The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning:

data updates, training and evaluation tools

Gaëlle Laperrière[1], Valentin Pelloin[2], Antoine Caubrière[1],
Salima Mdhaffar[1], Nathalie Camelin[2], Sahar Ghannay[3],
Bassam Jabaian[1], Yannick Estève[1]

[1] LIA - Avignon Université, France
[2] LIUM - Le Mans Université, France
[3] Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France

LABORATOIRE INFORMATIQUE D'AVIGNON

LIUM
Laboratoire d'Informatique
Université du Mans

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

LREC 2022
Marseille

Introduction

Context

- **MEDIA** : one of the most challenging accessible SLU dataset
- Multiple recent approaches : end-to-end, pre-training with self-supervision, ...

Goals

- Propose a complete MEDIA **recipe** integrated to **SpeechBrain** with :
  - data preparation
  - training of an end-to-end neural architecture
  - evaluation scripts
- Correct the initial **manual annotations** of MEDIA
- Highlight **evaluation tools** problems for MEDIA

The original MEDIA benchmark

Created by the Technolangue project, 2002, and distributed freely by ELRA for academic purposes.

“ Human-Machine dialogues of hotel reservation
with “Wizard-of-Oz” method,
for semantic extraction tasks from speech”

→ Telephone speech for a French hotel booking task ***[Bonneau-Maynard, et al. 2005]***

→ One of the most challenging SLU corpora ***[Béchet & Raymond, 2019]***

The original MEDIA benchmark

Dataset Specifications

- 1258 dialogues
- 250 speakers

	Nb. Utterances	Nb. Turns	Nb. Dialogues	Nb. Hours	Mean Turn Duration
train	13.7k	13.0k	727	16h56m	4.69s
dev	1.4k	1.3k	79	01h40m	4.77s
test	3.8k	3.5k	208	04h47m	4.89s
unused	4.0k	3.8k	244	05h35m	5.30s

The original MEDIA benchmark

Semantic annotation <semantic-concept [value] word support >

"I <task [reservation] would like to book >
a <room-type [double] double room > ..."

Relax scoring
83 possible concepts

...for <time [12h00] noon >."

Full scoring
1121 possible concepts

...for <time-**begin** [12h00] noon >."

The original MEDIA benchmark

Evaluation Metrics

- Concept Error Rate
- Concept Value Error Rate

$$\frac{\textit{Deletions + Insertions + Substitutions}}{\textit{Nb. of elements in the reference's semantic representation}}$$

→ Results of the French Evalda-Media evaluation campaign for literal understanding ***[Bonneau-Maynard, et al. 2006]***

→ Generative and Discriminative Algorithms for Spoken Language Understanding ***[Christian Raymond, et al. 2007]***

→ Comparing stochastic approaches to Spoken Language Understanding in multiple languages ***[Hahn Stefan, et al. 2010]***

Issues

Concept and Value normalization

- Use of human rules by looking at train and dev corpora
- Not all the possibilities are taken into account.
- **5.7%** of CVER on test reference

Error correction in manual annotation

- Semantic annotations errors due mainly to interpretations
- Audio segmentation problems

Unused data

→ Integrate it in the new MEDIA distribution by ELRA

Issues

r-CVER
rules-based CVER

u-CVER
unnormalized CVER

Reference : I <task [reservation] would like to book >...
Prediction : I <task would lik to book>...

→ Concept : task
→ normalized Value : reservation

→ Concept : task
→ unnormalized Value : would lik to book

Data Updates

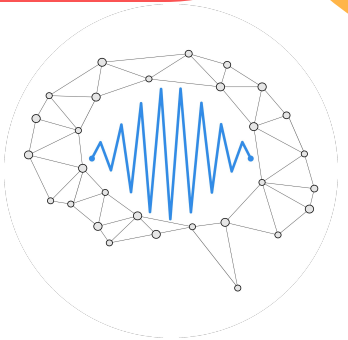
Correction of Manual Annotation

1. Normalization
   - Nouns uppercase, spelling...
2. Semantic annotations
3. Add informations
   - User's audio channel
   - Corrected user ID
4. Use the unused data
   - New test2 corpus

