TASK SPECIFIC SENTENCE EMBEDDINGS FOR ASR ERROR DETECTION

Sahar Ghannay, Yannick Estève and Nathalie Camelin
LIUM- Le Mans University, France

Introduction

Error detection

- Supervised machine learning task
- Detection of anomalies in automatic transcriptions:
  - from linguistic or semantic levels
  - from acoustic level

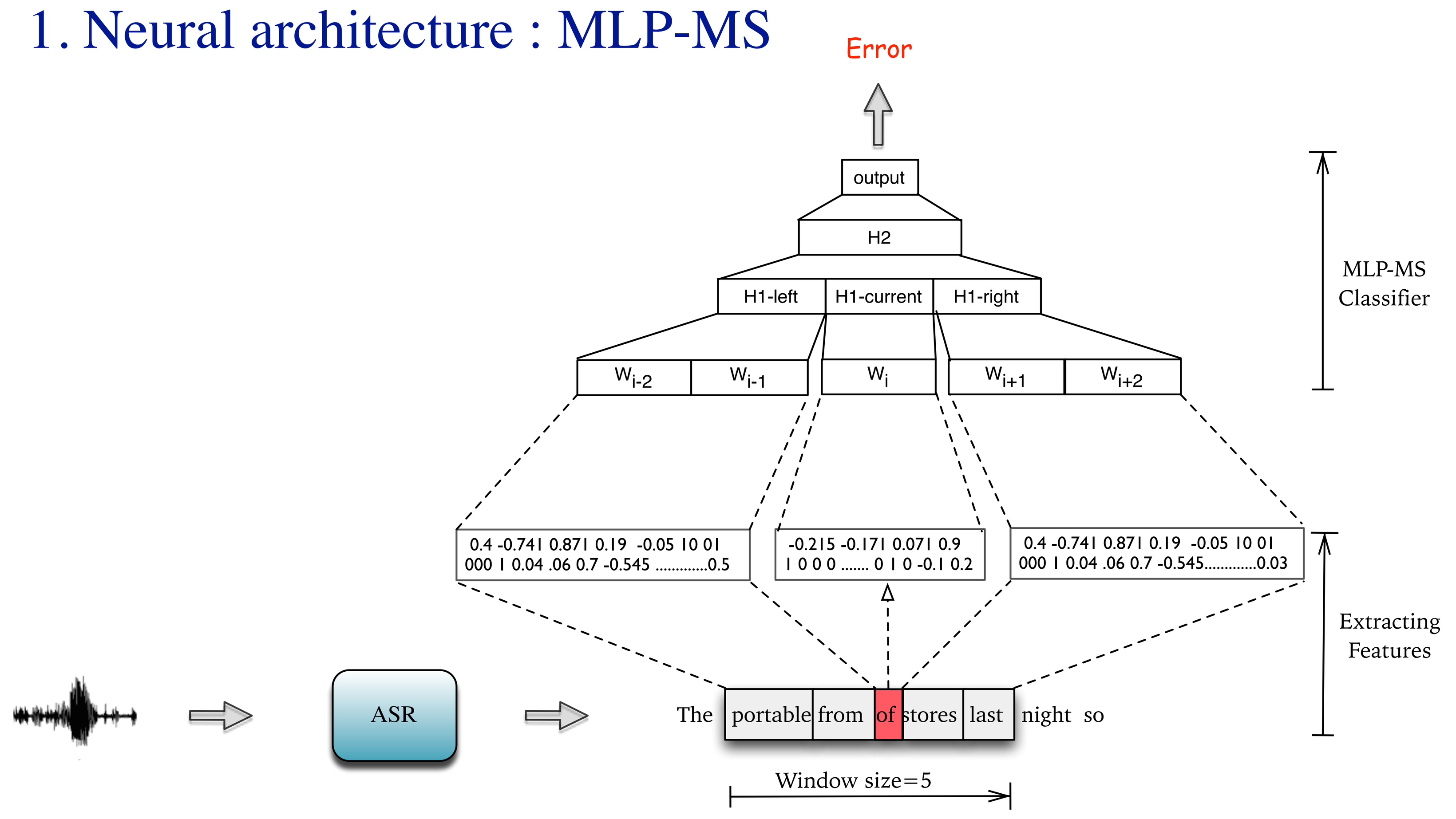
Main goal

- Modeling automatic speech recognition (ASR) errors at the sentence level through:
  - Continuous sentence representations (*embeddings*) specific to ASR error detection task
  - Probabilistic contextuel model

In addition, we will compare our approaches to bidirectional long short-term memory (BLSTM) architecture previously published.

