Which ASR errors are hard to detect?

Sahar Ghannay, Nathalie Camelin, Yannick Estève

LIUM, University of Le Mans France

ERRARE 2015, Errors produced and processed by humans and machines in multimedia, multimodal and multilingual data, Workshop

Introduction

MGB 2015 challenge results for ASR task on BBC data

	Best Sys	CRIM/ LIUM	Sys1	Sys2	Sys3	LIUM	Sys4	Sys5	Sys6	Sys7	Sys8	Sys9
Overall WER(%)	**23.7**	26.6	27.5	27.8	28.8	30.4	30.9	31.2	35.5	38.0	38.7	40.8

Introduction

MGB 2015 challenge result

Detailed performance of the best system

Show	CU
Daily Politics	10.4
Magnetic North	11.6
Dragons'Den	11.5
Eggheads	14.1
Athletics London	14.7
Point of View	13.5
Syd Barrett	21.3
Top Gear	21.8
Blue Peter	24.6
Legend of the Dragon	21.7
The North West 200	27.7
Holby City	32.1
The Wall	33.7
One Life Special Mum	35.3
Goodness Gracious ME	37.2
Oliver Twist	**41.4**
Overall WER(%)	***23.7***

Introduction

ASR errors have impact on applications:

- ✤ Information retrieval
- ✤ Speech to speech translation
- ✤ Spoken language understanding
- ✤ etc.

Introduction

ASR errors have impact on applications:

- Information retrieval
- Speech to speech translation
- Spoken language understanding
- etc.

➡ ASR error detection can help

Introduction

✓ Related work

- Approaches based on Conditional Random Field (CRF)
  - OOV detection [C. Parada *et al*. 2010]
    - contextual informations
  - Errors detection [F. Béchet & B. Favre 2013]
    - ASR based, lexical and syntactic informations
- Approach based on neural network
  - Errors detection [T. Yik-Cheung *et al*. 2014]
    - complementary ASR systems

Introduction

✓ Related work

- Approaches based on Conditional Random Field (CRF)
  - OOV detection [C. Parada *et al.* 2010]
    - contextual informations
  - Errors detection [F. Béchet & B. Favre 2013]
    - ASR based, lexical and syntactic informations
- Approach based on neural network
  - Errors detection [T. Yik-Cheung *et al.* 2014]
    - complementary ASR systems

✓ Contributions

- Neural approach
  - Effective word embeddings combination
  - New neural architecture
- Analysis of ASR error detection system outputs

Set of features

The features are inspired by [F. Béchet and B. Favre 2013]

- ✤ Posterior probabilities
- ✤ Lexical features
  - word length
  - existence 3-gram
- ✤ Syntactic features
  - POS tag
  - dependency labels
  - word governors
- ✤ Word

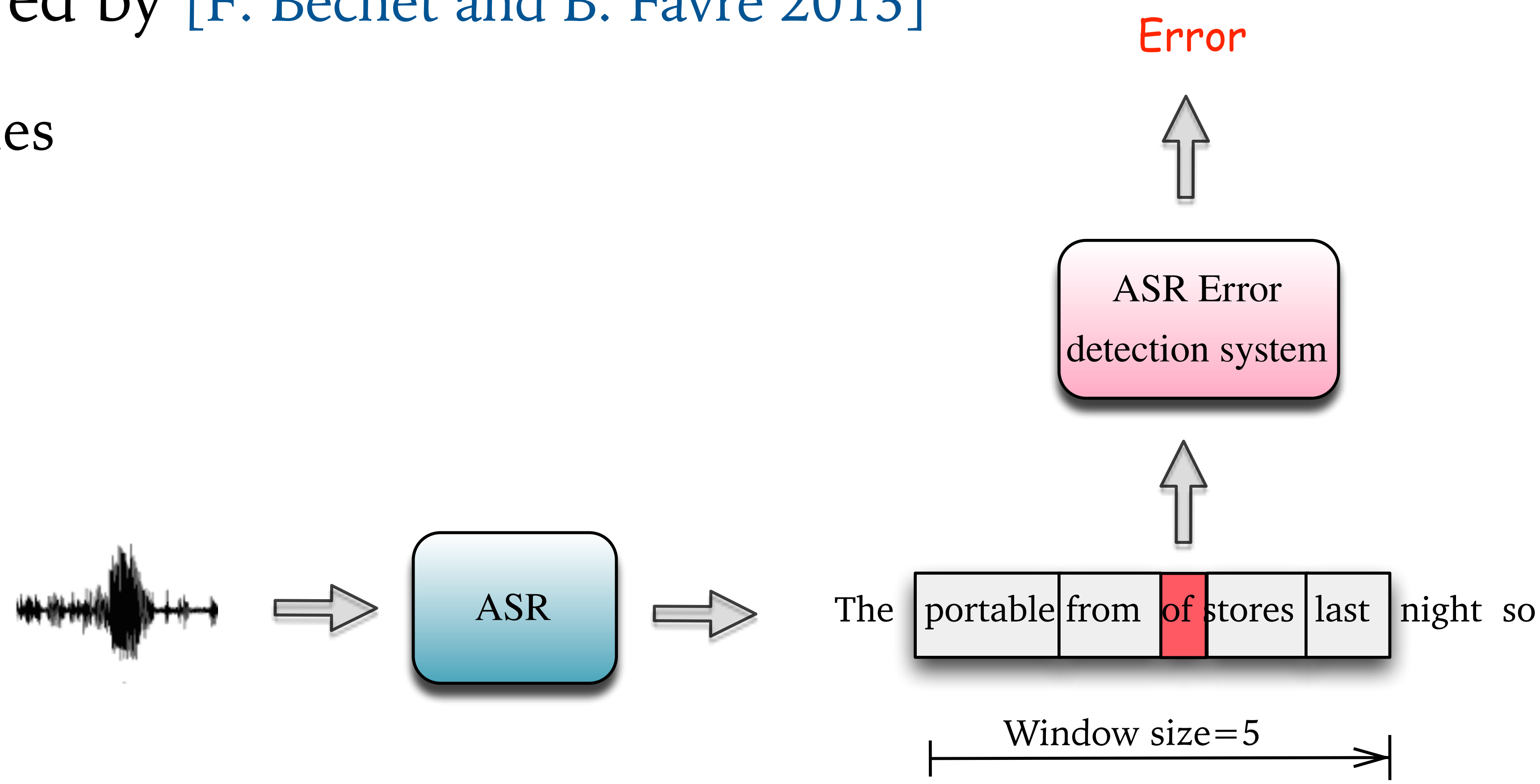


Figure 1: ASR error detection system

Set of features

The features are inspired by [F. Béchet and B. Favre 2013]

- ✤ Posterior probabilities
- ✤ Lexical features
  - word length
  - existence 3-gram
- ✤ Syntactic features
  - POS tag
  - dependency labels
  - word governors
- ✤ Word

