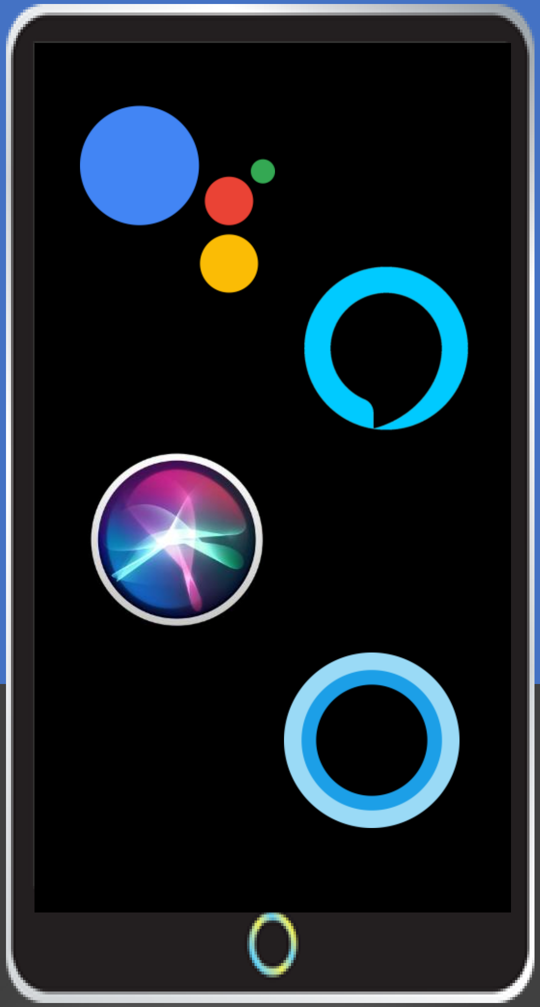
Robust and *Scalable* Conversational AI



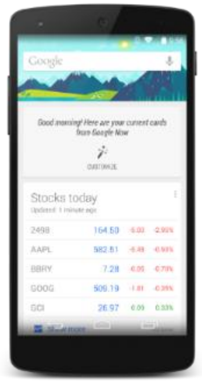
$\stackrel{?}{=}$

Yun-Nung (Vivian) Chen
Computer Science & Information Engineering
National Taiwan University

16th WORKSHOP ON SPOKEN DIALOGUE SYSTEMS FOR PHDS, POSTDOCS & NEW RESEARCHERS (YRRSDS 2020)

Language Empowering Intelligent Assistants

Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

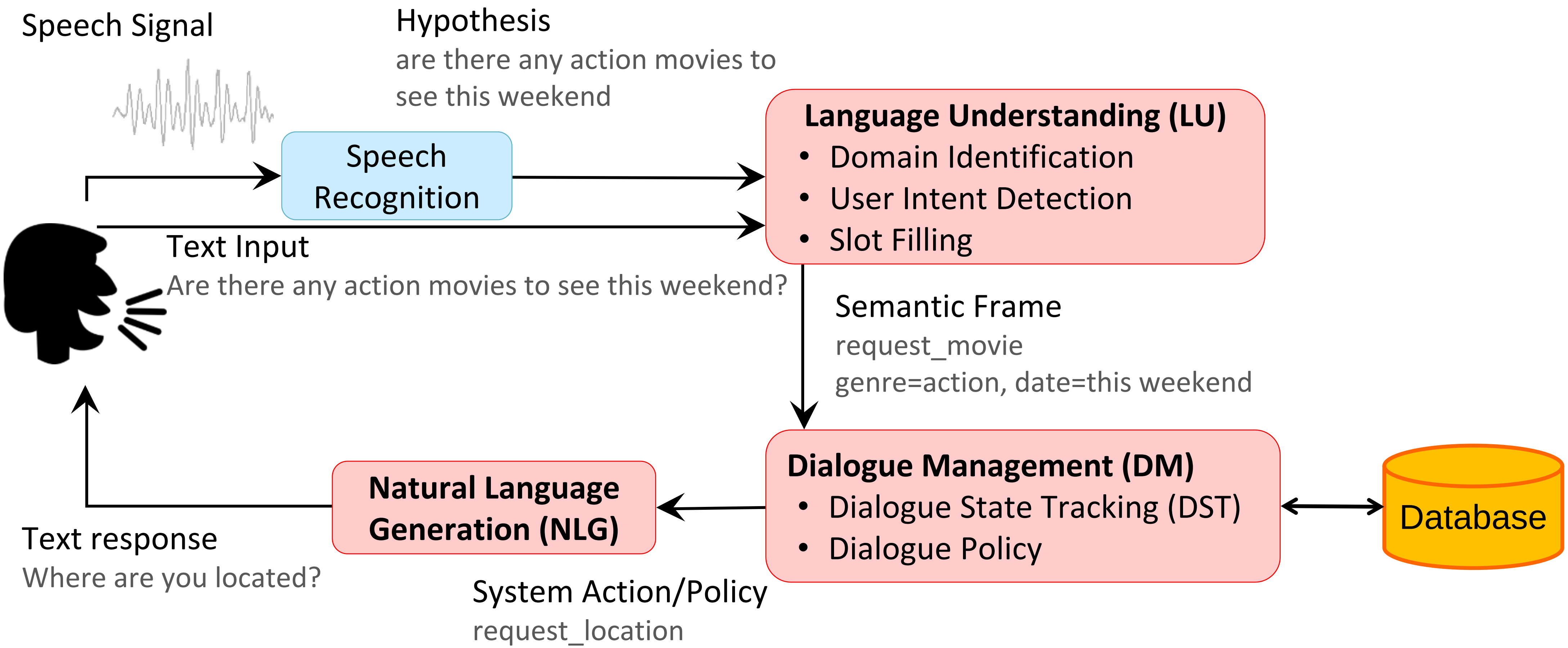
Amazon Alexa/Echo (2014)

Google Home (2016)

Apple HomePod (2017)

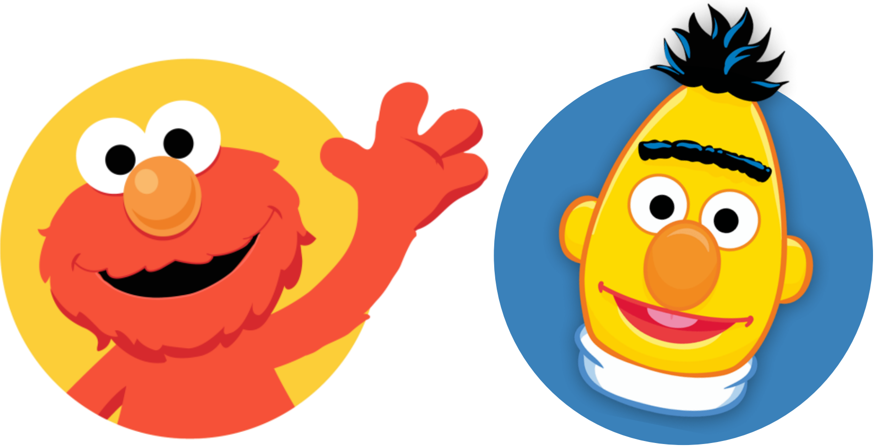
Facebook Portal (2019)

Task-Oriented Dialogue Systems (Young, 2000)

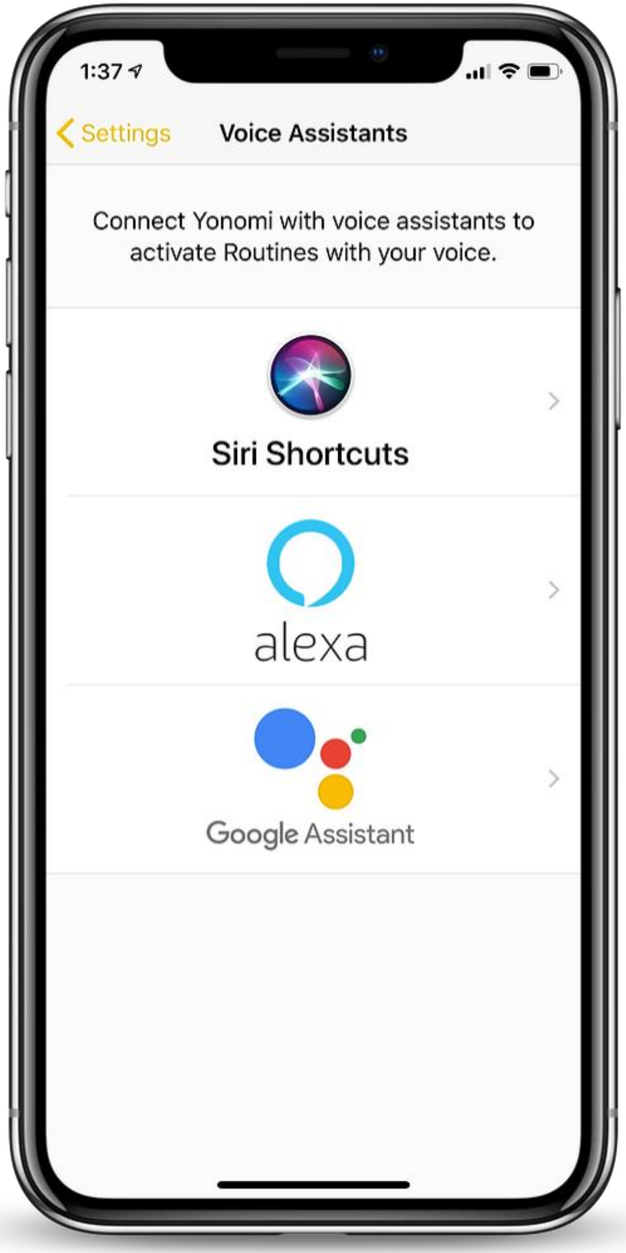


Recent Advances in NLP

- Contextual Embeddings (ELMo & BERT)
  - Boost many understanding performance with pre-trained language models



