Enriching confusion networks for post-processing

Sahar Ghannay, Yannick Estève, Nathalie Camelin

LIUM, IICC, Le Mans University

SLSP 2017, Le Mans, France

23/10/2017

Introduction

- Automatic speech recognition (ASR) errors are still unavoidable
- Impact of ASR errors
  - Information retrieval,
  - Speech to speech translation,
  - Spoken language understanding,
  - Subtitling
  - *Etc.*

INTRODUCTION

- Detection and correction of ASR errors
  - Improve recognition accuracy: using post processing of ASR outputs [S. Stoyanchev *et. al* 2012, E. Pincus *et. al* 2014]
  - Decrease word error rate using of confusion networks (CN) [L. Mangu *et. al* 2000]
  - Correct erroneous words in CNs [Y. Fusayasu *et. al* 2015]
  - Improve post-processing of ASR outputs using CNs
    - Propose alternative word hypotheses when ASR outputs are corrected by a human on post-edition

- CN bins don't have a fixed length and sometimes contain one or two words

- Number of alternatives to correct a misrecognized word is very low

CONTRIBUTIONS

- Approach of CN enrichment
  - Assumption: words in the same bin should be close in terms of acoustics and /or linguistics
  - New similarity measure computed from acoustic and linguistic word embeddings
- Evaluation
  - Predict potential ASR errors for rare words
  - Enrich CN to improve post-edition of automatic transcriptions
  - Propose semantically relevant alternative words to ASR outputs for Spoken Langage Understanding (SLU) system

Word embeddings

Acoustic embeddings

- f: speech segments → $\mathbb{R}^n$ is a function for mapping speech segments to low-dimensional vectors.

  → words that sound similar = neighbors in the continuous space

- Successfully used in:
  - Query-by-example search system [levin *et al*, 2013, kamper *et al*, 2015]
  - ASR lattice re-scoring system [S. Bengio and Heiglod 2014]
  - ASR Error detection [S. Ghannay *et al*, 2016]

