Using Self-Supervised Models for Code-Switching and Multilingual ASR
JSALT 2022

July 20th 2022

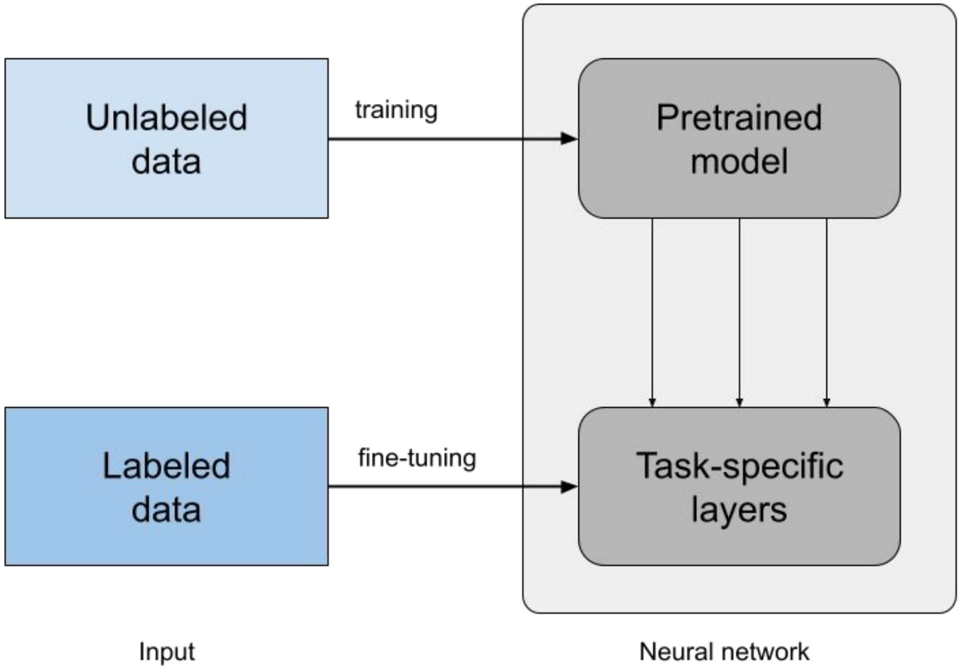
Léa-Marie LAM-YEE-MUI, Lucas ONDEL, Ondřej KLEJCH

Research questions

- Are self-supervised (SSL) models comparable to the traditional approaches for low-resource languages ?

- Can we use SSL models with semi-supervised learning ?

Self-supervised learning



CS data: corpus “soapies”

Code-switched South African languages [1]

- sesotho-english (3h)
- setswana-english (3h)
- xhosa-english (3h)
- zulu-english (5.5h)

Labeled audio: 15h of soap operas

Unlabeled audio: ~200 hours, all languages mixed, also soap operas

Few monolingual texts

[1] Barnard, Etienne, et al. "The NCHLT speech corpus of the South African languages." Workshop Spoken Language Technologies for Under-resourced Languages (SLTU), 2014.

