2021 - St Petersburg

Where are we in semantic concept extraction for Spoken Language Understanding?

Sahar Ghannay[1], Antoine Caubrière[2], Salima Mdhaffar[2], Gaëlle Laperrière[2]
Bassam Jabaian[2], Yannick Estève[2]

[1] *LISN - Paris-Saclay University, France*
[2] *LIA - Avignon University, France*

Introduction

Context

Spoken Language Understanding has seen a lot of progress recently

Emergence of the End-to-End (E2E) approach based on deep neural networks

Self-supervised training with unlabeled data open new perspectives

Our study is in the context of the challenging french MEDIA task

Introduction

Context

Spoken Language Understanding has seen a lot of progress recently

Emergence of the End-to-End (E2E) approach based on deep neural networks

Self-supervised training with unlabeled data open new perspectives

Our study is in the context of the challenging french MEDIA task

Goal

Observe the recent progress on MEDIA with both E2E and cascade approaches
Improve the state-of-the-art by using self-supervised pre-trained models

The French MEDIA task

The MEDIA corpus

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

Simulation of dialog system recorded with the “wizard-of-oz” method

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

The French MEDIA task

The MEDIA corpus

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

Simulation of dialog system recorded with the “wizard-of-oz” method

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

Corpus specification

Annotation according to 76 semantic concepts (*location-town*, *stay-nbNight*, *nb-reservation*, ...)

Data	Nb Words	Nb Utterances	Nb Concepts	Nb Hours
Train	94.2k	13.7k	31.7k	10h46
Dev	10.7k	1.3k	3.3k	01h13
Test	26.6k	3.7k	8.8k	02h59

The French MEDIA task

The MEDIA corpus

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

Simulation of dialog system recorded with the “wizard-of-oz” method

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

Corpus specification

Annotation according to 76 semantic concepts (*location-town*, *stay-nbNight*, *nb-reservation*, ...)

Data	Nb Words	Nb Utterances	Nb Concepts	Nb Hours
Train	94.2k	13.7k	31.7k	10h46
Dev	10.7k	1.3k	3.3k	01h13
Test	26.6k	3.7k	8.8k	02h59

