Joint Online Spoken Language Understanding and Language Modeling with Recurrent Neural Networks

Bing Liu, Ian Lane

Carnegie Mellon University
liubing@cmu.edu, lane@cmu.edu

Outline

- Background & Motivation
- Proposed Methods
- Experiments & Results
- Conclusions

Outline

- **Background & Motivation**
- Proposed Methods
- Experiments & Results
- Conclusions

Background

- Spoken Language Understanding (SLU) is an important component in spoken dialog systems.
- Main tasks in SLU:
  - Intent Detection
  - Slot Filling

Utterance	show	flights	from	Seattle	to	San	Diego	tomorrow
Slots	O	O	O	B-fromloc	O	B-toloc	I-toloc	B-depart_date
Intent	Flight							

Background

- Intent detection
  - Sequence classification
  - SVM, CNN[1], Recursive NN[2], etc.

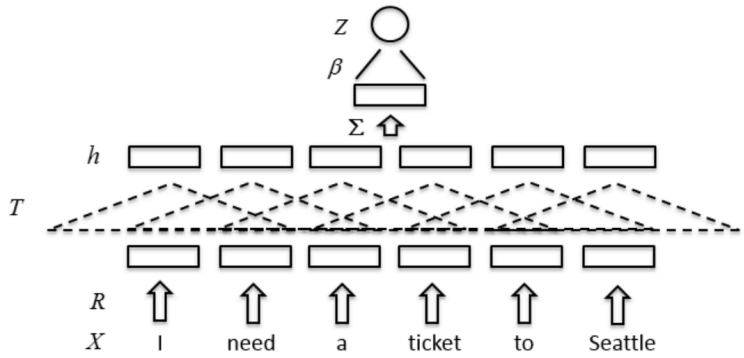


Fig 1. CNN[1] intent model

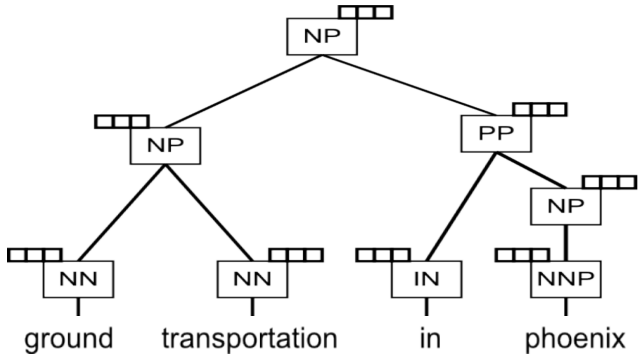


Fig 2. Recursive NN[2] intent model

[1] Xu, Puyang, and Ruhi Sarikaya. "Convolutional neural network based triangular crf for joint intent detection and slot filling." ASRU, 2013.
[2] Guo, Daniel, et al. "Joint semantic utterance classification and slot filling with recursive neural networks." SLT 2014.

Background

- Slot filling
  - Sequence labeling
  - MEMM, CRF, RNN[1, 2], etc.

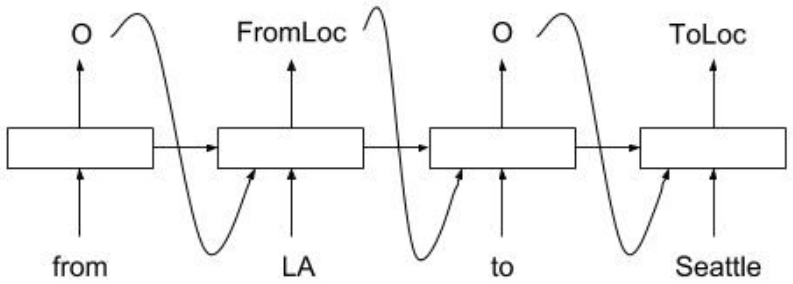


Fig. RNN slot filling model

[1] Mesnil, Grégoire, et al. "Using recurrent neural networks for slot filling in spoken language understanding." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015.
[2] Yao, Kaisheng, et al. "Spoken language understanding using long short-term memory neural networks." SLT, 2014.

Background

- **Joint** intent detection & slot filling
  - **Benefits:**
    - Simplifies the SLU systems
    - Improves the generalization performance of a task using the other related task
  - CNN[1], Recursive NN[2]

[1] Xu, Puyang, and Ruhi Sarikaya. "Convolutional neural network based triangular crf for joint intent detection and slot filling." ASRU, 2013.