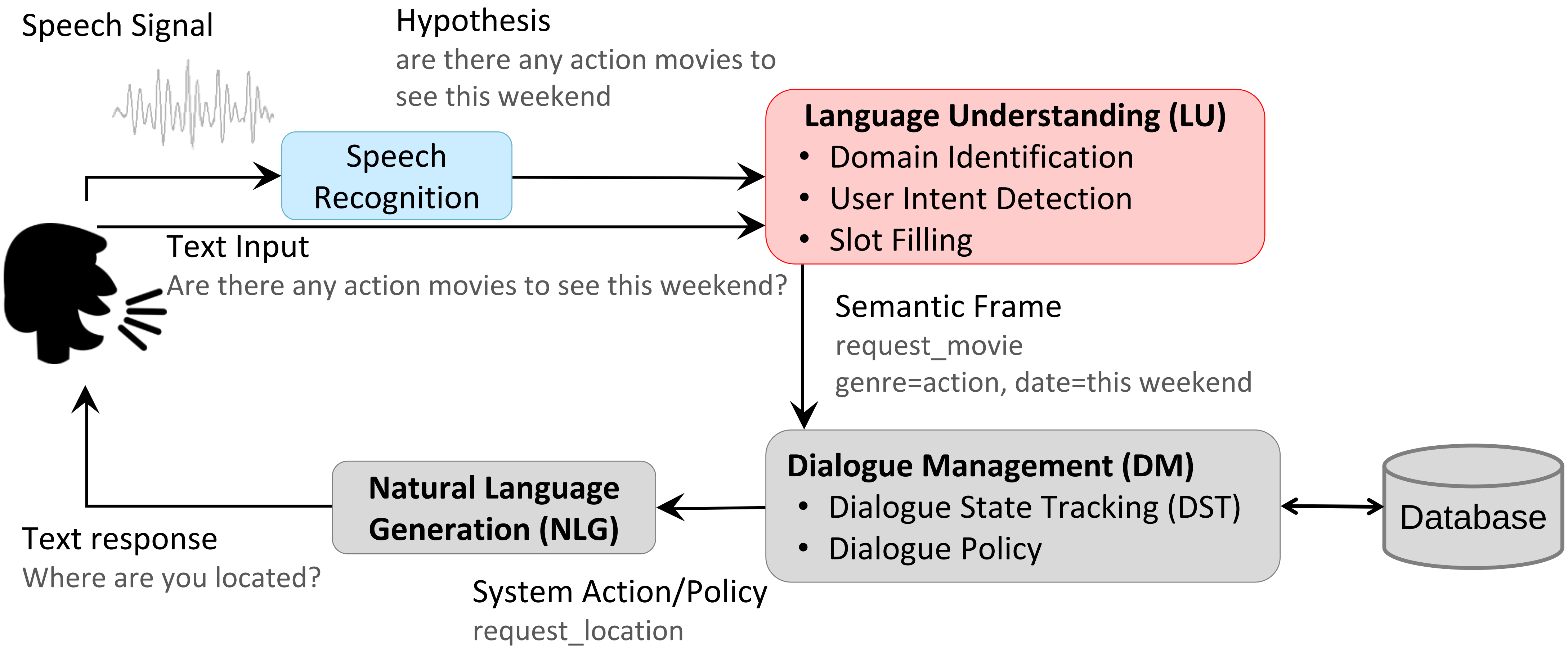
?




Call me ASAP


Lift all lights to Morocco
List all flights tomorrow
Matt, what can I assist you with?

Task-Oriented Dialogue Systems (Young, 2000)



Mismatch between Written and Spoken Languages

Training

- Written language

Testing

- Spoken language
  - Include recognition errors

- Goal: ASR-Robust Contextualized Embeddings
  - ✓ learning *spoken* contextualized word embeddings
  - ✓ better performance on *spoken* language understanding tasks

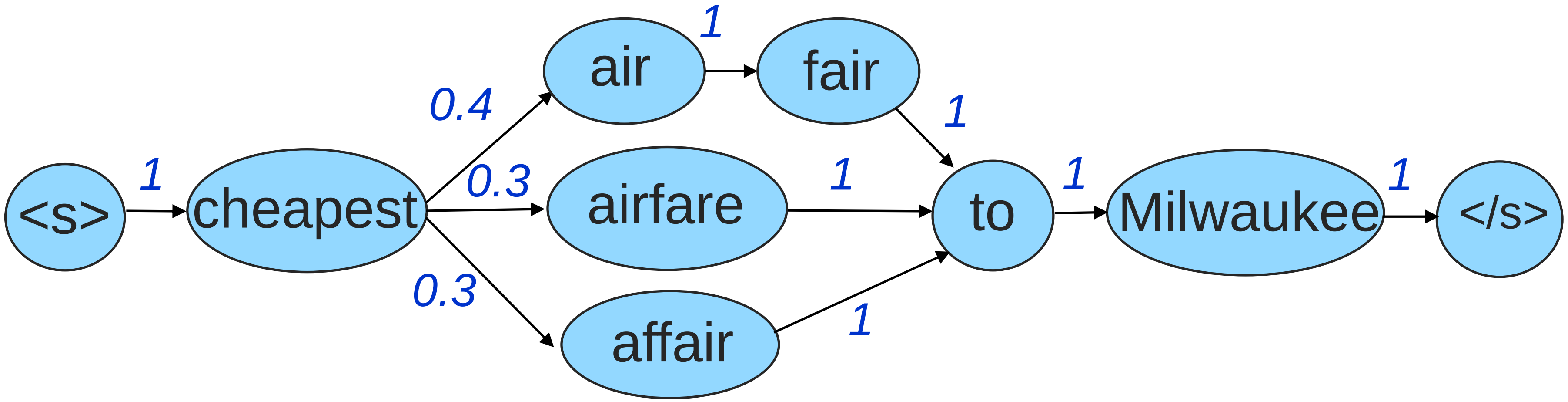
Solution: LatticeLM

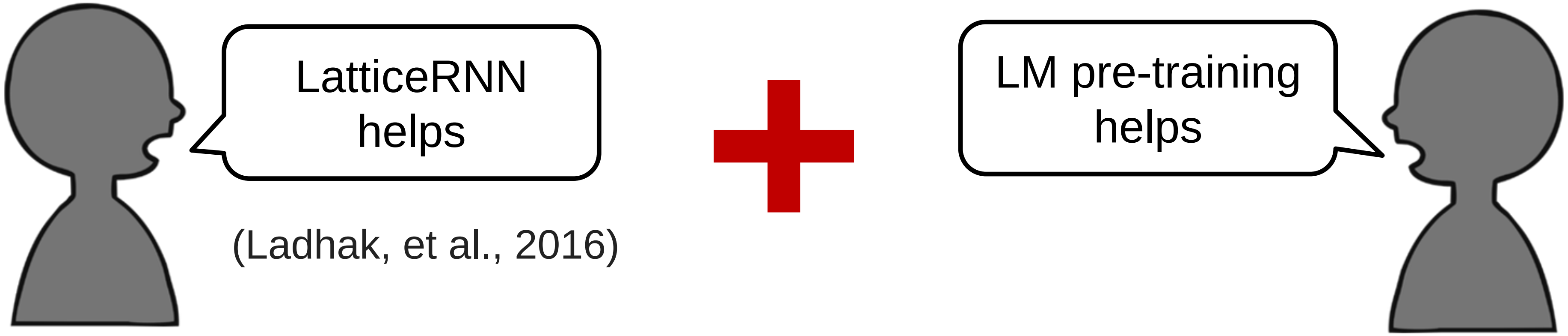
(Huang & Chen, ACL 2020)

Chao-Wei Huang and Yun-Nung Chen, "Learning Spoken Language Representations with Neural Lattice Language Modeling," in *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

ASR Lattices for Preserving Uncertainty

- Idea: lattices may include correct words



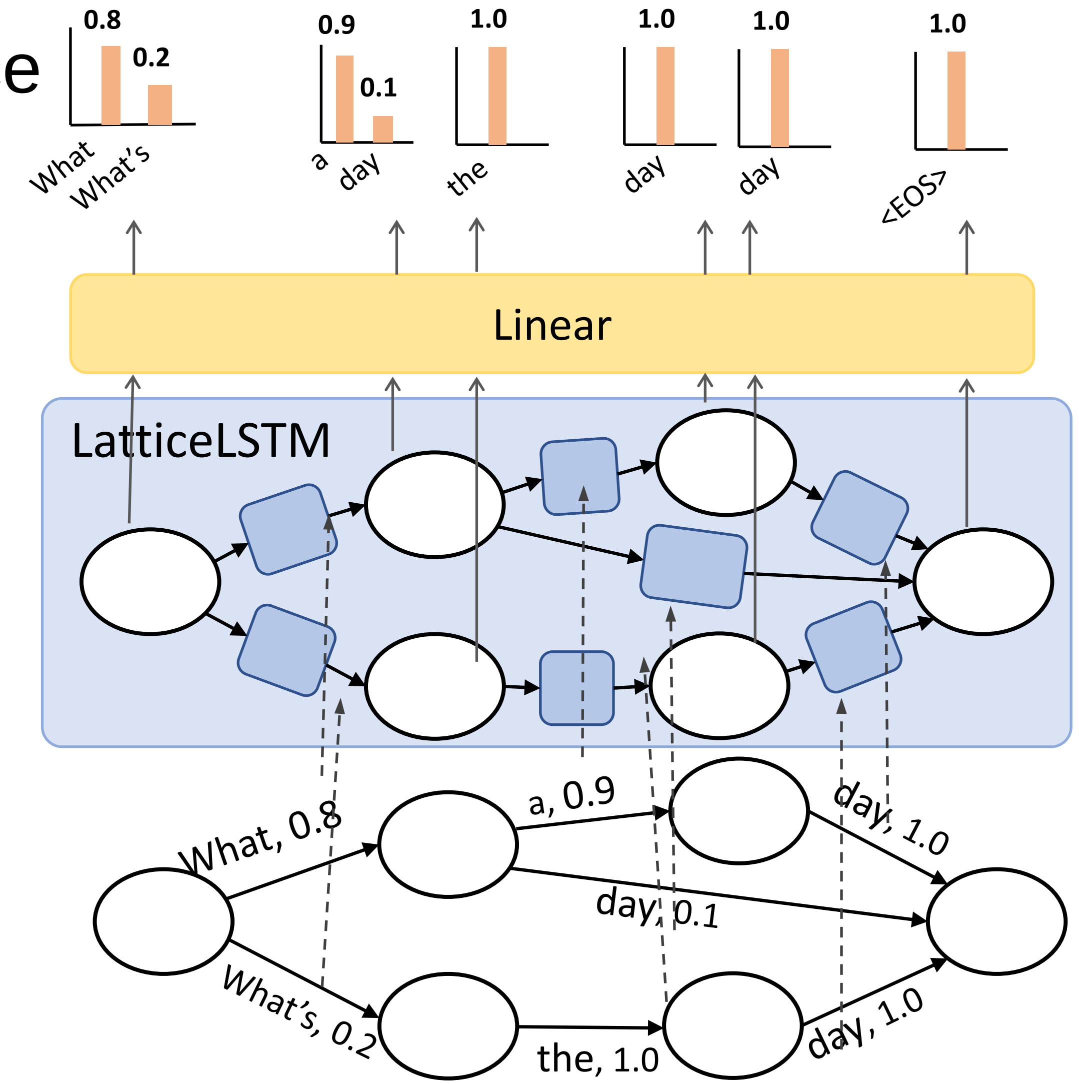


Chao-Wei Huang and Yun-Nung Chen, “Learning Spoken Language Representations with Neural Lattice Language Modeling,” in *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Lattice Language Modeling

