IEEE 2018 Workshop on Spoken Language Technology
« Spoken Language Technology in the Era of Deep Learning:
Challenges and Opportunities »

# End-to-end named entity and semantic concept extraction from speech

Sahar Ghannay[1, (*)], Antoine Caubrière[1], Y. Estève[2], Nathalie Camelin[1], Edwin Simonnet[1], Antoine Laurent[1], Emmanuel Morin[3]

[1] LIUM - University of Le Mans, [2] LIA - University of Avignon, [3] LS2N - University of Nantes — France

[(*)] now in LIMSI

## Abstract

Named entity recognition (NER) is among SLU tasks that usually extract semantic information from textual documents.
Until now, NER from speech is made through a pipeline process that consists in processing first an automatic speech recognition (ASR) on the audio and then processing a NER on the ASR outputs. Such approach has some disadvantages (error propagation, metric to tune ASR systems sub-optimal in regards to the final task, reduced space search at the ASR output level,...) and it is known that more integrated approaches outperform sequential ones, when they can be applied.
In this study, we explore an end-to-end approach that directly extracts named entities from speech, though a unique neural architecture.
We also explore this approach applied to semantic concept extraction, through a slot filling task known as a spoken language understanding problem.

## Named entity recognition through a speech to character sequence approach

word sequence ⇨ *le sculpteur césar est mort hier à paris à l' âge de soixante dix sept ans*

word sequence with EN tags ⇨ le sculpteur **<pers** césar **>** est mort **<time** hier **>** à **<loc** paris **>** à l' âge de **<amount** soixante dix sept ans **>**

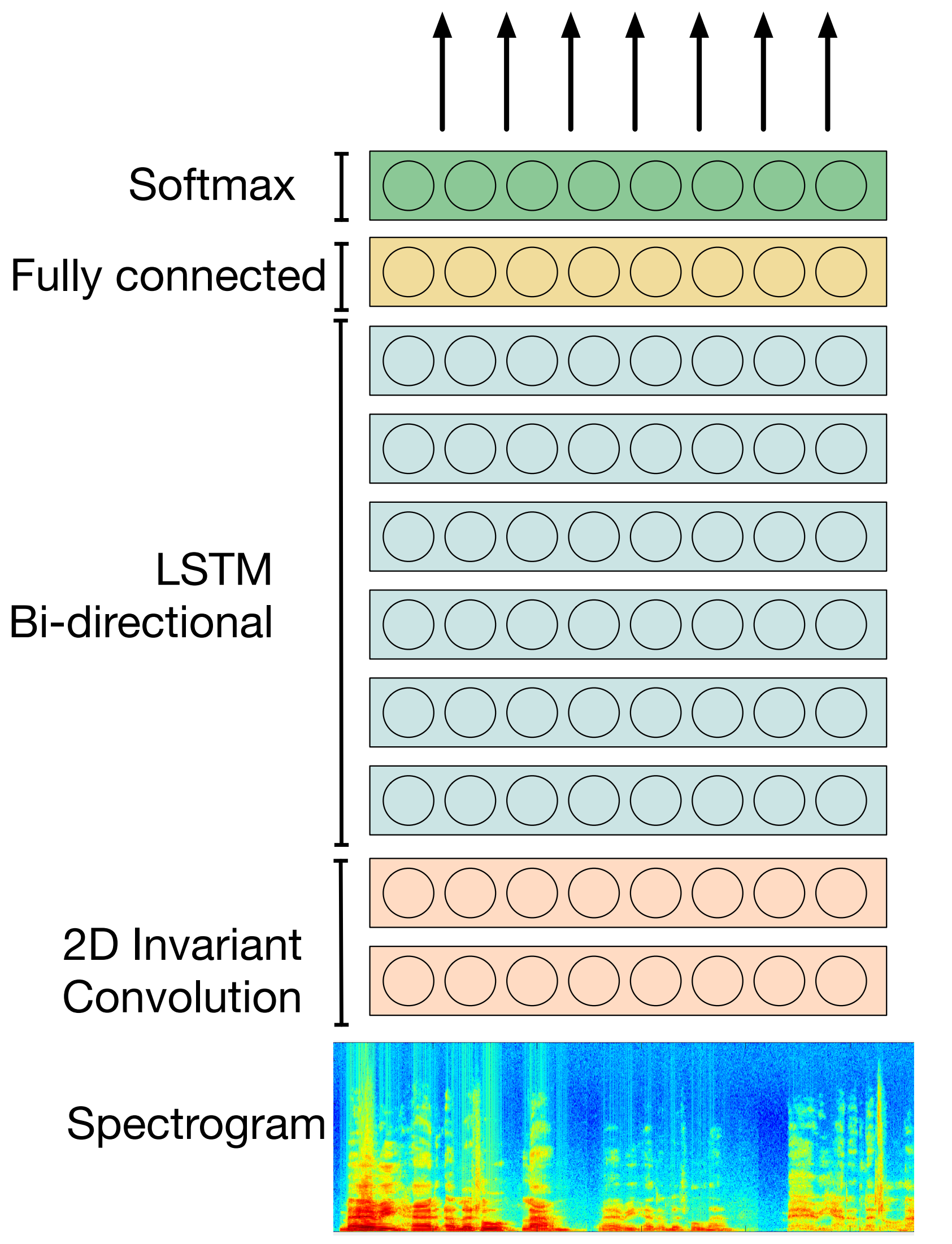
starred mode: only EN tags, EN values, and stars ⇨ * **<pers** césar **>** * **<time** hier **>** * **<loc** paris **>** * **<amount** soixante dix sept ans **>**