[2] Guo, Daniel, et al. "Joint semantic utterance classification and slot filling with recursive neural networks." SLT 2014.

Background

- **Limitations** of previous joint SLU models:
  - Conditioned on the entire word sequence
  - Not suitable for ***online*** tasks

Motivation

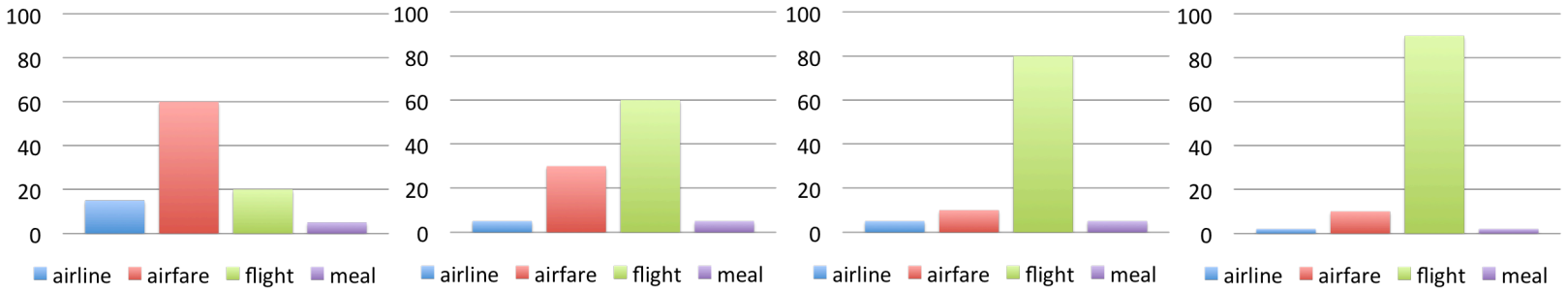
- Develop a model that performs online (incremental) SLU as the new word arrives.
- SLU results provide additional context for next word prediction in ASR online decoding.

➔ **Joint online (incremental) SLU + LM**

Query: First class flights from Phoenix to Seattle

First → class → flights → from → Phoenix → to → Seattle

Intent confidence scores



Next word probability from LM

Next Word	Prob
…	…
pittsburgh	1.1e-3
phone	0.7e-3
phoenix	1.4e-3
price	3.0e-3
…	…