1) LatticeLSTM encodes nodes of a lattice
2) The goal is to predict the outgoing transitions (words) given a node's representation

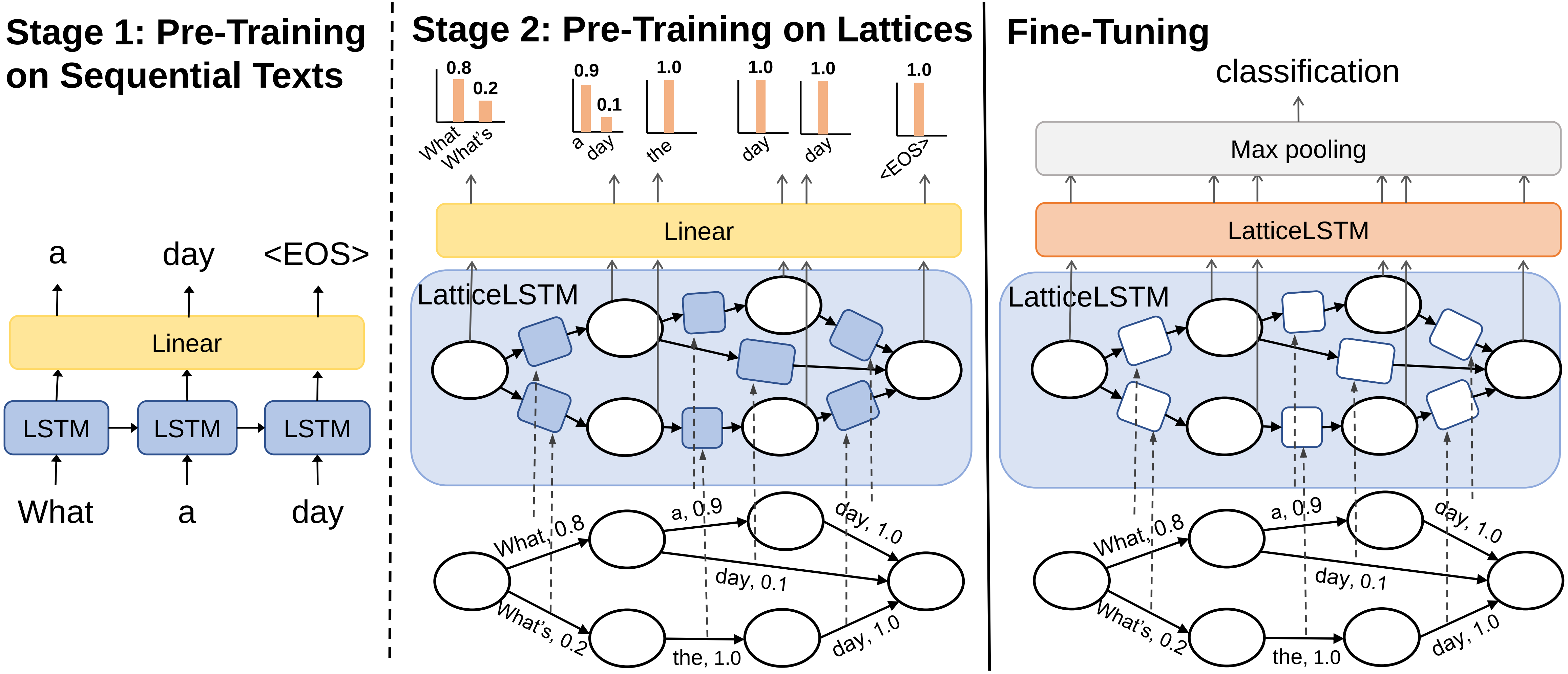
- The one-hypothesis lattice reduces to normal language modeling



Issue: LatticeLSTM runs prohibitively slow

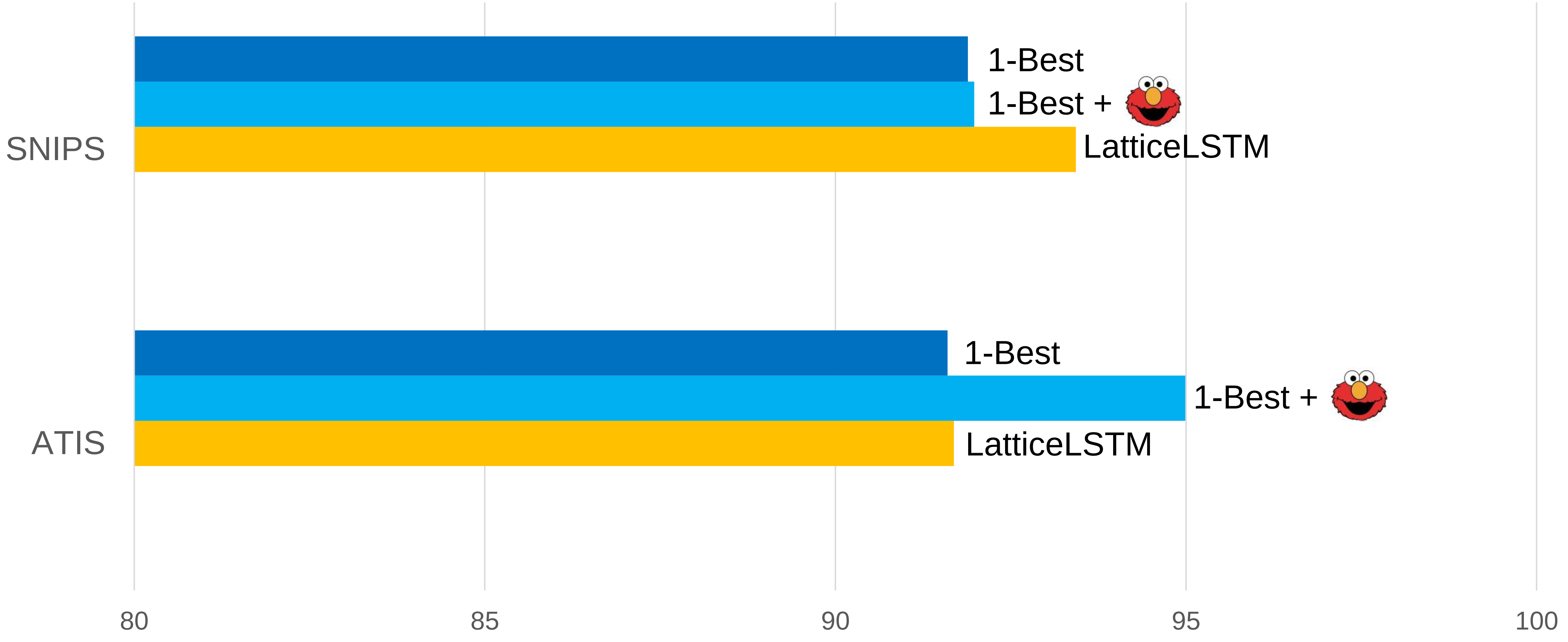
Chao-Wei Huang and Yun-Nung Chen, "Learning Spoken Language Representations with Neural Lattice Language Modeling," in *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Efficient Two-Stage Pre-Training