Word embeddings

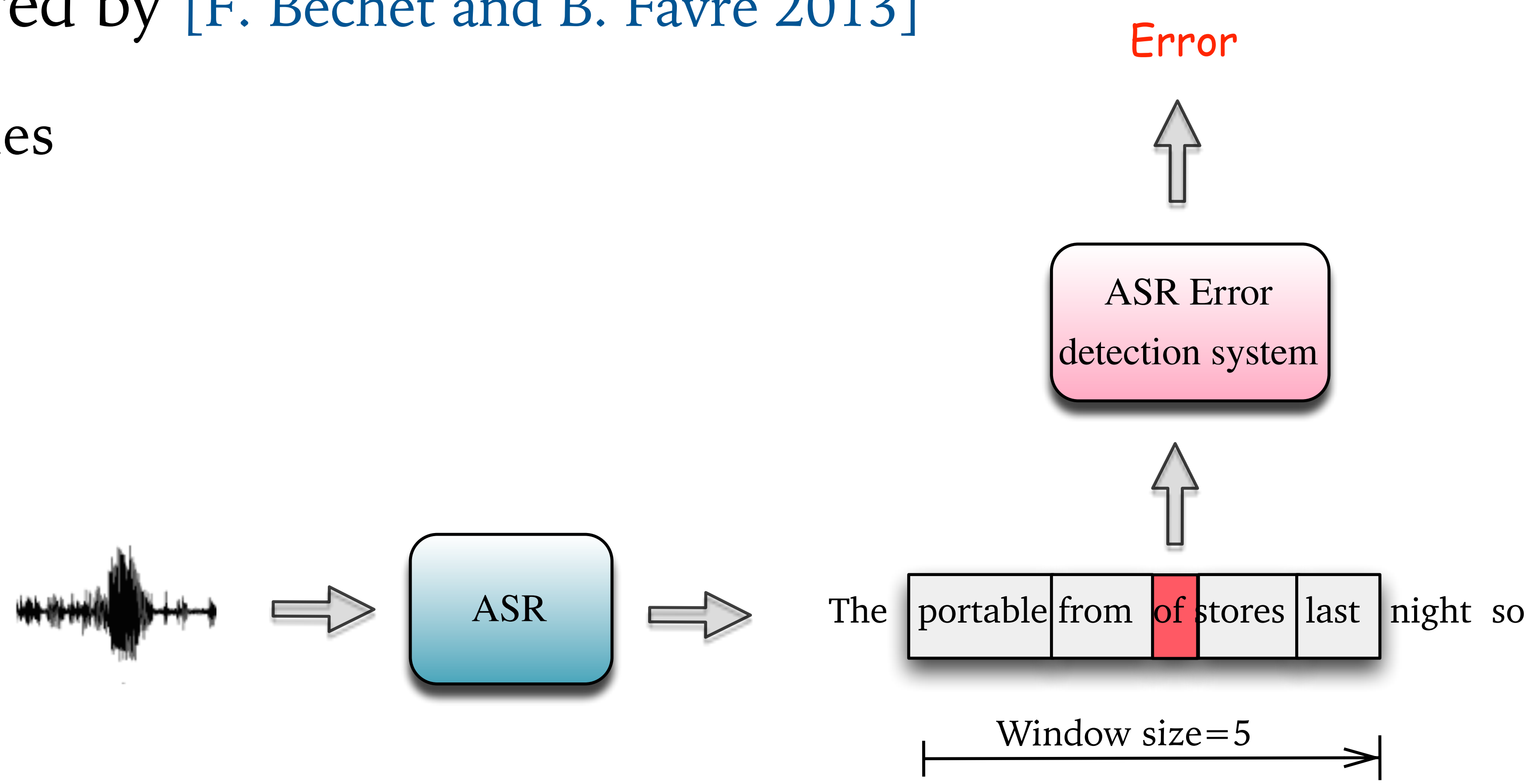


Figure 1: ASR error detection system

Word embeddings

Mapping words to high-dimensional vectors (e.g. 200 dimensions)

$$R : Words = \{W_1, ..., W_n\} \rightarrow Vectors = \{R(W_1), ..., R(W_n)\} \subset R^d$$

Distance between vectors indicates the relation between words

$$R(W_1) \approx R(W_n) \rightarrow W_1 \approx W_n$$

Word embeddings

Mapping words to high-dimensional vectors (e.g. 200 dimensions)

$$R : Words = \{W_1, ..., W_n\} \rightarrow Vectors = \{R(W_1), ..., R(W_n)\} \subset R^d$$

Distance between vectors indicates the relation between words

$$R(W_1) \approx R(W_n) \rightarrow W_1 \approx W_n$$

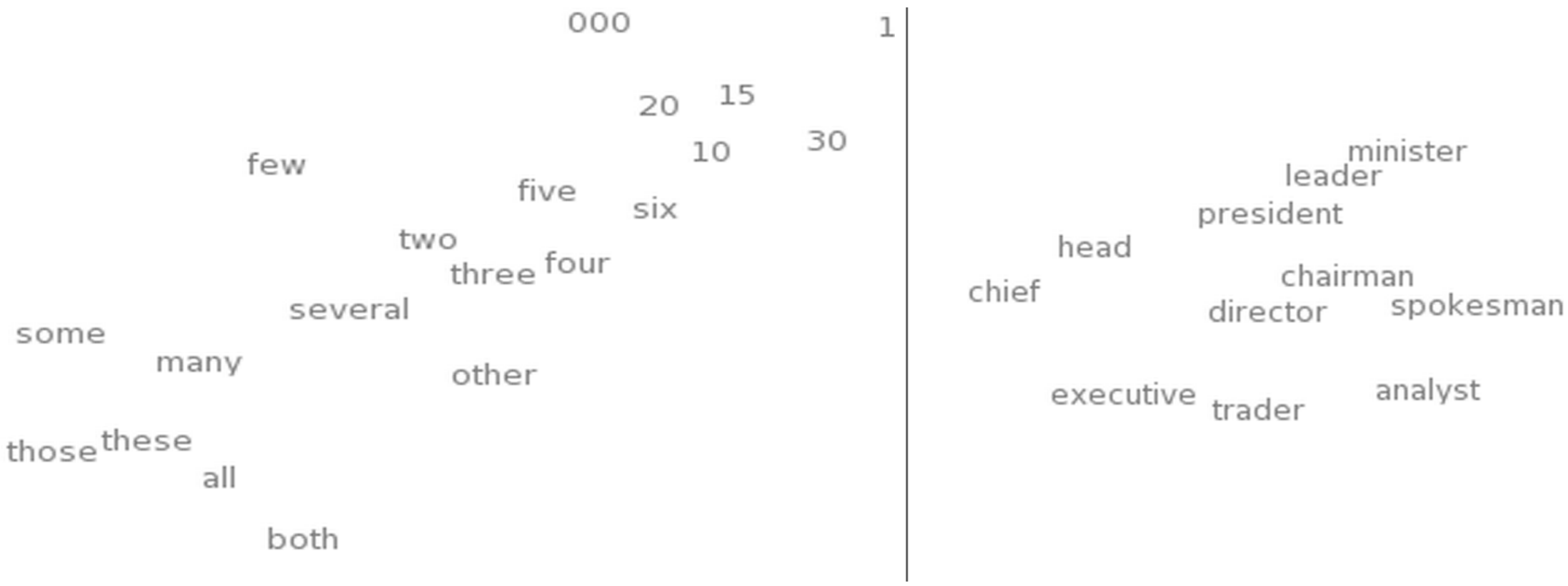


Figure 2: 2D t-SNE visualizations of word embeddings.
Left: Number Region; Right: Jobs Region [J.Turian *et al* . 2010]

Word embeddings

Mapping words to high-dimensional vectors (e.g. 200 dimensions)

$$R : Words = \{W_1, ..., W_n\} \rightarrow Vectors = \{R(W_1), ..., R(W_n)\} \subset R^d$$