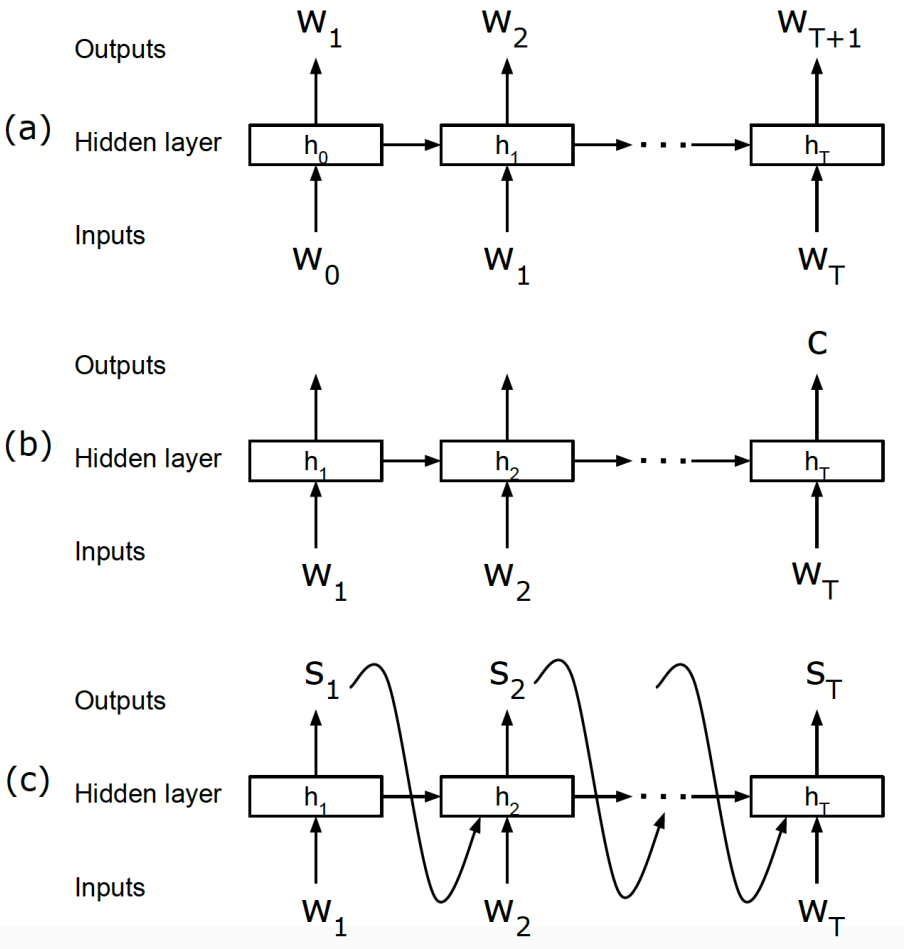
Prob
…
2.1e-3
0.7e-3
2.4e-3
1.8e-3
…

Prob
…
2.6e-3
0.7e-3
2.4e-3
1.2e-3
…

Outline

- Background & Motivation
- **Proposed Methods**
- Experiments & Results
- Conclusions