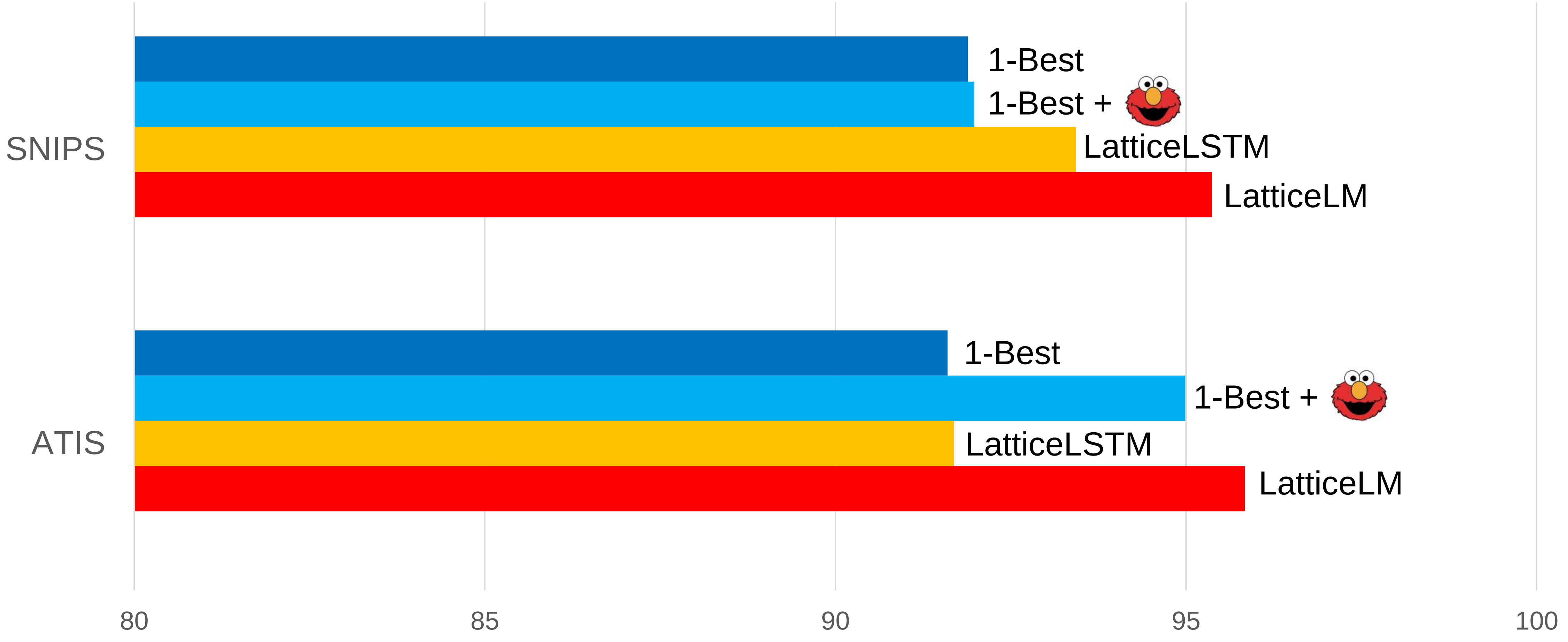
Spoken Language Understanding Results

- Intent Prediction
  - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



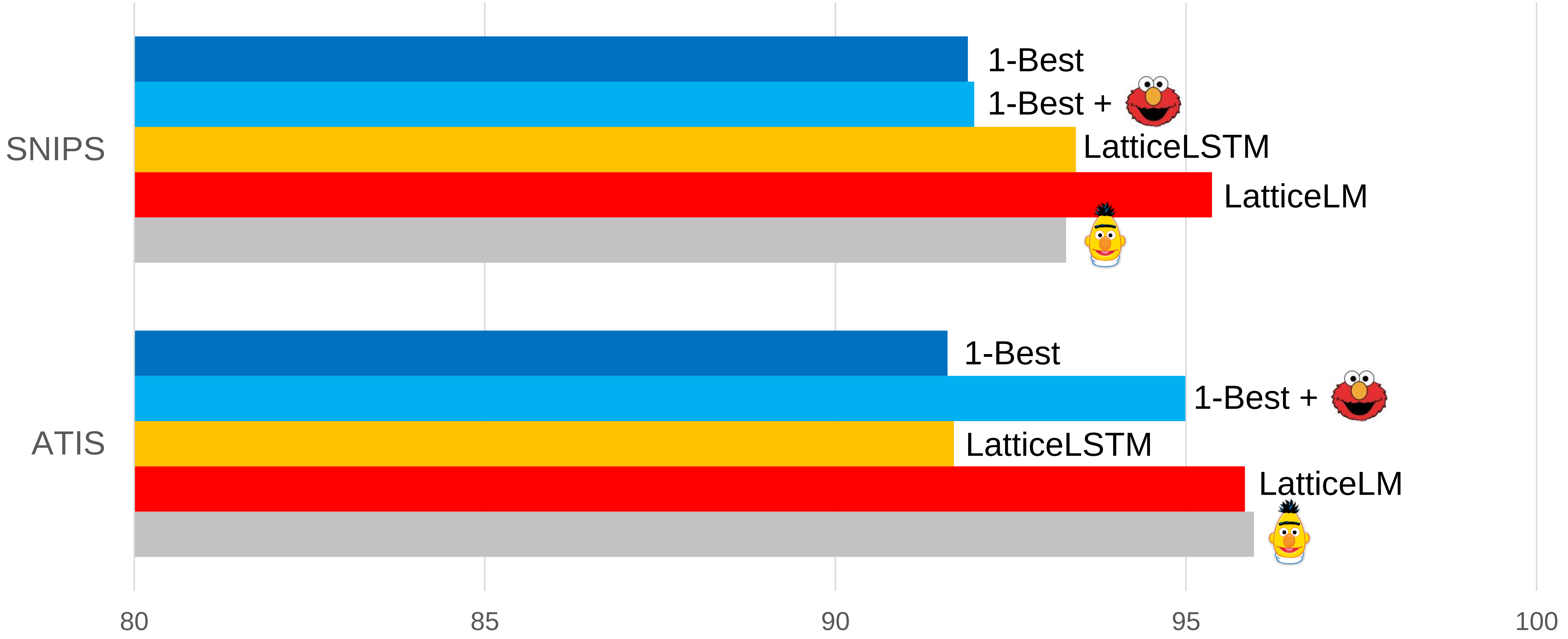
Spoken Language Understanding Results

- Intent Prediction
  - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



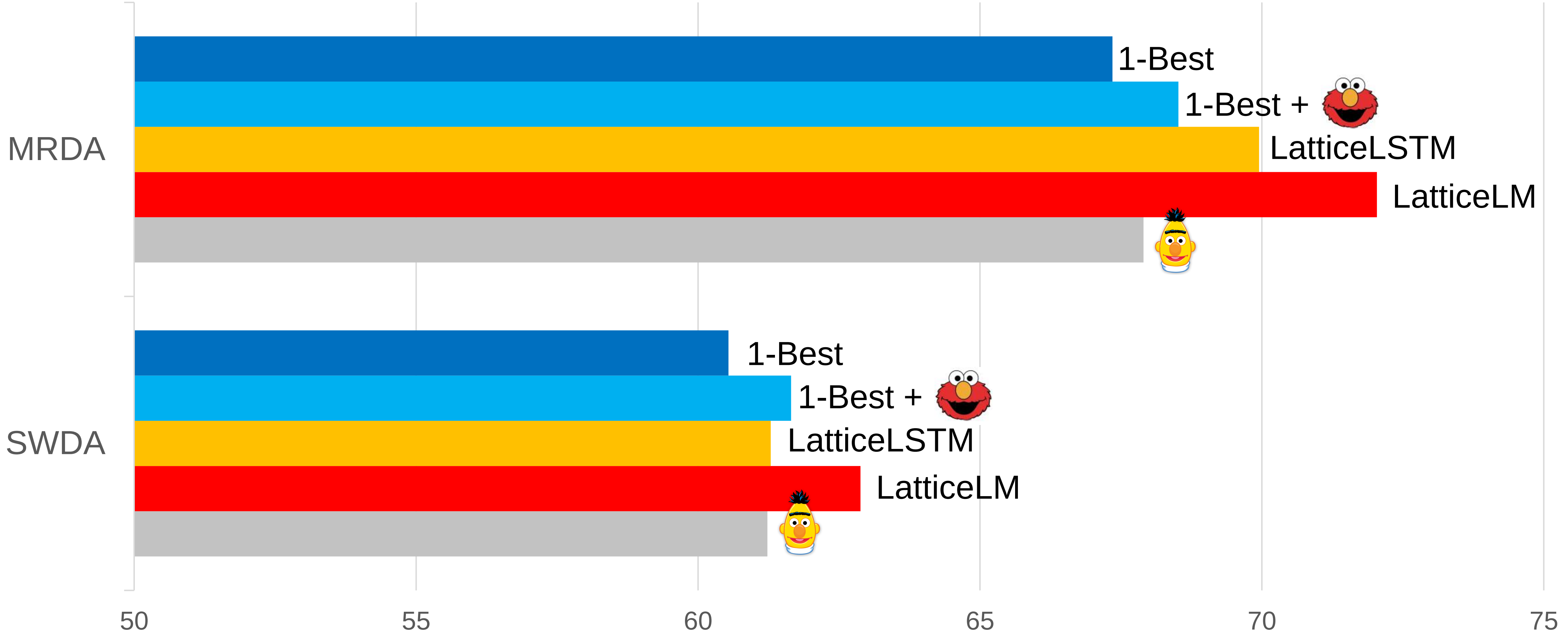
Spoken Language Understanding Results

- Intent Prediction
    - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



Spoken Language Understanding Results

- Dialogue Act Prediction
  - Word Error Rate: 32.0% (MRDA); 28.4% (SWDA)



What if we do not have ASR lattices?

Solution:
Learning ASR-Robust Embeddings

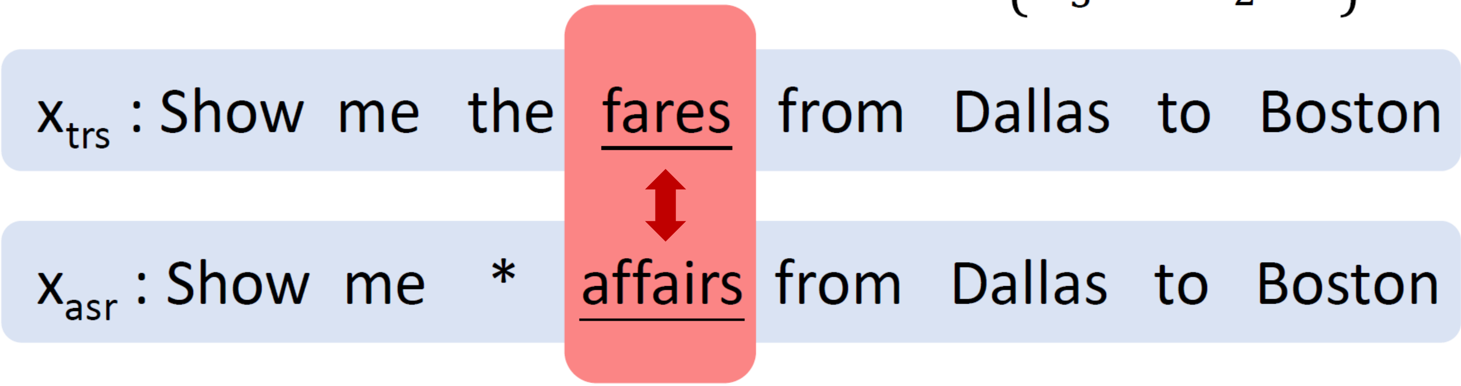
(Huang & Chen, ICASSP 2020)

Chao-Wei Huang and Yun-Nung Chen, “Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding,” in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

ASR-Robust Contextualized Embeddings

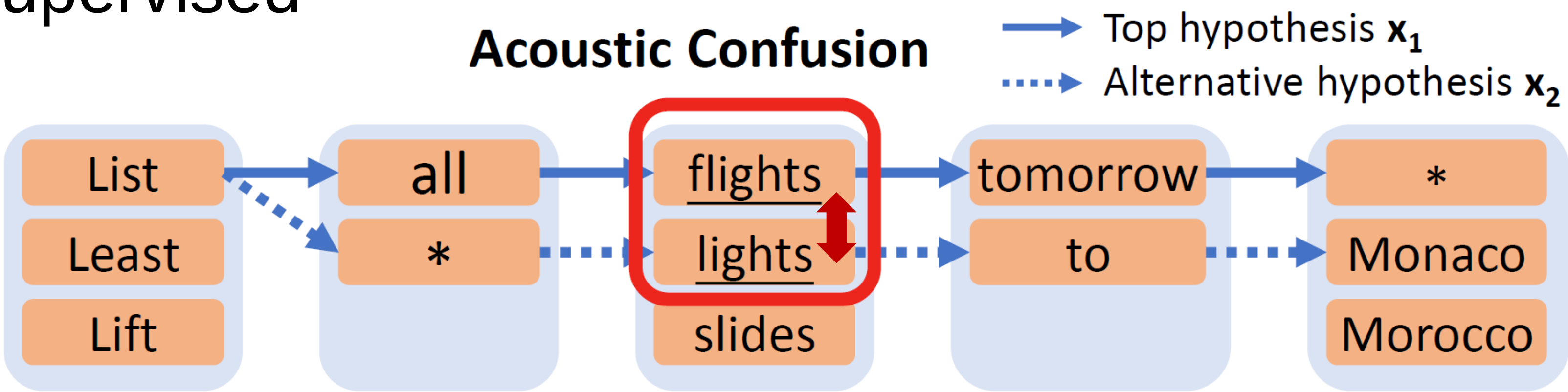
- Confusion-Aware Fine-Tuning
  - Supervised

