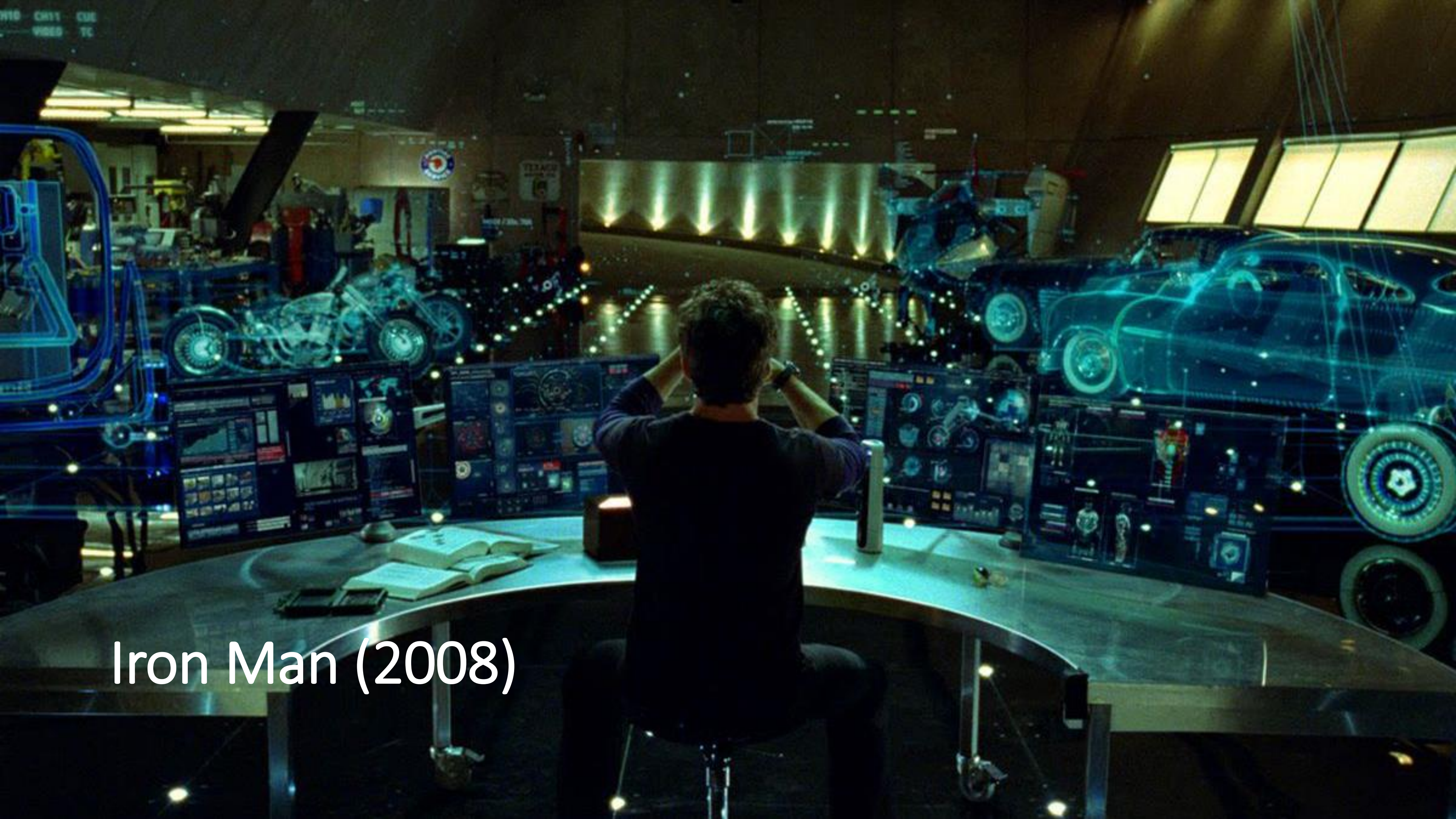
Can Current Conversational Assistants Satisfy Users?

