Simulating ASR errors for training SLU systems

Edwin Simonnet , Sahar Ghannay , Nathalie Camelin , Yannick Estève

LIUM, Le Mans University, France
(firstname.lastname@univ-lemans.fr)

Introduction

Subject:

Simulating **automatic speech recognition** (ASR) errors from manual transcriptions to improve **spoken language understanding** (SLU) systems performances

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

Problems:

◆ **Transition** from **Manual to ASR** transcriptions makes SLU performances **worse**
◆ SLU systems need to be **prepared to ASR errors** during their **training**
◆ Large **automatic** transcription corpora needed for **training and validation** are **not always available**

Error Simulation Approach

◆ **Substitution** of **correct** words by **similar ones** in manual transcriptions
• **Assumption:**
words **confusable** by ASR are **acoustically/linguistically** close

• Computing a **confusability measure** between **words *(x,y)*** from **cosine similarities** between **acoustic (*Asim*)** and **linguistic (*Lsim*)** word embeddings:

$$confus(x,y) = LASimInter(\lambda, x, y)$$

with

$$LASimInter(\lambda, x, y) = (1-\lambda) \times LSim(x,y) + \lambda \times ASim(x,y)$$

$$\lambda = argmin\ MSE(\forall(hyp, ref) : P(hyp|ref), LASimInter(\lambda, hyp, ref))$$

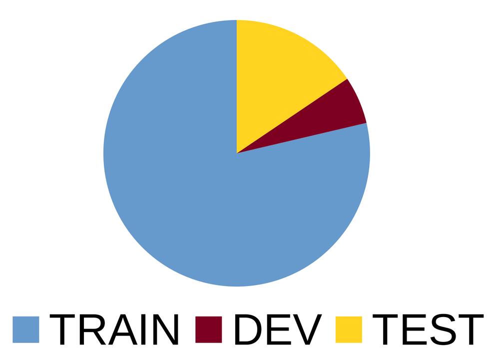
◆ Applying ***confus(x,y)*** in order to substitute **20%** *(cf. ASR WER)* of correct words **randomly** by one of its ***n*** **closest confusable words**
• Noised corpus **Noisy7** with n=7
• Noised corpus **Noisy10** with n=10
• Noised corpus **NoisyNaive** not taking confus(x,y) into account

◆ Confusability measure used as a **feature** like ASR confidence measure

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 or MS-MLP

SLU Architectures

Conditional Random Fields (CRF):

◆ **Discrete** values
◆ **Best performance** on MEDIA
◆ **Wapiti** toolkit
◆ Word with **context window**
◆ No need for **validation**

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**
◆ Proceed **validations** during training

Results on ASR TEST and conclusions

ASR SYSTEM AVAILABLE DURING TRAINING:

	NN-EDA		*CRF*	
TRAIN set	*CER*	*CVER*	*CER*	*CVER*
Manual	31.6	36.2	27.5	31.6
ASR	22.5	28.3	**19.9**	**25.1**
Noisy7	23.8	29	22.6	27.7
DoubleNoisy7	23.2	28.8	26.3	31.3
Manual+Noisy7	22.7	28.1	22.6	27.7
Manual+Noisy10	23.3	28.5	23.2	28.3
Manual+NoisyNaive	23.7	28.8	25	30.3
Manual+ASR	**20.7**	**25.8**	20.2	25.3
Manual+Noisy7+ASR	**20.2**	**26**	29.1	33.0

◆ For Both SLU systems:
→ Importance of getting *ASR* or *ASR simulated* transcriptions to get training data as close as possible to the test data
ASR > Noisy (acceptable simulation) **> Manual** (insufficient)

→ Performance on Manual+Noisy corpora: **Noisy7** > **Noisy10** > **NoisyNaive**
▪ Substituting correct words with globally more similar words increases the results
▪ Importance of an intelligently generated noise

◆ Neural system only (ASR DEV is used during validation) :
→ Benefits from training data augmentation
▪ Manual+Noisy as good as ASR
▪ **Manual+ASR+Noisy>ASR and Manual+ASR>ASR**

→ **Gap** between **CRF** and **NN-EDA** performances strongly **reduced**

ASR SYSTEM UNAVAILABLE DURING TRAINING:

		NN-EDA	
TRAIN set	*DEV set*	*CER*	*CVER*
Manual	Manual	33.9	38.2
Noisy7	Noisy7	23.5	28.6
Manual+Noisy7	Noisy7	23.1	28.5

◆ Significant improvement by applying ASR error simulation approach
▪ Manual transcriptions of training and development corpora are noised

With no ASR data but noisy data → very close results to ASR TRAIN/DEV