Independent task models

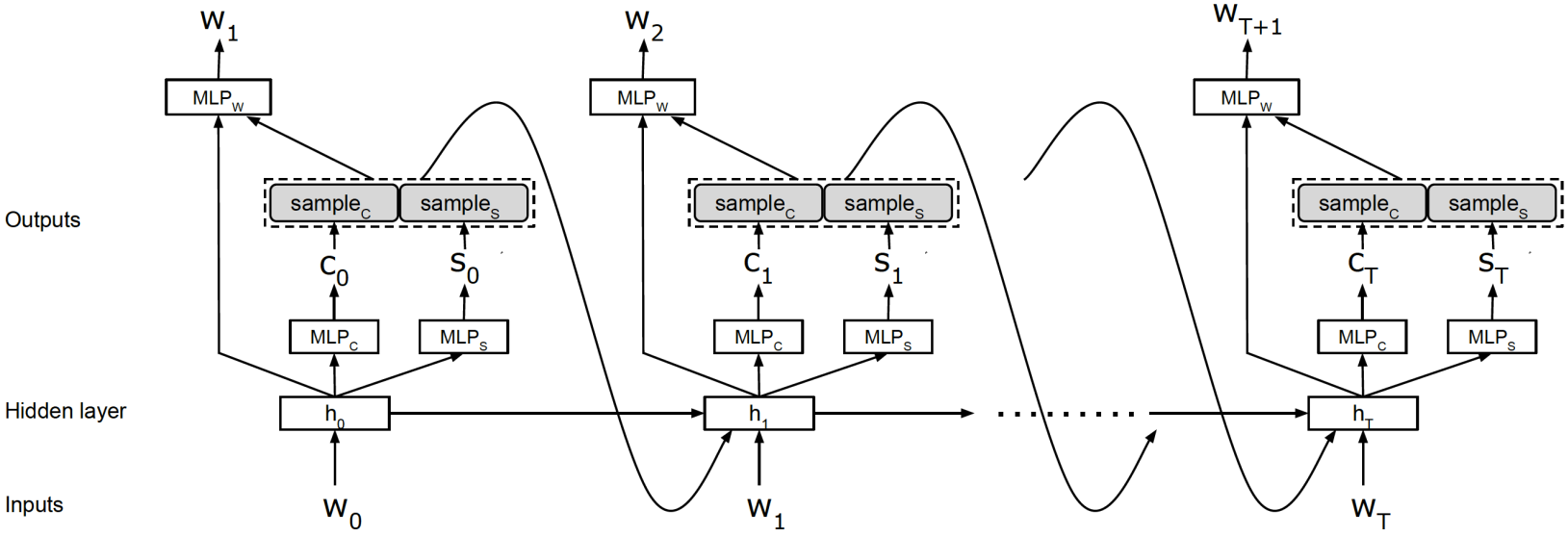


RNN Language Model

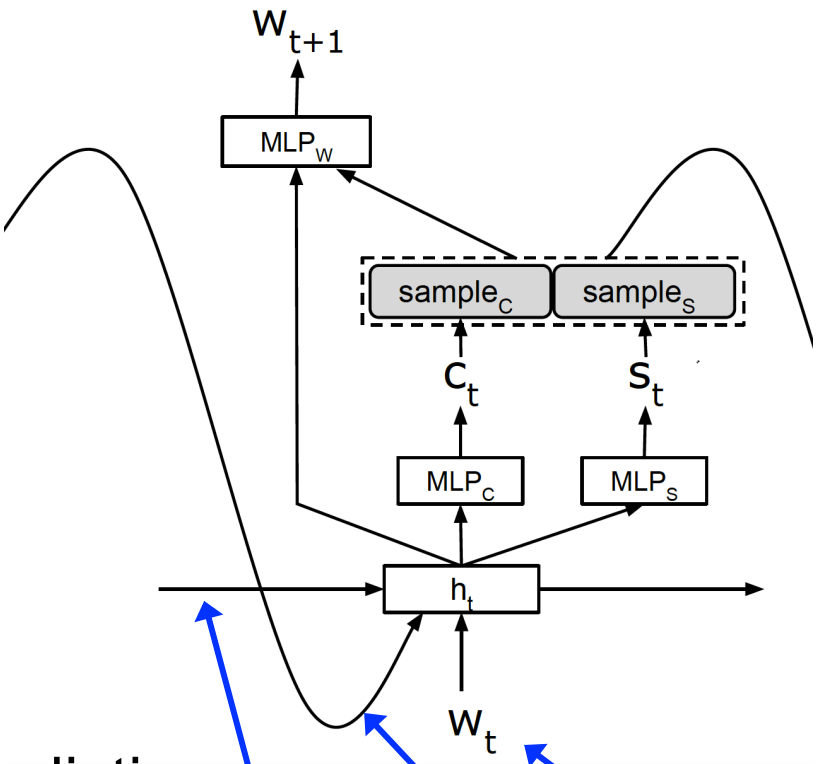
RNN Intent Detection Model

RNN Slot Filling Model

Joint model



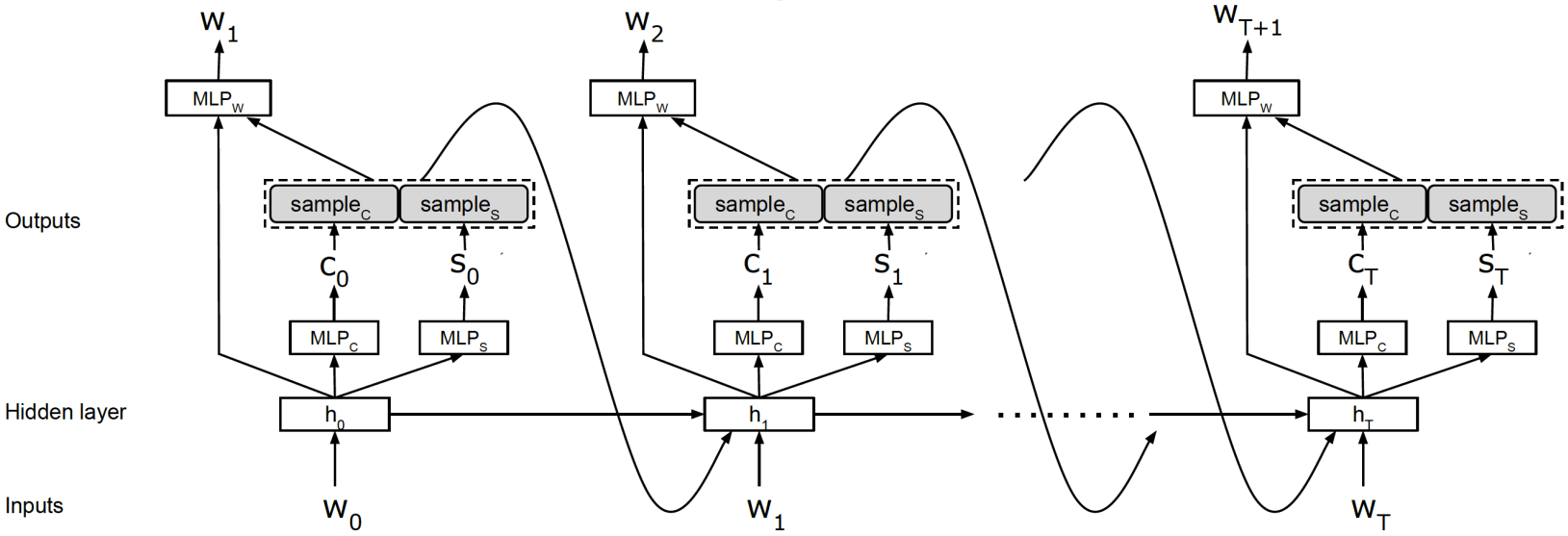
- Intent model: $P(c_T|\mathbf{w}) = P(c_T|w_{\leq T}, c_{<T}, s_{<T})$
- Slot filling model: $P(\mathbf{s}|\mathbf{w}) = P(s_0|w_0)\prod_{t=1}^{T} P(s_t|w_{\leq t}, c_{<t}, s_{<t})$
- Language model: $P(\mathbf{w}) = \prod_{t=0}^{T} P(w_{t+1}|w_{\leq t}, c_{\leq t}, s_{\leq t})$