Distance between vectors indicates the relation between words

$$R(W_1) \approx R(W_n) \rightarrow W_1 \approx W_n$$

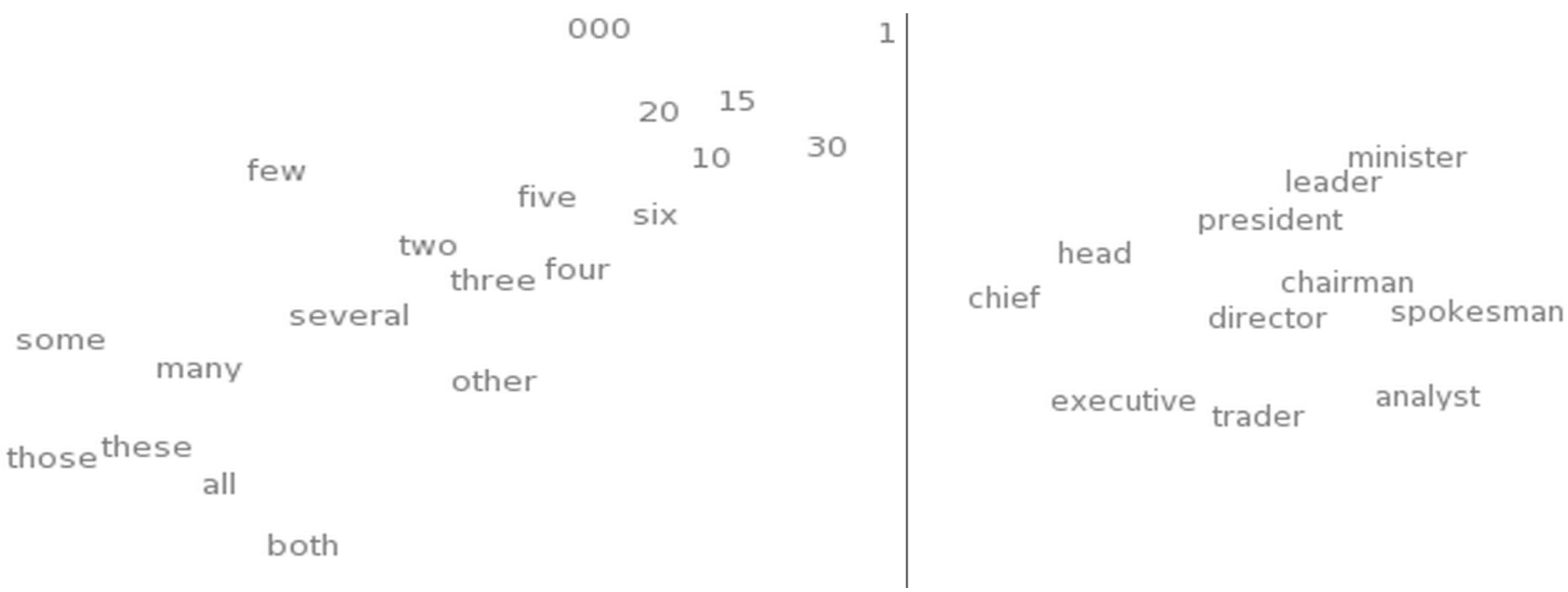


Figure 2: 2D t-SNE visualizations of word embeddings.
Left: Number Region; Right: Jobs Region [J.Turian *et al* . 2010]

FRANCE	JESUS	XBOX
AUSTRIA	GOD	AMIGA
BELGIUM	SATI	PLAYSTATION
GERMANY	CHRIST	MSX
ITALY	SATAN	IPOD
GREECE	KALI	SEGA
SWEDEN	INDRA	PSNUMBER
NORWAY	VISHNU	HD
EUROPE	ANANDA	DREAMCAST
HUNGARY	PARVATI	GEFORCE
SWITZERLAND	GRACE	CAPCOM

Figure 3: What words have embeddings closest to a given word? [R.Collobert *et al* . 2011]

Word embeddings approaches(1/3)

1. Tur: Collobert and Weston embeddings revised by Joseph Turian [J.Turian *et al.* 2010]

- ✤ Existence n-gram
- ✤ Training criterion: score (n-gram) > score (corrupted n-gram) + some margin

➡ Morpho-syntactic similarities

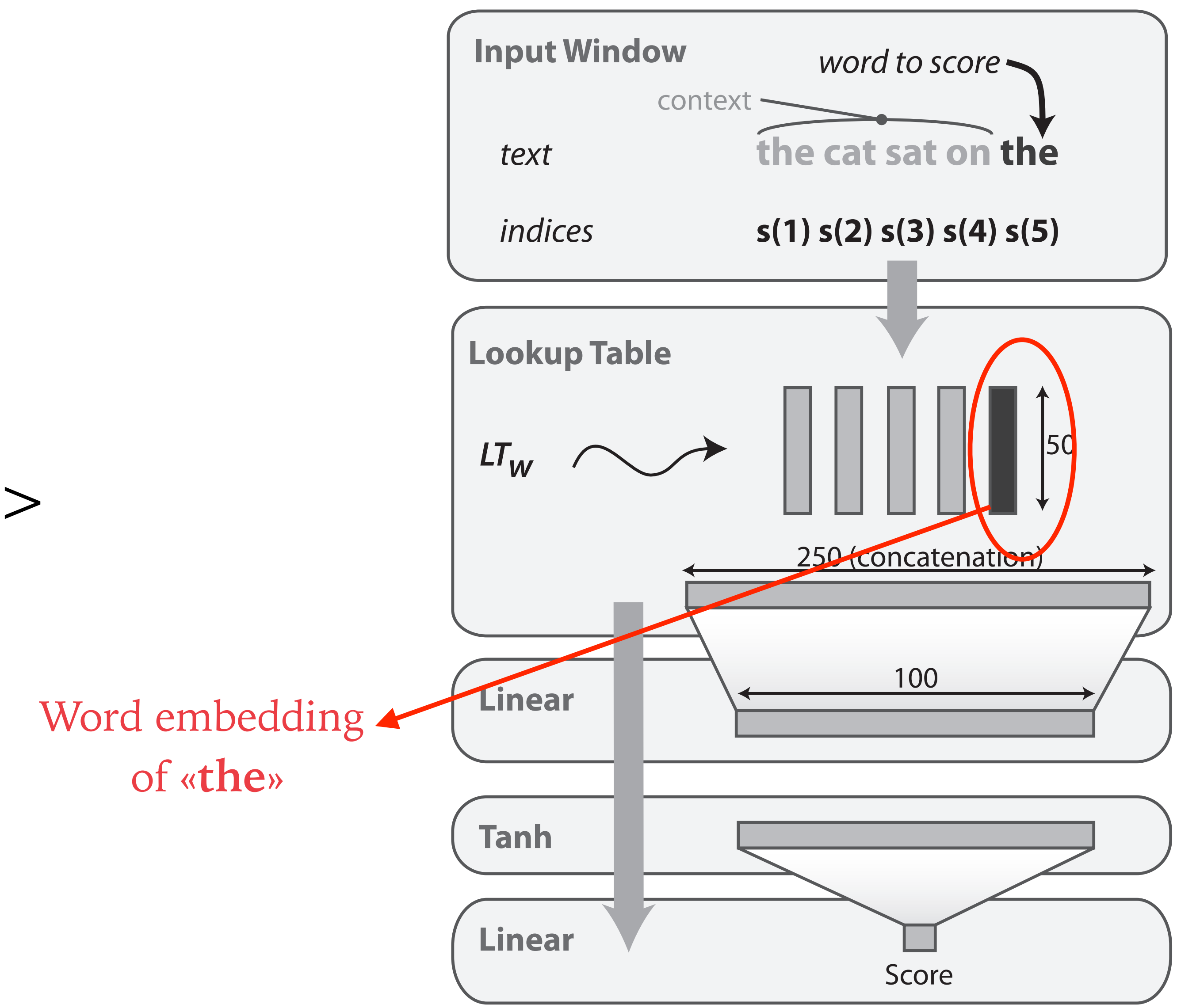


Figure 4: Neural architecture to compute 50 dimensional word embeddings

Word embeddings approaches(2/3)

2. Word2vec [T.Micolov *et al.* 2013]

- ✤ Continuous bag of words (CBOW)
  - ✦ predicting the current word based on its context

