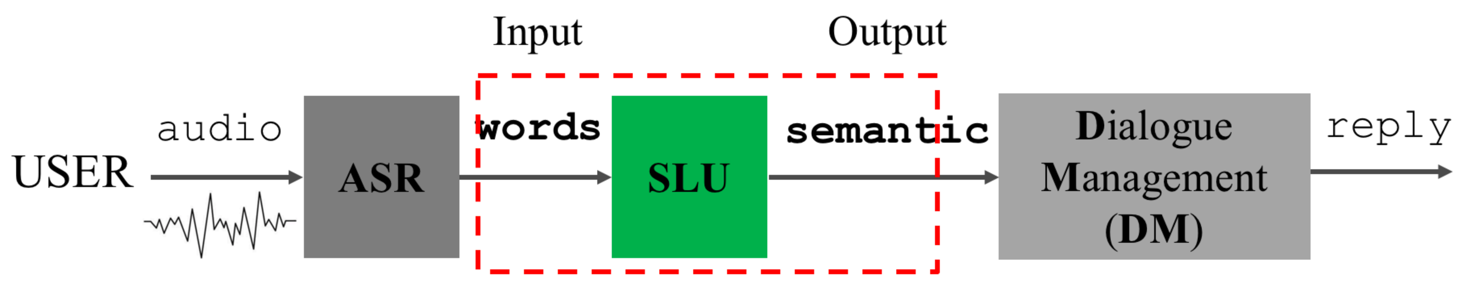
Robust Spoken Language Understanding with Unsupervised ASR-error Adaptation

Su Zhu, Ouyu Lan, Kai Yu
SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{paul2204,blue-0-0-,kai.yu}@sjtu.edu.cn

Summary

- Motivation: **Speech Recognition Errors & Robustness**
  - to improve robustness of SLU (Spoken Language Understanding).
- Our approach: **Parameter Partial-Sharing BLSTMs & Unsupervised ASR-error Adaptation**
  - Only speech recognized text is used for adaptation. There is no SLU annotation on the speech recognized text.
- Result:
  - Our method improves the robustness of SLU significantly.
  - There is no need of SLU annotation on the speech recognized text.

1. Introduction



SDS: Spoken Dialogue System ; **ASR** : Automatic Speech Recognition

Slot tagging task of SLU

Input: words	show	flights	from	Boston	to	New	York	today
Output: slots	O	O	O	B-FromCity	O	B-ToCity	I-ToCity	B-DepartDate

In/Out/Begin (IOB) representation

Robustness of SLU to ASR-error

- Inputs of SLU (e.g. slot tagging):
  1. **Manual transcription** (Oracle)
  2. **ASR output** (May contain errors)

  The actual input of SLU in SDS
- Target of SLU (e.g. slot tagging):
  - Human annotation based on the inputs of SLU. It aims to investigate SLU module independently.