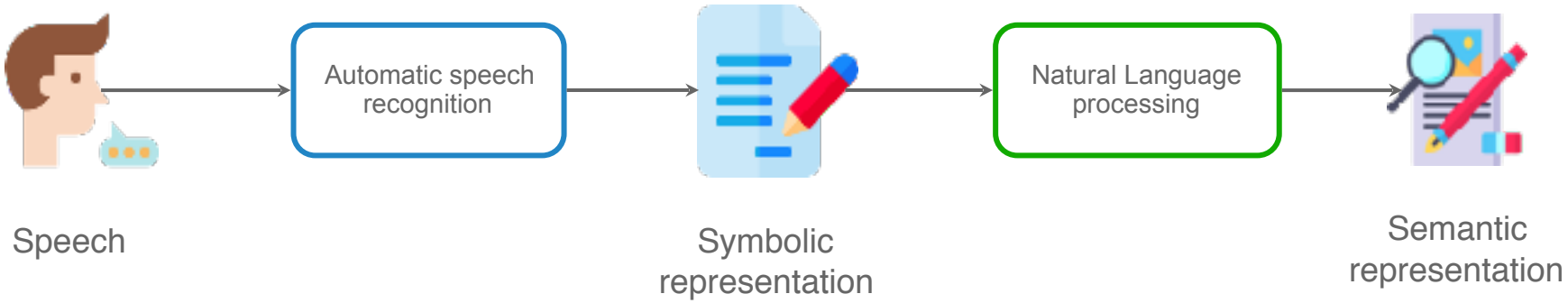
Evaluation metrics

CER : Evaluates concepts only

CVER : Evaluates concepts and value

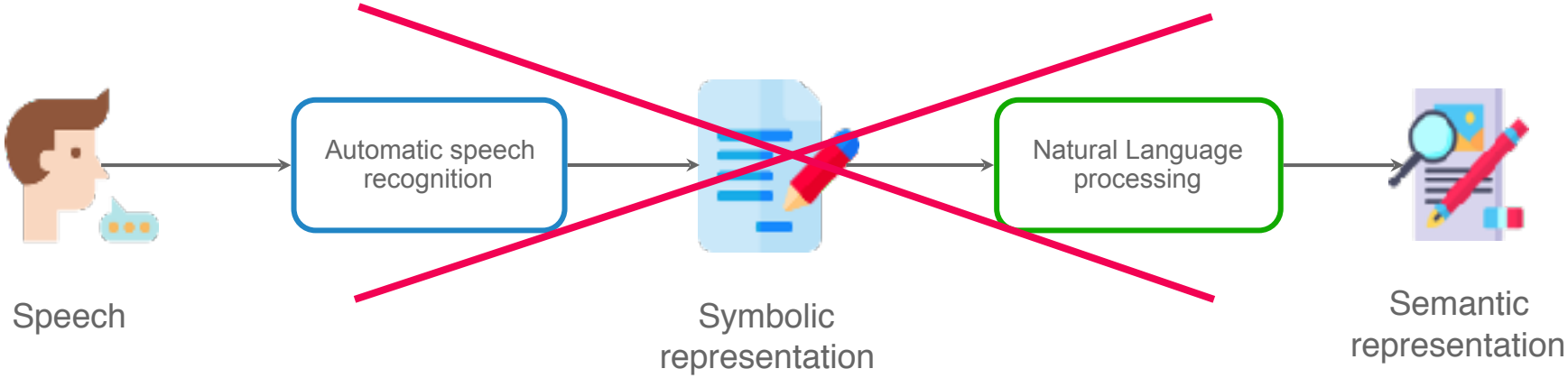
Cascade vs E2E approach

Cascade approach



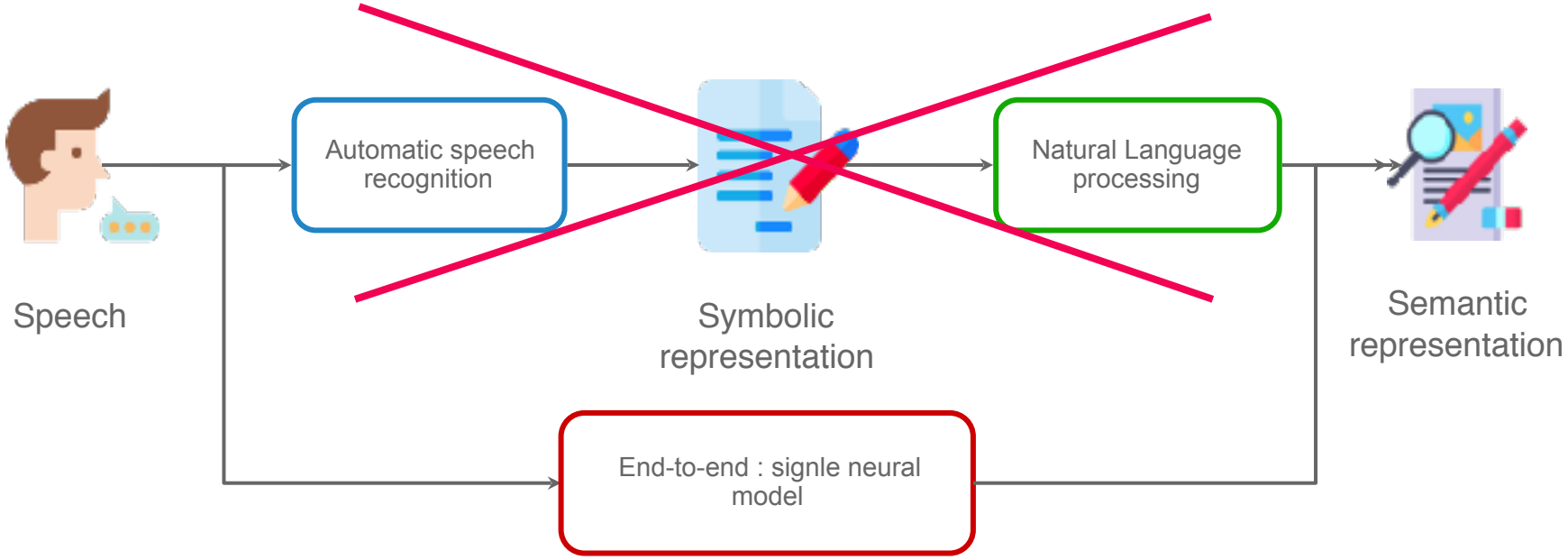
Cascade vs E2E approach

E2E approach



Cascade vs E2E approach

E2E approach



Cascade approach

Semantic labeling using BIO format

Hello **<command-task** i want to book **>** **<nbNight** a night **>**

A label predicted for each words