Yun-Nung Vivian Chen
http://vivianchen.idv.tw

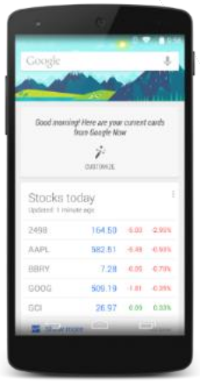
國立臺灣大學
National Taiwan University



Iron Man (2008)

Language Empowering Intelligent Assistants

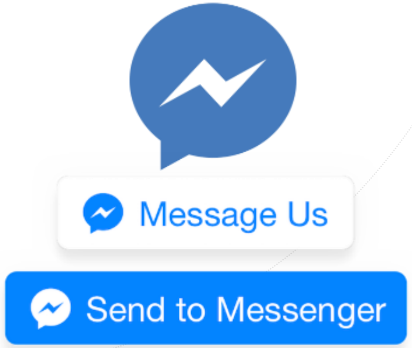
Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)

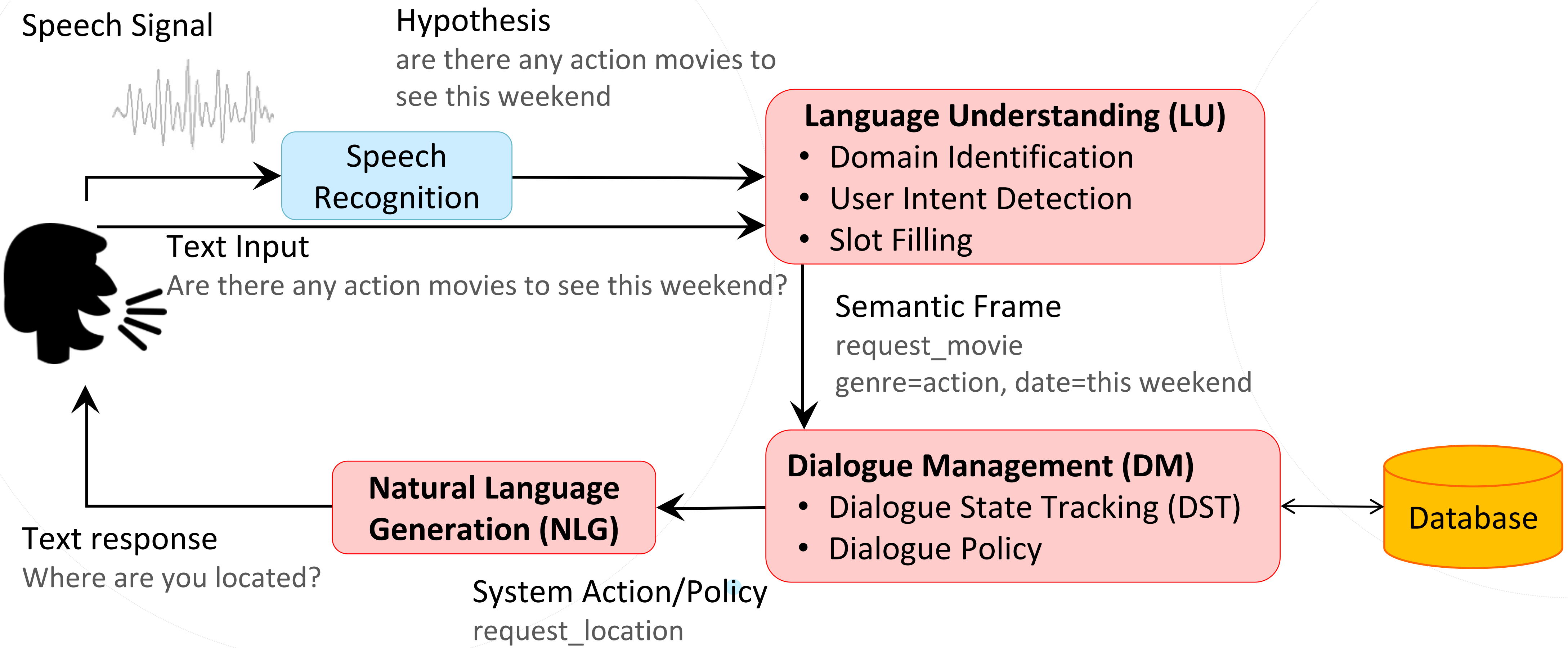


Facebook M & Bot (2015)

Google Home (2016)

Apple HomePod (2017)