Traditional Methods: Prepare Training Data

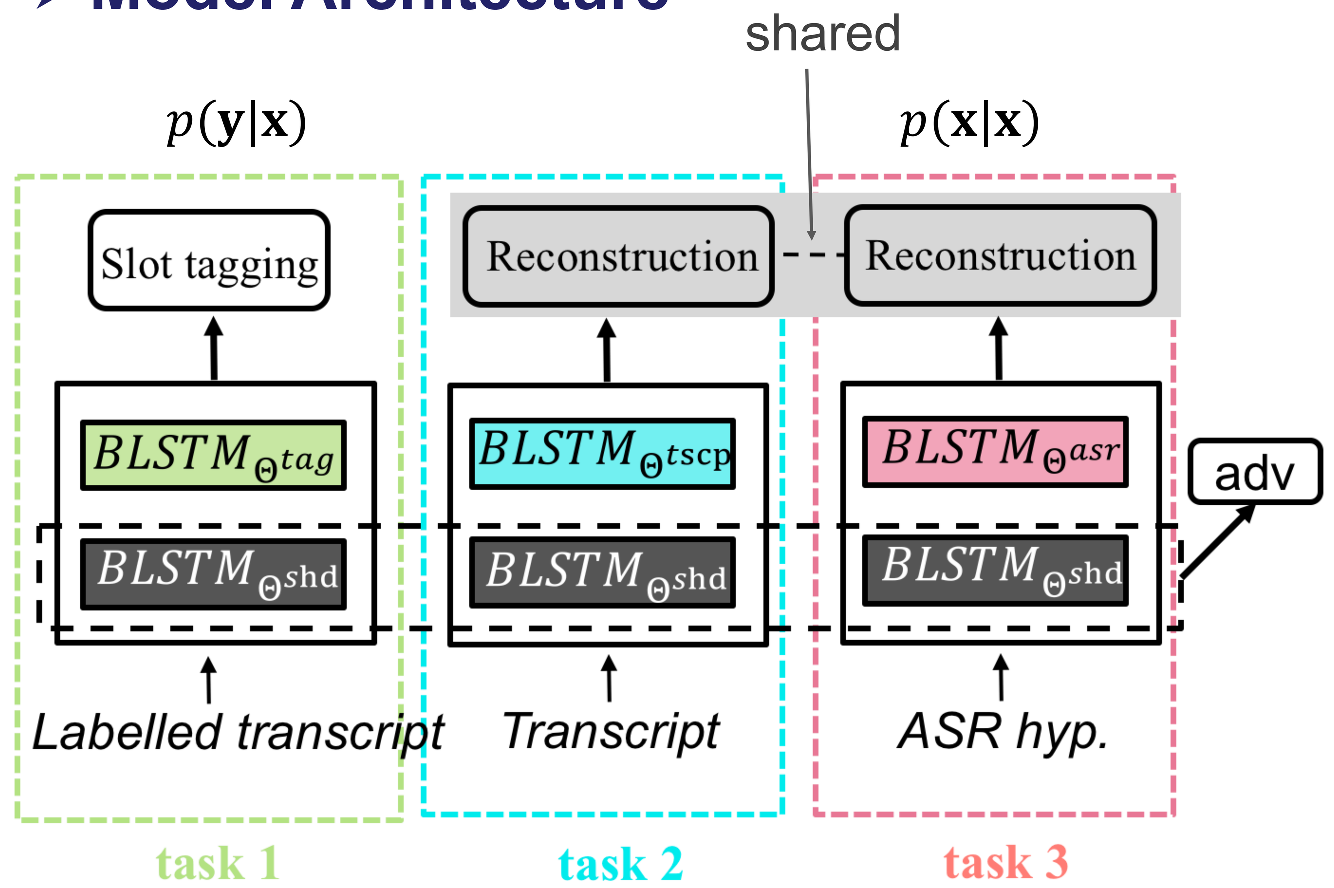
1. Human annotation on the manual transcription. ✘
   Manual transcription is **mismatched** with ASR output.
2. Human annotation on the ASR output. ✘
   What if the ASR system changes? (i.e. If ASR output is changed, we need to renew the semantic annotation.) **labor-intensive** & **time-consuming**

