

End-to-End Joint Learning of Natural Language Understanding and Dialogue Manager

XUESONG YANG



YUN-NUNG (VIVIAN) CHEN



DILEK HAKKANI-TÜR



PAUL CROOK

XIUJUN LI

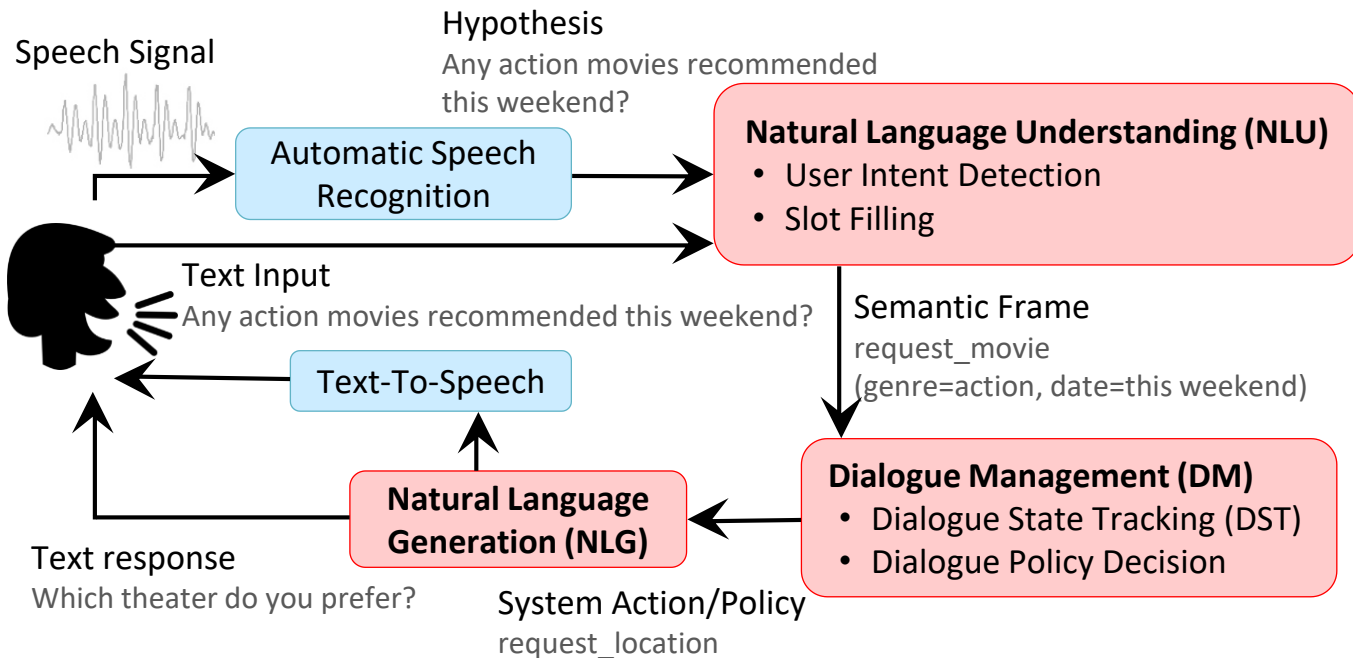
JIANFENG GAO

LI DENG



Pipelined Task-Oriented Dialogue System

2



Motivation: The pipelined system (NLU → DM) results in **error propagation** issues.

Proposed Approach

3

- End-to-end model
 - ▣ Mitigate the effects of noisy output from NLU
 - ▣ Refine NLU by supervised signals from DM
- Multi-task jointly learning
 - ▣ NLU - User intent classification
 - ▣ NLU - User slot tagging
 - ▣ DM - System action prediction
- Contextual understanding
 - ▣ Access to the user history
 - ▣ Monitor user behavior states over turns

Human-Human Dialogue Interaction

4

Hi, how may I help you?

Are there any **cheap rate hotels** to put my bags?

Do you want to have a backpack type of hotel?

Yes. Just gonna **leave our things** there and stay out the whole day.

So you don't mind if it is **not roomy**, right?

Yes.

Okay. These hotels are available for you: ...

Ok, thank you, bye!

Thanks, goodbye.