Acoustic Confusion $C = \{w_3^{x_{trs}}, w_2^{x_{asr}}\}$



  - Unsupervised

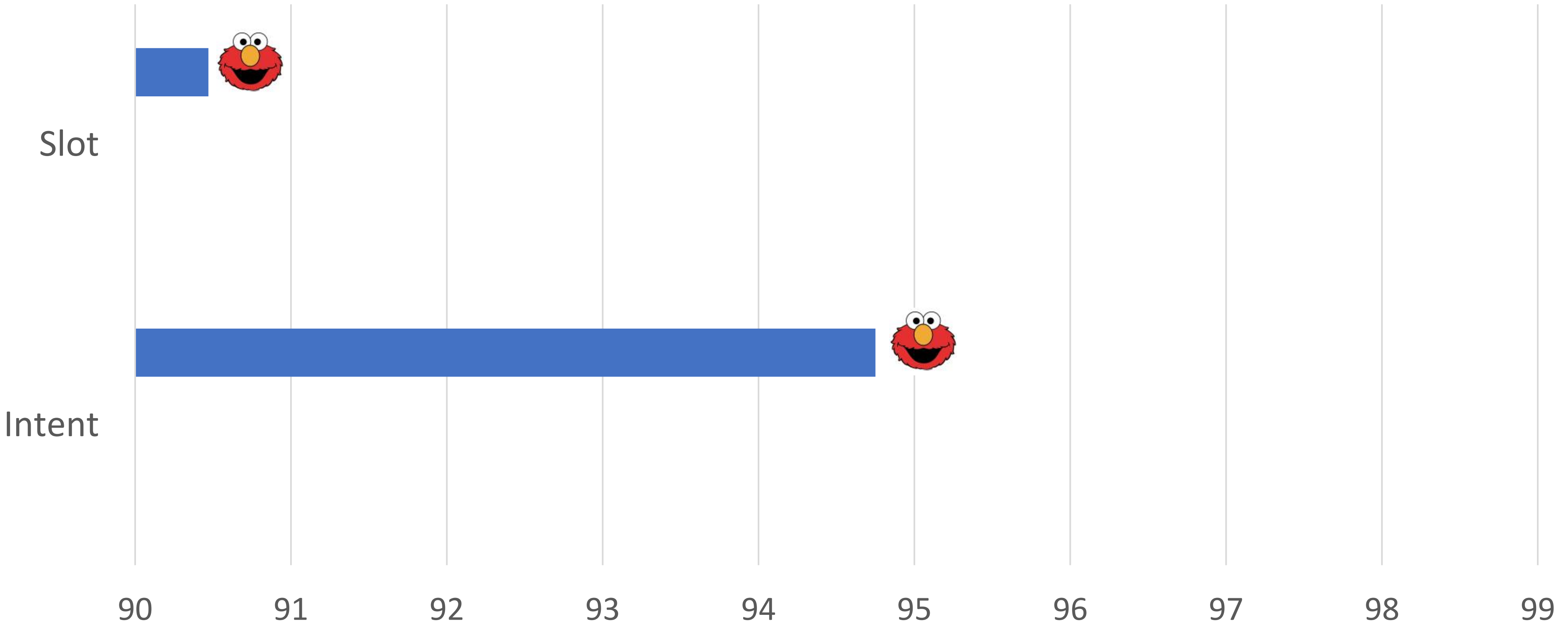
Acoustic Confusion



Chao-Wei Huang and Yun-Nung Chen, "Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding," in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Spoken Language Understanding Results

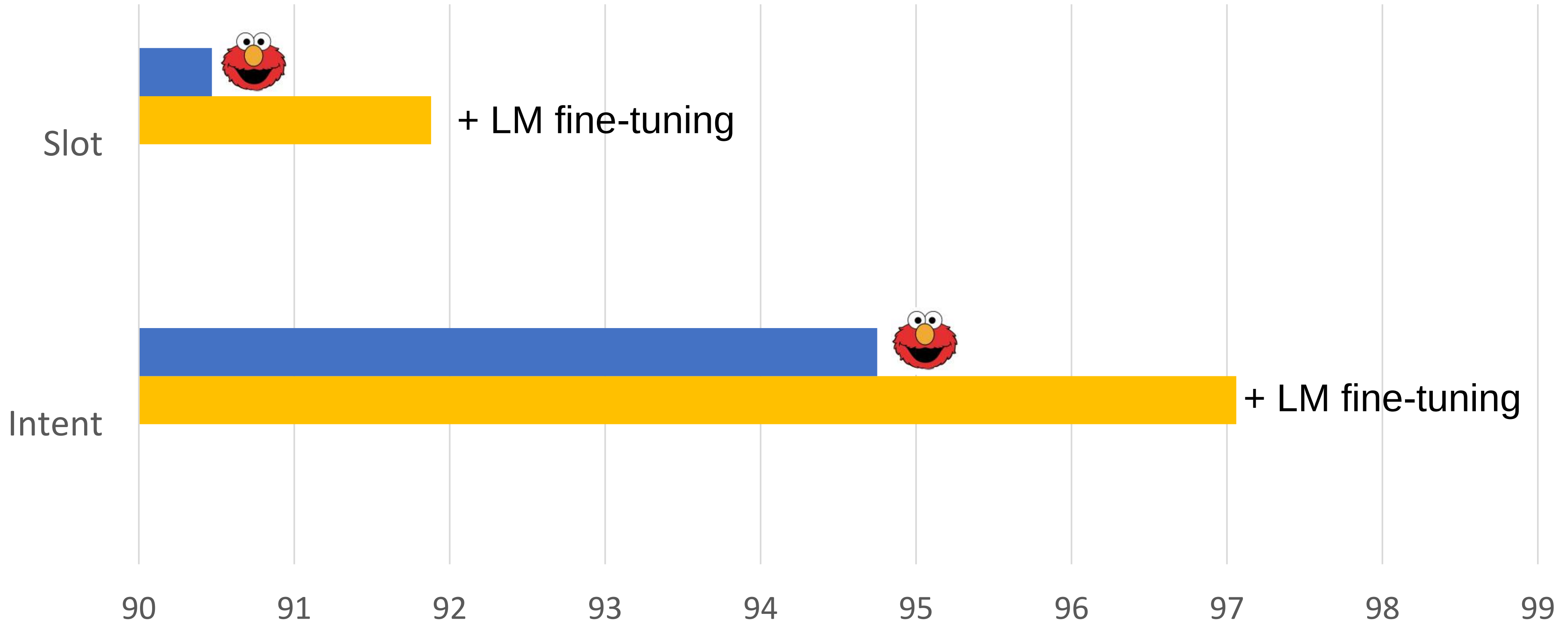
- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%



Chao-Wei Huang and Yun-Nung Chen, "Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding," in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Spoken Language Understanding Results

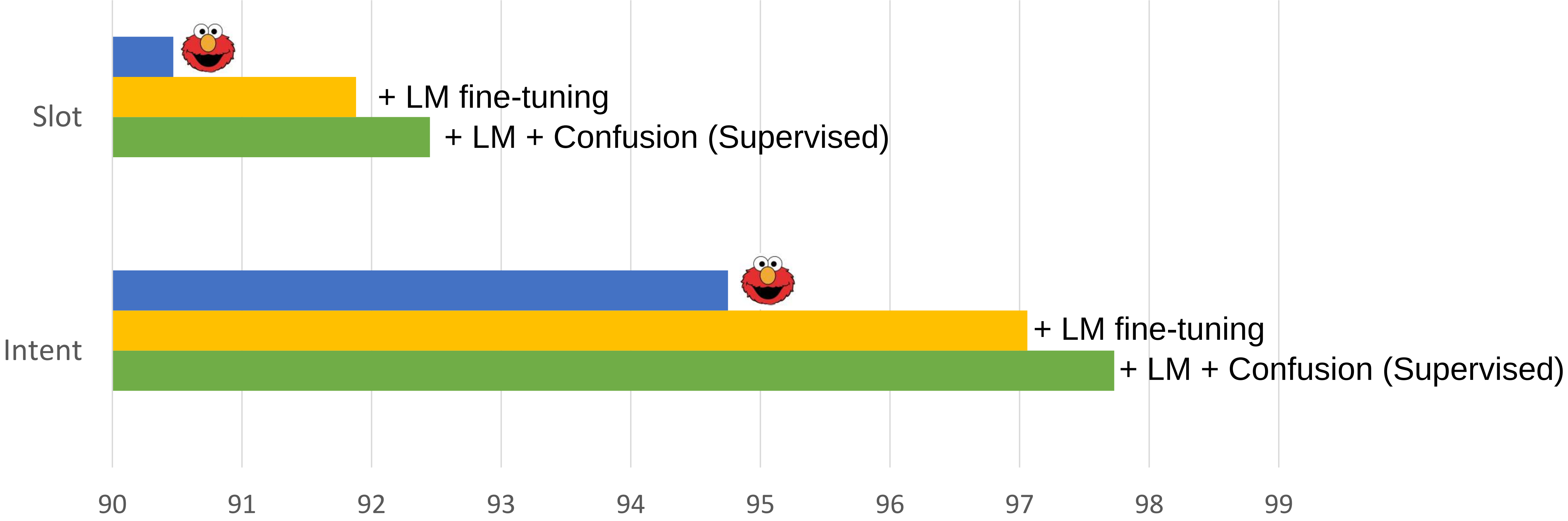
- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%



Chao-Wei Huang and Yun-Nung Chen, "Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding," in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Spoken Language Understanding Results

