Plugging a neural phoneme recognizer into a simple language model

A workflow for low-resource settings

Séverine Guillaume Guillaume Wisniewski Benjamin Galliot Minh-Châu Nguyên Maxime Fily Guillaume Jacques **Alexis Michaud**

Interspeech 2022

Take-Home Messages

- Neural phonemic recognizer for an "endangered" language: Japhug
- ASR for endangered languages raises many important scientific challenges
- Fine-tuning a pretrained model $\Rightarrow$ "good" recognizer from 3 hrs of data
- adding a simple count-based LM to `wav2vec` improves performance

Our Goal(s)

1. computer-aided linguistic documentation

   ↪ help fieldworkers document endangered/little-described languages

2. develop NLP tools for people speaking minority languages

Challenges

1. small amount of annotated data

   e.g. in the `Pangloss Collection`: 153 languages

   with less than 2h of annotated data

2. language with very specific structural features

   e.g. Japhug: impressive morphosyntactic complexity,

   Yongning Na: rich morphotonological systems

3. many loanwords

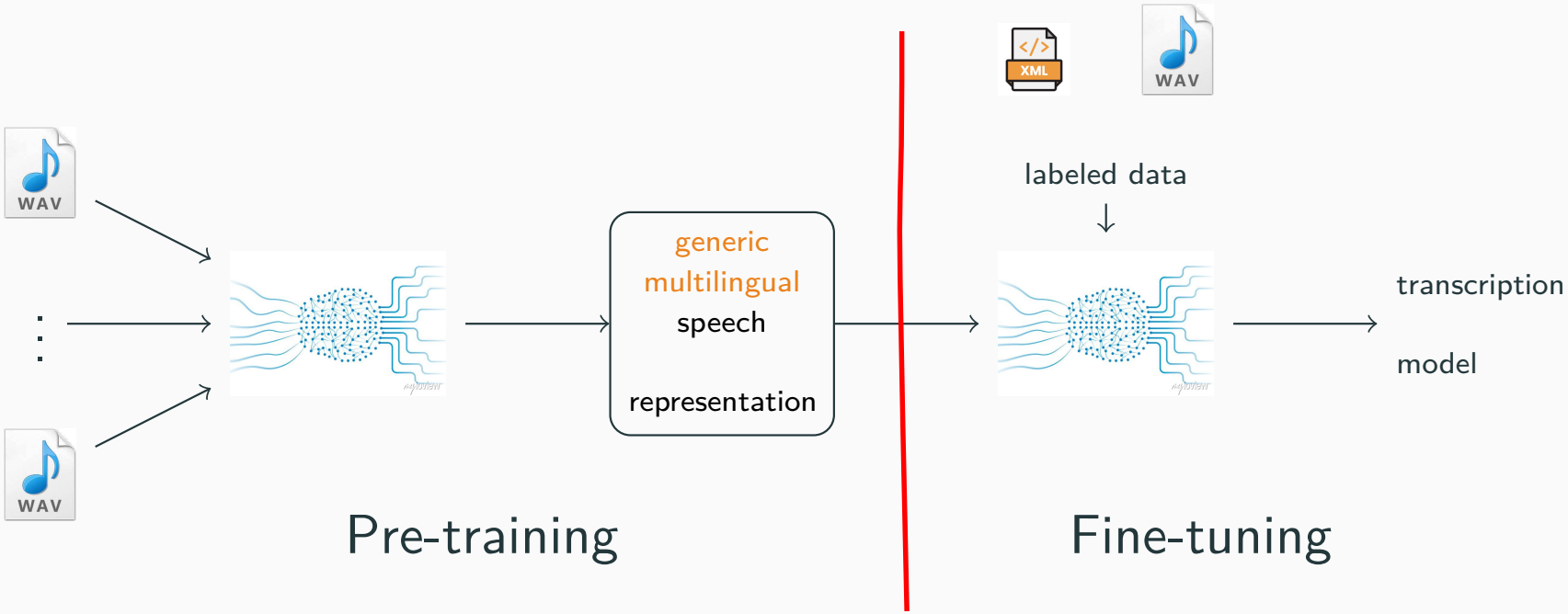
$\Rightarrow$ many scientific challenges and not a (complicated) engineering problem

A Simpler Context

ASR may be easier for linguistic documentation than for usual languages:

- (often) mono-speaker setting
- high degree of orthographic transparency

Pre-Trained Models: the Solution for Modeling Low-Resource Languages



↪ promise: good model from a small corpus of annotated data

↪ make training of a model for a new language / domain easy

now a standard approach in NLP

Application to Phoneme Transcription

- pretrained model : `XLS-R`
- fine-tuned on a corpus of phonemic transcriptions of Japhug
  - ↪ consider space (word separator) as "normal" character ⇒ will be predicted ⇒ prediction of words
  - ↪ keep punctuation

  ⇒ prediction as close as possible to the annotation of a fieldworker

Using an off-the-shelf `wav2vec` model for Japhug

		CER		WER	
	punctuation	✔	✖	✔	✖
training data	1.5 hrs	18.2	13.6	50.0	41.3
	3 hrs	13.6	8.6	35.7	26.7
	6 hrs	13.0	7.7	33.0	22.0

↪ very good results when using 3 hrs of annotated data

↪ we are getting closer to word-level ASR

Plugging a neural phoneme recognizer into a simple LM

Intuition

- low CER but high WER $\Rightarrow$ many words have a single wrong character $\Rightarrow$ use a lexicon/list of word to correct them
- main idea: use a LM to provide soft constraints

$\hookrightarrow$ "standard" approach in ASR: acoustic model $\oplus$ LM

Our Proposal

- simple count-based LM (2-gram with KN-smoothing)
- integrate the LM in CTC decoder with beam search
- different dataset for training the LM:
  - *small*: same data as the acoustic model
  - *large*: 8,400 "extra" sentences ($\simeq$ 10 hrs of annotated data)

Results on Japhug

With 1.5 hrs of annotated data:

	CER		WER	
punctuation	✔	✖	✔	✖
no LM	18.2	13.6	50.0	**41.3**
small LM	21.3	13.0	47.2	**35.7** $\Delta = -5.7$
large LM	20.6	12.7	46.4	**31.0** $\Delta = -10.3$

↪ better WER but at the cost of a worse CER

↪ smaller improvements for 3 hrs and 6 hrs of annotated data

↪ not all information is captured by the neural networks

Prospects: Facilitating multidisciplinary collaborations

Dataset 'pushed' to public attention

↪ on HFT: `https://huggingface.co/datasets/Lacito/pangloss/viewer`

↪ on Zenodo: `https://zenodo.org/record/5521112`

Tool to tailor downloads from the Pangloss Collection (190+ languages): `https://gitlab.com/lacito/outilspangloss`