Hello	**O**
i	**B-comm**
want	**I-comma**
to	**I-comma**
book	**I-comma**
a	**B-nbNig**
night	**I-nbNigh**

Recent advances

System	CER	CVER
HMM-DNN + Neural NLU *[Simonnet et al. 2018]*	20.2	26.0
HMM-DNN + CRF *[Simonnet et al. 2018]*	20.2	25.3
HMM-TDNN + CRF *[Caubrière et al. 2019]*	**16.1**	**20.4**

E2E approach

Semantic labeling using boundaries

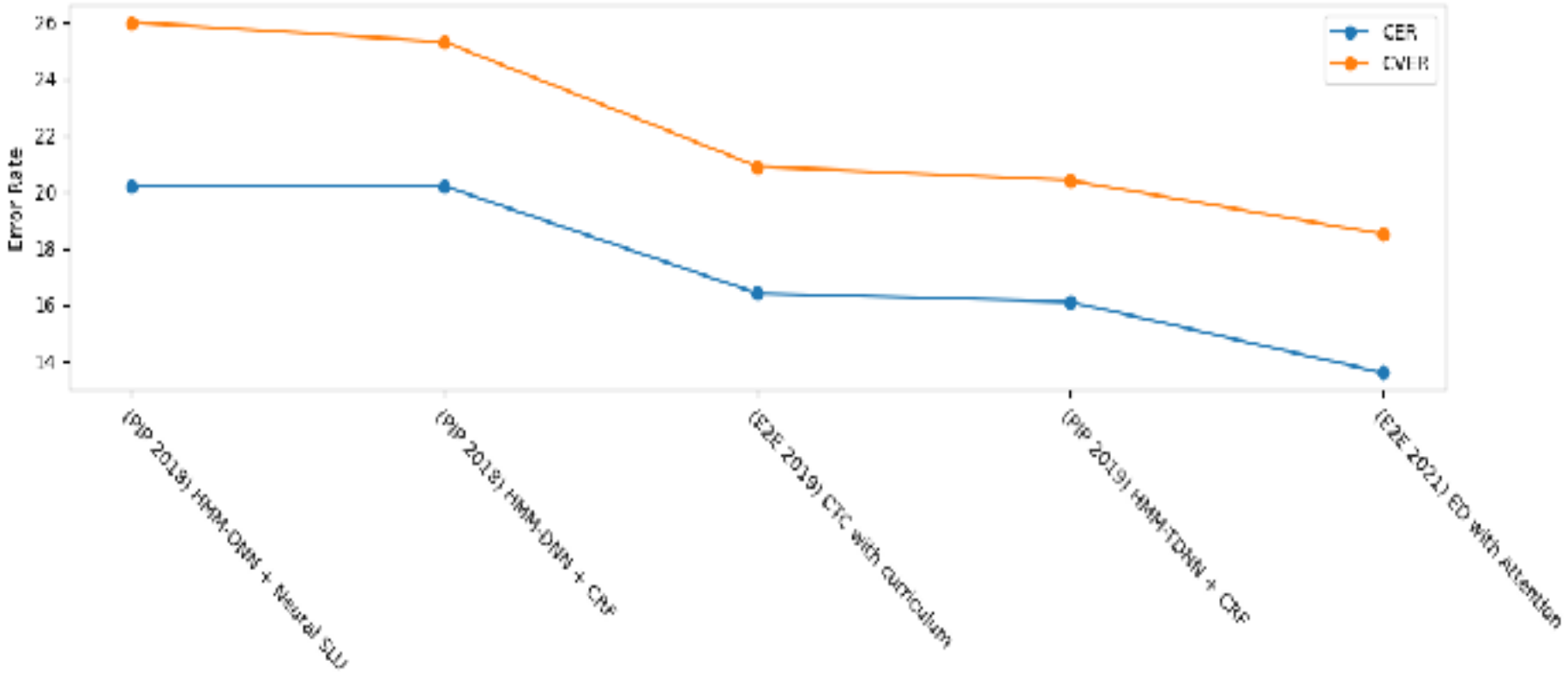
Hello **<command-task** i want to book **>** **<nbNight** a night **>**