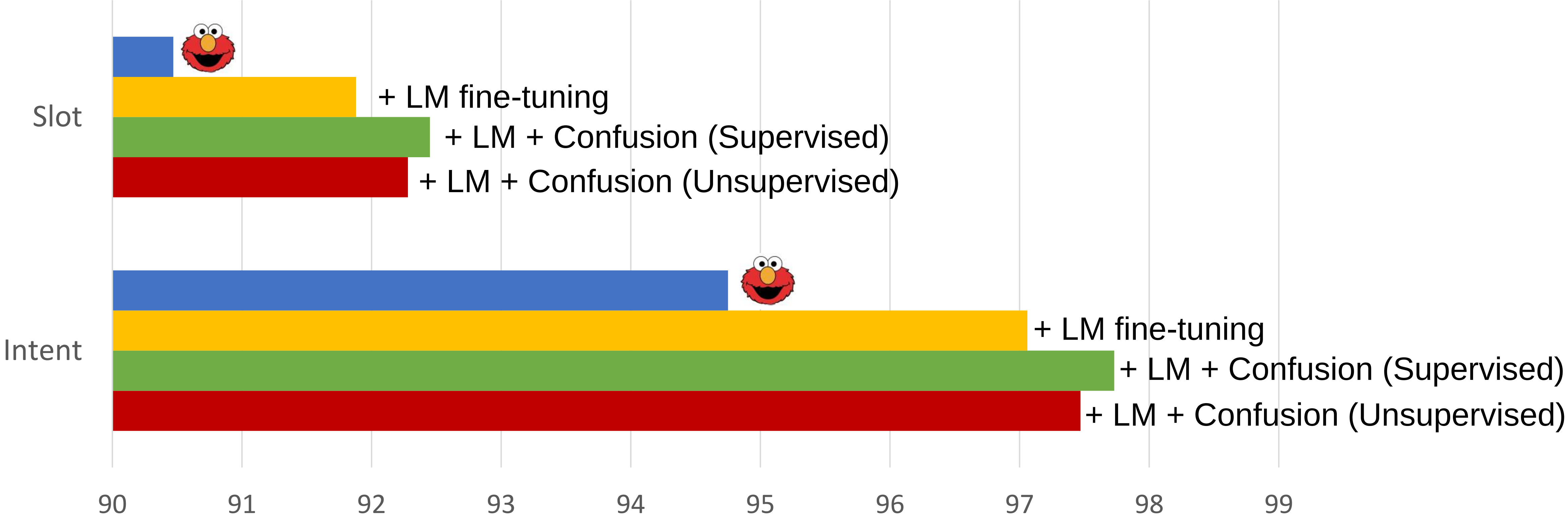
- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%



Chao-Wei Huang and Yun-Nung Chen, “Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding,” in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

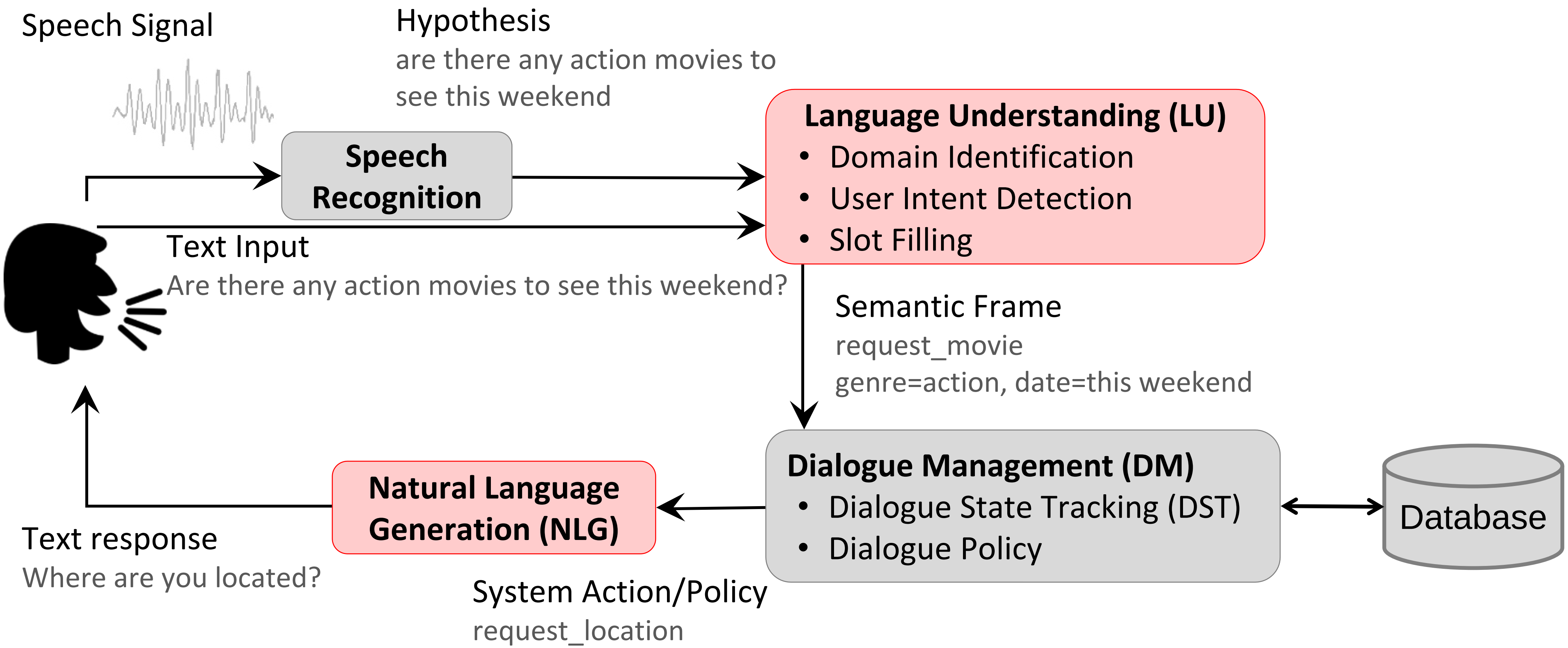
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%



Chao-Wei Huang and Yun-Nung Chen, “Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding,” in *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Task-Oriented Dialogue Systems (Young, 2000)



Natural Language Understanding (NLU)

- Parse **natural language** into **structured semantics**

NLU

Natural Language
McDonald's is a cheap restaurant nearby the station.

Semantic Frame
RESTAURANT="McDonald's"
PRICE="cheap"
LOCATION= "nearby the station"

Natural Language Generation (NLG)

- Construct **natural language** based on **structured semantics**

Natural Language
McDonald's is a cheap restaurant nearby the station.

Semantic Frame
RESTAURANT="McDonald's"
PRICE="cheap"
LOCATION= "nearby the station"

NLG

Duality between NLU and NLG

NLU

Natural Language
McDonald's is a cheap restaurant nearby the station.

Semantic Frame
RESTAURANT="McDonald's"
PRICE="cheap"
LOCATION= "nearby the station"

NLG

How can we leverage this dual relationship?

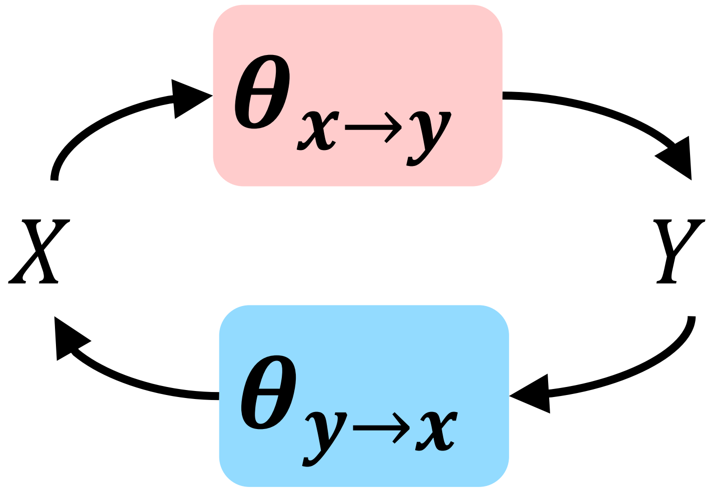
Solution:
Dual Supervised Learning for NLU & NLG

(Su et al., ACL 2019)

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, "Dual Supervised Learning for Natural Language Understanding and Generation," in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

DSL: Dual Supervised Learning (Xia et al., 2017)

- Proposed for machine translation
- Consider two domains X and Y, and two tasks $X \rightarrow Y$ and $Y \rightarrow X$



We have $P(x,y) = P(x \mid y)P(y) = P(y \mid x)P(x)$

Ideally $P(x,y) = P(x \mid y; \boldsymbol{\theta}_{y\rightarrow x})P(y) = P(y \mid x; \boldsymbol{\theta}_{x\rightarrow y})P(x)$

Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., & Liu, T. Y., "Dual supervised learning," in *Proc. of the 34th International Conference on Machine Learning*, 2017.

Dual Supervised Learning

- Exploit the duality by forcing models to follow the probabilistic constraint $P(x \mid y; \boldsymbol{\theta}_{\boldsymbol{y}\to\boldsymbol{x}})P(y) = P(y \mid x; \boldsymbol{\theta}_{\boldsymbol{x}\to\boldsymbol{y}})P(x)$

Objective function

$$\begin{cases} \min_{\theta_{x\to y}} \mathbb{E}\left[l_1\left(f\left(x;\theta_{x\to y}\right),y\right)\right] + \lambda_{x\to y}\, l_{duality} \\ \min_{\theta_{y\to x}} \mathbb{E}\left[l_2\left(g\left(y;\theta_{y\to x}\right),x\right)\right] + \lambda_{y\to x}\, l_{duality} \end{cases}$$

$$l_{duality} = \left(\boxed{\log \hat{P}(x)} + \log P(y \mid x; \theta_{x\to y}) - \boxed{\log \hat{P}(y)} - \log P(x \mid y; \theta_{y\to x})\right)^2$$

How to model the marginal distributions of X and Y?

Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., & Liu, T. Y., “Dual supervised learning,” in *Proc. of the 34th International Conference on Machine Learning*, 2017.

Dual Supervised Learning

- Let's go back to NLU and NLG

NLU

NLG

Natural Language

X

McDonald's is a cheap restaurant nearby the station.

