# ASR error management for improving spoken language understanding

*Edwin Simonnet [1], Sahar Ghannay [1], Nathalie Camelin [1], Yannick Estève [1], Renato De Mori [2 3]*

[1]**LIUM, Le Mans University, France** (firstname.lastname@univ-lemans.fr)
[2 3]**LIA, University of Avignon, France – McGill University, Montreal, Canada**
(rdemori@cs.mcgill.ca)

## Introduction

### Subject:

**Automatic speech recognition** (ASR) error detection for improving **spoken language understanding** (SLU)

### SLU task:

- Automatically extracting **semantic concepts** and **concept/values** pairs from ASR transcriptions
- **BI** (Begin,Inside) annotation : **delimits** utterances mentioning concepts
- **Evaluation** in **Concept Error Rate** (CER) and **Concept-Value Error Rate** (CVER)

| WORD | I | want | to | book | a | room |
|---|---|---|---|---|---|---|
| CONCEPT | command | | | | number | object |
| TAG | command-B | command-I | command-I | command-I | number-B | object-B |
| VALUE | booking | | | | 1 | room |

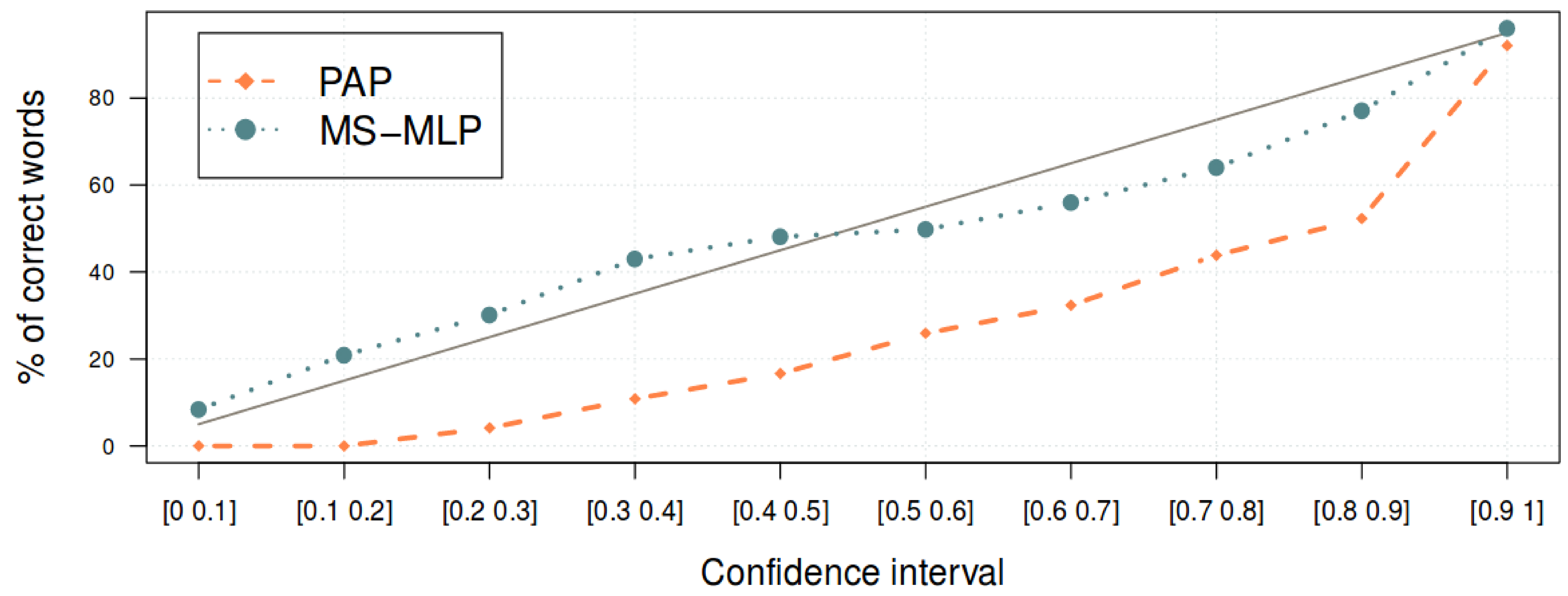
### Problem:

**ASR** still makes **errors** involving error-prone interactions between SLU and ASR:
- ASR errors may **affect** the mention of a **concept** and the **value** of a concept instance.
- context features may be **insufficient** or cause **interpretation errors** due to ASR errors

