End2End Acoustic to Semantic Transduction

Valentin Pelloin[1] Nathalie Camelin[1] Antoine Laurent[1] Renato De Mori[2,3]
Antoine Caubrière[1] Yannick Estève[2] Sylvain Meignier[1]

[1]LIUM - Le Mans University - France [2]LIA - Avignon University - France [3]McGill University - Montréal, Canada

Paper ID: 3182

Introduction

We tackle the complex task of Spoken Language Understanding (SLU) in a **human-computer dialogue** system, based on a frame language to represent semantic domains:

- **domain specific frames** (e.g. *HOTEL*, *ROOM*)
- **domain independent frames** (e.g. logical connectors)

We propose an end-to-end sequence-to-sequence system for the French MEDIA

- using an **attention mechanism** to focus on relevant audio semantic contexts
- **transduce directly** pairs of **concepts/values** from acoustic representations

Spoken Language Understanding (SLU)

How to obtain concepts and values pairs ?

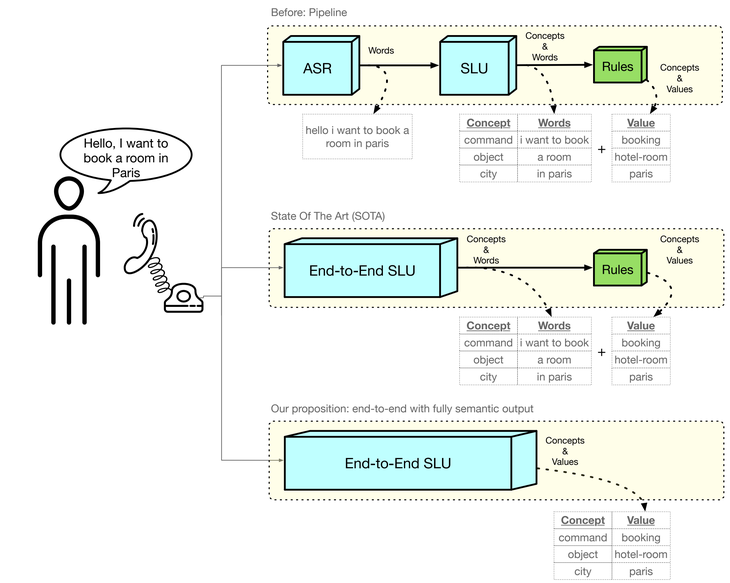

Figure 1. Spoken Language Understanding using a pipeline chain, an end-to-end method with human-rules OR an end-to-end method to fully extract semantic content

Different outputs:

AllWords-C	`hello <command i want to book > <object a room > <city in paris >`
SupWords-C	`* <command i want to book > <object a room > <city in paris >`
NormValues-C	`* <command booking > <object hotel-room > <city paris >`

End-to-End SLU model

- **Attention based encoder-decoder** model architecture (Figure 2).
  - inputs: MelFBanks
  - outputs: sequence of characters, with special characters representing the concepts
- Adapted from the WSJ ASR task recipe provided in *Espresso*[2].
- **Shallow-fusion** is used to incorporate an additional **LSTM-based Language Model**.
  - word-based LM
  - lookahead-probabilities to provide word and character probabilities during decoding

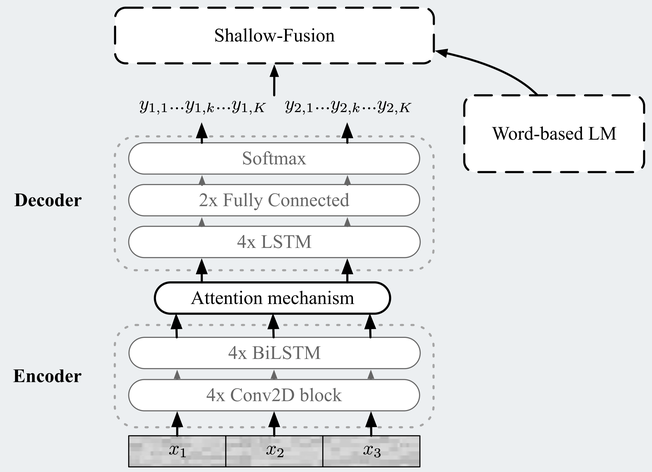

Figure 2. Our speech encoder-decoder model with it's Language Model used during decoding

Experimental Protocol

1. Datasets
   - French ASR generic corpora: mainly broadcast news. Total: 414 hours of audio ASR training
   - PORTMEDIA (PM): French SLU telephone calls about theater shows. 7h of audio SLU training
   - **MEDIA** (M): French SLU telephone conversations, requests and lodging booking. 76 different semantic concepts. Total of audio training for SLU: 17h

	Train	Dev	Test
Number of words	95 022	10 870	26 820
Number of frame instances	31 670	3 333	8 788

Table 1. MEDIA datasets statistics

2. Training chain – inspired by the *curriculum-based tranfer learning* procedure proposed in [1]:
   AllWords-C: ASR → ASR M+PM → SLU M+PM → SLU M
   SupWords-C: ASR → ASR M+PM → SLU M+PM → SLU M*
   NormValues-C: ASR → ASR M+PM → SLU M+PM → SLU Norm M*

Results

- Evaluation metrics
  - **Concept Error Rate (CER)**: only the sequence of concepts is considered
  - **Concept-Value Error Rate (CVER)**: the sequence of pairs concept/value is considered. Both concept AND value must be correct.
- Performances on the French MEDIA dataset

	Dev		Test	
%	CER	CVER	CER	CVER
Without a Language Model				
AllWords-C [1]	–	–	21.6	27.7
AllWords-C	18.1	22.5	15.6	**20.4**
SupWords-C	17.3	22.0	15.6	20.5
NormValues-C	16.0	21.9	**15.4**	21.7
With a Language Model				
AllWords-C [1]	–	–	18.1	22.1
SupWords-C [1]	–	–	16.4	20.9
AllWords-C	16.1	20.4	**13.6**	**18.5**
SupWords-C	17.6	22.5	15.5	20.5
NormValues-C	16.1	22.0	15.4	21.6

Table 2. Results obtained on the MEDIA Dev and Test Corpora by our models compared to the State of the art [1].

- **Attention mechanism**: great results improvements comparing to SOTA (2.8% CER and 2.4% CVER absolute gains)
- **Direct transduction from audio features to semantic pairs concept/value**: good CER results, CVER results are still quite good (better than SOTA), but no gain with LM

Conclusion and perspectives

- **Suitable end-to-End model** with attention for the complex MEDIA SLU task
- **Good results without hand-made rules**. Needs further investigation to:
  - find a Language Model that improves these results
  - improve value recognition
- Combining **multiple decoders** might improve the value recognition
- Building **new attention mechanisms** to select suitable semantic sound spans

References

[1] A. Caubrière, N. Tomashenko, A. Laurent, E. Morin, N. Camelin, and Y. Estève. Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. *INTERSPEECH*, pages 1198–1202, 2019.

[2] Y. Wang, T. Chen, H. Xu, S. Ding, H. Lv, Y. Shao, N. Peng, L. Xie, S. Watanabe, and S. Khudanpur. Espresso: A fast end-to-end neural speech recognition toolkit. In *ASRU*, pages 136–143, 2019.

This work has been funded by the AISSPER project supported by the French National Research Agency (ANR) under contract ANR-19-CE23-0004-01.