1-2-11-SLP

On the Use of Semantically-Aligned Speech Representations for Spoken Language Understanding

Gaëlle Laperrière, Valentin Pelloin, Mickaël Rouvier, Themos Stafylakis, Yannick Estève

Esperanto
Exchanges for SPEech ReseArch aNd TechnOlogies
Horizon 2020 project
SELMA

SAMU-XLSR for SLU

Spoken Language Understanding refers to **automatic natural language processing** tasks related to the **extraction of semantic information** from **speech signal**.

The French dataset **MEDIA**, *considered as a very challenging benchmark*, and the Italian dataset **PortMEDIA** propose a task aiming at the understanding of spoken language through a **rich and complex semantic annotation**.

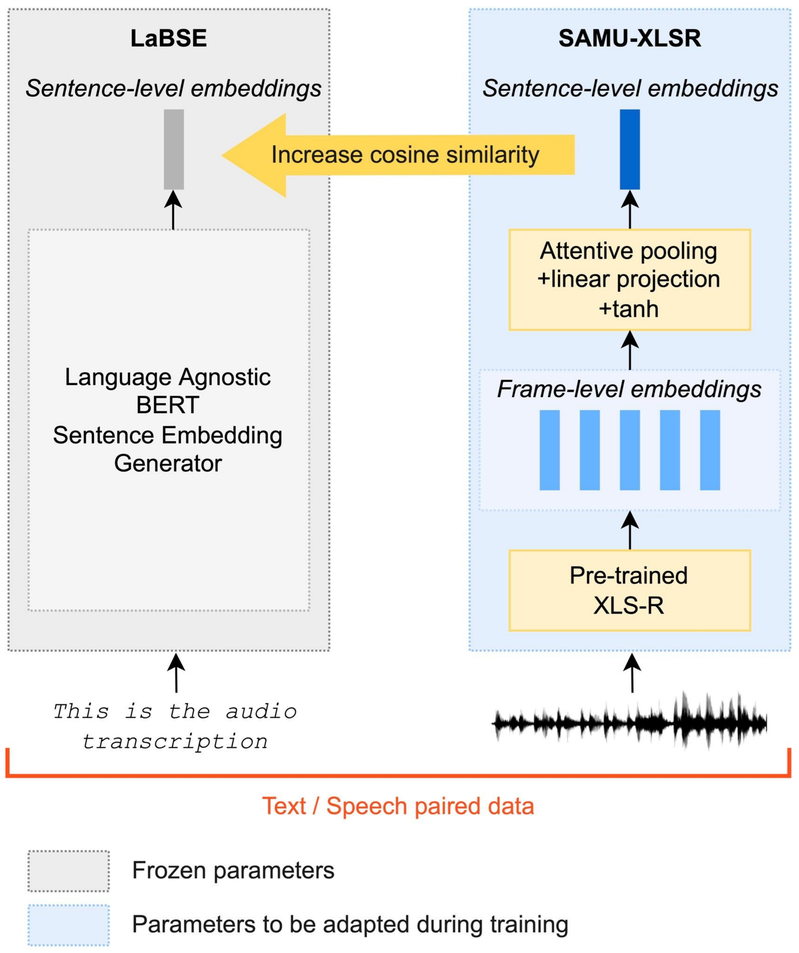
How does SAMU-XLSR's speech embeddings perform on cross-lingual semantics extraction ?

...compared to XLS-R's speech embeddings?

- ➢ Layer-wise analysis
- ➢ Language portability

...compared to LaBSE's text embeddings?

- ➢ Semantic analysis



MEDIA (FR) & PortMEDIA (IT)

This benchmark defines a protocol for evaluating SLU modules with a task of semantic extraction from speech.
Created in 2002 with the French governmental project Technolangue, it is freely distributed by ELRA for academic research since 2005.

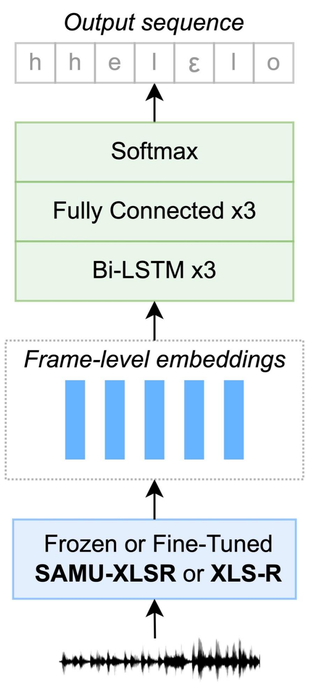
" Human-Machine dialogues of hotel reservation collected through the "Wizard-of-Oz" method, for semantic extraction tasks from speech "

		train	dev	test
Hours	**MEDIA**	10h52m	01h13m	03h01m
	PortMEDIA	07h18m	02h32m	04h51m
Words	**MEDIA**	94.5k	10.8k	26.6k
	PortMEDIA	21.7k	7.7k	14.7k

<concept> [value] word-support >

I <task> [reservation] would like to book >
a <room-type> [double] double room >
in <location-city> [Paris] paris >.