➡ Syntactic modeling

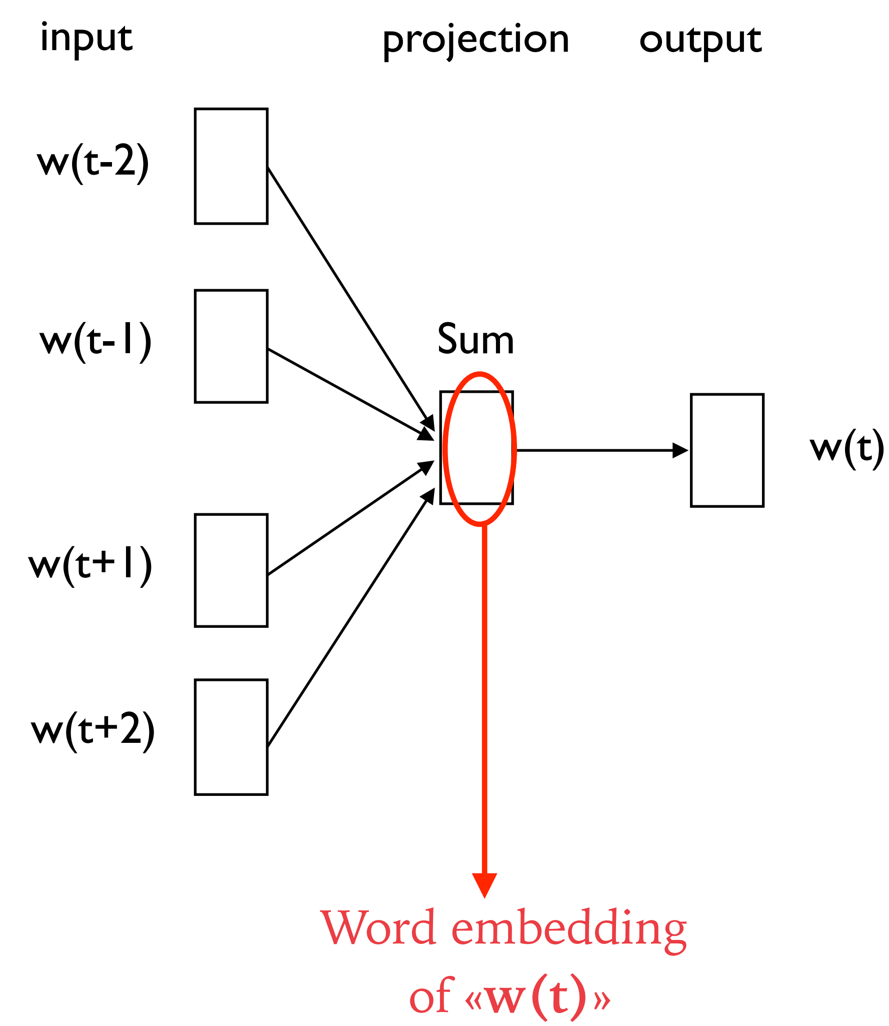


Figure 5: CBOW architecture

Word embeddings approaches(3/3)

3. Glove: global vector for word representation [J.Pennington *et al.* 2014]

- ✤ Analysis of co-occurrences of words in a window
  - ✦ building a co-occurrence matrix
  - ✦ estimating continuous representations of the words

➡ Semantic similarities

Word embeddings combination

Combine word embeddings using denoising auto-encoder

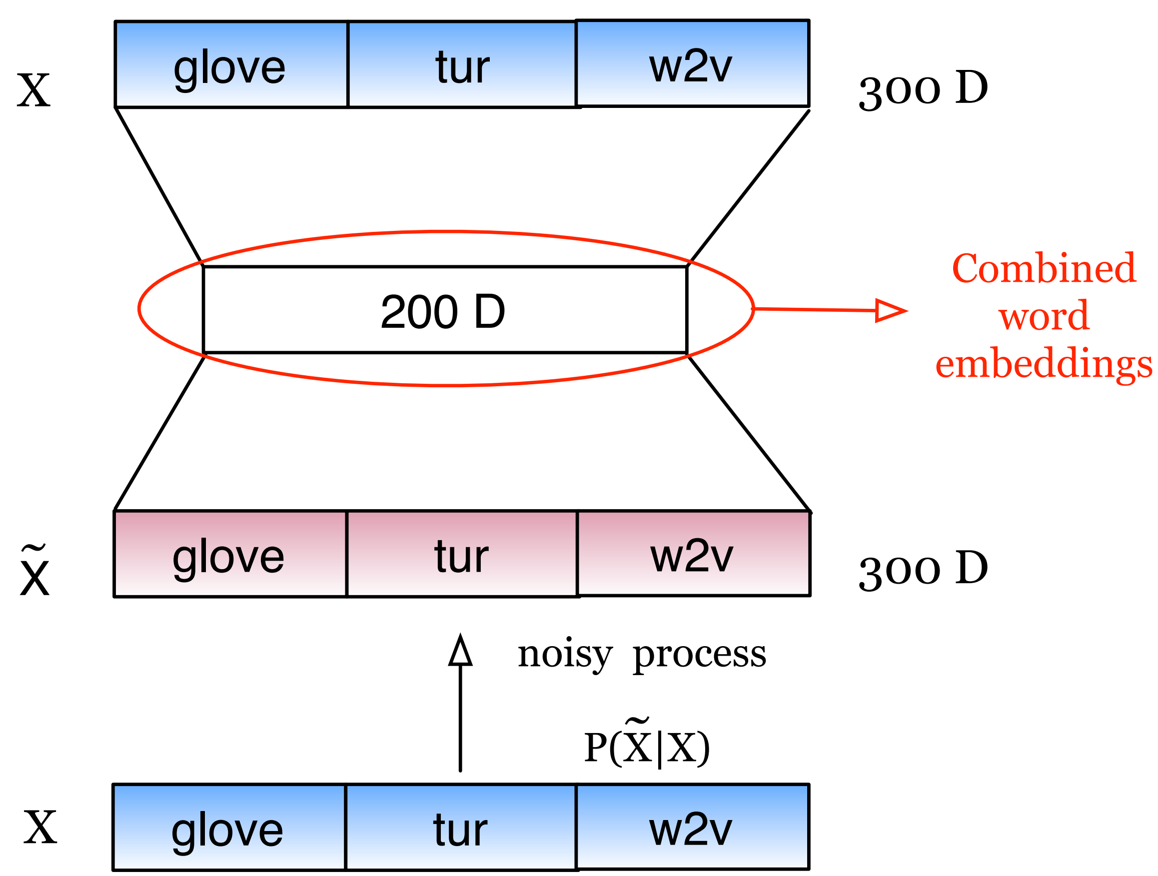


Figure 6: Using denoising auto-encoder to combine word embeddings

Neural architecture: MLP-Multi-Stream

[S. Ghannay *et al.* 2015]

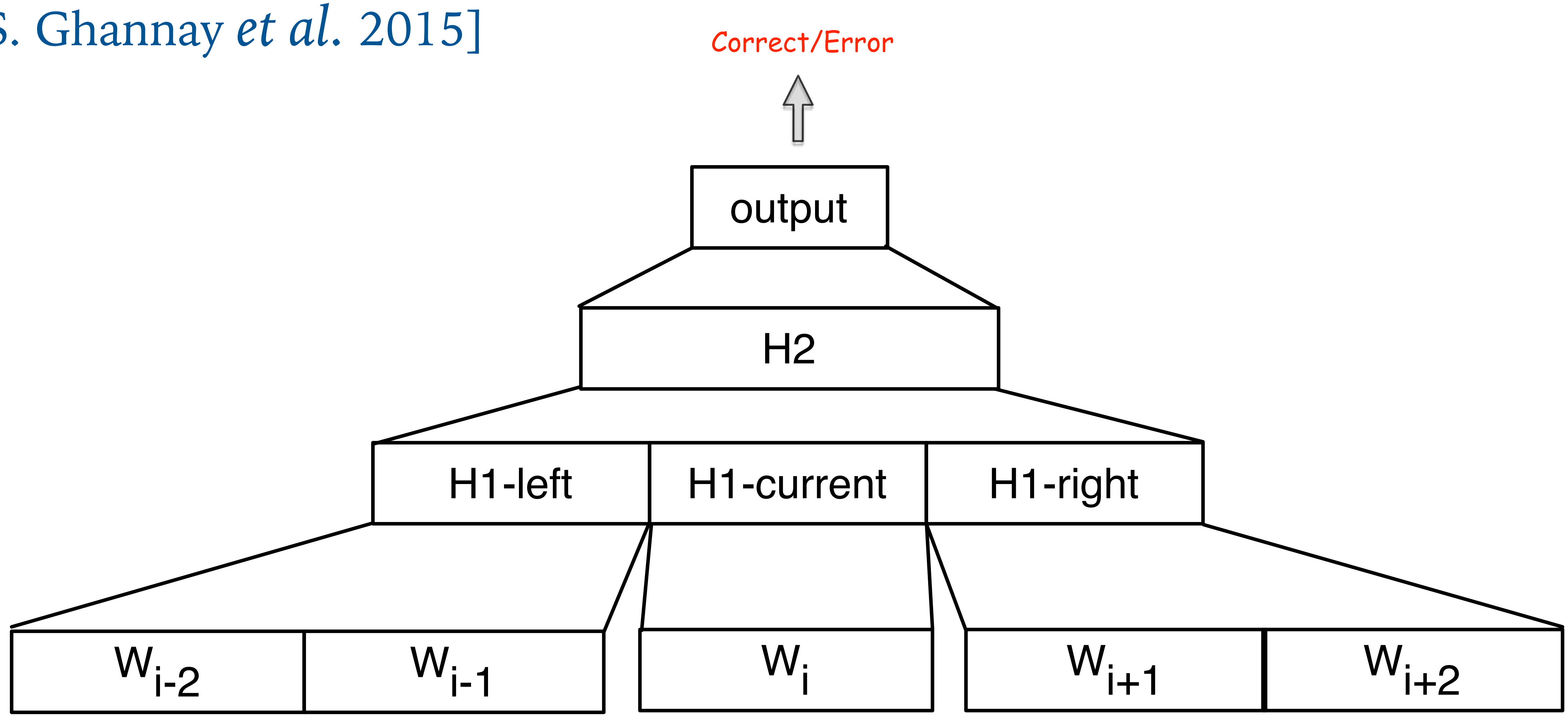


Figure 7: MLP-MS architecture for ASR error detection task

Experimental data

Training of the neural system:

Automatic transcriptions of the ETAPE Corpus, generated by:

- ASR 1: CMU Sphinx decoder
  - acoustic models: GMM/HMM

ASR	Name	#words REF	#words HYP	WER
Sphinx GMM	Train	349K	316K	25.9
	Dev	54K	50K	25.2
	Test	58K	53K	22.5

Table 1: Composition of the experimental corpus

Training data of the word embeddings:

Corpus composed of 2 billions of words:

- Articles of the French newspaper ”Le Monde”,
- French Gigaword corpus,
- Articles provided by Google News,
- Manual transcriptions: 400 hours of French broadcast news.

Evaluation results

- Neural architecture vs. CRF
- Evaluation metrics:
  - Error label: Recall, Precision and F-measure
  - Overall classification: CER

Experimental results

Comparison of different word representations

			Label error	Global
Approach	Corpus	representation	F-measure	CER
Neural (MLP-MS)	Dev	glove	58.83	10.66
		w2v	61.81	10.54
		tur	59.11	10.56
		Auto-encoder-200	**62.47**	**9.99**

Table 2: Comparison on Dev of different types of word embeddings used as additional features in MLP-MS error detection system.

Experimental results

2.Performance of MLP-MS on Test corpus

	Label error	Global
Approach	F-measure	CER
CRF(baseline)	57.6	8.78
MLP-MS	**61.4**	**8.43**

Table 3: Error detection results on Test corpus

Analysis of the ASR error detection system outputs

- **Ground truth:** alignement of the reference with the automatic transcriptions
- **Predictions:** classifier outputs
- **Correct predictions:** label predictions = label ground truth

Ground truth	C	C	C	E	E	E	E
Predictions	E	C	E	E	E	C	C
Correct predictions	E	C	E	E	E	C	C

- **Span:** contiguous errors segment correctly detected.

Ground truth	C	C	C	E	E	E	E
Predictions	E	C	E	E	E	C	C
Correct predictions	E	C	E	E	E	C	C

Analysis of the ASR error detection system outputs

1.Word length analysis

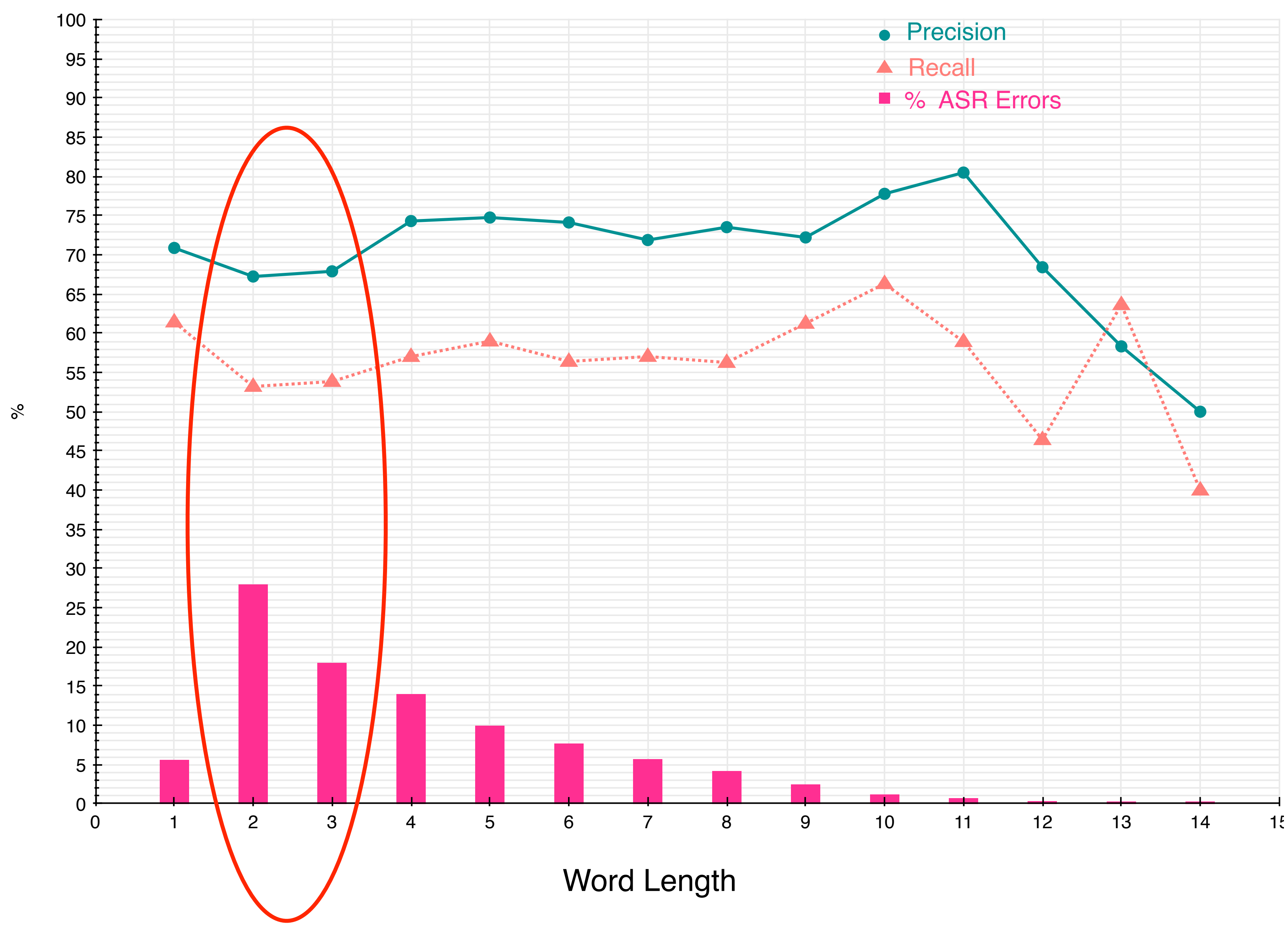


Figure 9: Recall and precision for the erroneous word prediction and the percentage of erroneous words by word length on Dev corpus

Analysis of the ASR error detection system outputs

2.Function and non function words analysis

- Function words
  - stop list of 160 words
  - average length: 2.8 letters
- Non function words
  - average length: 6.3 letters

	Label error	
Words	Precision	Recall
Non function	**75.1**	**61.0**
Function	66.9	51.7

Table 5: Function and non function words analysis on Dev corpus

Analysis of the ASR error detection system outputs

2.Function and non function words analysis

- Function words
  - stop list of 160 words
  - average length: 2.8 letters
- Non function words
  - average length: 6.3 letters

	Label error	
Words	Precision	Recall
Non function	**75.1**	**61.0**
Function	66.9	51.7

Table 5: Function and non function words analysis on Dev corpus

➡ 75.65% of erroneous function words are of length 2 or 3

Analysis of the ASR error detection system outputs

3.Average error segment size (average span) analysis

	Corpus	Average span	Standard deviation
Ground truth	Train Dev	3.03 **3.24**	1.72 2.15
Predictions	Dev	2.92	2.82
Correct predictions	Dev	2.67	1.17
CRF	Dev	**3.29**	1.81

Table 6: The average span and the standard deviation for the ground truth, the predictions, the correct predictions and the CRF outputs.