	Nb. words		Nb. concepts		
	Occurrences	Lexicon	Occurrences	Lexicon	
				Full	Relax
train	92.6k	2.3k	31.7k	144	73
dev	10.5k	0.8k	3.3k	104	63
test	26.0k	1.4k	8.8k	125	71
test2	28.0k	1.3k	9.4k	129	71

MEDIA SpeechBrain recipe

https://github.com/speechbrain/speechbrain/tree/develop/recipes/MEDIA
→ for ASR and SLU



Data preparation

- Removed **special characters** except chevrons, hyphens and apostrophes
- Strict respect of the manual annotations about **audio segmentation**
- Removed **hyphens between numbers**
- Put all in **uppercase**
- Process **disfluencies**
- Conversion to SpeechBrain **format**

	Nb. Hours	Mean Turn Duration
train	10h52m	2.85s
dev	01h13m	3.23s
test	03h01m	2.88s
test2	03h16m	2.94s

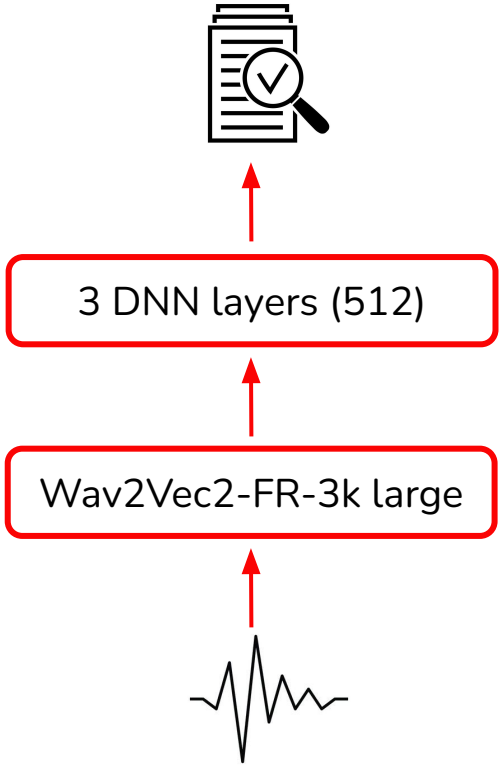
MEDIA SpeechBrain Recipe

Neural Architecture

→ Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark ***[Solène Evain, et al. 2021]***

- LeakyReLU
- SoftMax
- Adam / AdaDelta

- DNN weights init : Random
- W2V weights init : Pre-trained
  - *3K hours of speech / broadcast news (LeBenchmark)*
  - *450 hours of speech (CommonVoice FR)*



First Experimental Results

	Model	test		
		CER	u-CVER	r-CVER
Relax	**media-base**	21.8	34.1	29.4

➢ u-CVER is more reliable but **stricter** by 5 points than r-CVER

First Experimental Results

	Model	test		
		CER	u-CVER	r-CVER
Relax	**media-base**	21.8	34.1	29.4
	media-comvoice	16.3	27.7	23.7

- **media-comvoice** strongly better than media-base
- recipe operational, only needs tuning to enhance the results to the state-of-the-art

First Experimental Results

	Model	test2			
		ChER	CER	u-CVER	r-CVER
Full	**media-comvoice**	6.7	21.1	30.9	-
Relax	**media-comvoice**	6.4	16.4	27.1	21.0

➢ consistency of the new corpus results

Conclusion

- First results on the new sub-corpus
- Raised the interrogations about the human rules-based CVER
- Hopefully brought back interest for the Full and Relax scorings of MEDIA
- Shared a user-friendly recipe on a maintained and reliable toolkit to increase SLU researches

Goals

- Update the initial **manual annotations** of MEDIA ✓
- Propose a complete MEDIA **recipe** integrated to **SpeechBrain** with : ✓
  - data preparation
  - training of an end-to-end neural architecture
  - evaluation scripts
- Highlight **evaluation tools** problems for MEDIA ✓