Idea: predicting the next system action given the current user utterance together with the aggregated observations



Guide Agent



Tourist User

Natural Language Understanding (NLU)

5

Utterance: BOS % um how much is a taxi cab there ? EOS

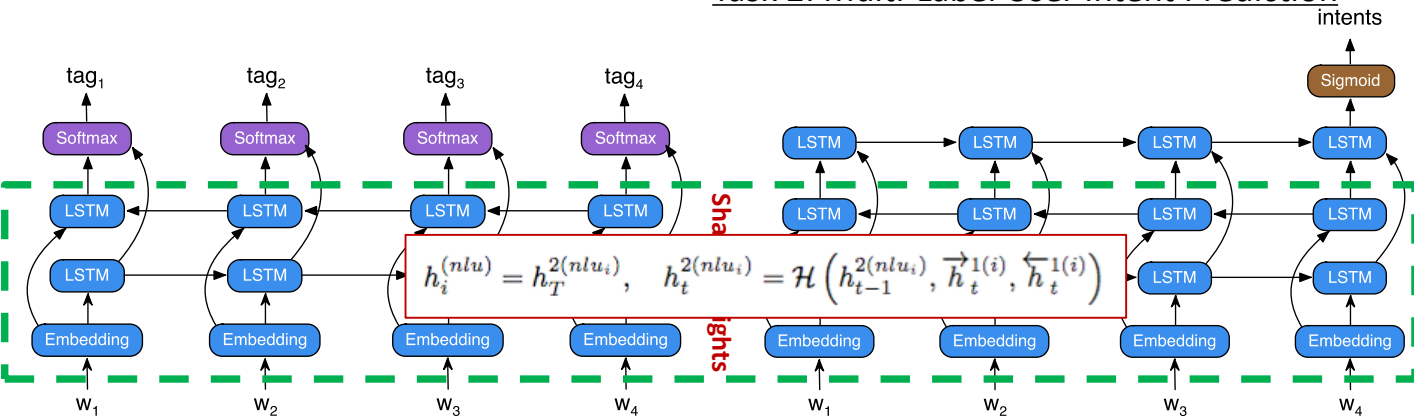
Slot Tags: O O O B-det_PRICE I-det_PRICE O O B-trsp_TYPE I-trsp_TYPE B-area_CITY O O

User Intents: QST_HOW_MUCH; QST_INFO

System Actions: RES_EXPLAIN; RES_INFO; FOL_EXPLAIN; FOL_HOW_MUCH; FOL_INFO

Task 1: Slot Tagging

Task 2: Multi-Label User Intent Prediction



$$\vec{h}_t^{1(i)} = \mathcal{H}(w_t, \vec{h}_{t-1}^{1(i)}), \quad \overleftarrow{h}_t^{1(i)} = \mathcal{H}(w_t, \overleftarrow{h}_{t+1}^{1(i)})$$

$$\hat{y}_t^{(tag_i)} = \arg \max \left(\text{softmax} \left(\vec{W}_{hy}^{(tag)} \vec{h}_t^{1(i)} + \overleftarrow{W}_{hy}^{(tag)} \overleftarrow{h}_t^{1(i)} \right) \right)$$

$$\vec{h}_t^{2(int_i)} = \mathcal{H}(h_{t-1}^{2(int_i)}, \vec{h}_t^{1(i)}, \overleftarrow{h}_t^{1(i)})$$

$$p_n^{(int_i)} = \text{sigm} \left(W_{hy}^{2(int_i)} h_T^{2(int_i)} \right)$$

$$\hat{y}_n^{(int_i)} = \begin{cases} 1, & p_n^{(int_i)} \geq \text{threshold} \\ 0, & \text{otherwise} \end{cases}$$

NLU+DM 1: Pipelined BLSTMs

6

Utterance: BOS % um how much is a taxi cab there ? EOS

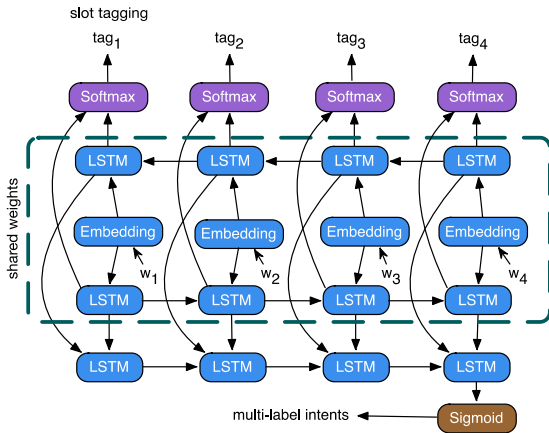
Slot Tags: O O O B-det_PRICE I-det_PRICE O O B-trsp_TYPE I-trsp_TYPE B-area_CITY O O