- ➔ Word Error Rate (WER)
- ➔ Concept Error Rate (CER)

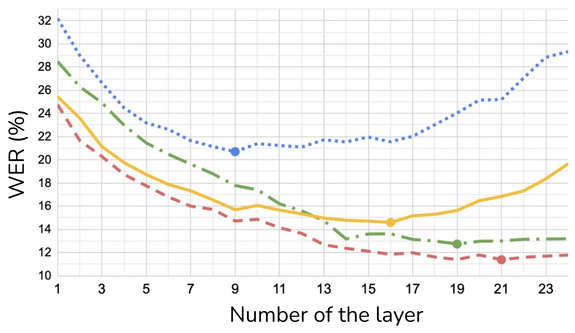


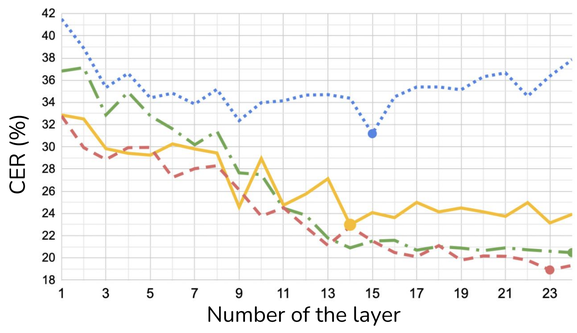
Layer-wise analysis

To make a layer-wise analysis of both speech encoders, we removed the upper layers of SAMU-XLSR or XLS-R, one by one, and extracted our speech embeddings.

- **Frozen:** the encoder kept layers are frozen with their initial weights
- **Fine-tuned:** the encoder kept layers are fine-tuned by supervision to solve the MEDIA French SLU task

Figures illustrate how the **linguistic** (WER) and **semantic** (CER) information is encoded through each layer of both speech encoders.





- ➔ SAMU-XLSR **outperforms** XLS-R for the MEDIA ASR task
- ➔ Best WER achieved by **higher layers** with SAMU-XLSR than XLS-R
- ➔ XLS-R lost **almost 7 pts** between layers 15 and 24
- ➔ SAMU-XLSR lost **less than 1 pt** between layers 14 and 24
- ➔ *SAMU-XLSR* ***captures and encodes the semantics until its top layer***

Language Portability

1) **Zero-shot:** evaluate the multilingual portability of SAMU-XLSR compared to XLS-R
2) **Low resource:** measure how fine-tuning the speech encoder on French impacts language portability capabilities

1) Zero-shot

	XLS-R		SAMU-XLSR	
	WER	**CER**	**WER**	**CER**
Frozen	129.08	88.24	100.13	54.62
Fine-tuned	123.94	85.36	124.49	83.45

Training on MEDIA (FR) and Testing on PortMEDIA (IT) %

<answer> si grazie >
↓
<answer> oui merzie >

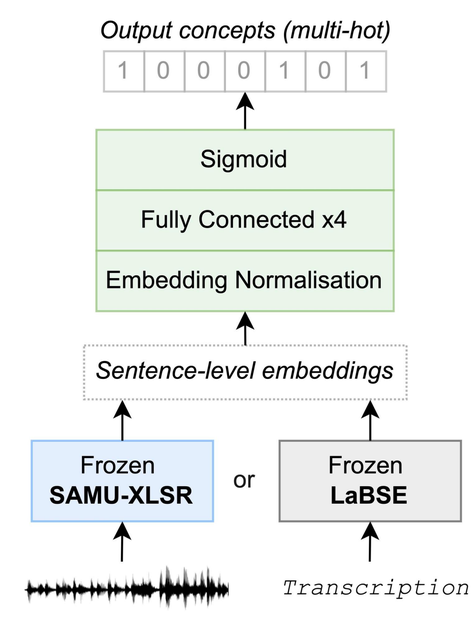
- ➔ SAMU-XLSR **is designed** to extract semantics better than XLS-R
- ➔ SAMU-XLSR **is not designed** to produce language-dependent embeddings
- ➔ **Multilinguality loss**: Fine-tuning SAMU-XLSR on French degrades Italian embeddings quality

2) Low resource

		XLS-R		SAMU-XLSR	
		WER	**CER**	**WER**	**CER**
Frozen	**IT**	36.90	42.66	27.92	33.01
	FR→IT	32.41	35.39	25.09	26.90
Fine-tuned	**IT**	37.02	42.72	16.59	30.66
	FR→IT	20.01	26.92	17.81	26.18

Training on PortMEDIA (IT) or MEDIA then PortMEDIA (FR→IT) and Testing on PortMEDIA (IT) %

- ➔ **SAMU-XLSR outperforms XLS-R** semantic extraction and speech transcription
- ➔ Fine-tuning XLS-R on both French then Italian enhances Italian performances by far
- ➔ Fine-tuning SAMU-XLSR on both French then Italian leads to **better Italian semantic extraction** but **reduces speech transcription performances**



Sentence-level semantic analysis

	Test Encoder	Train Encoder	
		SAMU-XLSR	**LaBSE**
FR	SAMU-XLSR	77.52	71.77
	LaBSE	78.04	82.15
IT	SAMU-XLSR	68.55	65.14
	LaBSE	62.05	69.58

Training on MEDIA (FR) and Testing on MEDIA (FR) and PortMEDIA (IT) *F1-scores %*

- ➔ Model trained on speech (SAMU-XLSR) embeddings can process text (LaBSE) embeddings: **both embeddings are close**
- ➔ Model trained with **text embeddings perform better** than the model trained with **speech embeddings**
- ➔ Good results on **never-seen** Italian data

Conclusion

From investigating the capacity of the recently introduced SAMU-XLSR in addressing a challenging SLU task, we can affirm that:

- ➔ SAMU-XLSR **outperforms** XLS-R semantic extraction and speech transcription by encoding the semantics until its top layer
- ➔ SAMU-XLSR's higher to lower resource pre-training **degrades speech transcription** but **enhances semantics extraction**
- ➔ SAMU-XLSR is building a sentence-level embedding able to **highlight the semantic information of the task**

Omilia LIUM