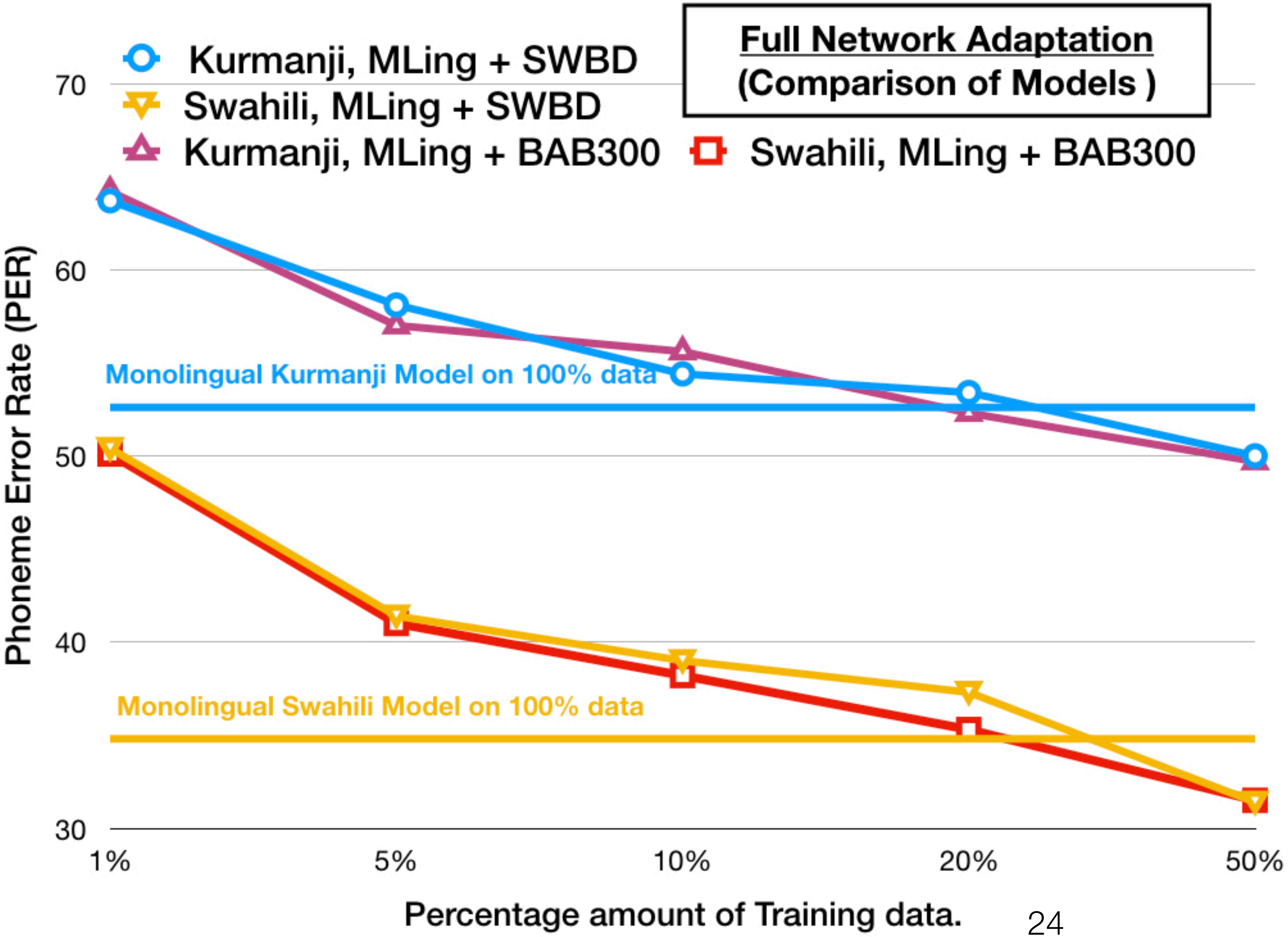
Larger gap while adapting to an unrelated language, in this case Swahili.