*This sentence means, in English and case sensitive: « the sculptor **Caesar** died **yesterday** in **Paris** at the age of **seventy-seven years** »*

1 The sequence-to-sequence architecture used in this study is very close to the Deep Speech2 neural ASR system proposed by Baidu.
The system is trained end-to-end using the CTC loss function, in order to predict a sequence of characters from the input audio.

2 We would like to evaluate if a such neural architecture is able to capture high level semantic information that allow it to recognize named entities. For that, we propose to modify the character sequence that the neural network has to produce: information about named entities are added in the initial character sequence.
Usually, the named entity recognition task is to assign a named entity tag to every word in a sentence. A single named entity could concern several words within a sentence. For this reason, the word-level labels begin-inside-outside (BIO) encoding is very often adopted.

3 Instead of applying a BIO approach, we propose to add some tag characters in this sequence to delimit named entities boundaries, but also their category.
We are interested to eight NE categories that are: person, function, organization, location, production, amount, time and event.



*Two CNN layers and six bidirectional recurrent layers with batch normalization are used in the experiments*

### 4 Multi-task training

To compensate the lack of data (audio + manual annotations of named entities), we apply a multi-task learning approach to train the neural network.

1. First, train it only for the ASR task on all the audio recordings available with their manual transcriptions (~ 300 hours of speech).

2. The softmax layer is reinitialized to take into consideration the named entity tag markers, and a new training process is realized, restricted to training data with manual annotations of named entities (~160 hours).

### 5 Data augmentation (+)

1. We apply a named entity recognition system dedicated to text data in order to tag the manual transcriptions used to train the ASR neural network.

2. These manual transcriptions automatically annotated with named entities are injected in the training data used to train the neural network to extract named entities from speech.

### 6 Starred mode (*)

Since the CTC loss gives the same importance to each character, we propose to modify the character sequence that the neural network must emit to give more importance to named entities.

This proposition is interesting to better understand how the CTC loss behaves on this case, and consists in replacing by a star * all character subsequences that do not contain a named entity.

## Experimental results on ETAPE and QUAERO test data (French broadcast news)

| System | Detection | Precision | Recall | F-measure |
|---|---|---|---|---|
| E2E | category | **0.83** | 0.52 | 0.64 |
| E2E* | category | 0.82 | 0.57 | 0.67 |
| E2E+ | category | 0.82 | 0.57 | 0.67 |
| E2E+* | category | 0.76 | **0.63** | **0.69** |
| E2E | cat+value | **0.55** | 0.36 | 0.44 |
| E2E* | cat+value | 0.47 | 0.38 | 0.42 |
| E2E+ | cat+value | **0.55** | 0.40 | 0.46 |
| E2E+* | cat+value | 0.49 | **0.41** | **0.47** |

## Comparison to a classical pipeline approach (two steps: ASR first, NER last)

*End-to-end ASR
WER =19.95%
CER = 7.68%*

*Neural NER system:
NeuroNLP, also used for
data augmentation (+)*

| System | Detection | Precision | Recall | F-measure |
|---|---|---|---|---|
| Pip | category | **0.75** | 0.56 | 0.64 |
| Pip+POS | category | 0.74 | **0.58** | **0.65** |
| Pip | cat+value | **0.58** | 0.43 | 0.49 |
| Pip+POS | cat+value | 0.57 | **0.45** | **0.50** |