WORD EMBEDDINGS

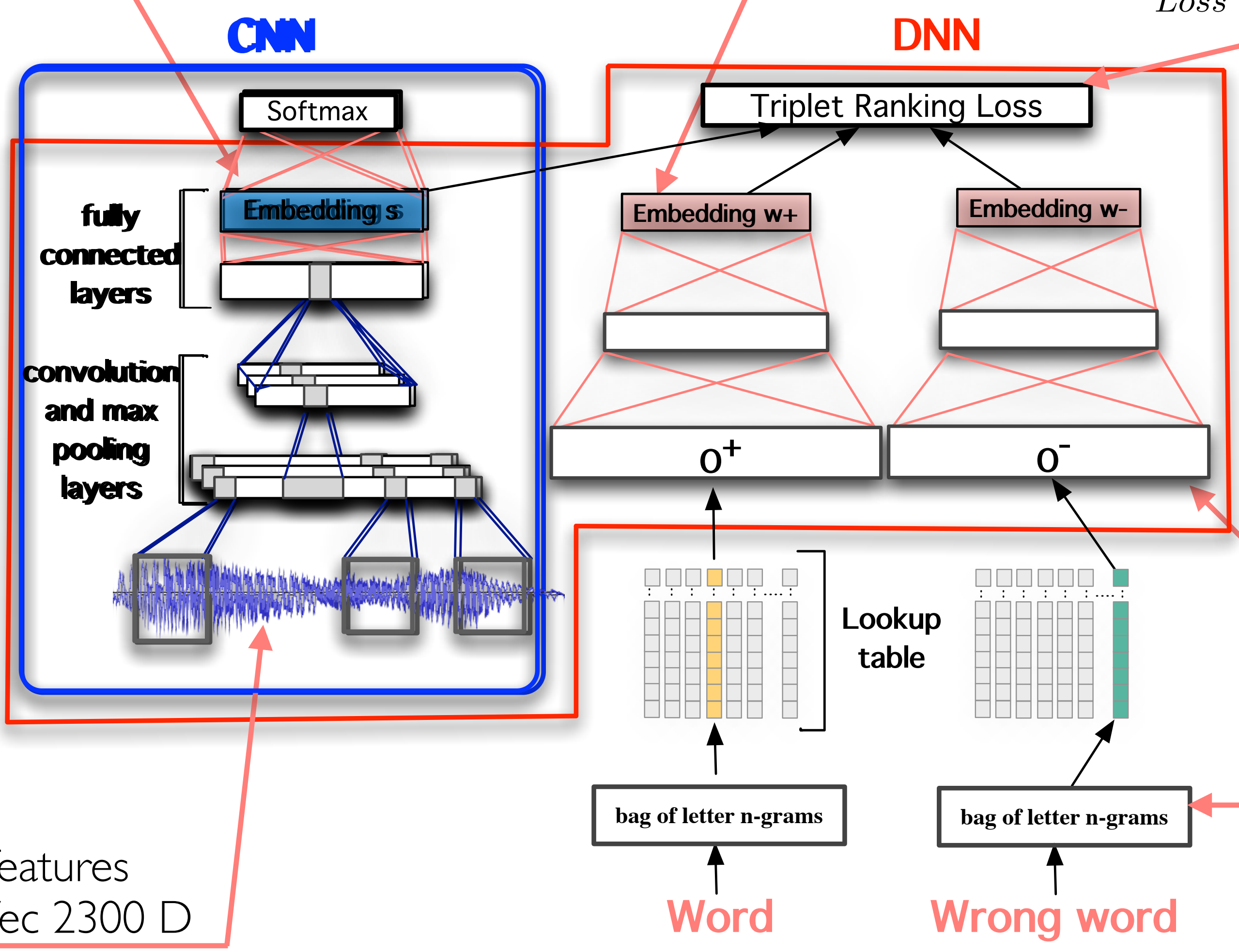
ACOUSTIC EMBEDDINGS-ARCHITECTURE

Approach inspired by [Bengio and Heiglod 2014]

acoustic signal embedding (**s**)

acoustic word embedding (**a**)

$Loss = \max(0, m - Sim_{dot}(s, w^{+}) + Sim_{dot}(s, w^{-}))$



Orthographic embedding (**o**)

filter bank features
1 word = Vec 2300 D

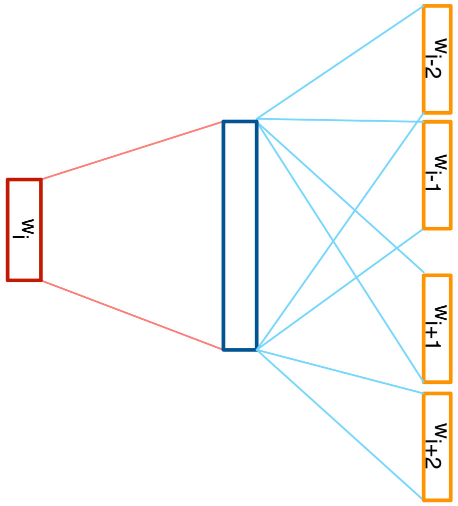
Orthographic representation

bag of letter n-grams=
10222 tri-bi-1-grammes

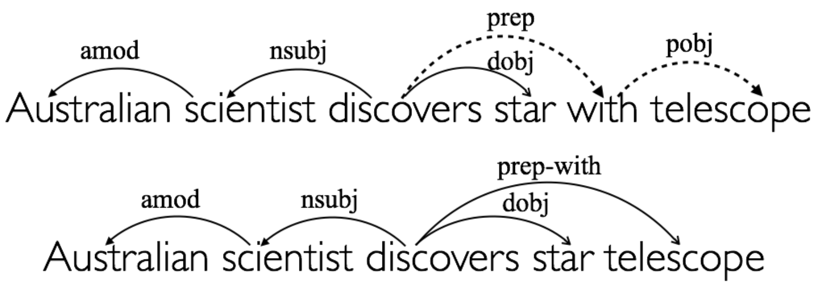
LINGUISTIC EMBEDDINGS

COMBINED WORD EMBEDDINGS

Skip-gram [T. Mikolov *et al.* 2013]



