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

- Unlabeled data
- Labeled data
- Pretrained model
- Task-specific layers

Training and fine-tuning flows through the diagram.
CS data: corpus “soapies”

Code-switched South African languages
- 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

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

<table>
<thead>
<tr>
<th>Languages</th>
<th>TDNN (LF-MMI training) + weak LM - PyChain</th>
<th>XLSR (CTC fine-tuning - no LM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>CER</td>
</tr>
<tr>
<td>sesotho-english</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tetswana-english</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xhosa-english</td>
<td>96.34</td>
<td>71.30</td>
</tr>
<tr>
<td>zulu-english</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HuBERT vs XLSR-53 fine-tuning on South African corpus

- **HuBERT-CER**
- **XLSR-53-CER**
- **HuBERT-WER**
- **XLSR-53-WER**

The diagram compares the performance of HuBERT and XLSR-53 on the South African language corpora, specifically showing the CER (Character Error Rate) and WER (Word Error Rate) for four languages: Sesotho-English, Tswana-English, Xhosa-English, and 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