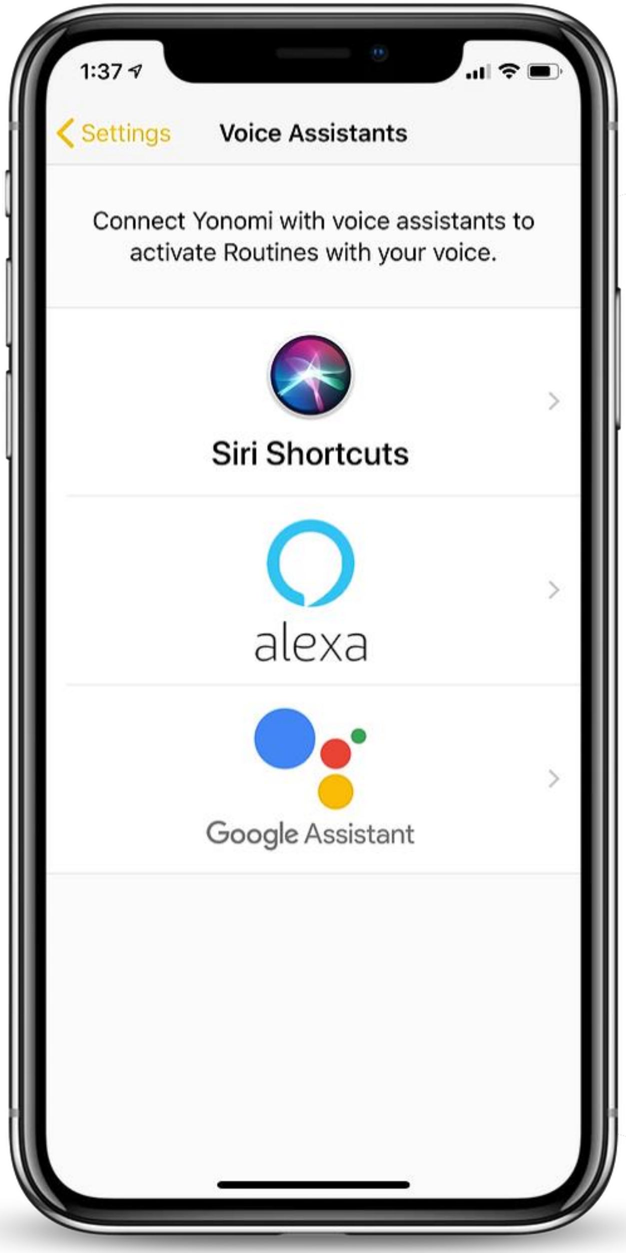
Task-Oriented Dialogue Systems (Young, 2000)



Recent Advances in NLP

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

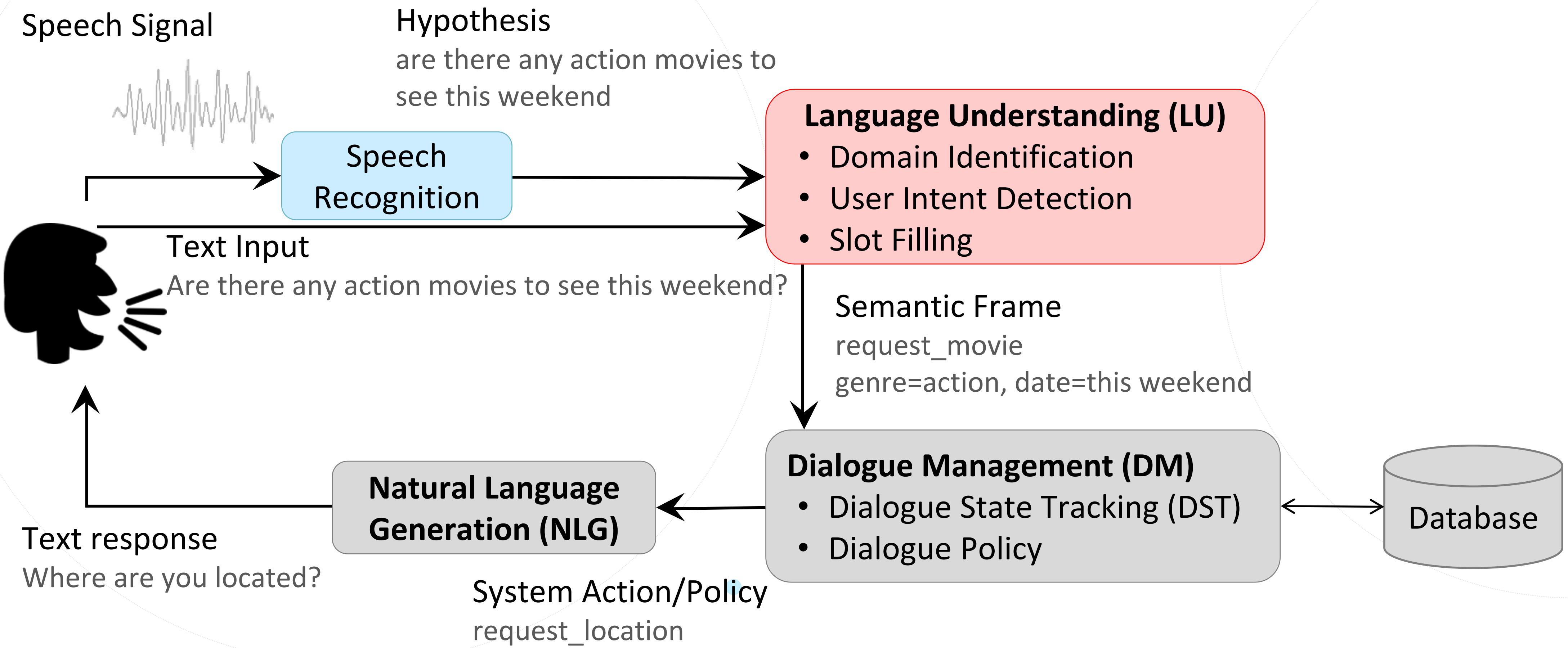
?