- Next step prediction:

$$h_t = \text{LSTM}(h_{t-1}, [c_{t-1}, s_{t-1}, w_t])$$
$$P(c_t | w_{\leq t}, c_{<t}, s_{<t}) = \text{IntentDist}(h_t)$$
$$P(s_t | w_{\leq t}, c_{<t}, s_{<t}) = \text{SlotLabelDist}(h_t)$$
$$P(w_{t+1} | w_{\leq t}, c_{\leq t}, s_{\leq t}) = \text{WordDist}(h_t, c_t, s_t)$$

Joint model training



- Training: linear interpolation of the cost for each task:

Intent

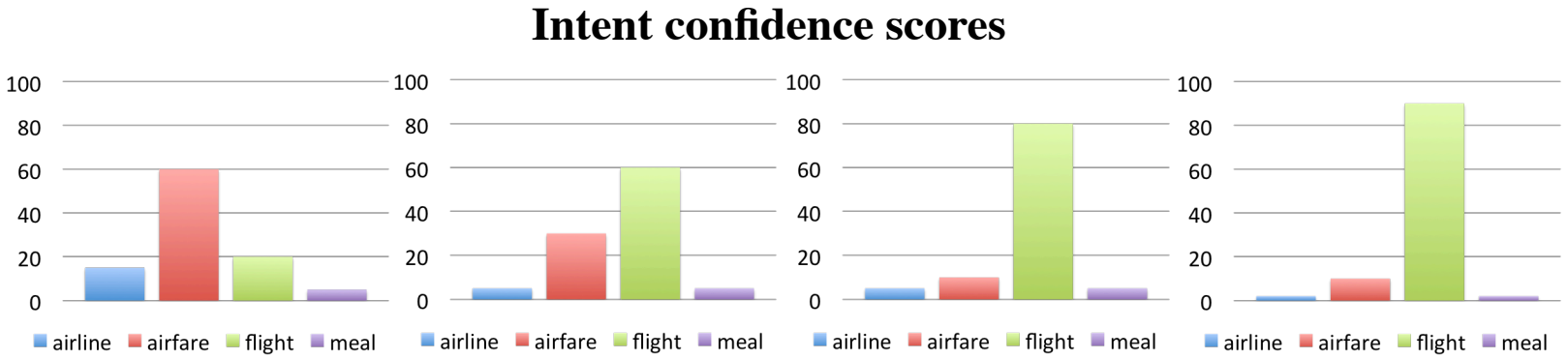
Slot filling

$$\max_{\theta} \sum_{t=0}^{T} \Big[\alpha_c \log P(c^* | w_{\leq t}, c_{<t}, s_{<t}; \theta) + \alpha_s \log P(s_t^* | w_{\leq t}, c_{<t}, s_{<t}; \theta) + \alpha_w \log P(w_{t+1} | w_{\leq t}, c_{\leq t}, s_{\leq t}; \theta) \Big] - \lambda R(\theta)$$