Examples

sesotho-english JA JA WELL I MEAN HO HONA HO TLA MO LERATONG

tetswana-english DILO TSE NKA GO BOLELLANG TSONE KA FAMILY ELE

xhosa-english NDAMXELELA NAY'USIBUSISO KODWA KE UDINEO ZANGA AFUN'UKU PRESS THE CHARGES SO

zulu-english ODWA NEMVUNULO NAYO NJE IVEZA YONKE INTO OBALA

https://www.youtube.com/watch?v=CHhm8zrj2BA

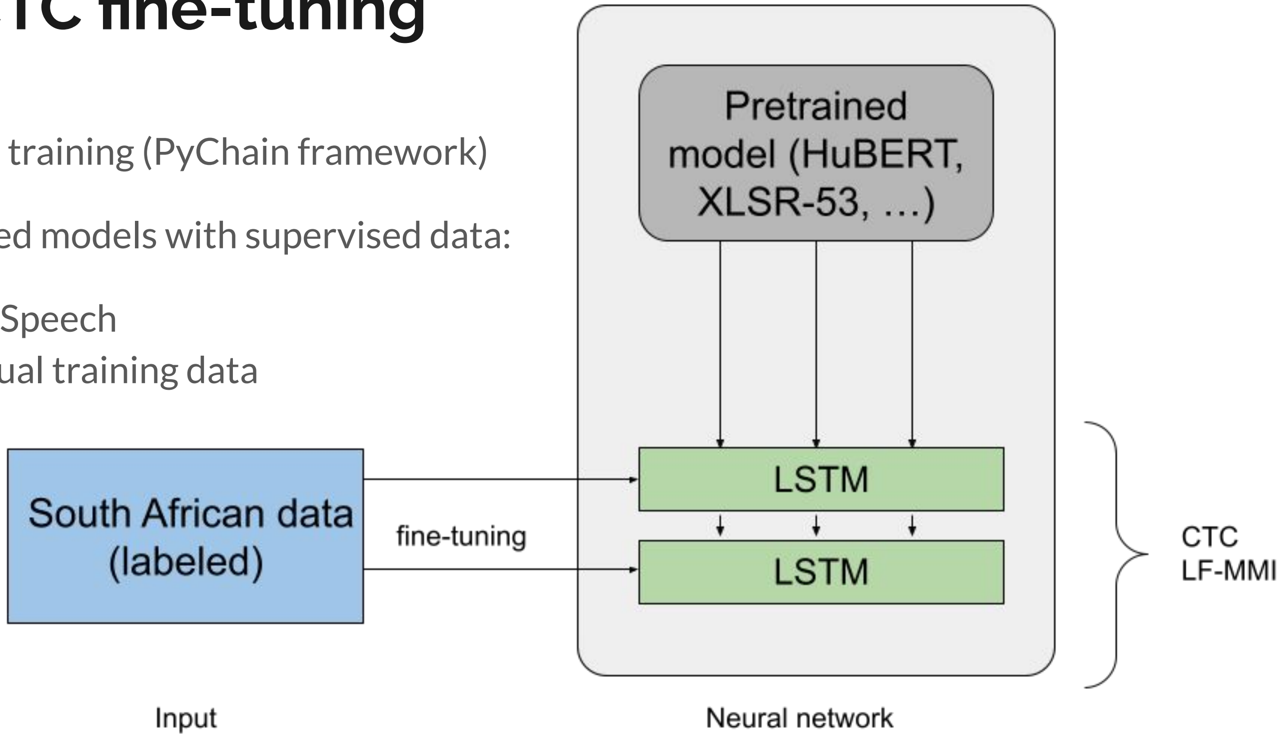
Experiments: CTC fine-tuning

Baselines: TDNN + LF-MMI training (PyChain framework)

Adaptation of self-supervised models with supervised data:

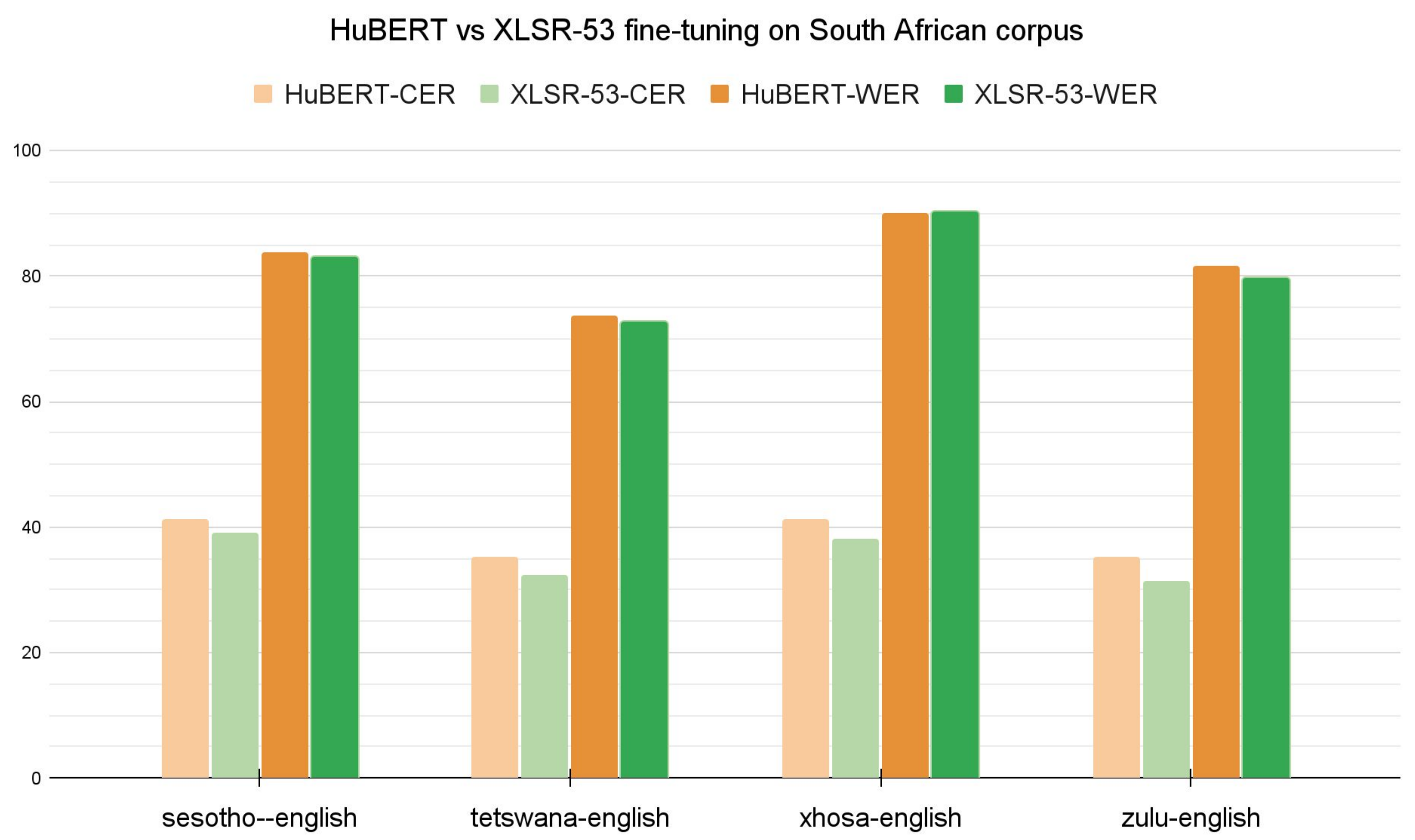
- HuBERT: english LibriSpeech
- or XLSR-53: multilingual training data

No LM to decode



SSL results on South African languages

	TDNN (LF-MMI training) + weak LM - PyChain		XLSR (CTC fine-tuning - no LM)	
Languages	WER	CER	WER	CER
sesotho-english			83.17	39.05
tetswana-english			72.68	32.40
xhosa-english	96.34	71.30	**90.27**	**38.13**
zulu-english			79.69	31.37



HuBERT vs XLSR-53 fine-tuning on South African corpus
HuBERT-CER
XLSR-53-CER
HuBERT-WER
XLSR-53-WER
100
80
60
40
20
0
sesotho--english
tetswana-english
xhosa-english
zulu-english

Results summary

1) SSL in low-resource settings seems promising.
2) XLSR-53 vs HuBERT: XLSR-53 works better for low-resource

What’s next

- decode the fine-tuned models with a LM (with k2)
- make use of the phonetic information available: using a phonetic dictionary and fine-tuning SSL models with LF-MMI

- do continued pretraining on unlabeled in-domain data