Call me ASAP

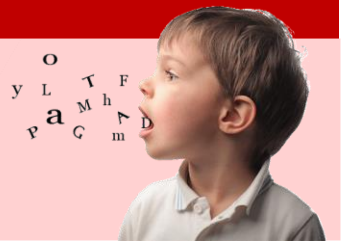

SAMSUNG
6:00
Listening...
Lift all lights ~~to~~ Morocco
List all flights tomorrow

Task-Oriented Dialogue Systems (Young, 2000)



Mismatch between Written and Spoken Languages

Training	Testing
• Written language	• Spoken language • Include recognition errors

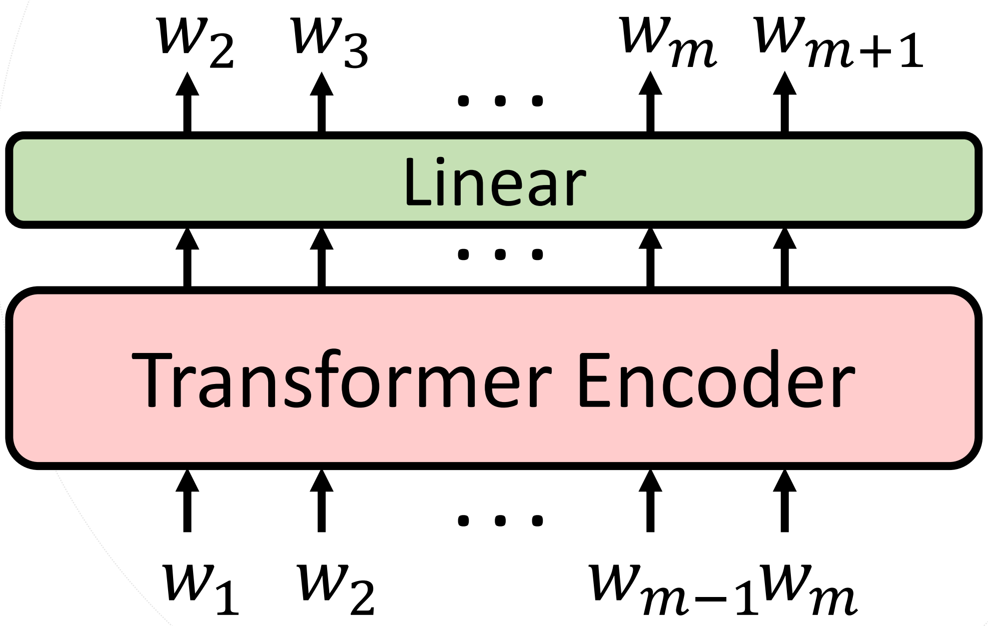


- Goal: ASR-Robust Contextualized Embeddings
  - ✓learning contextualized word embeddings specifically for spoken language
  - ✓achieves better performance on *spoken* language understanding tasks
    - – shows better results on ASR transcripts
    - – maintain similar results on manual transcripts