w2vf-deps [O. Levy *et al.* 2014]


amod
nsubj
prep
dobj
pobj
Australian scientist discovers star with telescope
amod
nsubj
prep-with
dobj
Australian scientist discovers star telescope

GloVe [J. Pennington *et al.* 2014]

- building a co-occurrence matrix
- estimating continuous representations of the words

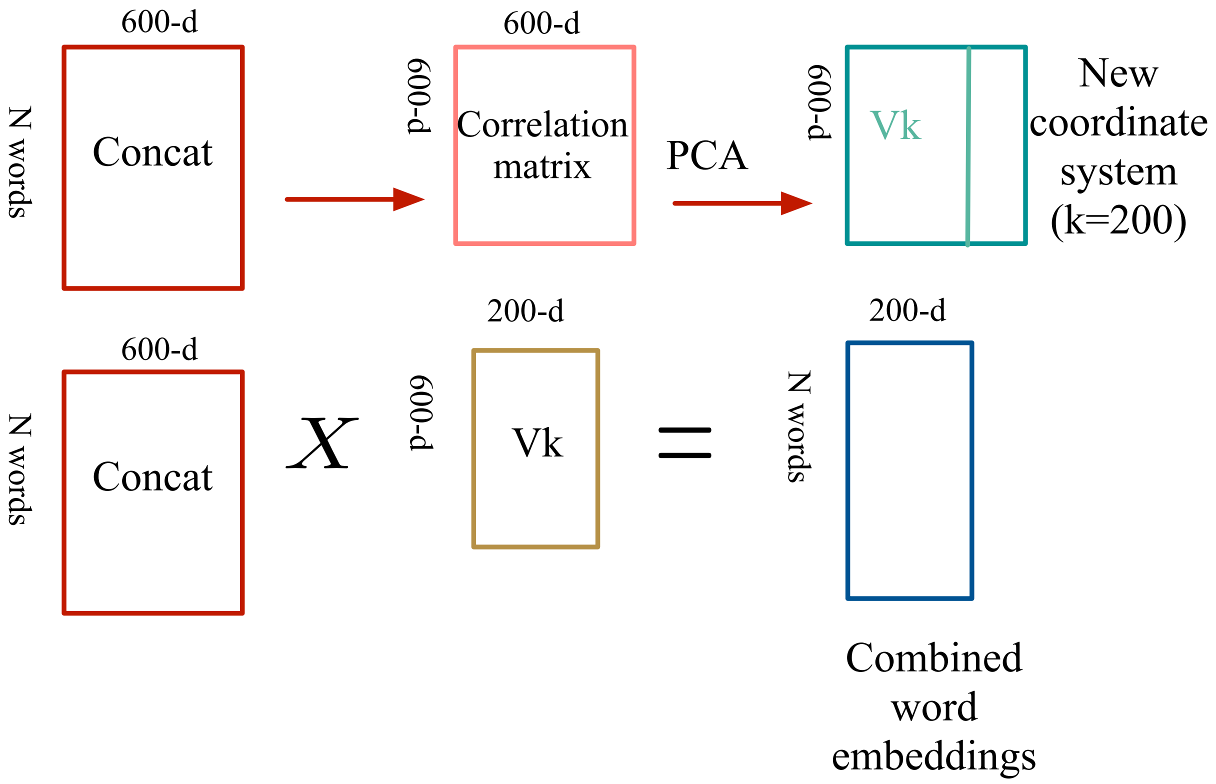
Evaluation and combination of word embeddings