Cross-lingual Explorations



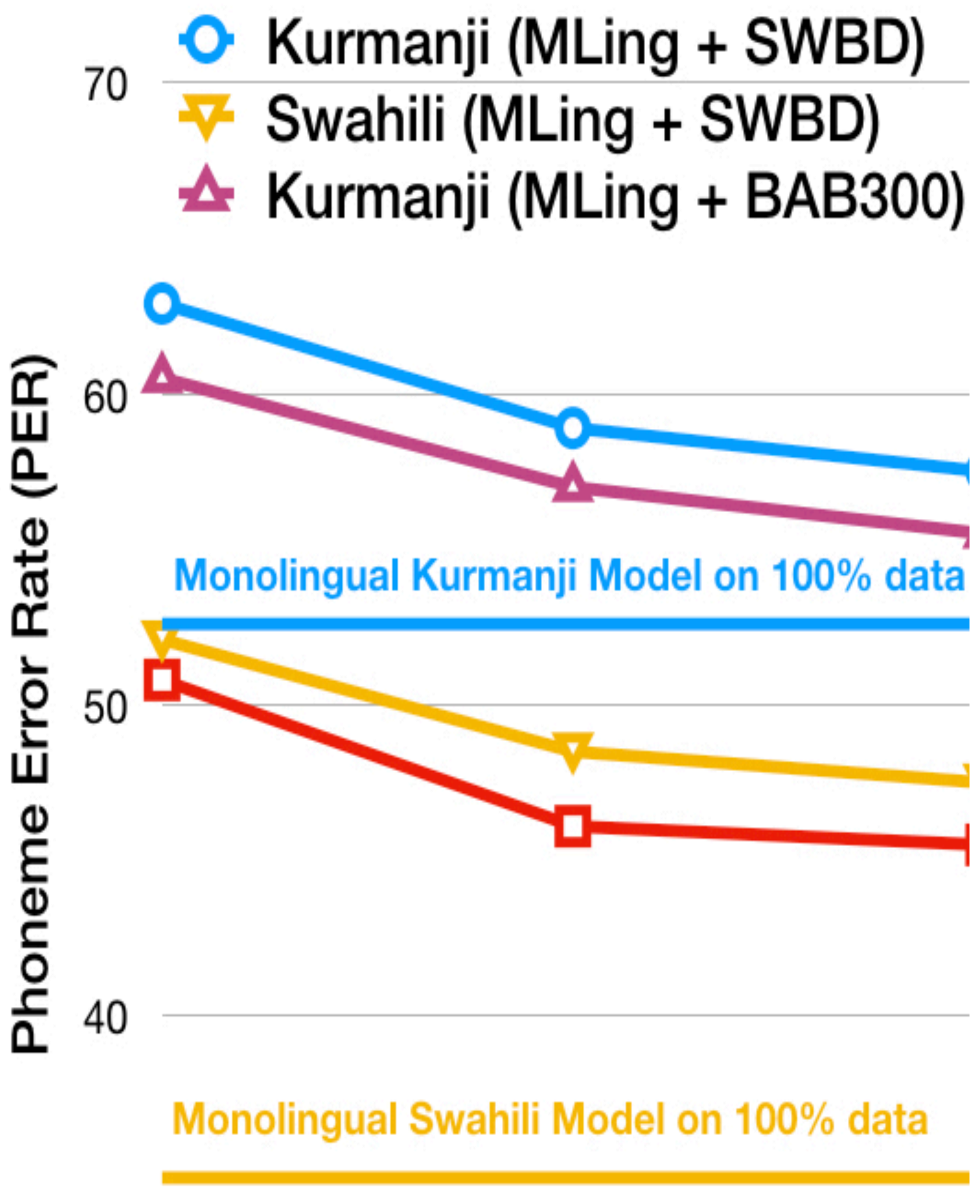
Initialization with many languages (MLing + BAB300) is beneficial!