Analysis of the ASR error detection system outputs

3.Average error segment size (average span) analysis

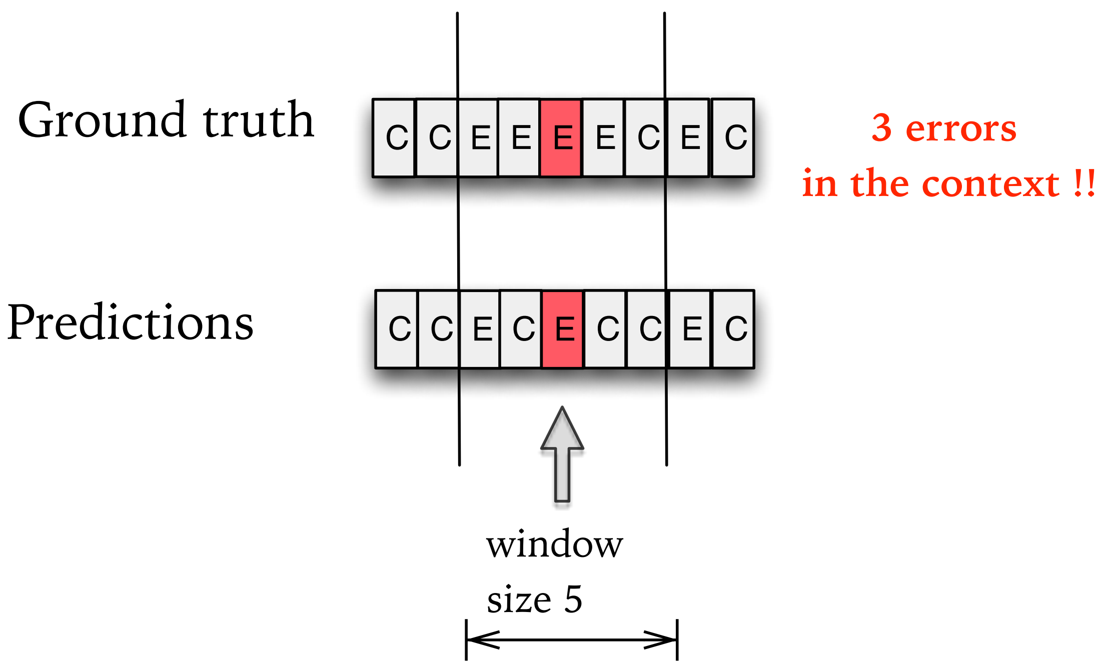
	Corpus	Average span	Standard deviation
Ground truth	Train Dev	3.03 **3.24**	1.72 2.15
Predictions	Dev	2.92	2.82
Correct predictions	Dev	2.67	1.17
CRF	Dev	**3.29**	1.81

Table 6: The average span and the standard deviation for the ground truth, the predictions, the correct predictions and the CRF outputs.

➡ MLP-MS takes local decisions

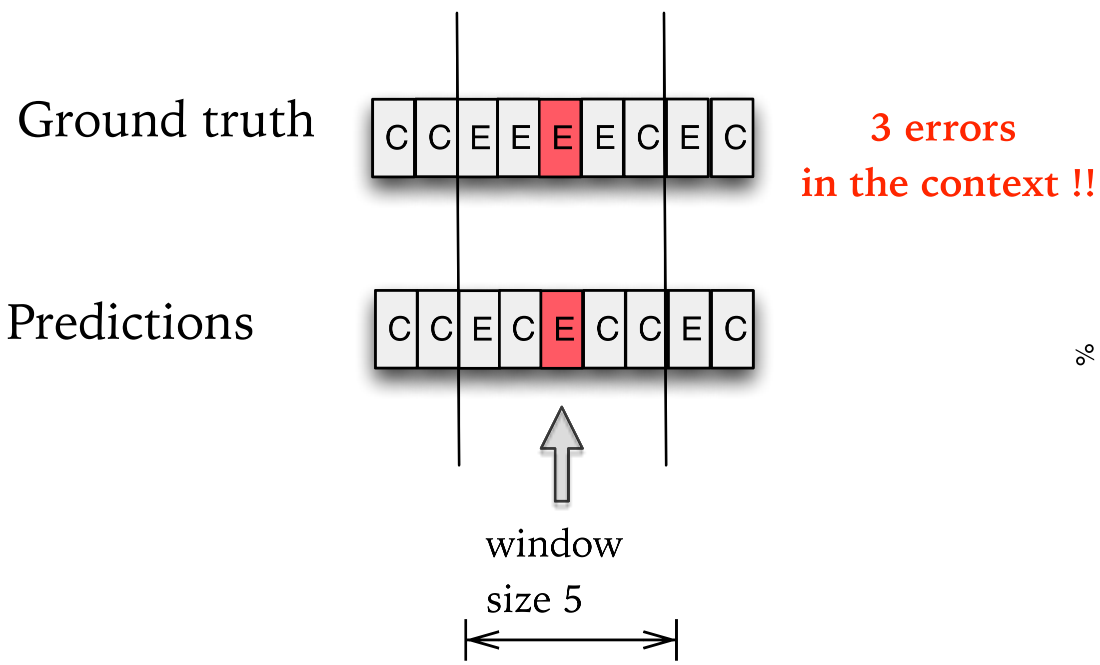
Analysis of the ASR error detection system outputs