LM

Query: First class flights from Phoenix to Seattle

First → class → flights → from → Phoenix → to → Seattle



→ Intent estimation might be unstable at the beginning of the sequence

Adjusted / Scaled intent context

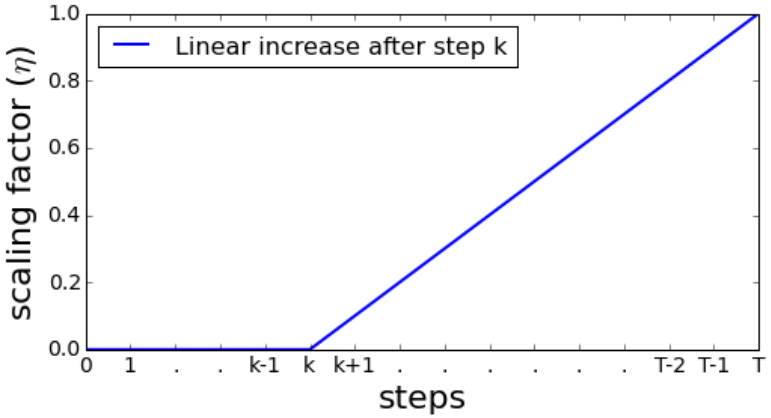
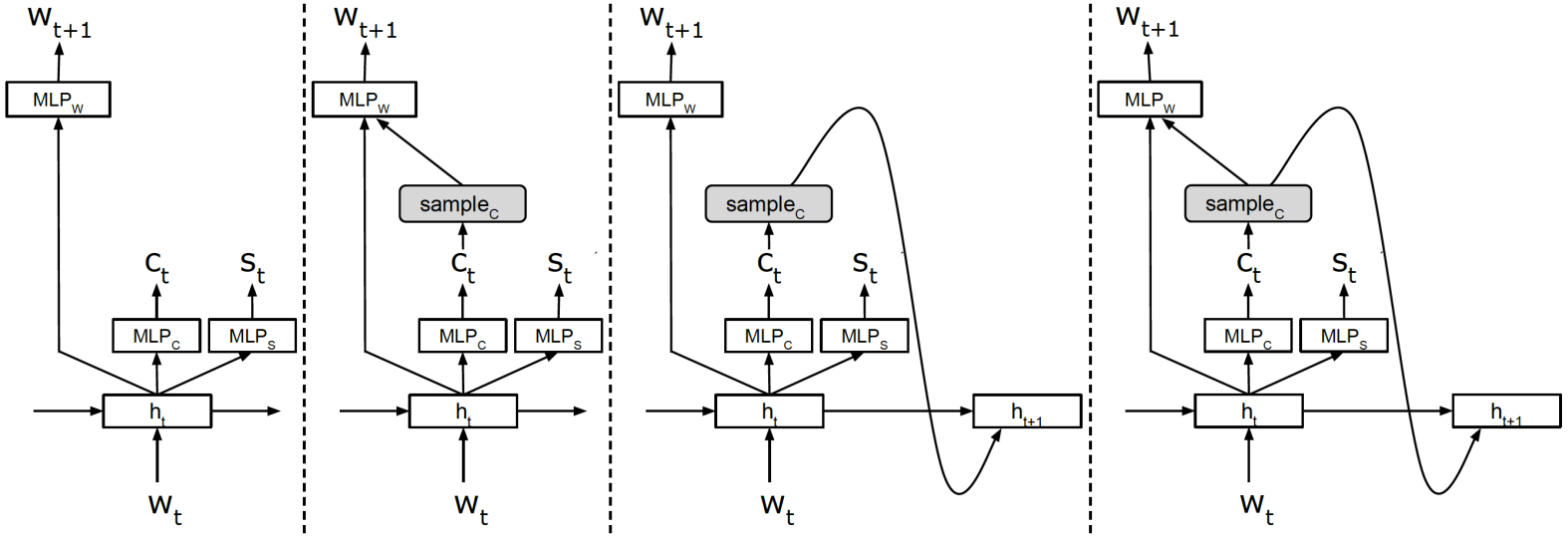


Fig. Schedule of increasing intent contribution to the context vector along with the growing input word sequence.

Joint model variations



(a) Basic joint model (b) Model with local context (c) Model with recurrent context (d) Model with local and recurrent context

Outline

- Background & Motivation
- Proposed Methods
- **Experiments & Results**
- Conclusions

Data set

- ATIS (Airline Travel Information System)
  - Intent
    - 18 intent classes
    - evaluated on classification error rate.
  - Slot Filling
    - 127 slot labels
    - evaluated on F1 score.