User Intents: QST_HOW_MUCH; QST_INFO

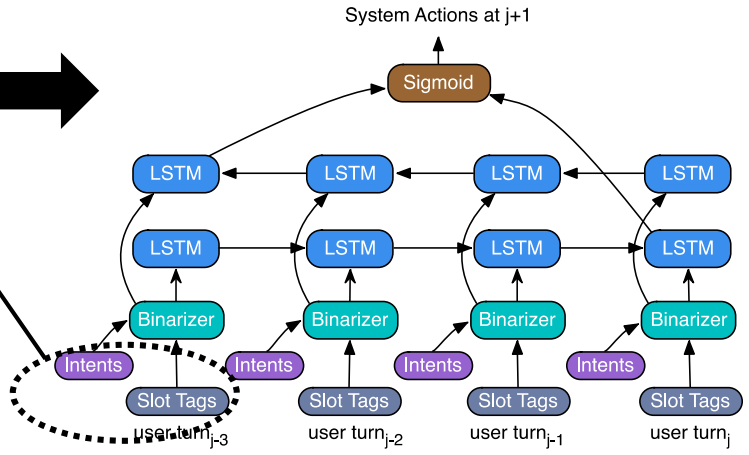
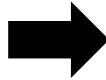
System Actions: RES_EXPLAIN; RES_INFO; FOL_EXPLAIN; FOL_HOW_MUCH; FOL_INFO

Task 1+2: Natural Language Understanding

Task 3: Multi-Label System Action Prediction



single user turn



current user turn w/ contextual history

NLU+DM 2: End-to-End Model (JointModel)

7

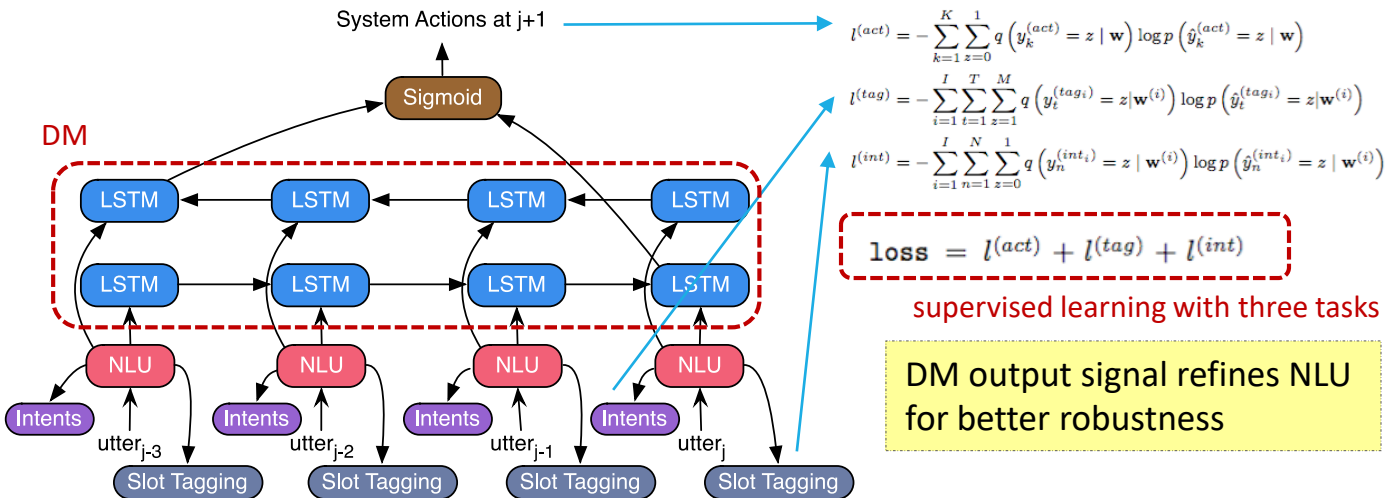
Utterance: BOS % um how much is a taxi cab there ? EOS

Slot Tags: O O O B-det_PRICE I-det_PRICE O O B-trsp_TYPE I-trsp_TYPE B-area_CITY O O

User Intents: QST_HOW_MUCH; QST_INFO

System Actions: RES_EXPLAIN; RES_INFO; FOL_EXPLAIN; FOL_HOW_MUCH; FOL_INFO

Task 1+2+3: End-to-End Joint NLU+DM



Data

8

- Dialogue State Tracking Challenge 4
 - Human-human dialogues: 21-hour dialogue sessions on touristic information collected via Skype between tour guides and tourists

Domains	Speech Act	Speech Act Attributes			
Accommodation	QST (QUESTION)	ACK	CLOSING	COMMIT	THANK
Attraction	RES (RESPONSE)	CANCEL	CONFIRM	ENOUGH	WHAT
Food	INI (INITIATIVE)	EXPLAIN	HOW_MUCH	HOW_TO	WHEN
Shopping	FOL (FOLLOW)	INFO	NEGATIVE	OPENING	WHERE
Transportation		POSITIVE	PERFERENCE	RECOMMEND	WHO

User Intent = Speech Act + Attributes
System Action = Speech Act + Attributes

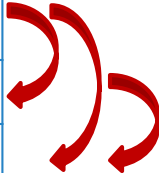
	Train	Dev	Test
#utter	5,648	1,939	3,178
#intent	68	54	58