[S.Ghannay *et al.* SLSP 2015, LREC 2016]

- ASR error detection
- NLP tasks
- Analogical and similarity tasks

➡ Combination of word embeddings through PCA yields good results on analogical and similarity task

Principal Component Analysis


600-d
N words
Concat
600-d
600-d
Correlation matrix
PCA
600-d
Vk
New coordinate system (k=200)
600-d
N words
Concat
X
200-d
600-d
Vk
=
200-d
N words
Combined word embeddings

SIMILARITY MEASURE TO ENRICH CONFUSION NETWORKS (1/2)

- Enriching confusing network by adding nearest neighbors
  - Based on cosine similarities (A_{Sim}, L_{Sim}) of acoustic and linguistic embeddings

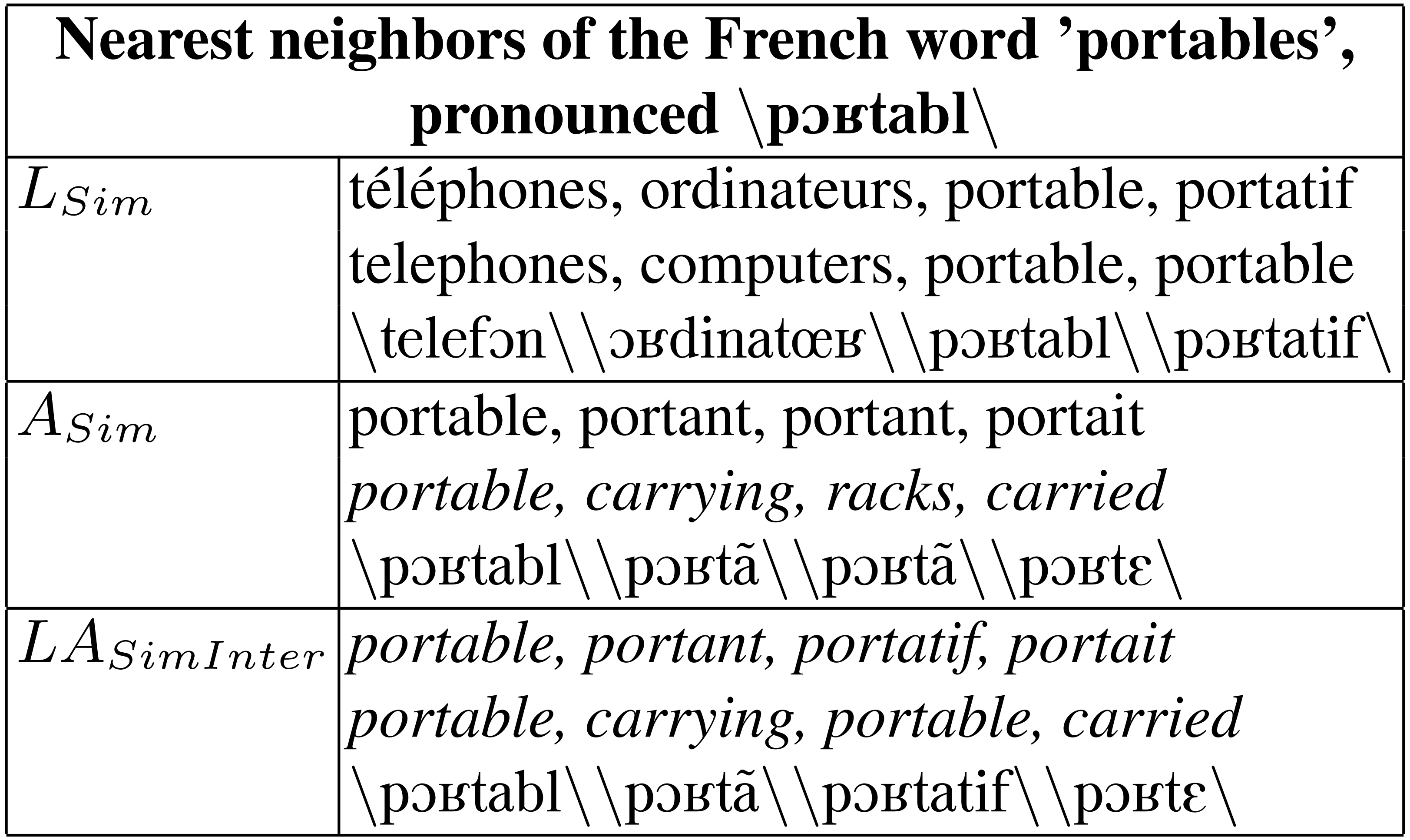
$$LA_{SimInter}(\lambda, x, y) = (1-\lambda) \times L_{Sim}(x, y) + \lambda \times A_{Sim}(x, y)$$