## Approach

- **Enriching** the set of semantic labels with **ASR error** labels
  - erroneous hypothesized word supporting a **concept** → **ERROR-C**
  - otherwise (**null**) → **ERROR-N**
  - then **replaced** by **null** (usual SLU MEDIA evaluation protocol)
- **ASR confidence measures** used as additional **SLU features** for localizing ASR errors
  1) Word **posterior probability (pap)** computed with confusion networks
  2) Acoustic word embeddings for ASR error detection computed with a **Multi-Stream Multi-Layer Perceptron (MS-MLP)** architecture

*[S. Ghannay, INTERSPEECH 2016, Acoustic word embeddings for asr error detection]*

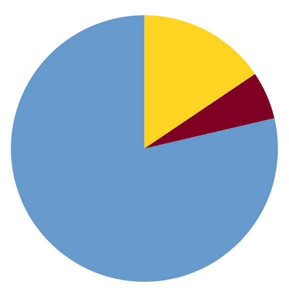



*ASR error prediction capabilities on TEST*

## Experimental Protocol

### MEDIA corpus:

- **Touristic** information system
- **French** corpus
- **22,5k** telephone utterances
- **74** concept labels




### LIUM ASR system dedicated to MEDIA:

- **Winner** on **last evaluation campaign** (REPERE) on French language
- **Kaldispeech recognition toolkit** based
- **Trained** on **145,781** speech segments
- **DNN** model

| | train | dev. | test. |
|---|---|---|---|
| ASR WER | 23.7% | 23.4% | 23.6% |

### Set of features:

Word dependent features → **improve understanding** performance
- **Semantic**
  - MEDIA specific (cities, hotels…) or more general (figures, months …)
- **Syntactic**
  - lemma, POS tag, word governor and relation with the current word
- **Morphological**
  - first and last letters ngrams
- **ASR confidence measures**
  - pap and MS-MLP

## SLU Architectures

### Conditional Random Fields (CRF):

- **Discrete** values
- **Best performance** on MEDIA
- **Wapiti** toolkit
- Word with **context window**

### Encoder-Decoder Bidirectional Neural Network with a Mechanism of Attention (NN-EDA):

- **Continuous** values
- **nmtpy** framework
- Inspired from **machine translation:**
  - words → semantic concept tags
- Encoding**:**
  - **bidirectional NN** encodes the sentence
- Decoding**:**
  - **attention mechanism** gives more weight to **relevant information**

## Results on TEST and conclusions

*[baseline refers to state of the art CRF baseline issued from S. Hahn, 2011 Comparing stochastic approaches to spoken language understanding in multiple languages]*

### Standard SLU task (no error detection):

| | Concept | | | Concept-Value | | |
|---|---|---|---|---|---|---|
| | %Error | P | R | %Error | P | R |
| baseline | 23.8 | - | - | 27.3 | - | - |
| NN-EDA | 22.3 | 0.88 | 0.84 | 28.8 | 0.81 | 0.77 |
| CRF | 19.9 | 0.90 | 0.85 | 25.1 | 0.85 | 0.80 |

➢ **CRF outperformed** NN-EDA with significant improvement **over the baseline**

### Impact of the Confidence Measure (CM):

| | without CM | | +pap | | +pap +MS-MLP | |
|---|---|---|---|---|---|---|
| | C | CV | C | CV | C | CV |
| CRF | 20.9 | 26.0 | 20.5 | 25.7 | 19.9 | 25.1 |

➢ Confidence and input features **contribute to error reductions**

### Joint SLU and ASR error detection tasks (standard SLU evaluation):

| | Concept | | | Concept-Value | | |
|---|---|---|---|---|---|---|
| | %Error | P | R | %Error | P | R |
| NN-EDA | 22.1 | **0.90** | 0.82 | 27.8 | **0.84** | 0.77 |
| CRF | 20.6 | **0.91** | 0.84 | 25.4 | **0.86** | 0.79 |

➢ Similar to standard SLU task but **better precision**

### Consensus among CRF and neural systems and their combination:

| | Concept | | | Concept-Value | | |
|---|---|---|---|---|---|---|
| | %Err. | P | R | %Err. | P | R |
| baseline combination | 23.1 | - | - | 27.0 | - | - |
| CRF+NN combination | **19.3** | 0.91 | 0.85 | **24.5** | 0.86 | 0.80 |
| CRF+NN consensus | - | **0.96** | 0.72 | - | **0.89** | 0.68 |

➢ **Combination: weighted vote** between best systems
- Provides a **significant error reduction**

➢ **Consensus: agreement** among systems (null otherwise)
- provides significantly **higher precision** and a restrained **recall reduction**
- identifies **confidence islands** and **uncertain semantic output segments**