ASR error detection

1. Neural architecture : MLP-MS



2. List of features

- **Posterior probabilities**
- **Lexical features:** word length, existence of 3-grams in the ML
- **Syntactic features**: POS tag, dependency label, word governor
- **Prosodic features:** number and average duration of phonemes, duration of previous and next pause, average f0 of the word, *etc.*
- **Word**: linguistic and acoustic embeddings

Experiments

1. Experimental data

Automatic transcriptions of Etape Corpus

Name	#words ref	#words hyp	WER
Train	349K	316K	25.3
Dev	54K	50K	24.6
Test	58K	53K	21.9

2. Baseline results

Sys1: all features described above excepting the prosodic ones

Sys2: all features

Corpus	System	Label *Error*			Global
		P	R	F	CER
Dev	*Sys1*	0.71	0.58	0.64	9.53
	Sys2	0.71	0.60	0.65	**9.38**
Test	*Sys1*	0.70	0.59	0.64	7.94
	Sys2	0.70	0.61	0.65	**7.75**

+ prosodic features improve all the results

Average ASR error span analysis of Sys2 outputs:

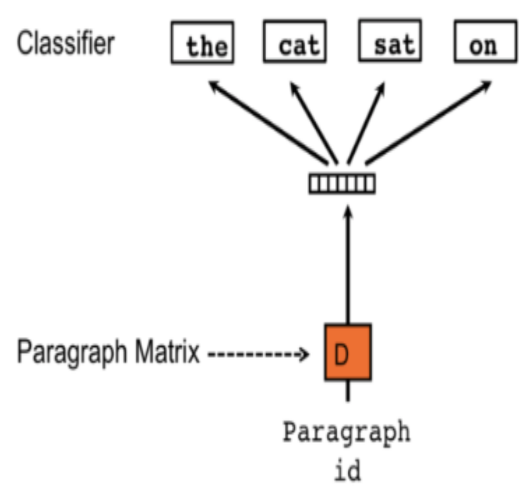
Corpus		Average span	Standard deviation
Train	Ground truth	3.03	1.72
Dev		3.24	2.15
Dev	Predictions	2.82	1.28
	Correct predictions	2.66	1.05