Experiments

- RNN model settings
  - LSTM Cell
  - Mini batch training
  - Adam optimization method
  - Dropout & L2 regularization

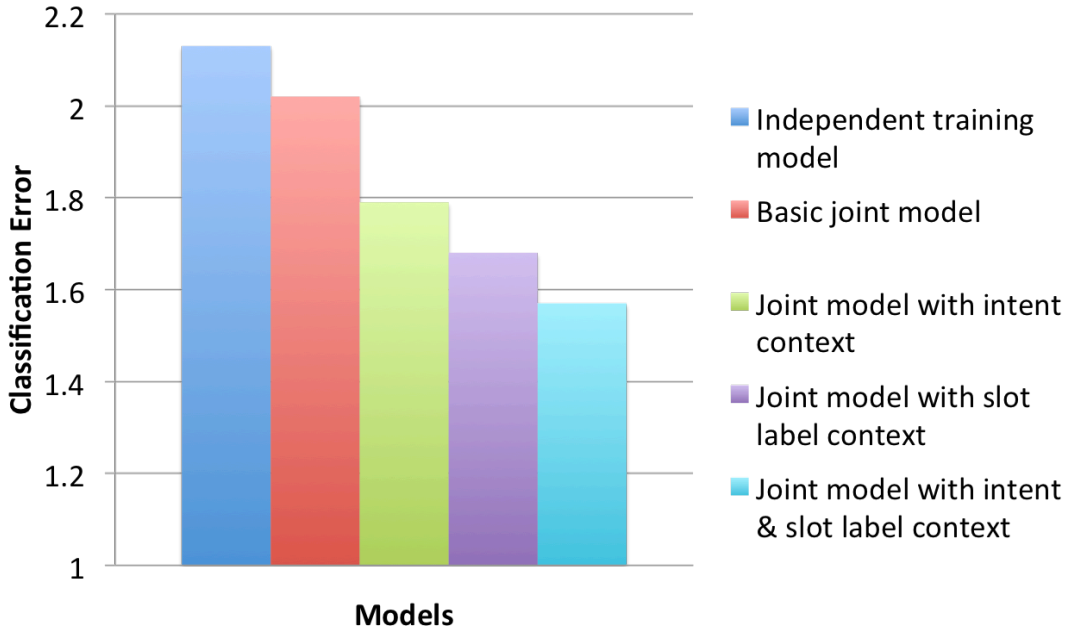
- ASR model settings
  - AM: LibriSpeech AM
  - LM: trained on ATIS corpus

Experiments

- Inputs:
  - True text input
  - Speech input with simulated noise
- Models:
  - Independent training model
  - Basic joint model
  - Joint model with intent context
  - Joint model with slot label context
  - Joint model with intent & slot label context
- Tasks:
  - Intent detection; Slot filling; Language modeling