$\log \hat{P}(x)$

Semantic Frame

Y

RESTAURANT="McDonald's"
PRICE="cheap"
LOCATION= "nearby the station"

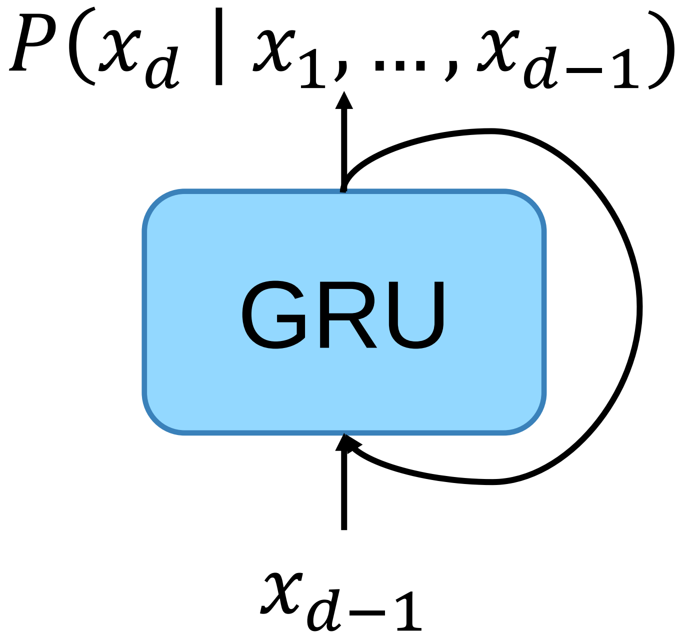
$\log \hat{P}(y)$

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, "Dual Supervised Learning for Natural Language Understanding and Generation," in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL),* 2019.

Natural Language $\log \hat{P}(x)$

- Language modeling

$$p(x) = \prod_{d}^{D} p(x_d \mid x_1, \ldots, x_{d-1})$$



Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, "Dual Supervised Learning for Natural Language Understanding and Generation," in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL),* 2019.

Semantic Frame $\log \hat{P}(y)$

- We treat NLU as a multi-label classification problem
- Each label is a slot-value pair

RESTAURANT=“McDonald’s”
PRICE=“cheap”
LOCATION= “nearby the station”

0
1
.
.
.
0
1

How to model the marginal distributions of y?

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, “Dual Supervised Learning for Natural Language Understanding and Generation,” in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL),* 2019.

Semantic Frame $\log \hat{P}(y)$

- Naïve approach
  - Calculate prior probability for each label $\hat{P}(y_i)$ on the training set.
  - $\hat{P}(y) = \prod \hat{P}(y_i)$

Assumption: labels are independent

Restaurant: “McDonald’s”	Price: “cheap”	Food: “Pizza”
Restaurant: “KFC”	Price: “expensive”	Food: “Hamburger”
Restaurant: “PizzaHut”		Food:”Chinese”

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, “Dual Supervised Learning for Natural Language Understanding and Generation,” in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Semantic Frame $\log \hat{P}(y)$

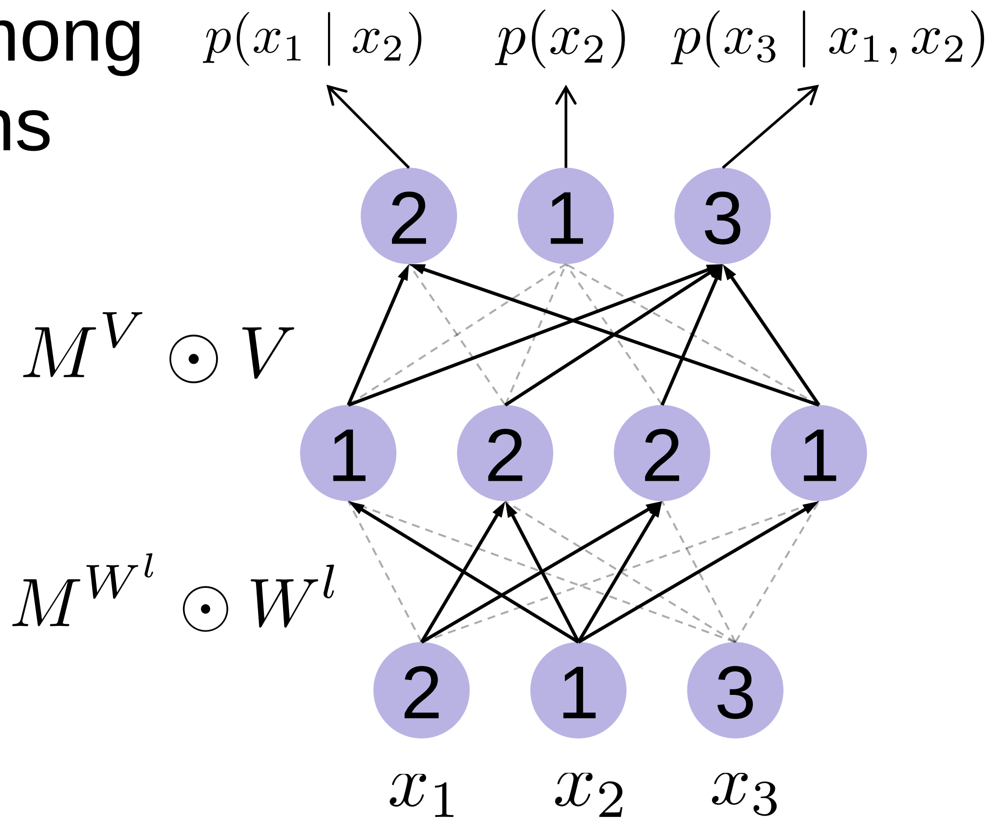
- Masked autoencoder for distribution estimation (MADE)

Introduce sequential dependency among labels by masking certain connections

$$M = \begin{cases} 1 & \text{if } m^l(k') \geq m^{l-1}(k) \text{ or } m^L(d) > m^{L-1}(k) \\ 0 & \text{otherwise} \end{cases}$$

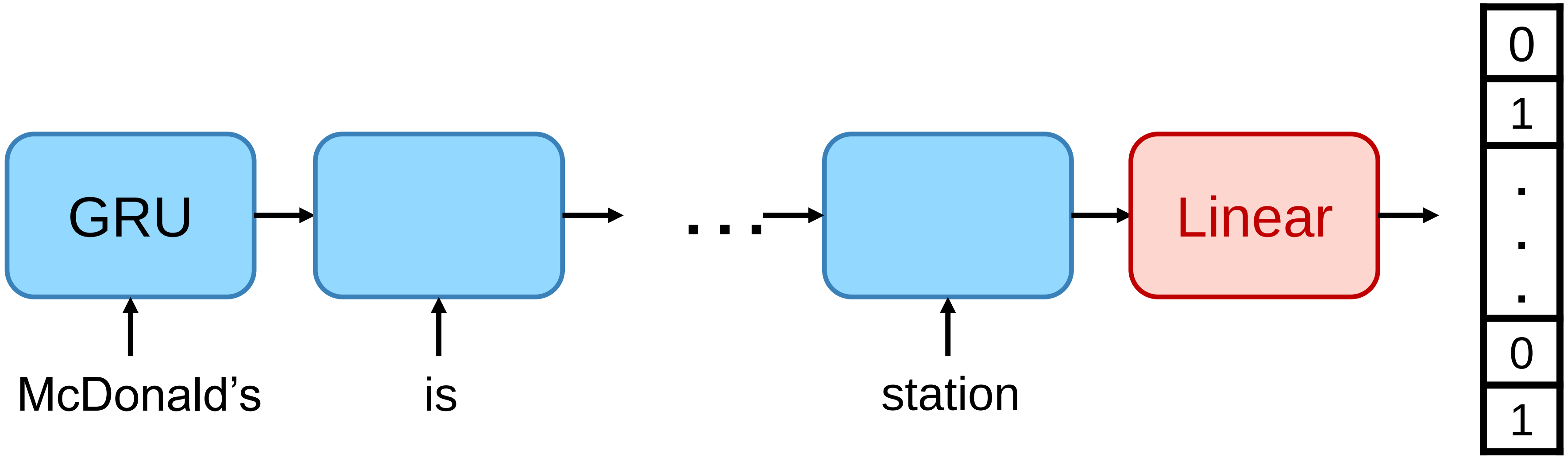
$$p(x) = \prod_{d}^{D} p(x_d \mid S_d)$$

→ marginal distribution of y

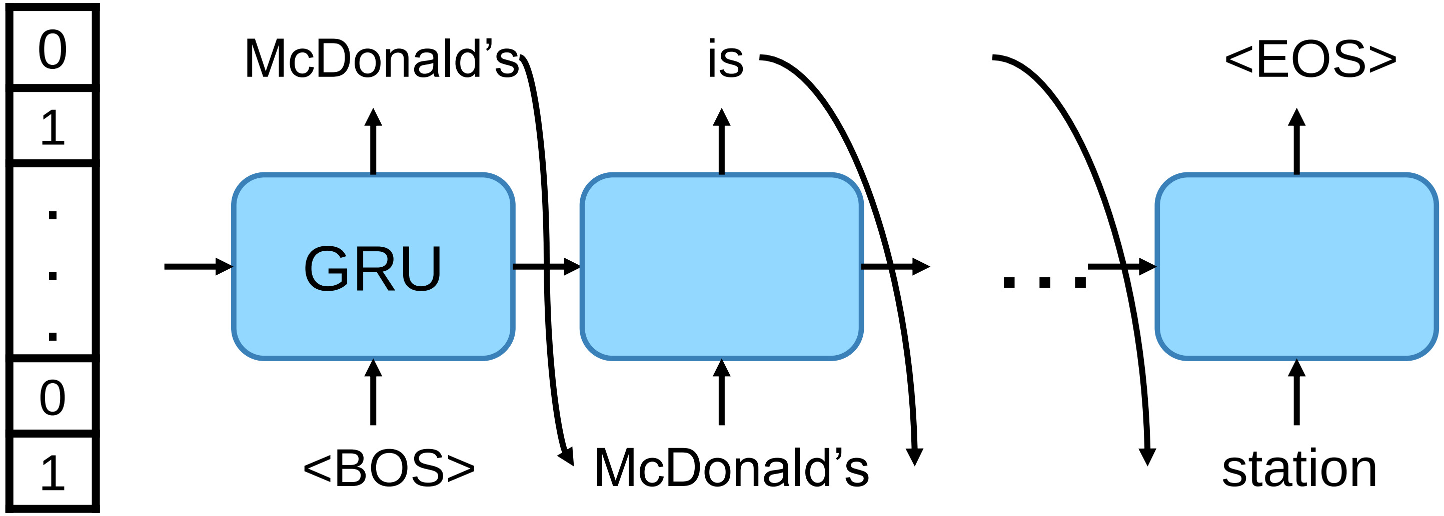


Germain, M., Gregor, K., Murray, I., & Larochelle, H., "MADE: Masked autoencoder for distribution estimation," in *Proceedings of International Conference on Machine Learning*, 2015.