Cross-lingual Explorations

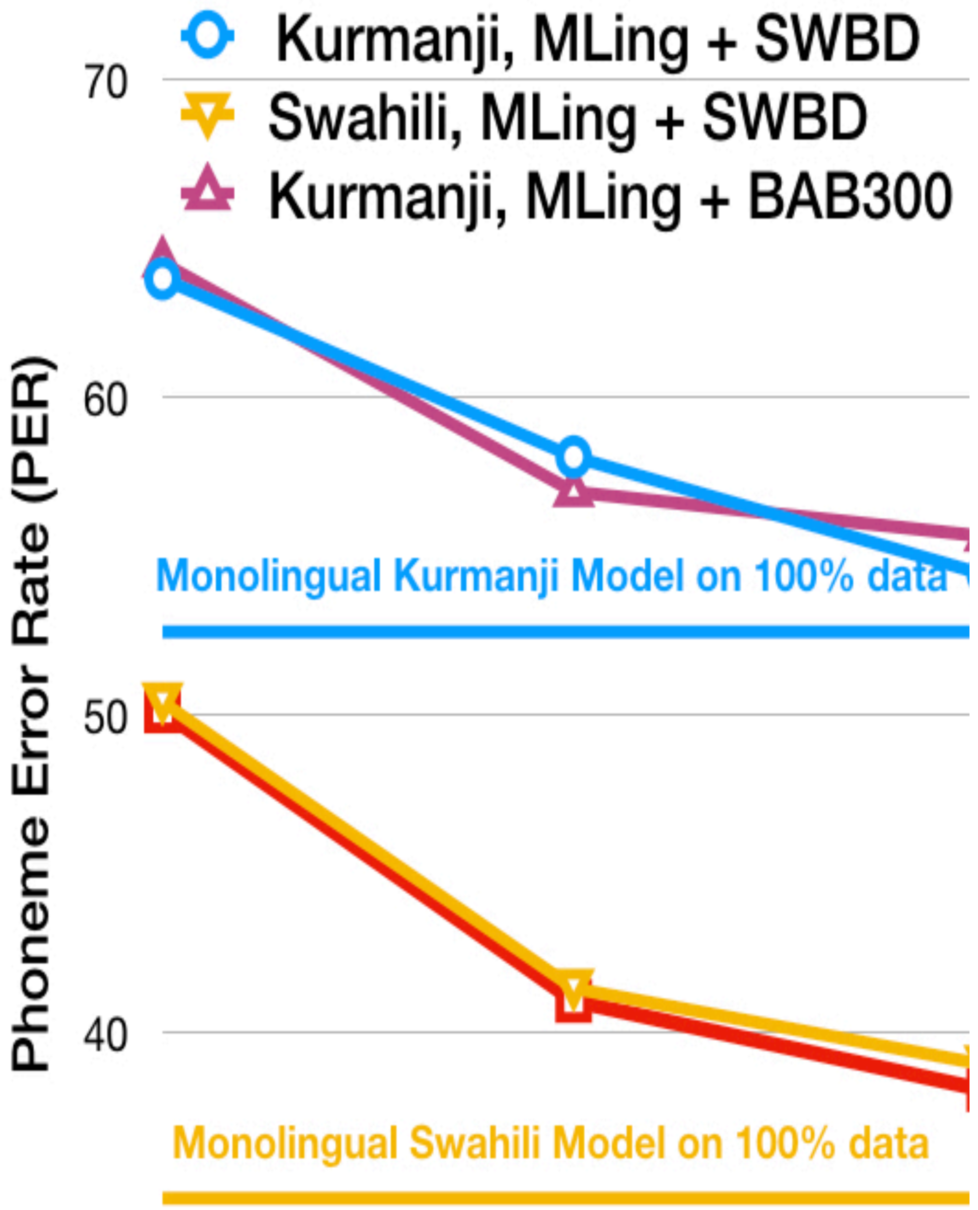


When the entire network can be retrained -
Starting with (MLing + SWBD) or (MLing + BAB300) perform almost equally well!