  - Optimisation of **λ** value:

$$\hat{\lambda} = argmin_{\lambda} MSE(\forall(h, \overline{r}) : P(h|\overline{r}), LA_{SimInter}(\lambda, h, \overline{r}))$$

SIMILARITY MEASURE TO ENRICH CONFUSION NETWORKS (2/2)

- Nearest neighbors of the hypothesis word ***portables***



Nearest neighbors of the French word ’portables’, pronounced \pɔʁtabl\	
L_{Sim}	téléphones, ordinateurs, portable, portatif telephones, computers, portable, portable \telefɔn\\\ɔʁdinatœʁ\\\pɔʁtabl\\\pɔʁtatif\
A_{Sim}	portable, portant, portant, portait *portable, carrying, racks, carried* \pɔʁtabl\\\pɔʁtã\\\pɔʁtã\\\pɔʁtɛ\
$LA_{SimInter}$	*portable, portant, portatif, portait* *portable, carrying, portable, carried* \pɔʁtabl\\\pɔʁtã\\\pɔʁtatif\\\pɔʁtɛ\

EXPERIMENTS

EXPERIMENTAL SETUP

- Training data of acoustic embeddings
  - 488 hours of French Broadcast news (ESTER1, ESTER2 et EPAC)
  - Vocabulary : 45k words and classes of homophones
  - Occurrences : 5.75 millions

- Training data of the linguistic word embeddings

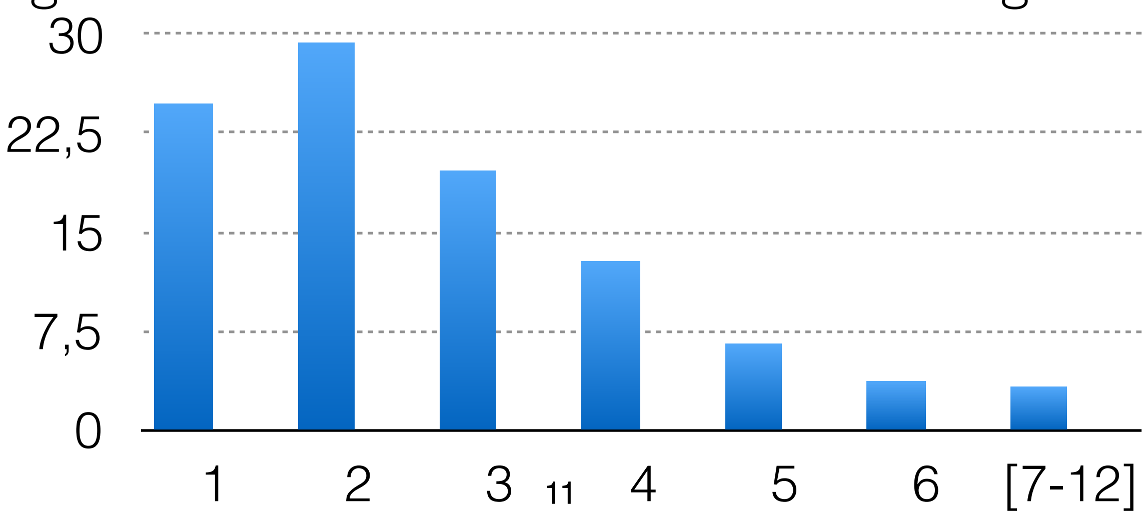
Corpus composed of 2 billions of words:

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

EXPERIMENTS

EXPERIMENTAL SETUP

- Experimental data
  - ETAPE corpus of French broadcast news shows
    - Enriched with automatic transcriptions generated by the LIUM ASR system
  - List of substitution errors:
    - $\text{Sub}_{\text{Train}}$: estimate the interpolation coefficient
    - Sub_{Test}: evaluate the performance of the Confusion Network (CN) enrichment approach
    - CN bins: Percentage of confusion network bins according to their sizes



Name	WER	Sub.Err.	#sub. Error pairs (ref, hyp)
Train	25.3	10.3	30678
Test	21.9	8.3	4678

Description of the experimental corpus

EXPERIMENTS

TASKS AND EVALUATION SCORE

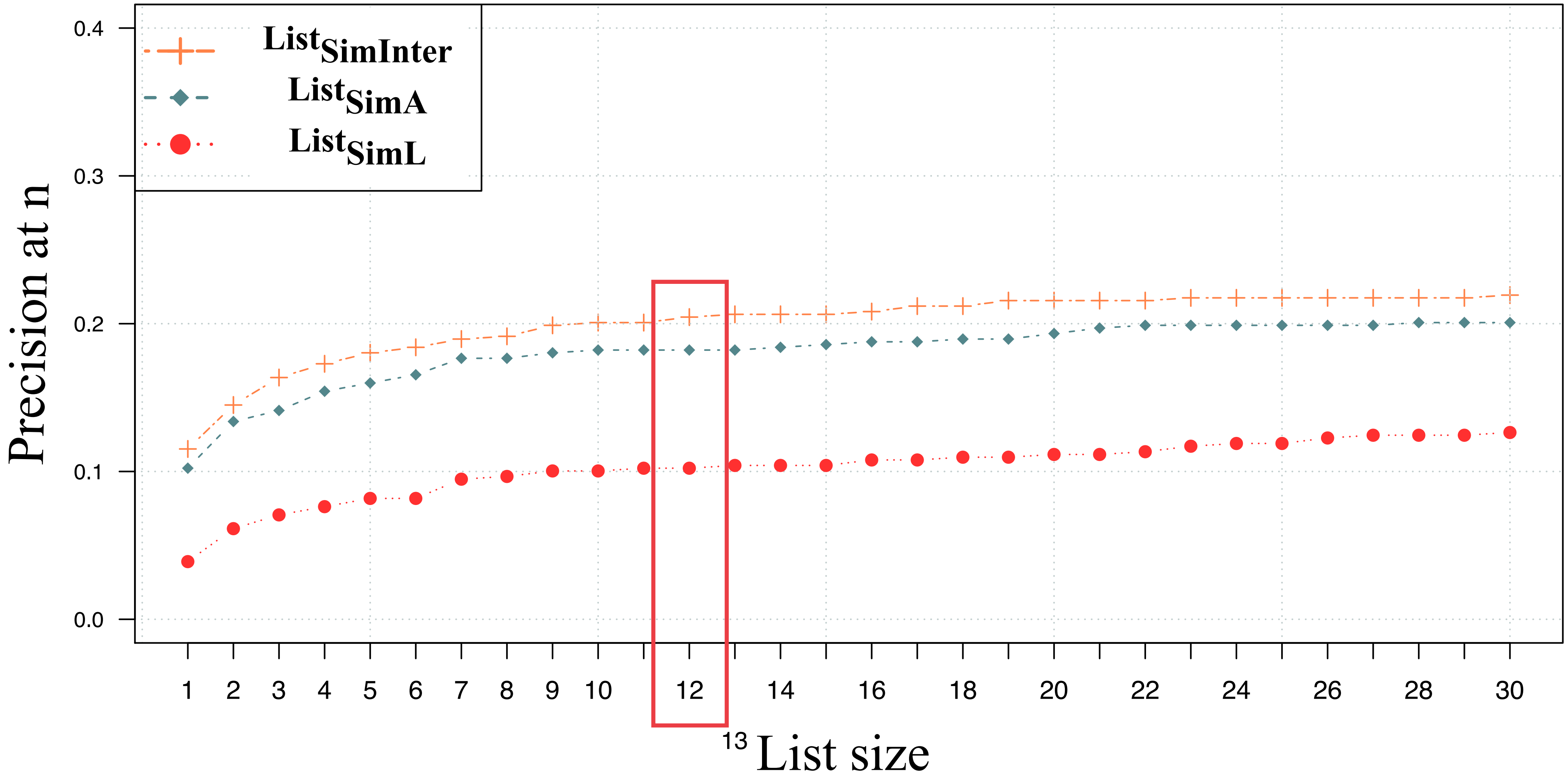
- Two Evaluation tasks
  - Task 1: prediction of errors for rare words (a = ref, b = hyp)
  - Task 2: post processing of ASR errors (a = hyp, b = ref)
  - Given a word pair (a,b) in a list L of m substitution errors
  - looking for b in list N of the n nearest words of a based on the similarity measure Γ: $A_{Sim,}$ or $L_{Sim,}$ or $LA_{SimInter}$
- Evaluation score: $S(\Gamma, n) = \frac{\sum_{i=1}^{m} f(i, \Gamma, n) \times \#(a_i, b_i)}{\sum_{i=1}^{m} \#(a_i, b_i)}$