→ Average span of Sys2 is too small in comparison to the ground truth

Experiments: Integration of global information

1. Continuous sentence representations

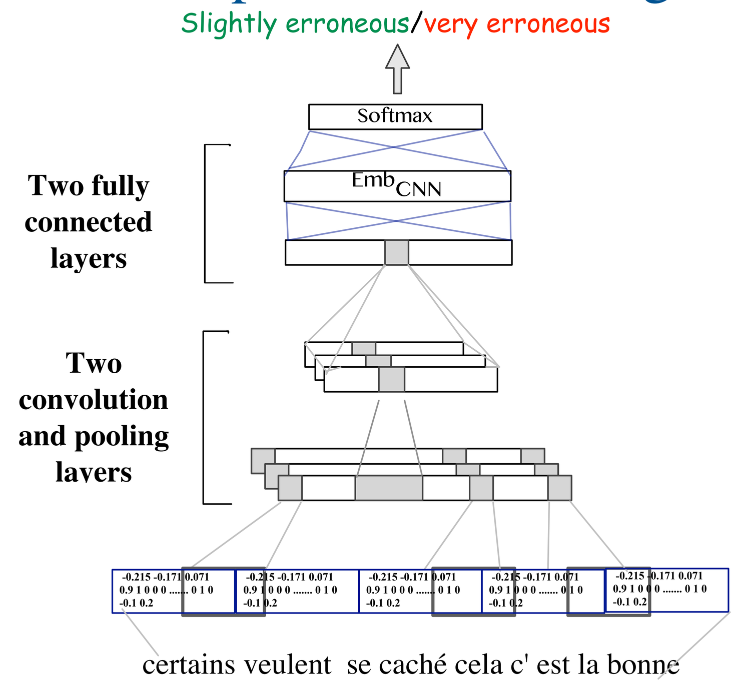
General embeddings



$\mathrm{Emb}_{\mathrm{DBOW}}$

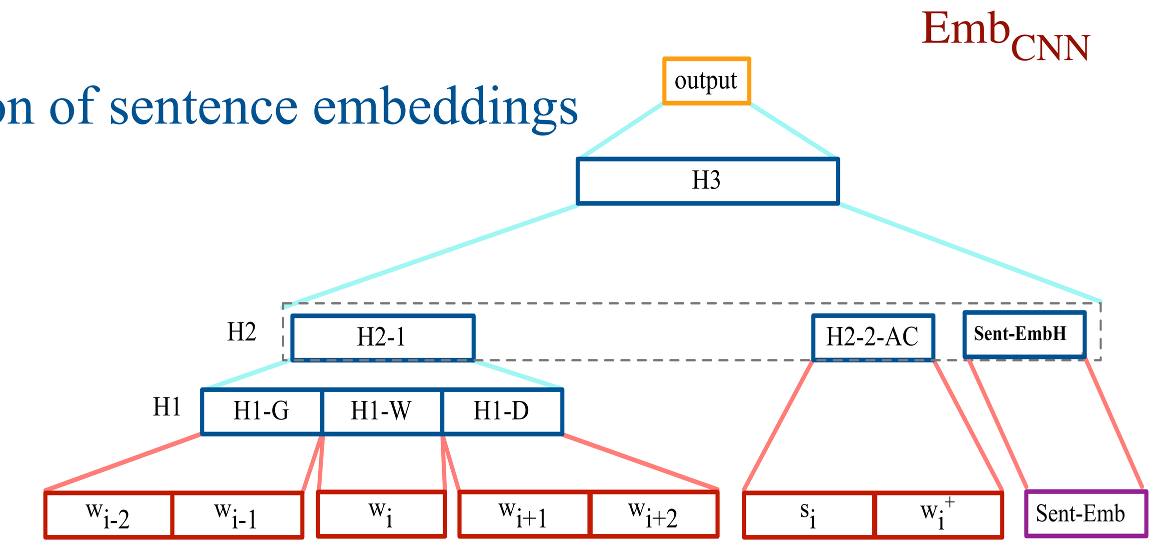
Task specific embeddings

Slightly erroneous/very erroneous



$\mathrm{Emb}_{\mathrm{CNN}}$

Integration of sentence embeddings



Performance of sentence embeddings

Corpus	Sentence Embed.	Label *Error*			Global
		P	R	F	CER
Dev	- (*Sys2*)	0.72	0.60	0.65	9.38
	Emb_{DBOW}	**0.73**	0.58	0.65	9.36
	Emb_{CNN}	0.72	0.60	0.65	**9.26**
Test	- (*Sys2*)	0.70	0.61	0.65	7.75
	Emb_{DBOW}	**0.72**	0.57	0.64	7.72
	Emb_{CNN}	**0.72**	0.58	0.64	**7.69**

+ Integration of sentence embeddings improves the results
+ Task specific sentence embeddings outperforms generic ones

2. Probabilistic contextual model (PCM)

- Smoothing of the classification results at the sentence level
- Re-scoring of a graph of labels by applying an n-order probabilistic model of error distribution
- Find the sequence label S that maximizes:

$$\overline{S} = \arg\max_{e} \prod_{i=1}^{n} c(e_i)^{\lambda} \times P(e_i|e_{i-2}, e_{i-1}, e_{i+1}, e_{i+2})$$

Sys3: Sys2 features + task specific sentence embeddings ($\mathrm{Emb}_{\mathrm{CNN}}$)

Corpus	*System*	Label *error*			Global
		P	R	F	CER
Dev	*Sys2-PCM*	**0.73**	0.56	0.65	**9.31**
	Sys3-PCM	0.73	0.60	0.65	**9.23**
Test	*Sys2-PCM*	0.72	0.59	0.65	**7.67**
	Sys3-PCM	0.73	0.57	0.64	7.69

+ PCM improves Sys2 results: global information is brought
+ Sys3 already contains global information: no real gain with PCM

3. BLSTM architecture

- BLSTM composed of two hidden layers (512 hidden units each)
- Includes the Sys2 features

Corpus	*System*	Label *Error*			
		P	R	F	CER
Dev	BLSTM	0.70	0.63	0.67	9.28
Test	BLSTM	0.69	0.63	0.66	7.83

BLSTM *vs.* Sys2

+ Better on Dev

→ Not generalized on Test, too few training data?

Conclusions

Effective integration of global information about the sentence into ASR error detection system to improve local decision

+ The task specific sentence embeddings $\mathrm{Emb}_{\mathrm{CNN}}$ perform better than generic embeddings $\mathrm{Emb}_{\mathrm{DBOW}}$
+ The probabilistic contextual model improves the results when no global information is included in the features