Add concepts boundaries in the sequences to be produced

Recent advances

System	CER	CVER
CTC approach with curriculum *[Caubrière et al. 2019]*	16.4	20.9
Encoder-decoder with Attention *[Pelloin et al. 2021]*	**13.6**	**18.5**

SoA in time



Confidence Interval

Confidence degree: 95%

Confidence margin: CER = 0.7% ; CVER = 0.8%

Improving the SoA

Our proposal

Use of pre-trained models with a large amount of data

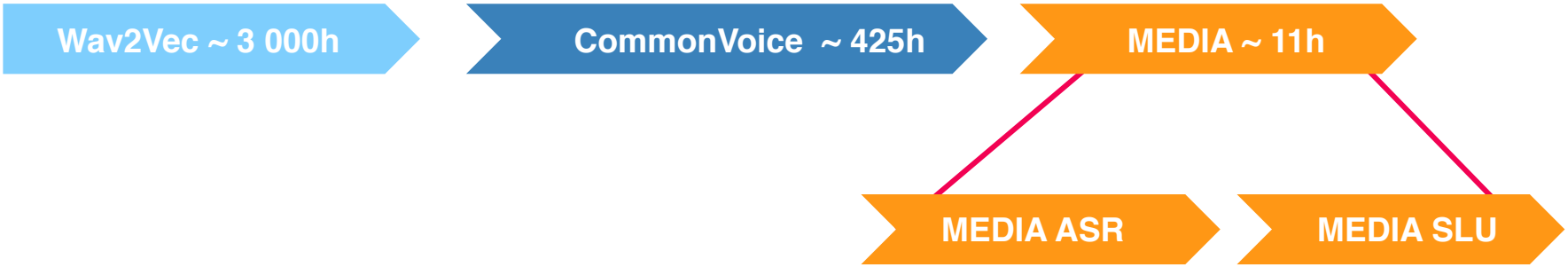
Compare E2E and cascade approach

E2E approach with Wav2Vec

Use a french self-supervised pre-trained Wav2Vec 2.0 model ***[Evain et al. 2021]***

Finetune the model with first French Common Voice and then MEDIA task

Split the MEDIA task into the two subtasks ASR and SLU



E2E Wav2Vec Results

Beam search decoding

5-gram language model trained with MEDIA manual transcription

System	CER	CVER
W2V • M-slu	18.8	23.6
W2V • common Voice • M-slu	15.8	20.4
W2V • common Voice • M-asr • M-slu	**14.5**	**18.8**

Cascade with CamemBert

ASR component performance

System	WER
Last pipeline ASR (HMM-TDNN) ***[Caubrière et al. 2019]***	9.3
W2V • common Voice • M-asr	**8.5**