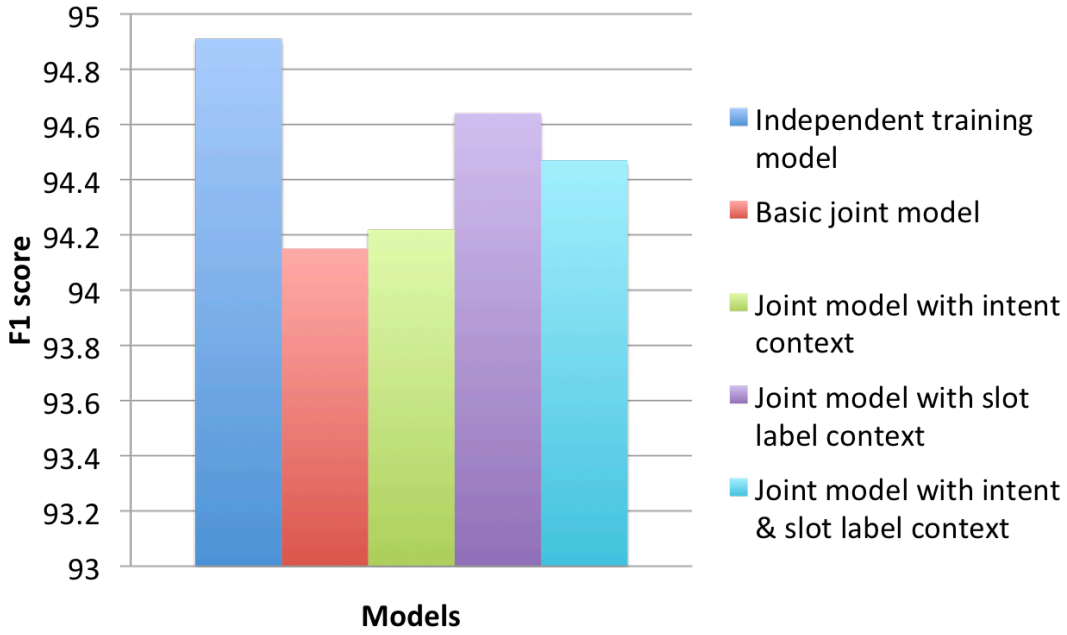
Experiment Results

- True text input – Intent detection
  - 0.56% absolute (26.3% relative) error reduction over independent training intent model



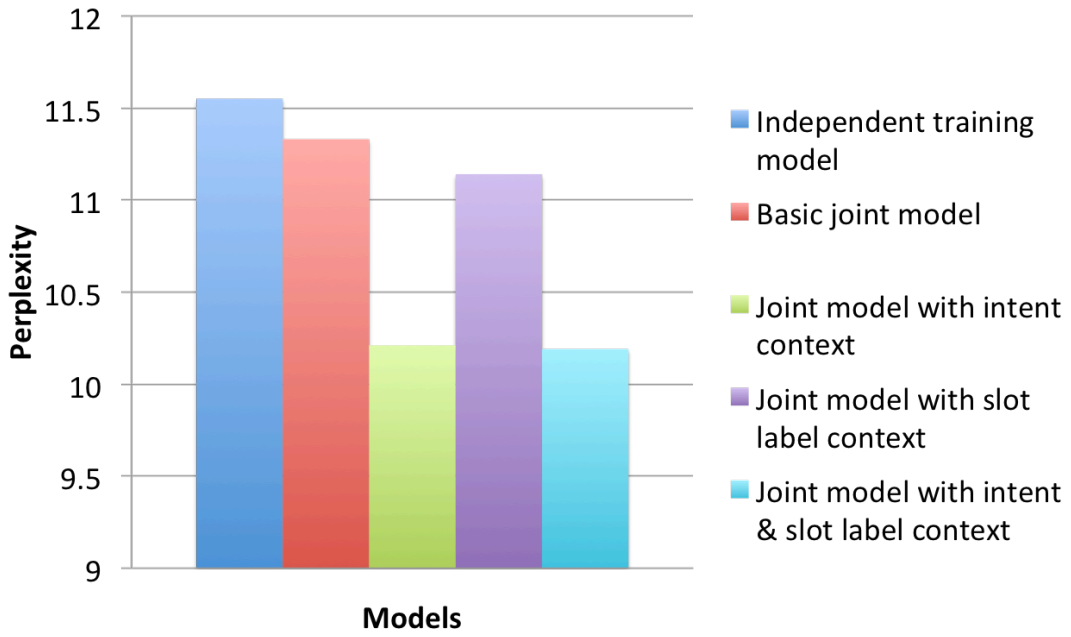
Experiment Results

- True text input – Slot filling
  - Slight degradation on slot filling F1 score comparing to independent training slot filling model.



Experiment Results

- True text input – Language modeling
  - 11.8% relative reduction on perplexity comparing to the independent training language model



Experiment Results

- Noisy speech input & ASR output

ASR Settings	WER	Intent Error	F1 Score
Decoding: LibriSpeech AM & 2-gram LM	14.51	4.63	84.46
Decoding: LibriSpeech AM & 2-gram LM Rescoring: 5-gram LM	13.66	5.02	85.08
Decoding: LibriSpeech AM & 2-gram LM Rescoring: **Independent training** RNNLM	12.95	4.63	85.43
Decoding: LibriSpeech AM & 2-gram LM Rescoring: **Joint training** RNNLM	**12.59**	**4.44**	**86.87**

Outline

- Background & Motivation
- Proposed Methods
- Experiments & Results
- **Conclusions**

Conclusions

- We proposed an RNN model for joint online (incremental) SLU and LM.
- Improved performance on intent detection and LM, with slight degradation on slot filling.
- Consistent performance gain over independent training model with noisy speech input.

Thanks & Questions