DM Result – System Action Prediction (SAP)

9

- Metric: frame-level accuracy (FrmAcc)

Model	FrmAcc
Baseline (CRF+SVMs)	7.7
Pipeline-BLSTM	12.0
JointModel	22.8



Human-human conversations are complicated, so predicting system actions for DM is difficult

Pipeline-BLSTM and **JointModel** outperform the baseline

JointModel improves **Pipeline-BLSTM** about 10% accuracy, indicating the importance of mitigating downside of pipeline

DM Result – System Action Prediction (SAP)

10

- Metric: frame-level accuracy (FrmAcc)

Model	FrmAcc
Baseline (CRF+SVMs)	7.7
Pipeline-BLSTM	12.0
JointModel	22.8
Oracle-SAP (SVM)	7.7
Oracle-SAP (BLSTM)	19.7

Fully correct NLU output

Oracle models show the upper-bound of the SAP performance, since it transfers the errors from NLU to SAP

Contextual user turns make significant contribution to DM performance

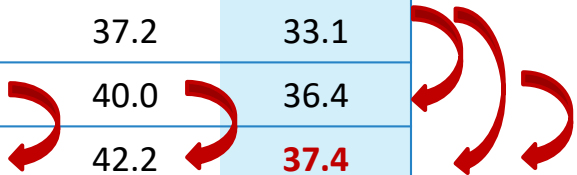
JointModel achieves the best DM performance (FrmAcc) with richer latent representations

NLU Result – Slot Filling & Intent Prediction

11

- Metrics: frame-level accuracy (FrmAcc)

Models	Slot	Intent	NLU
NLU-Baseline (CRF+SVM)	77.3	37.2	33.1
NLU-Pipeline-BLSTM	76.8	40.0	36.4
NLU-JointModel	76.5	42.2	37.4



CRF+SVMs baseline maintains strong NLU performance with 33.1%

Pipeline-BLSTM and **JointModel** outperformed the baseline

Extra supervised DM signal helps refine the NLU by back-propagating the associated errors

DM signal (system action prediction) helps more on user intent prediction than slot filling, and NLU is significantly improved

Conclusion

12

- First propose an **end-to-end** deep hierarchical model for **joint NLU and DM** with limited contextual dialogue memory
- Leverage **multi-task learning** using three supervised signals
 - ▣ NLU: User intent classification
 - ▣ NLU: Slot tagging
 - ▣ DM: System action prediction
- Outperform the state-of-the-art pipelined NLU and DM models
 - ▣ Better DM due to the contextual dialogue memory
 - ▣ Robust NLU fine-tuned by supervised signal from DM

13

Thanks for Your Attention! 😊

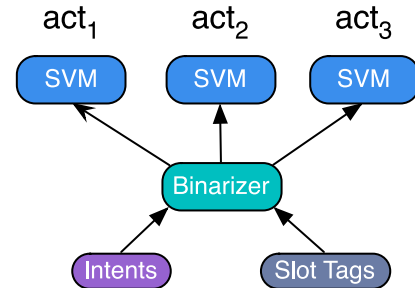
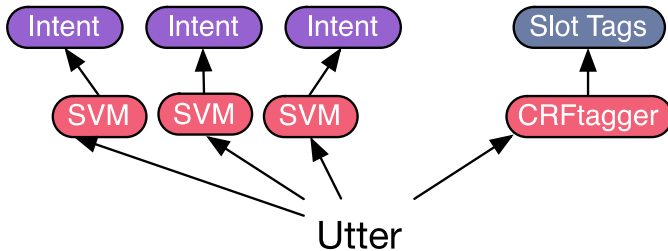
Code Available at

https://github.com/XuesongYang/end2end_dialog

Appendix - Baseline Model

Predicting system actions at the next turn as responses to the current user behaviors by pipelining NLU and SAP together

NLU: CRF for slot tagging, and
One-Vs-All SVMs for intent classification
SAP: One-Vs-All SVMs



[1] Raymond, Christian, and Giuseppe Riccardi. "Generative and discriminative algorithms for spoken language understanding." In *INTERSPEECH*, pp. 1605-1608. 2007.

Appendix: Configuration

15

- ❑ Optimizer: a mini-batch stochastic gradient descent method Adam
- ❑ Contextual history: five user turns
- ❑ Dimension of word embedding: 512
- ❑ Dropout ratio: 0.5
- ❑ No early stopping, but use 300 training epochs
- ❑ Best models for three tasks are selected individually under different metrics
 - ▣ Token-level micro-average F1 score is used for slot filling
 - ▣ frame-level accuracy (it counts only when the whole frame parse is correct) is used for both user intent prediction and system action prediction
 - ▣ the decision thresholds are tuned on dev set