4.Current word context analysis



Analysis of the ASR error detection system outputs

4.Current word context analysis



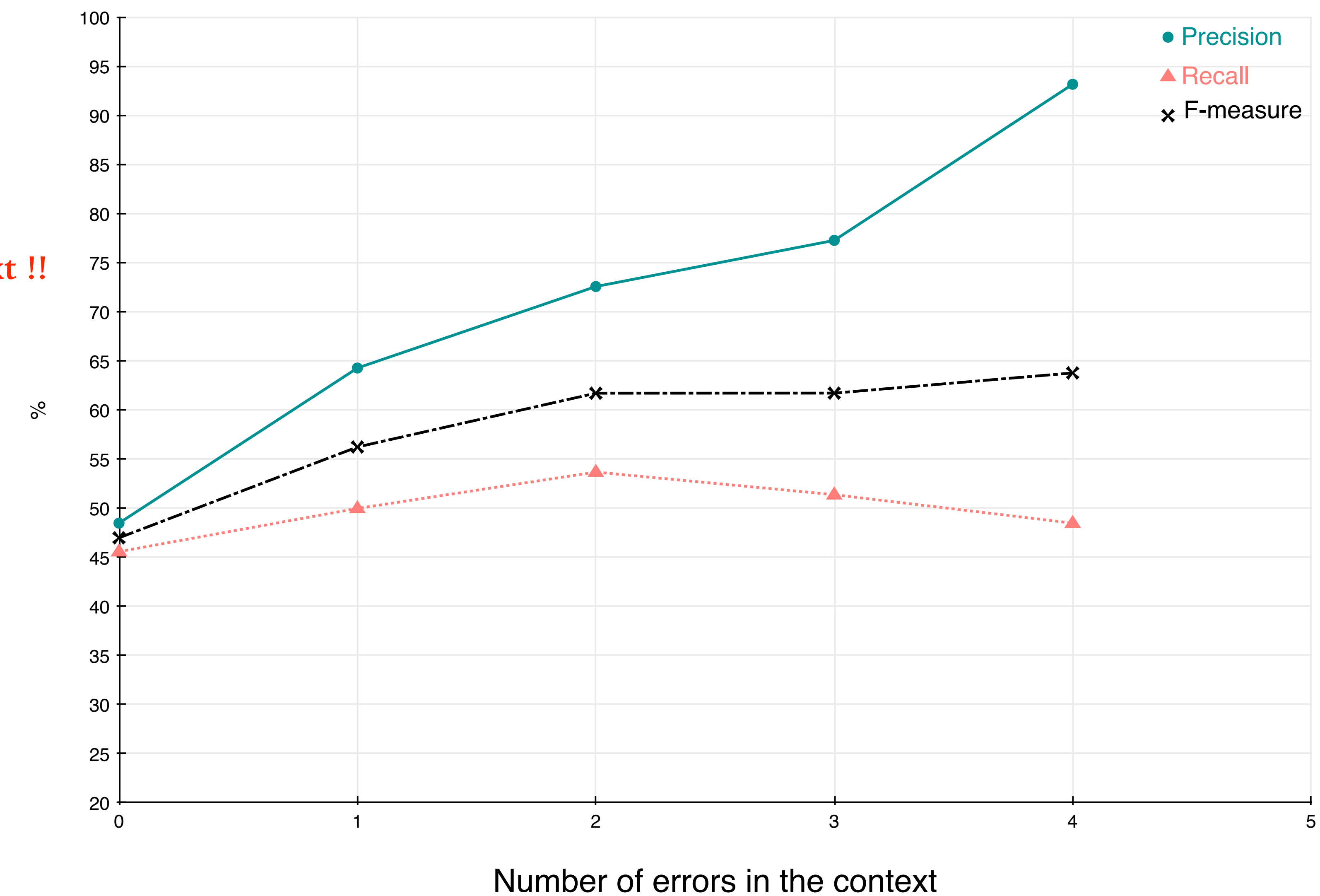
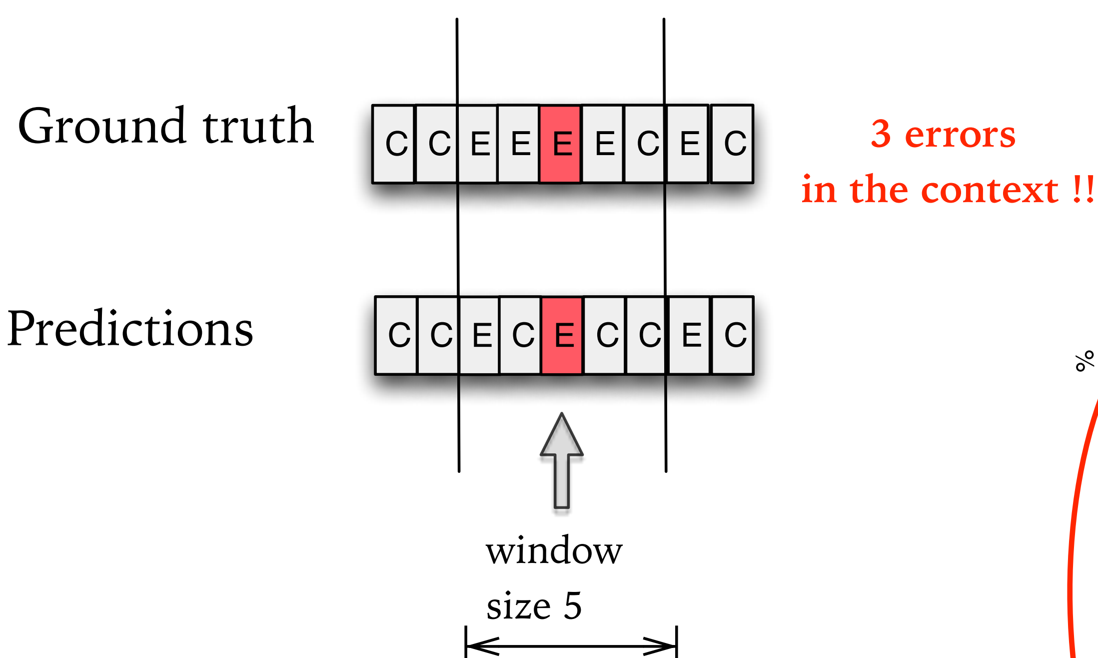
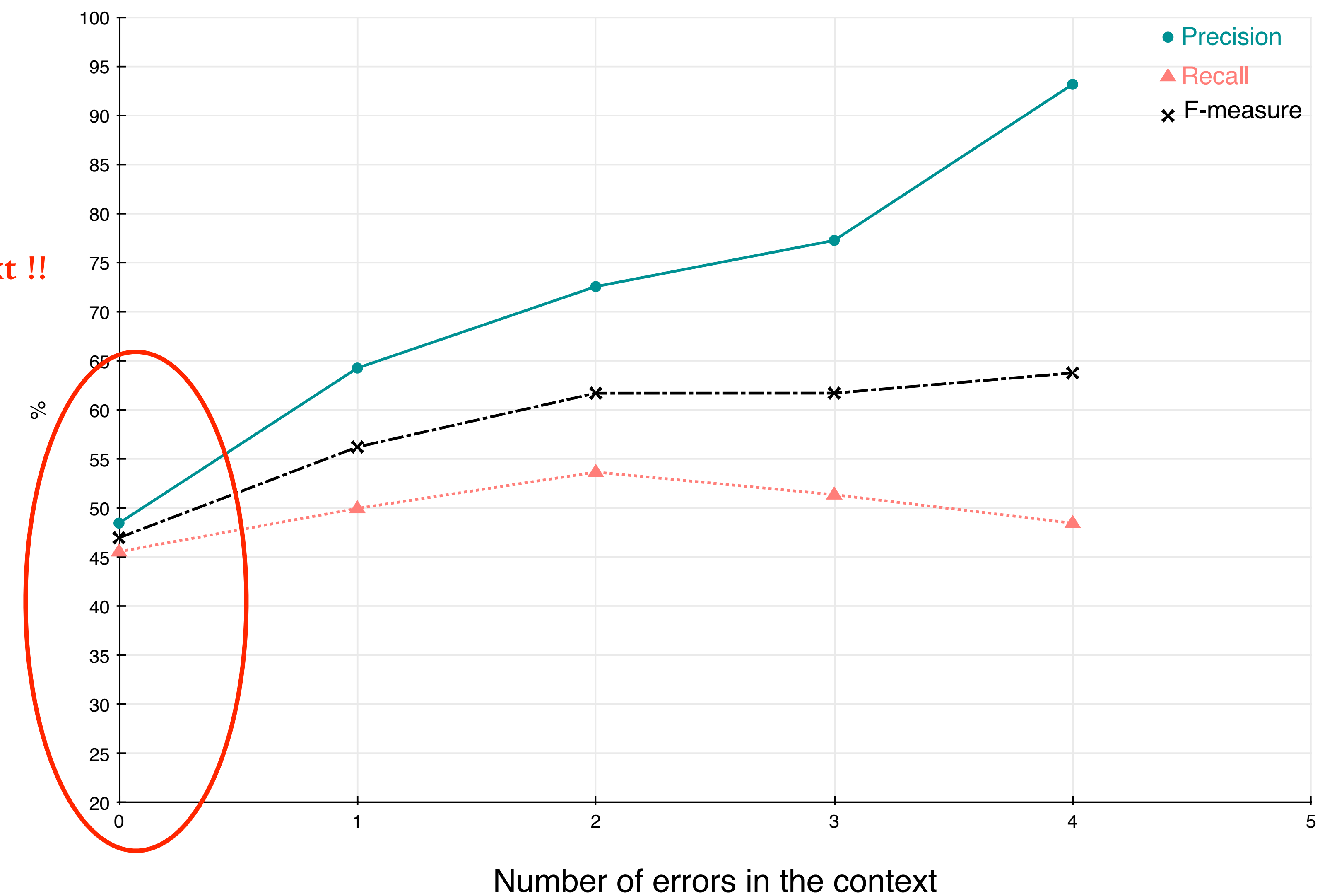


Figure 10: Precision, recall and F-measures for the erroneous word prediction relative to the number of errors in its context.

Analysis of the ASR error detection system outputs

4.Current word context analysis





- ➡ Difficulty to detect isolated errors
- ➡ Isolated errors don't trigger a significant linguistic rupture

Figure 10: Precision, recall and F-measures for the erroneous word prediction relative to the number of errors in its context.