Cross-lingual Explorations



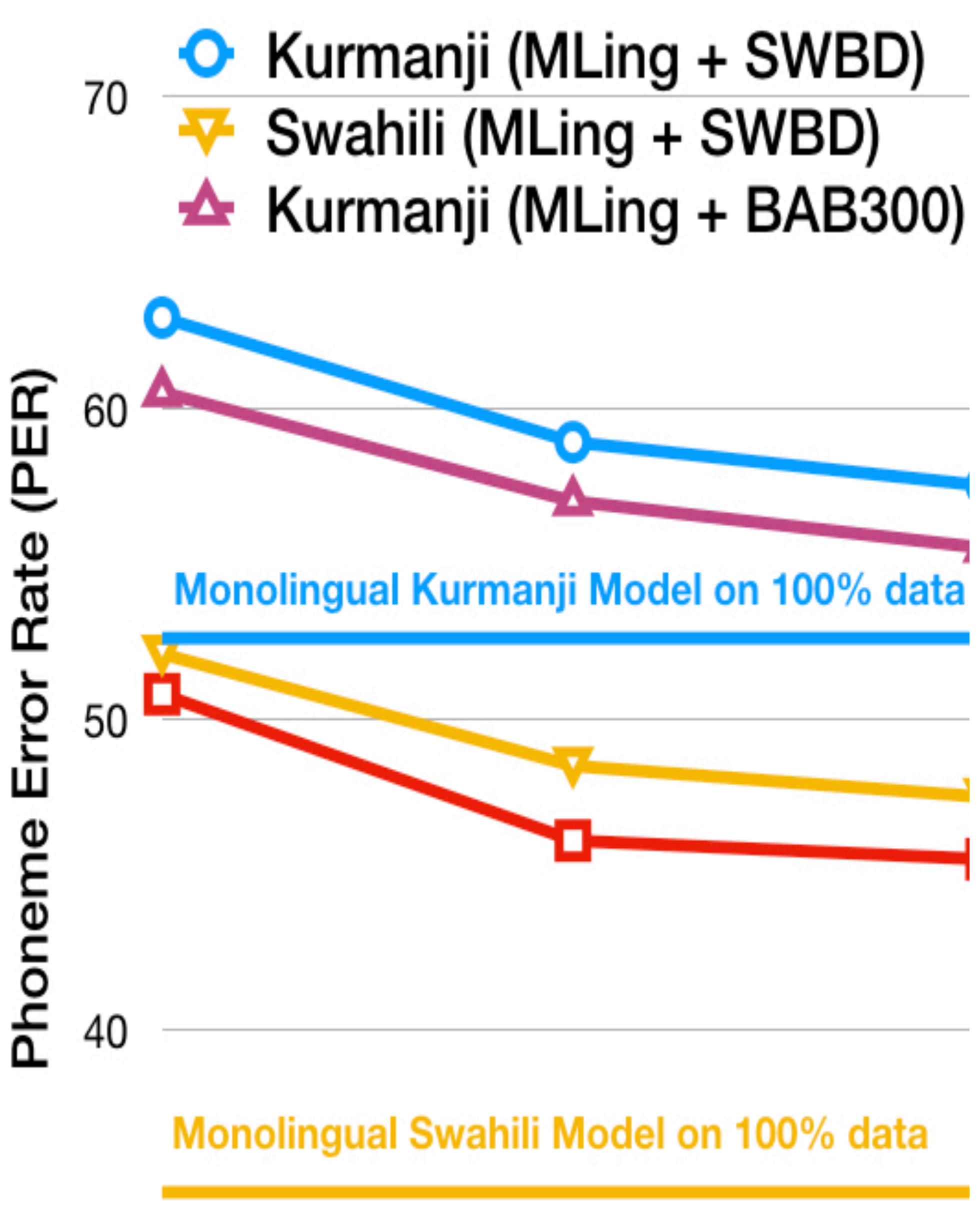
Softmax Adaptation (Comparison of Models)