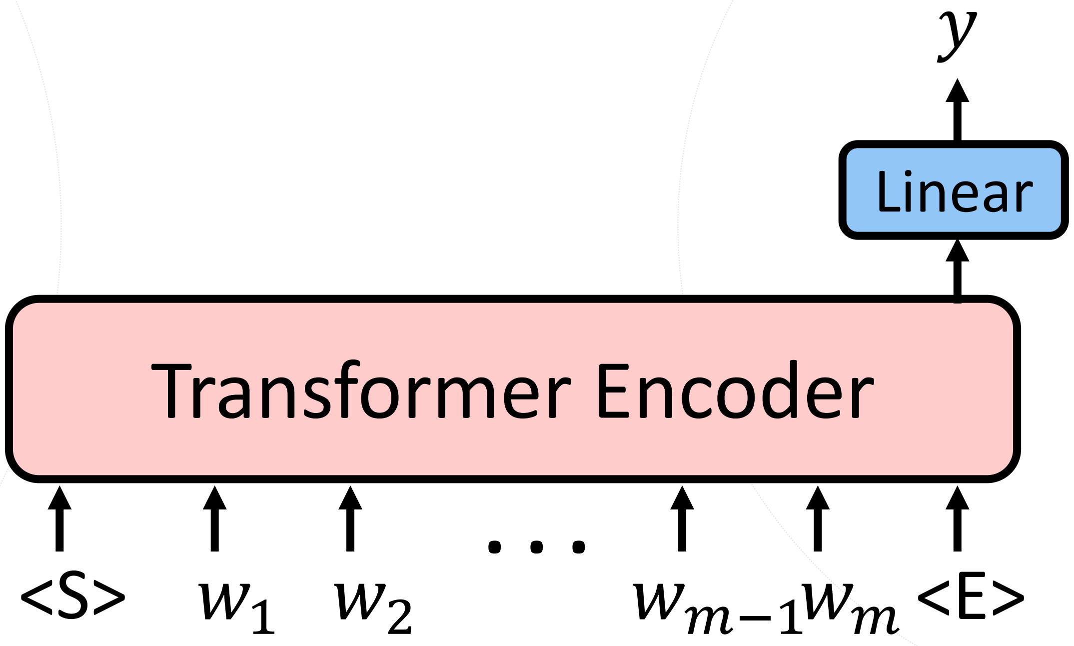
Solution 1:
Adapting Transformer to ASR Lattices