Analysis of the ASR error detection system outputs

5.Syntactic role analysis

- EQ: $POS_{Hyp} = POS_{Ref}$
- DIFF: POS_{Hyp} != POS_{Ref}

	Label error	
POS	Precision	Recall
EQ	29.01	51.51
DIFF	**95.57**	56.82

Table 7: Error analysis results on Dev corpus according to the part of speech tag of the automatic transcriptions and reference transcriptions

Analysis of the ASR error detection system outputs

5.Syntactic role analysis

- EQ: $POS_{Hyp} = POS_{Ref}$
- DIFF: POS_{Hyp} != POS_{Ref}

	Label error	
POS	Precision	Recall
EQ	29.01	51.51
DIFF	**95.57**	56.82

➡ Weak linguistic disruption makes ASR errors hard to detect

Table 7: Error analysis results on Dev corpus according to the part of speech tag of the automatic transcriptions and reference transcriptions

Conclusions

ASR error detection system

- Combined word embeddings
- MLP-MS architecture

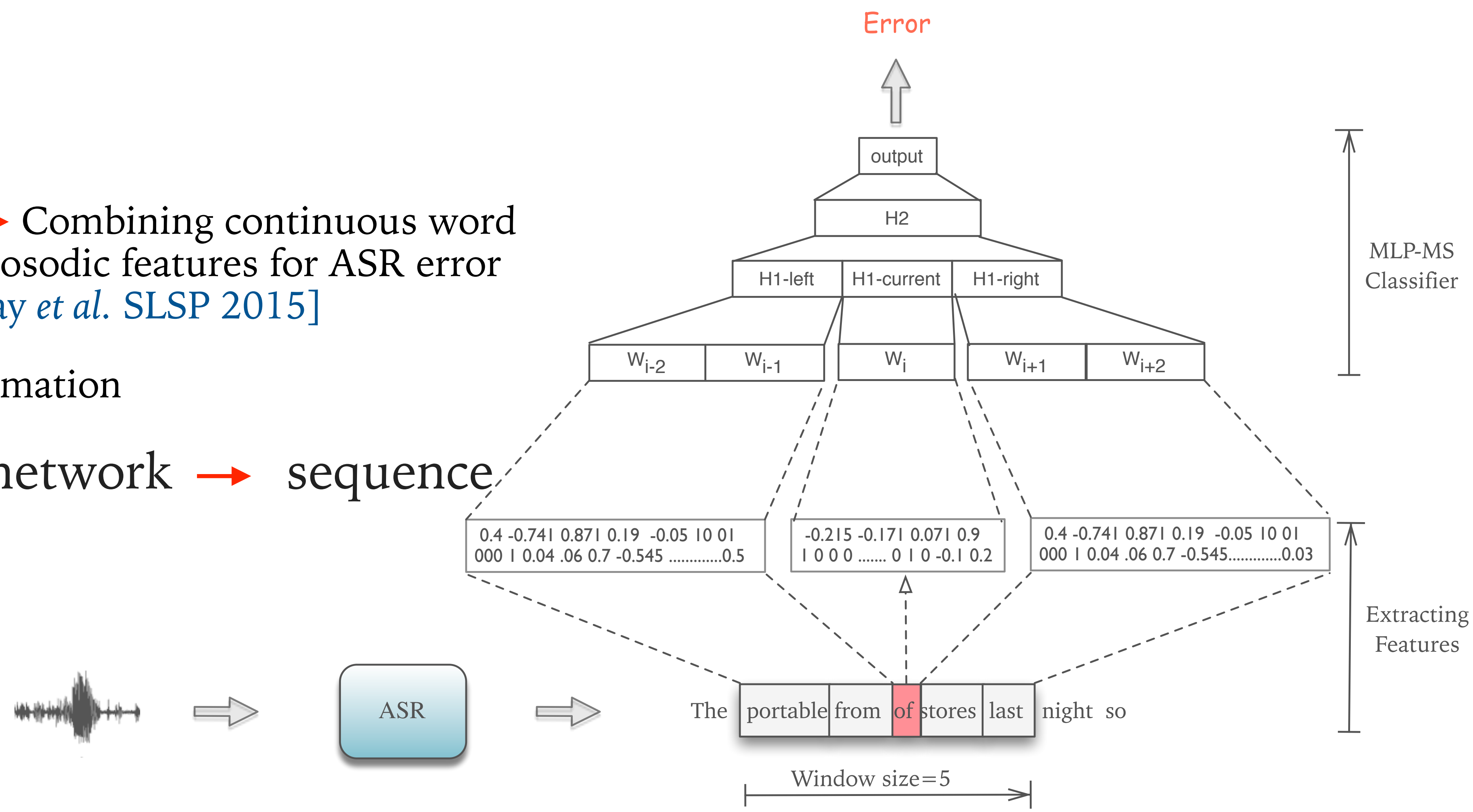
ASR errors hard to detect:

- words of length 2 and 3 (letters)
- function words
- isolated errors
- errors in a context slightly linguistically disrupted

Conclusions

Perspectives:

- New features:
  - Prosodic features → Combining continuous word representation and prosodic features for ASR error prediction [S.Ghannay *et al.* SLSP 2015]
  - Global semantic information
- Recurrent neural network → sequence prediction



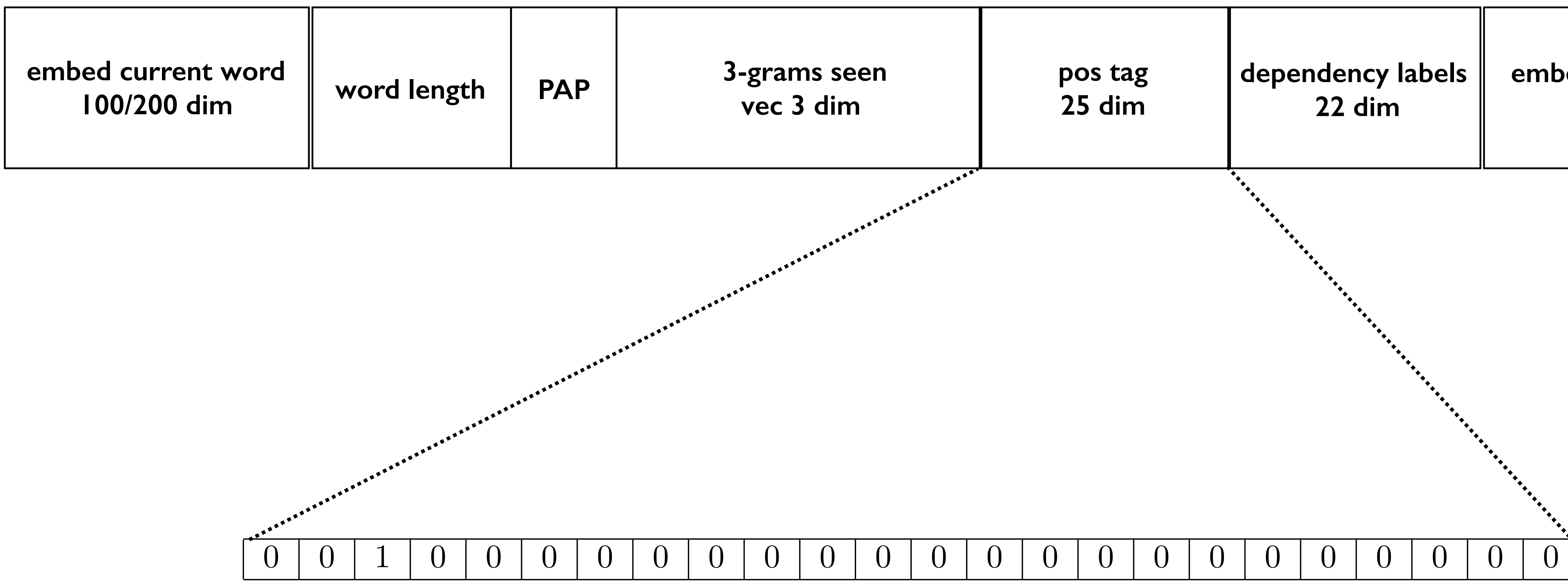
Thank you

Neural network input feature vector format

embed current word 100/200 dim	word length	PAP	3-grams seen vec 3 dim	pos tag 25 dim	dependency labels 22 dim	embed word governor 100/200 dim

Figure 2 : Neural network input feature vector format (152/252 D)

Neural network input feature vector format



Example: 25 POS tags, 3rd POS tag

Figure 2 : Neural network input feature vector format (152/252 D)