## Extension to slot filling task (semantic concept extraction)

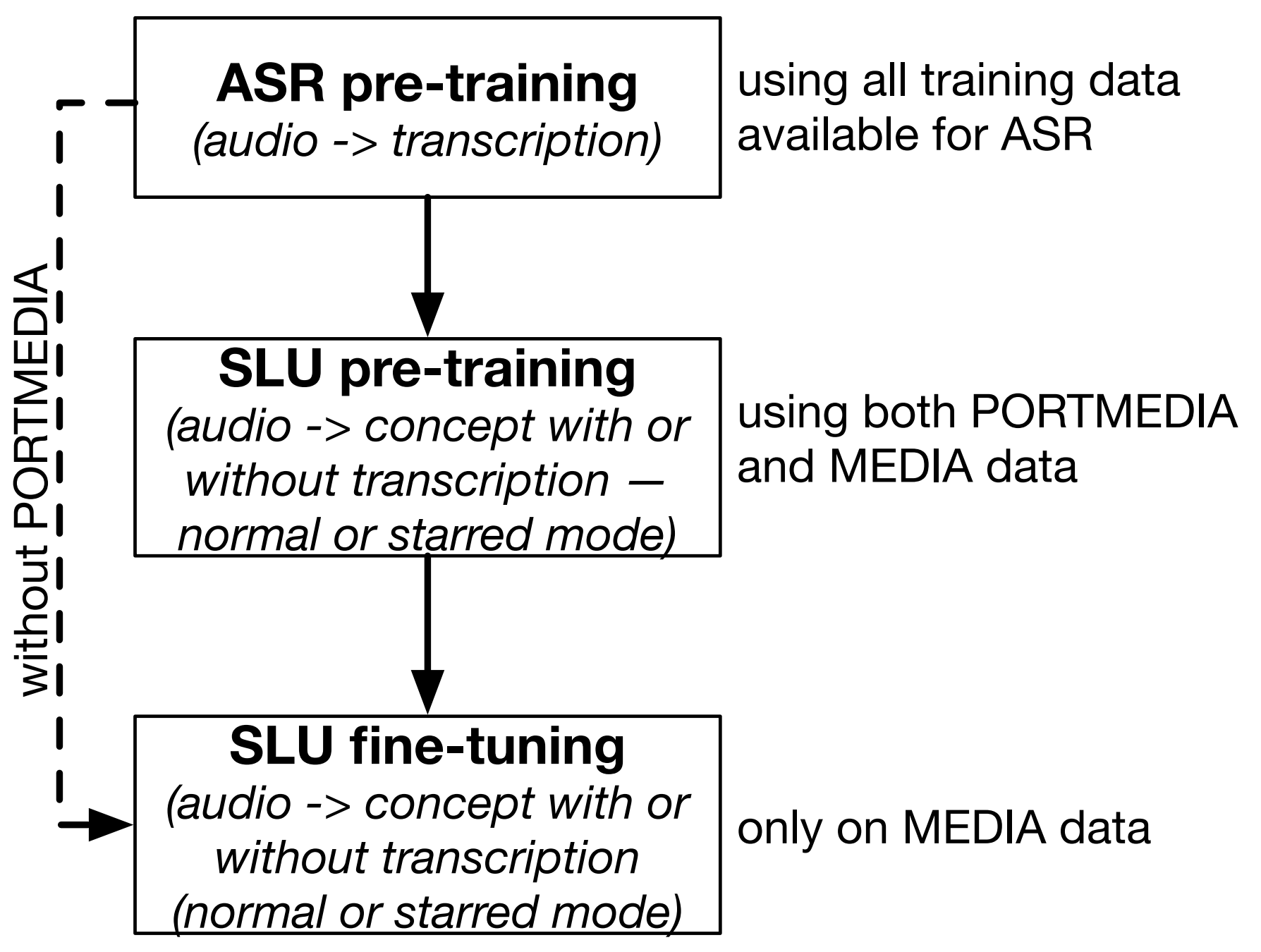
Named entities are replaced by semantic concepts to be detected in a human/machine spoken dialogue scenario.

The main target is the MEDIA corpus, dedicated to hotel booking.

76 semantic tags have to be recognized, *e.g. « number of room », « hotel name », « localization », « month », « room equipment »…*

A secondary target is the PORT-MEDIA corpus, dedicated to reservation of theater tickets.

Both are composed of telephone conversations, and were collected through a wizard of Oz approach, in which a human plays the role of the machine.



**ASR pre-training**
*(audio -> transcription)*
using all training data available for ASR

**SLU pre-training**
*(audio -> concept with or without transcription — normal or starred mode)*
using both PORTMEDIA and MEDIA data

**SLU fine-tuning**
*(audio -> concept with or without transcription (normal or starred mode)*
only on MEDIA data

## Experimental results on MEDIA

| System | Concept Error Rate |
|---|---|
| Pip-SLU | 32.0% |
| E2E without PORTMEDIA | 29.3% |
| E2E pretrained with PORTMEDIA | 28.1% |
| E2E* pretrained with PORTMEDIA | **27.0%** |

Pipeline approach (Pip-SLU)

*End-to-end ASR
WER = 20.4%*
✚
*Encoder-decoder with attention mechanism for concept recognition*