BERT/GPT Pre-Training & Fine-Tuning

- Pre-Training

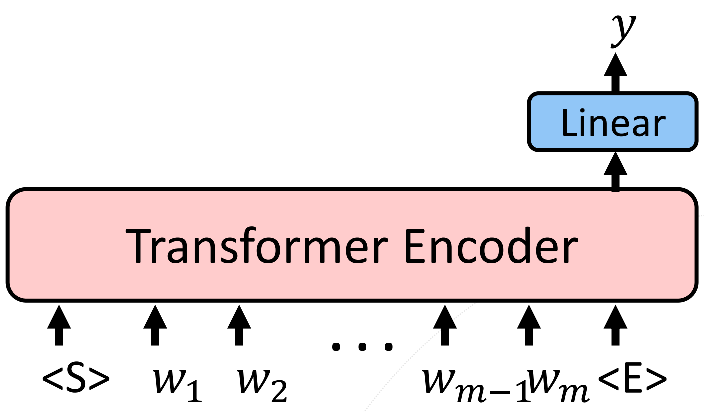


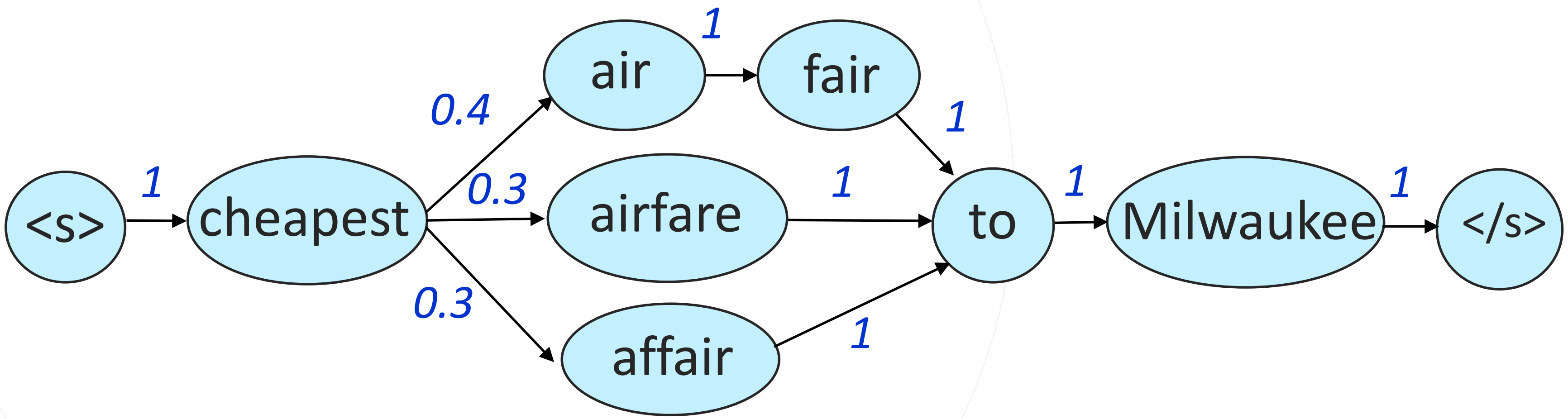
- Fine-Tuning



ASR Lattices

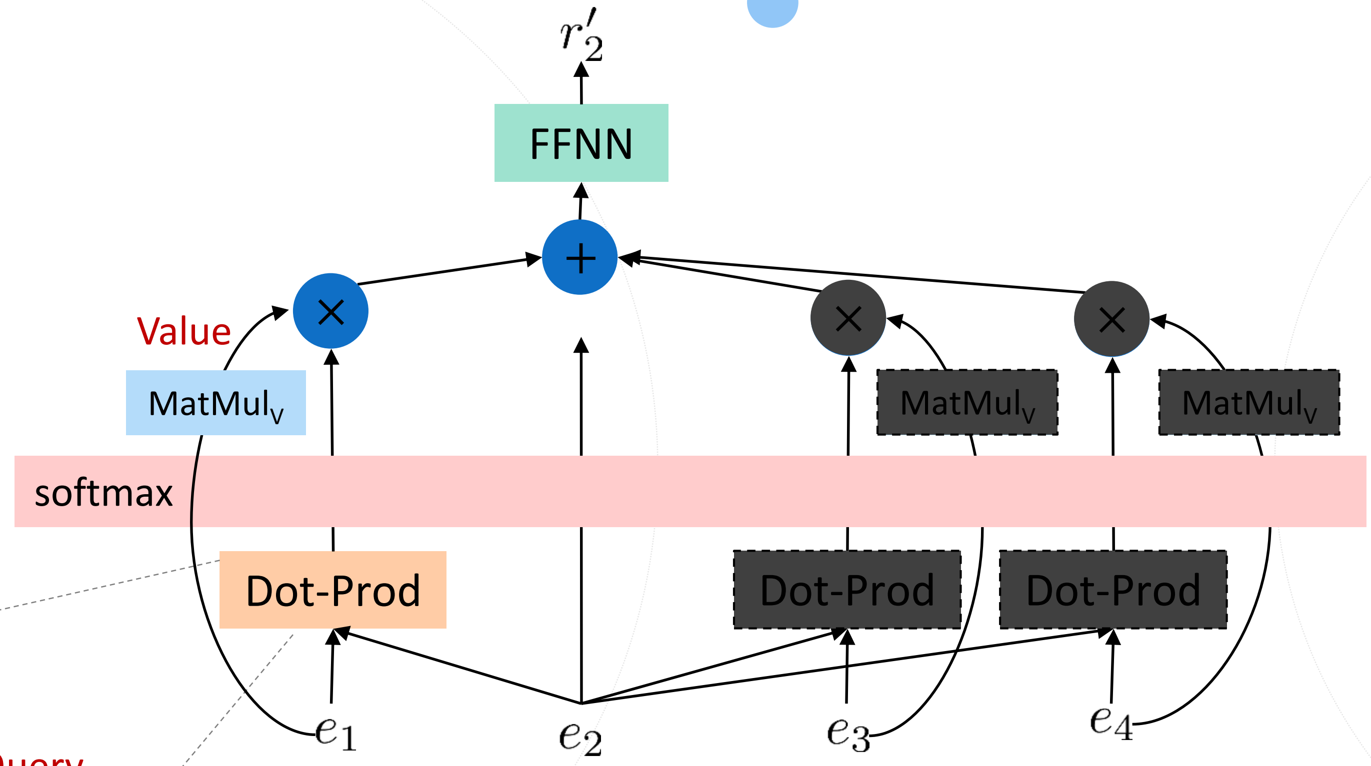
- Idea: lattices may include correct words
- Goal: feed lattices into Transformer

