The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools

Gaëlle Laperrière¹, Valentin Pelloin², Antoine Caubrière¹, Salima Mdhaffar¹, Nathalie Camelin², Sahar Ghannay³, Bassam Jabaian¹, Yannick Estève¹

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

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

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

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

→ Concept: task
→ normalized Value: reservation

u-CVER
unnormalized CVER

→ 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

<table>
<thead>
<tr>
<th></th>
<th>Nb. words</th>
<th>Nb. concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Occurrences</td>
<td>Lexicon</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train</td>
<td>92.6k</td>
<td>2.3k</td>
</tr>
<tr>
<td>dev</td>
<td>10.5k</td>
<td>0.8k</td>
</tr>
<tr>
<td>test</td>
<td>26.0k</td>
<td>1.4k</td>
</tr>
<tr>
<td>test2</td>
<td>28.0k</td>
<td>1.3k</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th></th>
<th>Nb. Hours</th>
<th>Mean Turn Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>train</strong></td>
<td>10h52m</td>
<td>2.85s</td>
</tr>
<tr>
<td><strong>dev</strong></td>
<td>01h13m</td>
<td>3.23s</td>
</tr>
<tr>
<td><strong>test</strong></td>
<td>03h01m</td>
<td>2.88s</td>
</tr>
<tr>
<td><strong>test2</strong></td>
<td>03h16m</td>
<td>2.94s</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CER</td>
</tr>
<tr>
<td>Relax</td>
<td>21.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CER</td>
</tr>
<tr>
<td>media-base</td>
<td>21.8</td>
</tr>
<tr>
<td>media-comvoice</td>
<td>16.3</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Model</th>
<th>test2</th>
<th>media-comvoice</th>
<th>ChER</th>
<th>CER</th>
<th>u-CVER</th>
<th>r-CVER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td></td>
<td>media-comvoice</td>
<td>6.7</td>
<td>21.1</td>
<td>30.9</td>
<td>-</td>
</tr>
<tr>
<td>Relax</td>
<td></td>
<td>media-comvoice</td>
<td>6.4</td>
<td>16.4</td>
<td>27.1</td>
<td>21.0</td>
</tr>
</tbody>
</table>

➢ 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 ✅