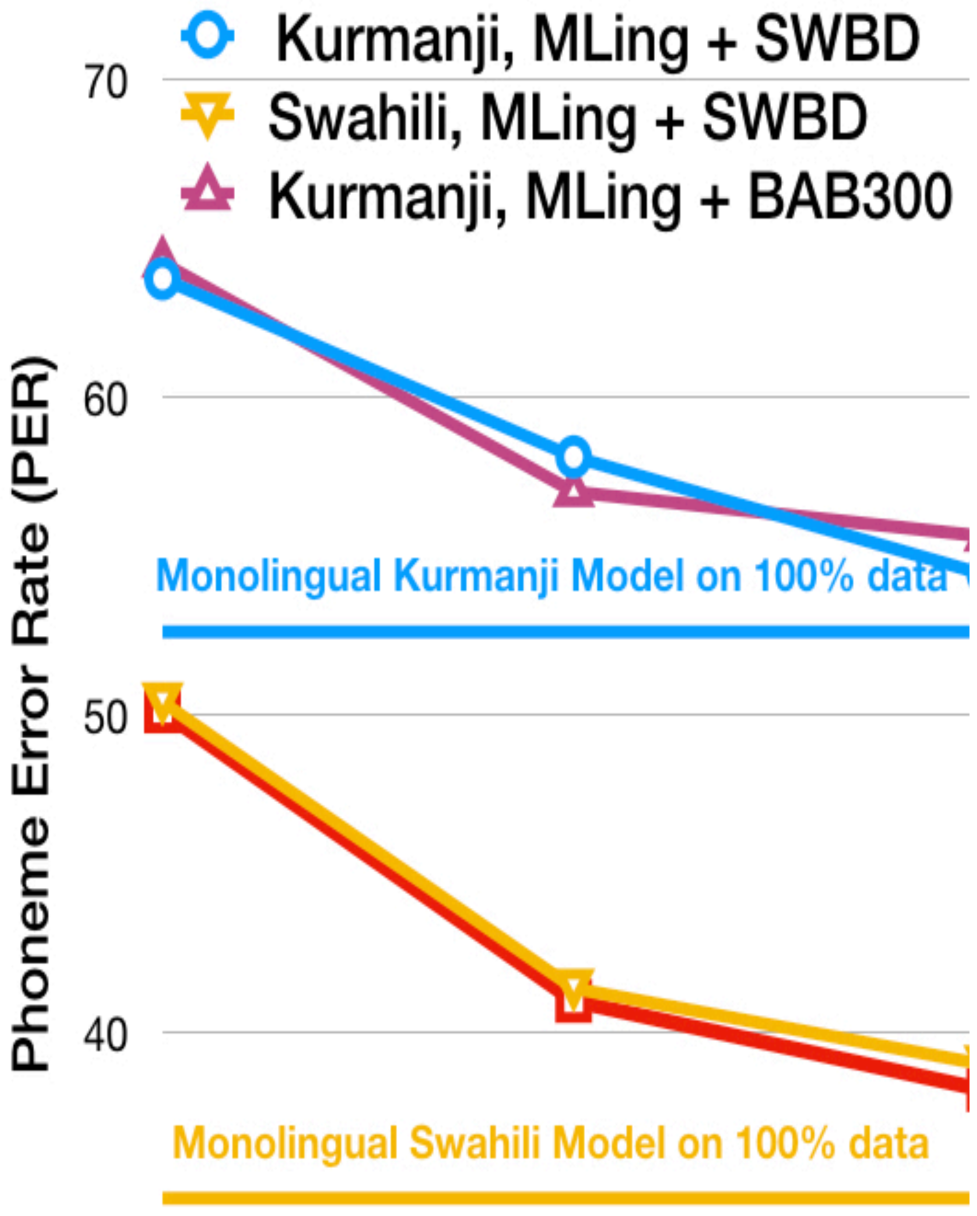
Full Network Adaptation (Comparison of Models)

Full Network adaptation (on the right) outperforms Softmax adaptation (on the left) as soon as 2-4 h of data become available.

Cross-lingual Explorations



Softmax Adaptation (Comparison of Models)



Full Network Adaptation (Comparison of Models)

In very low resource cross-lingual scenarios, it is probably better to adapt a model to an unseen language by re-training the softmax layer.

Conclusion

- It is possible to train multi-lingual and cross-lingual acoustic models directly on phone sequences.
- These models can learn a language independent representation.
- In multi-lingual settings, it seems beneficial to train on related languages only, or on large amounts of clean data.

Conclusion

- In very low resource cross-lingual scenarios, training on related languages help, as does training on many languages, rather than large amounts of single language.

- The effect of the choice of languages disappears as more and more data is available and the whole network can be retrained.

Steps Ahead

- Can we do ASR on a language without any training data?
  - Use a language universal recognizer (shared softmax layer).
  - Decode using a phoneme based neural language models trained on nonparallel text.
  - Thereby facilitating us to do zero-resource speech recognition!

Thank you!

- Code available in - https://github.com/srvk/eesen/tree/tf_clean/asr_egs/babel/105_201_302_401
- Contact us - {sdalmia,ramons,fmetze,awb}@cs.cmu.edu

Questions?