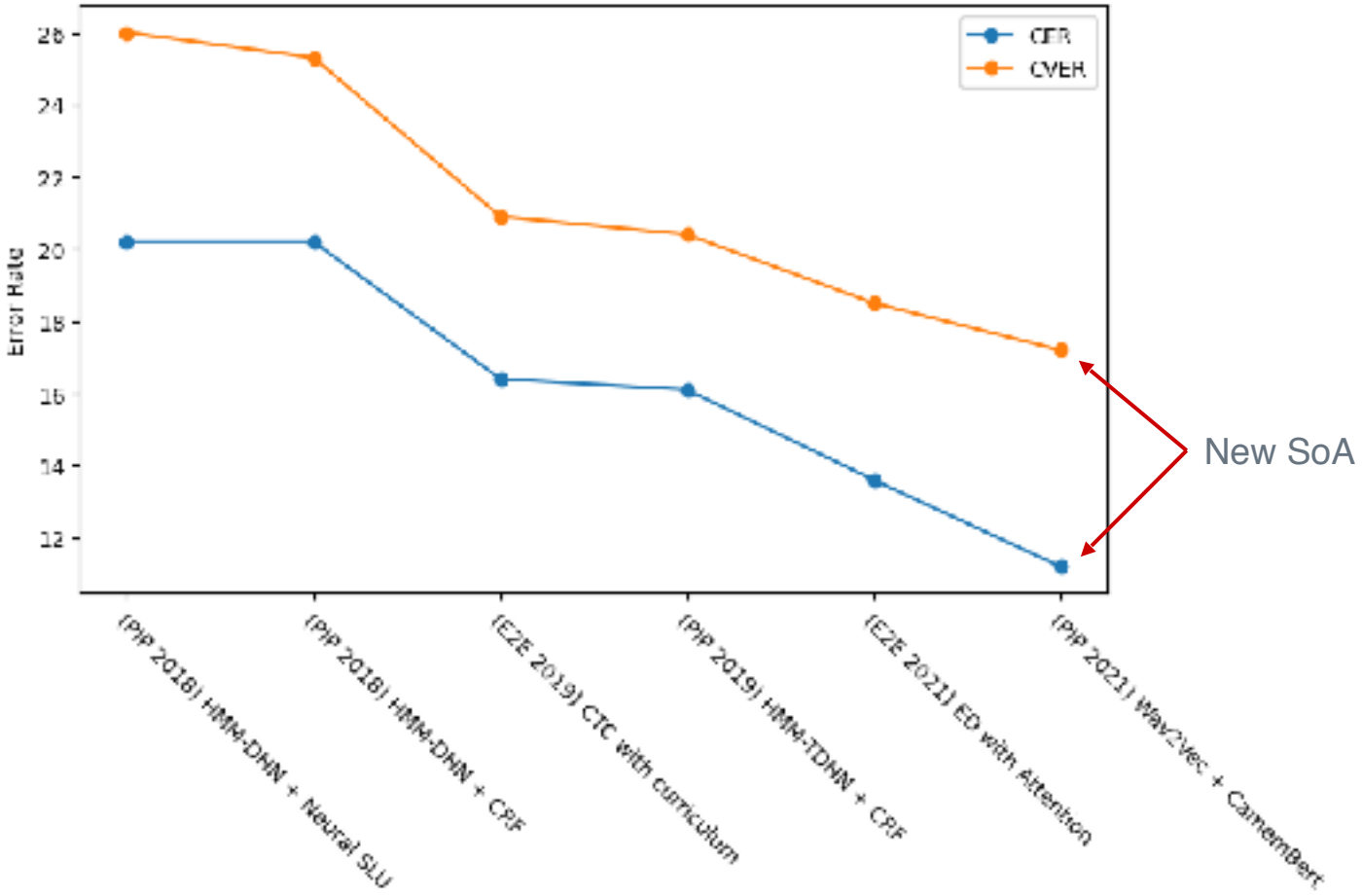
NLU component: CamemBert

Pretrained on the French CCnet corpus composed of 135 GB of raw text ***[Martin et al. 2020]***

Finetuning on the manual transcription of MEDIA

System	CER	CVER
W2V • common Voice • M-asr + CamemBert	**11.2**	**17.2**
Manual transcription + CamemBert ***[Ghannay et al. 2020]***	**7.56**	**X**

Cascade with CamemBert



Conclusion

We presented an overview of recent advances on the French SLU task: MEDIA

We compare both End-to-End and cascade approaches

Recently E2E approaches get very good results on MEDIA (CER 13.6%) ***[Pelloin et al. 2021]***

We proposed a cascade approach based on components pre-trained with unlabelled data

We combine Wav2Vec as ASR and CamemBert as NLU systems

We significantly outperformed the last E2E approach by reach a CER of 11.2

Thank you

The goal of this work

Observe recent advances for the MEDIA task

Improve the state of the art with self-supervised pretrained models