$$f(i, \mathit{\Gamma}, n) = \begin{cases} 1 & \mathit{if}\, b_i \subset N(a_i, \mathit{\Gamma}, n) \\ 0 & \mathit{otherwise} \end{cases}$$

EXPERIMENTS

EXPERIMENTAL RESULTS

- Prediction of potential error for rare words
  - List of rare words : 538 pairs of substitution errors
  - Lists: $\text{List}_{\text{SimL}}$, $\text{List}_{\text{SimA}}$, $\text{List}_{\text{SimInter}}$ of nearest neighbors to the reference word (r)



Experiments

Experimental results

- The similarity $LA_{SimInter}$ is used to:
  - Enrich confusion networks bins with nearest neighbors of hypothesis (hyp) word
    - Evaluation on post processing of automatic transcriptions

	$List_{CN}$	$List_{ErichCN}$
P@6	0,17	0,21 (+23,5%)

EXPERIMENTS

EXPERIMENTAL RESULTS

- The similarity $LA_{SimInter}$ is used to:
  - Expand the automatic transcriptions (1-best) provided for a spoken language understanding (SLU) system -> build confusion networks
    - Task: correction of semantically relevant erroneous word
    - Data: French MEDIA corpus (1257 dialogues for hotel reservation)
      - Evaluation corpus: 1204 occurrences of semantically relevant erroneous words

	Enrich 1-best
P@6	0,206

Conclusion

- Take benefit from linguistic and acoustic embeddings:
  - Enrich confusion networks (CN)
  - ➡ Improve post-processing
- Compute a similarity function $LA_{SimInter}$ optimized to ASR error correction
  - Relevant lists of nearest neighbors linguistically and acoustically
  - Enrich CN and increase the potential correction of erroneous words by 23%
  - Propose 6 alternative words to 1-best hypotheses carrying on semantics to be exploited by the SLU module
    - ➡ These alternatives contain the correct words in 20.6% of the cases

Thank you !

Contact

sahar.ghannay@univ-lemans.fr