IMPACT ANALYSIS OF THE USE OF SPEECH AND LANGUAGE MODELS PRETRAINED BY SELF-SUPERSIVION FOR SPOKEN LANGUAGE UNDERSTANDING

Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperriere, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, Yannick Estève

LABORATOIRE INFORMATIQUE D'AVIGNON

LIUM Laboratoire d'Informatique Université du Mans

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

Context

Self-supervised learning have been recently introduced for both acoustic and language modeling.
Pretrained models have shown their great potential by improving the state-of-the-art performances on **Spoken Language Understanding (SLU)**.
In this paper we present an **error analysis** reached by the use of pretrained models for SLU on the French MEDIA benchmark dataset.

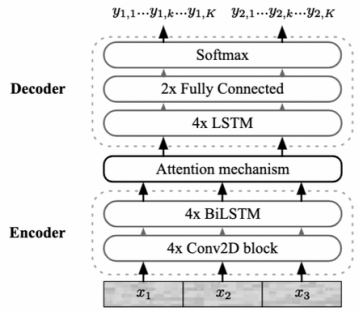
MEDIA dataset

Data	Nb. words	Nb. utterances	Nb. concepts	Nb. hours
train	94.2k	13.7k	31.7k	10h 46m
dev	10.7k	1.3k	3.3k	01h 13m
test	26.6k	3.7k	8.8 k	02h 59m

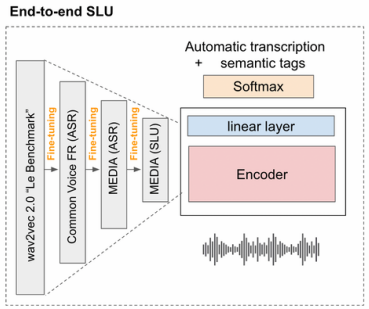
MEDIA corpus [BonneauMaynard2005 et al. (2005)]:
- Telephone dialogue recordings with manual transcriptions and semantic annotations.
- User/woz dialogues about hotel reservations
- The most challenging SLU benchmark available [Béchet and Raymond (2019)]

Comparaison of three Systems

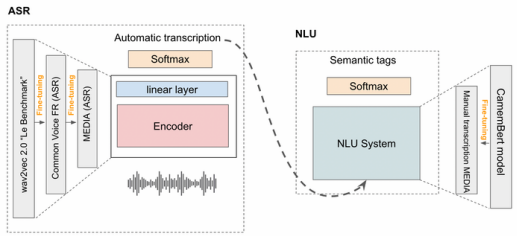
End-to-End Encoder-Decoder Approach with Attention Mechanism



End-to-End fine-tuned wav2vec2.0 for SLU



Cascade system with fine-tuned wav2vec2.0 for ASR and fine-tuned CamemBert for NLU



Systems Performance

Concept Error Rate (CER) / Concept Value Error Rate (CVER)

Model	Dev		Test	
	CER	CVER	CER	CVER
Enc-Dec/AM (Pelloin et al., 2021)	16.1 (±1.2)	20.4 (±1.3)	13.6 (±0.7)	18.5 (±0.8)
wav2vec 2.0 fine-tuned for SLU	15.2 (±1.2)	19.6 (±1.3)	14.5 (±0.7)	18.8 (±0.8)
wav2vec 2.0 + CamemBERT	**12.2** (±1.1)	**16.7** (±1.2)	**11.2** (±0.7)	**17.2** (±0.8)
manual transcription + CamemBERT	9.2 (±1.0)	13.2 (±1.1)	7.5 (±0.6)	12.2 (±0.7)

=> Best system : cascade system wav2vec2.0 model for ASR and CamemBert model for NLU

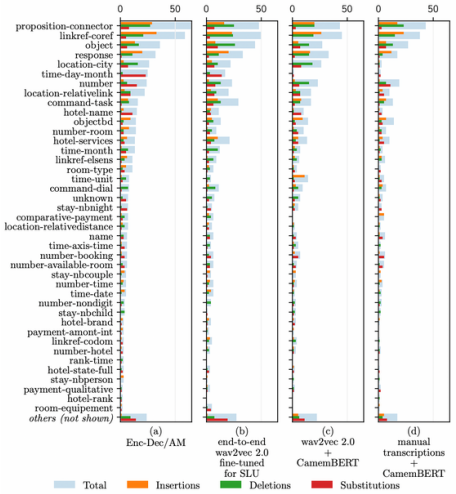
Detailed CER in terms of insertions, substitutions and deletions

Model	Dev			Test		
	Ins	Sub	Del	Ins	Sub	Del
Enc-Dec/AM (Pelloin et al., 2021)	5.3	4.9	5.9	4.3	4.3	4.9
wav2vec 2.0 fine-tuned for SLU	4.1	4.1	7.1	3.8	3.8	6.9
wav2vec 2.0 + CamemBERT	4.1	2.9	5.1	3.4	2.8	4.9
manual transcription + CamemBERT	3.5	2.5	3.2	2.8	2.1	2.6

Major error type: **deletions** for all the systems.
—> Transcriptions errors may prevent the capture of concept
—> Less deletions in the cascade system
—> NLU applied to manual transcription confirms that there is less deletions when transcription is correct

Error Analysis on development dataset

Error Distribution



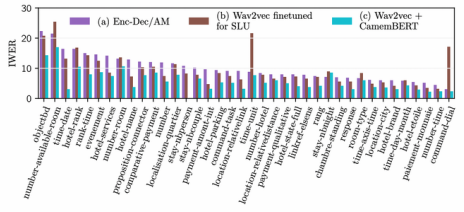
"propostition-connector" and "linkref-coref": **most challenging concepts** in MEDIA
—> Cascade system reduce errors for this two concepts
—> Some concepts are hard to recognize by the cascade system like "location-city"
—> Cascade system is more effective to extract concepts related to date : "time-day-month" , "time-date"

Transcription errors

Word Error Rate (WER) / Individual Word Error Rate (IWER)

Model	Global	Support words
Enc-Dec/AM (Pelloin et al., 2021)	12.37	13.66
wav2vec 2.0 fine-tuned for SLU	12	13.5
wav2vec 2.0 + CamemBERT	7.7	9.27

Individual Word Error Rate for support words by concept



wav2vec 2.0 models used in the cascade and end-to-end approaches are very close
—> Best system in terms of WER: cascade system
—> During the **fine-tuning of the wav2vec 2.0 model on the SLU data**, model forgot some of its automatic speech recognition abilities.
—> The increase of the number of token output (number of characters in ASR + 76 symbols of semantics concepts) can increases the difficulty for the model.

Generalisation capability

Unseen Concept/Value Pairs

Unseen Concept Value (UCV) pairs : concept/value pairs seen in the MEDIA development dataset which do not appear in the training dataset.
Number of UCV pairs on the MEDIA development dataset : **543**

Model	C✓ + V✓	C✗ + V✓
Enc-Dec/AM (Pelloin et al., 2021)	168	32
wav2vec 2.0 fine-tuned for SLU	158	47
wav2vec 2.0 + CamemBERT	242	16
manual transcription + CamemBERT	375	29

Conclusion

Error analysis of three systems to study the impact of pretrained model for spoken language understanding:
- *End-to-End Encoder-Decoder Approach with Attention Mechanism*
- *End-to-End fine-tuned wav2vec 2.0 for SLU*
- *Cascade system with fine-tuned wav2vec2.0 for ASR and fine-tuned CamemBert for NLU*

This paper was partially funded by the European Commission through the SELMA project and by the AISSPER project supported by the French National Research Agency .

AISSPER SELMA