✚ Unlabeled ASR output for adaptation. ✔ **(Our method)**

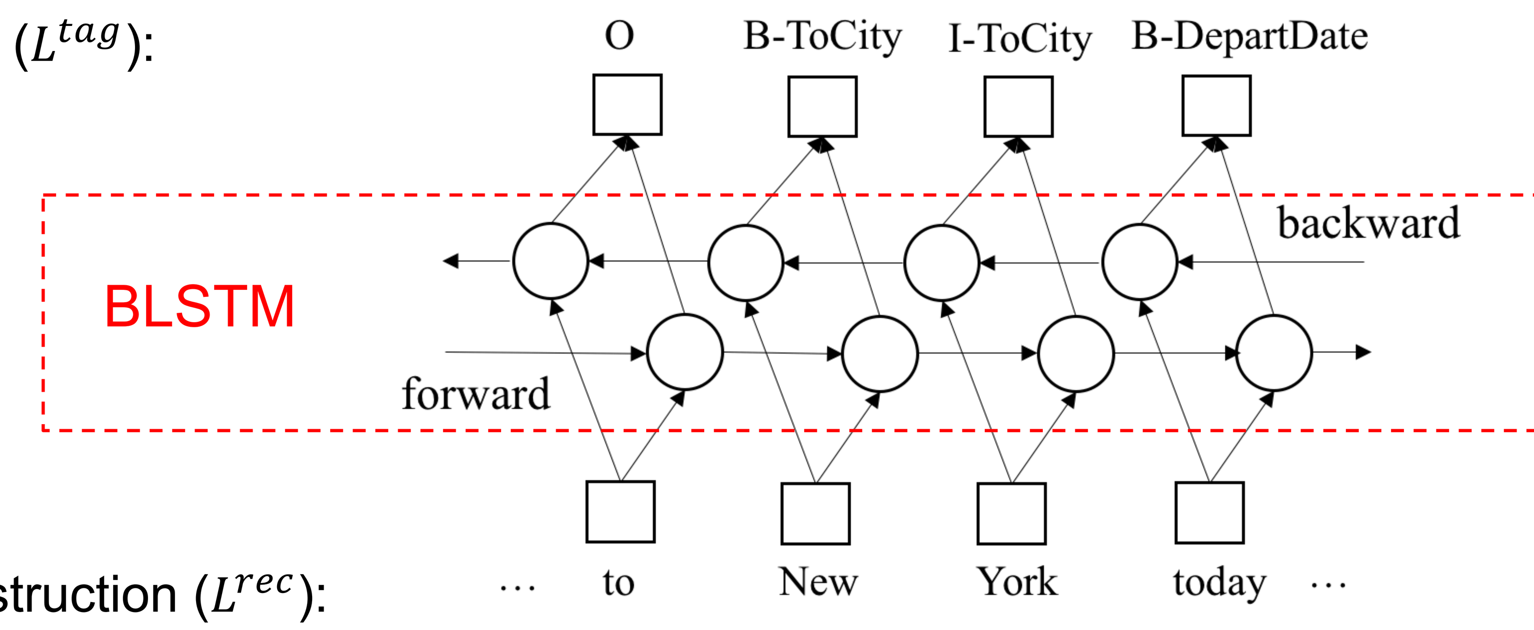
2. Unsupervised ASR-error Adaptation

Types of Data samples:
1. *tag*: manual transcription && semantic annotation of slot-tags;
2. *tscp*: manual transcription
3. *hyp*: ASR output (1-best, unlabeled);

Model Architecture



- Four BLSTMs:
  Shared: $BLSTM_{\Theta shd}$ **Private:** $BLSTM_{\Theta tag}$ $BLSTM_{\Theta tscp}$ $BLSTM_{\Theta asr}$
- Slot tagging (L^{tag}):



- Input reconstruction (L^{rec}):
  - Word to word (W2W): $p(\mathbf{x}|\mathbf{x}) = \sum_i p(\mathrm{x}_i|\mathbf{x})$ ← a naive try
  - Sequence to sequence (S2S): $p(\mathbf{x}|\mathbf{x}) = \sum_i p(\mathrm{x}_i|\mathrm{x}_{0:i-1};\mathbf{x})$
  - Bidirectional language model (BLM): $p(\mathbf{x}|\mathbf{x}) = \sum_i p(\mathrm{x}_{i+1}|\mathrm{x}_{0:i}) + \sum_i p(\mathrm{x}_{i-1}|\mathrm{x}_{i:T+1})$
- Adversarial task classification loss (L^{adv}):
  - We use random prediction training (Kim, 2017) to force the shared encoder more task-invariant, i.e. the label of task classifier is randomly set to task 1/2/3 with equal probability.

3. Experiments

- **Dataset**: collected from a Chinese commercial SDS.
  - Domain: car navigation (13 different slots).

Data partitions		#sentence	CER (Character Error Rate)
train+valid	labelled transcripts (*tag*)	7205	21.52
	Transcripts (*tscp*)	7205	
	ASR top-hyp. (*asr*)	7205	
test	labelled ASR top-hyp.	1803	23.47
	labelled transcripts	1803	

- **Systems**:
  - Oracle$_1$: It is trained on the data of **ASR output** with SLU annotation.
  - Oracle$_2$: It is trained on the data of both **manual transcription** and **ASR output** with SLU annotation.
  - Baseline$_1$: It is trained on **manual transcription** with SLU annotation.
  - Baseline$_2$: **+ ASR output** with SLU auto-annotation. (word **alignment** between transcript and ASR 1-best)
  - Domain adaptation (Kim, 2017): $BLSTM_{\Theta tscp} = BLSTM_{\Theta asr}$
  - $L^{tag} + L^{rec}$: Training is driven by slot tagging and reconstruction.
  - $L^{tag} + L^{rec} + L^{adv}$: Additional adversarial task classification.

- **Results** of slot tagging on ASR output and manual transcriptions.

System	Recon-struction	F1-score on (F1-score of slot segmentation): ASR-output	manual transcript
Oracle$_1$	----	84.65	88.01
Oracle$_2$	---	85.64	89.82
Baseline$_1$	----	81.90	88.63
Baseline$_2$	----	78.71	84.94
Domain adaptation	S2S	82.52	87.44
$L^{tag} + L^{rec}$	W2W	82.82	88.00
$L^{tag} + L^{rec}$	S2S	83.31	88.54
$L^{tag} + L^{rec}$	BLM	84.87	89.16
$L^{tag} + L^{rec}$	BLM^{sep}	84.02	89.77
$L^{tag} + L^{rec} + L^{adv}$	BLM	85.11	88.99

L^{tag} (Oracle$_1$–Baseline$_2$); upper bound (Oracle$_2$); **GAP** caused by ASR-errors (Baseline$_1$); BLM is most suited.

Reconstruction networks are separated (not shared).

- Bidirectional language model is a good choice for input reconstruction.
- Adv. performs better (but not significantly).
- Our method become very close to the upper bound.
- We need more data to verify our method.

[1] Young-Bum Kim, Karl Stratos, and Dongchan Kim, "Adversarial adaptation of synthetic or stale data," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1297–1307.