NLU



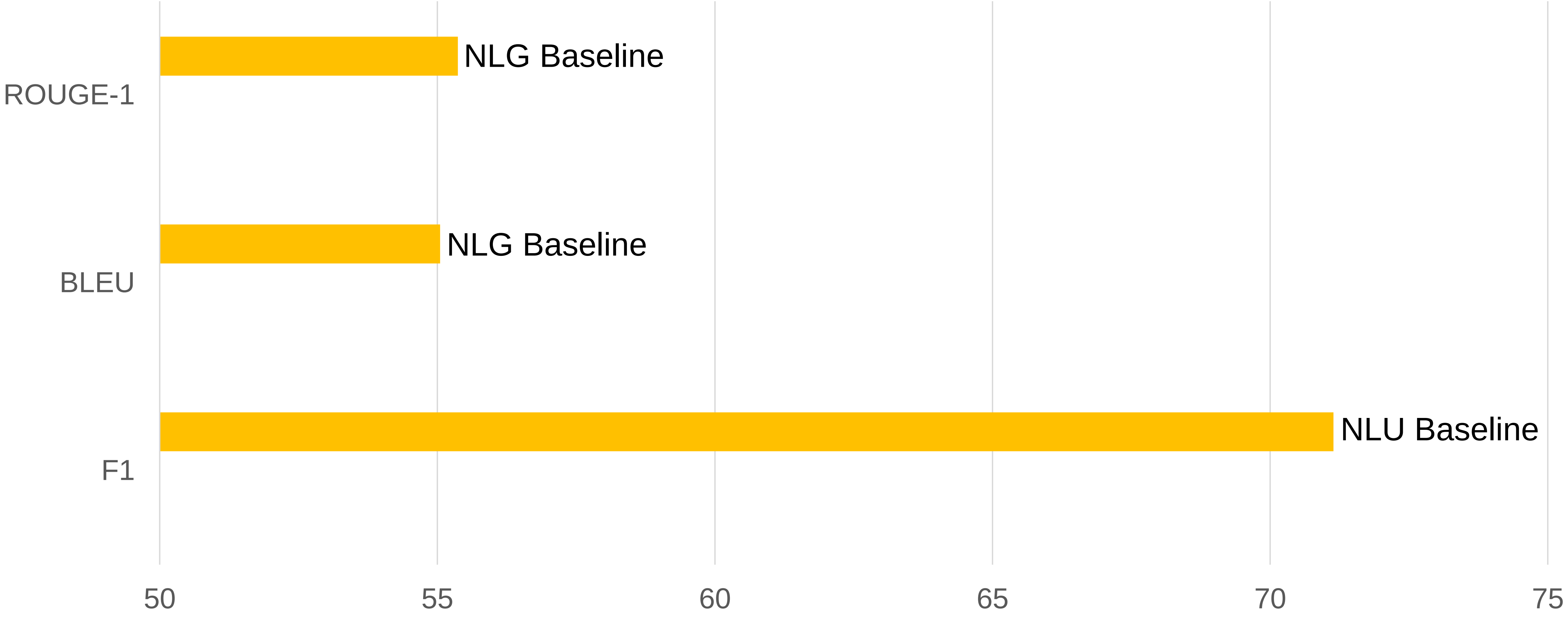
NLG



Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, “Dual Supervised Learning for Natural Language Understanding and Generation,” in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

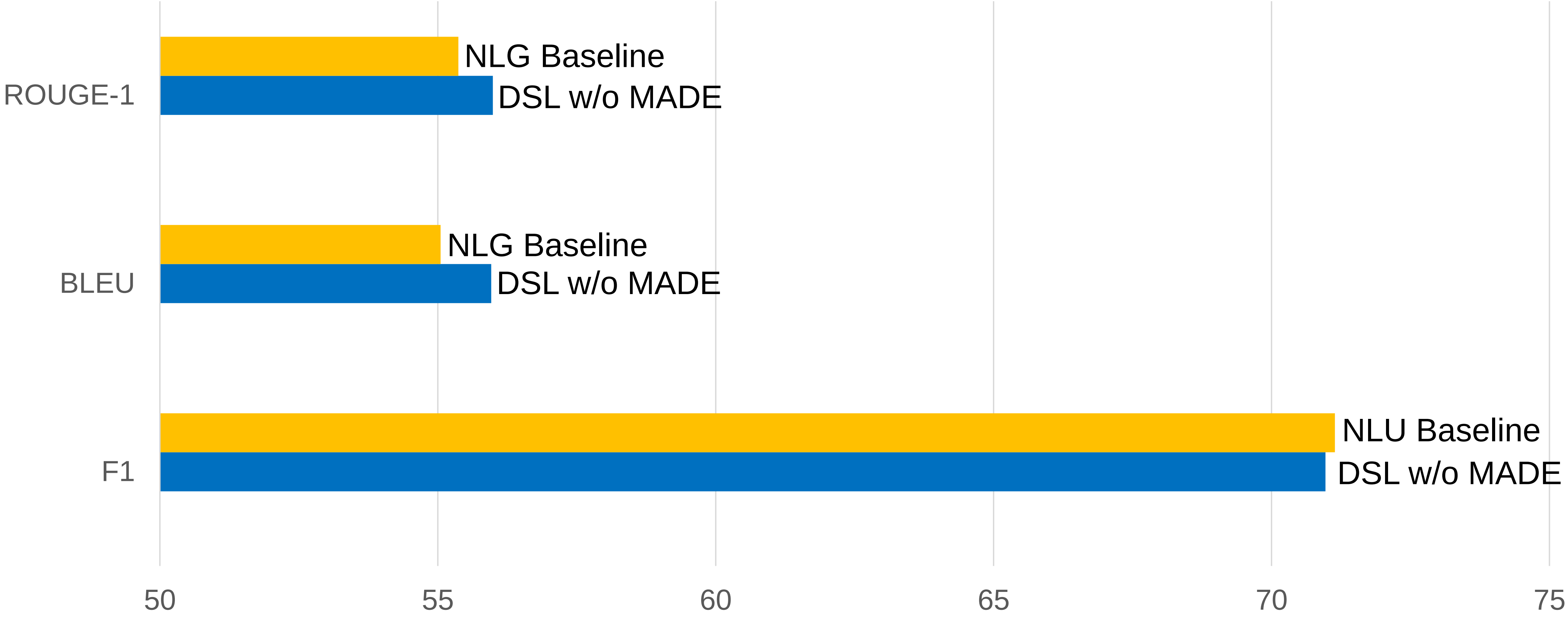
NLU/NLG Results

- E2E NLG data: 50k examples in the restaurant domain
- NLU: F-1 score; NLG: BLEU, ROUGE



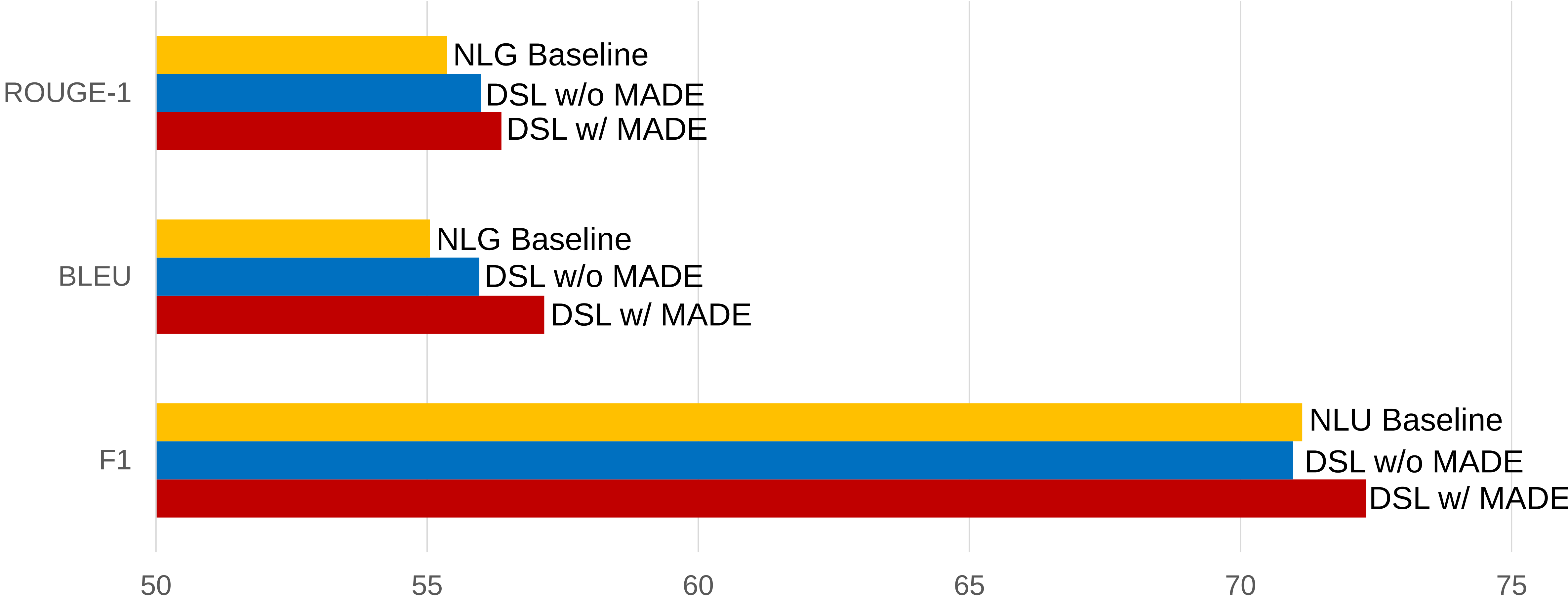
NLU/NLG Results

- E2E NLG data: 50k examples in the restaurant domain
- NLU: F-1 score; NLG: BLEU, ROUGE



NLU/NLG Results

- E2E NLG data: 50k examples in the restaurant domain
- NLU: F-1 score; NLG: BLEU, ROUGE



Summary

- **Robustness: spoken language embeddings** are needed for better conversational AI
  - Written texts enough for pre-training embeddings
  - Mismatch when applying to spoken language

  1) LatticeLM for preserving uncertainty
  2) Adapting contextualized embeddings robust to misrecognition

- **Scalability**: leveraging the **duality** of NLU and NLG
  - Apply dual learning to leverage the duality
  - Data distribution property is important
  - Better performance and flexibility for diverse NLU/NLG models

- Yun-Nung (Vivian) Chen
- Assistant Professor, National Taiwan University
- y.v.chen@ieee.org / http://vivianchen.idv.tw