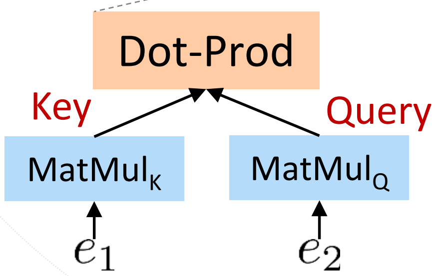




1) Linearize
2) Binary mask
3) Probabilistic mask

Self-Attention (Vaswani+, 2017)



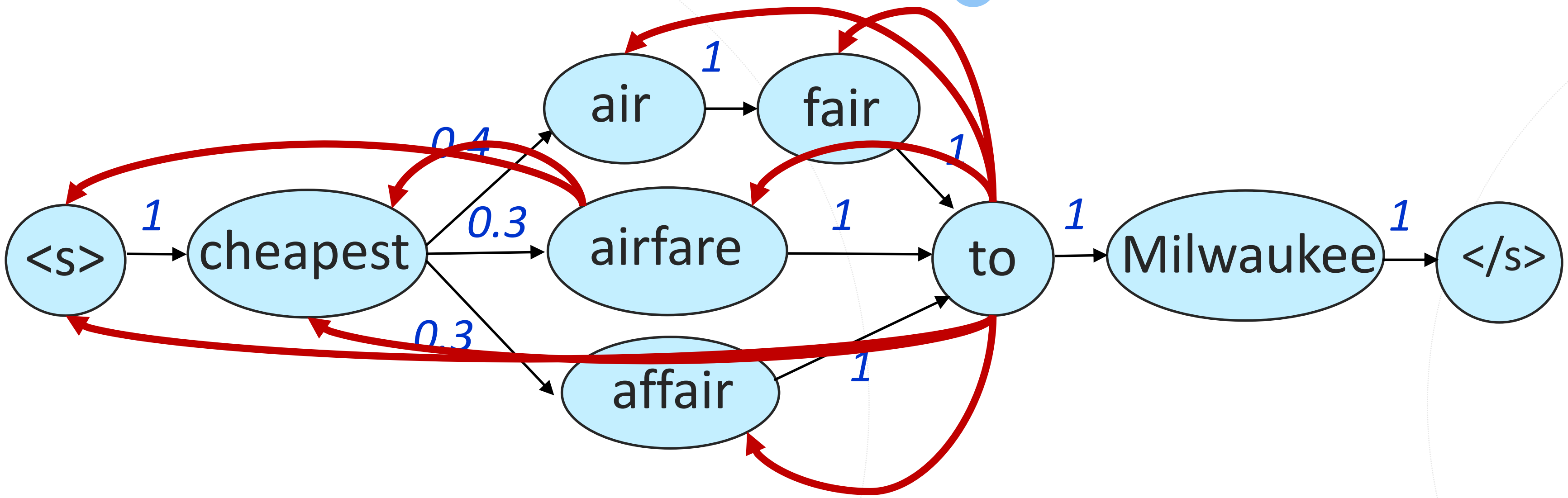


$$A(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}}\cancel{)V}+M)V$$

Vaswani et al., "Attention Is All You Need", in *NIPS*, 2017.

Attention Masks

$$A(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \boxed{M}\right)V$$



- Binary masks

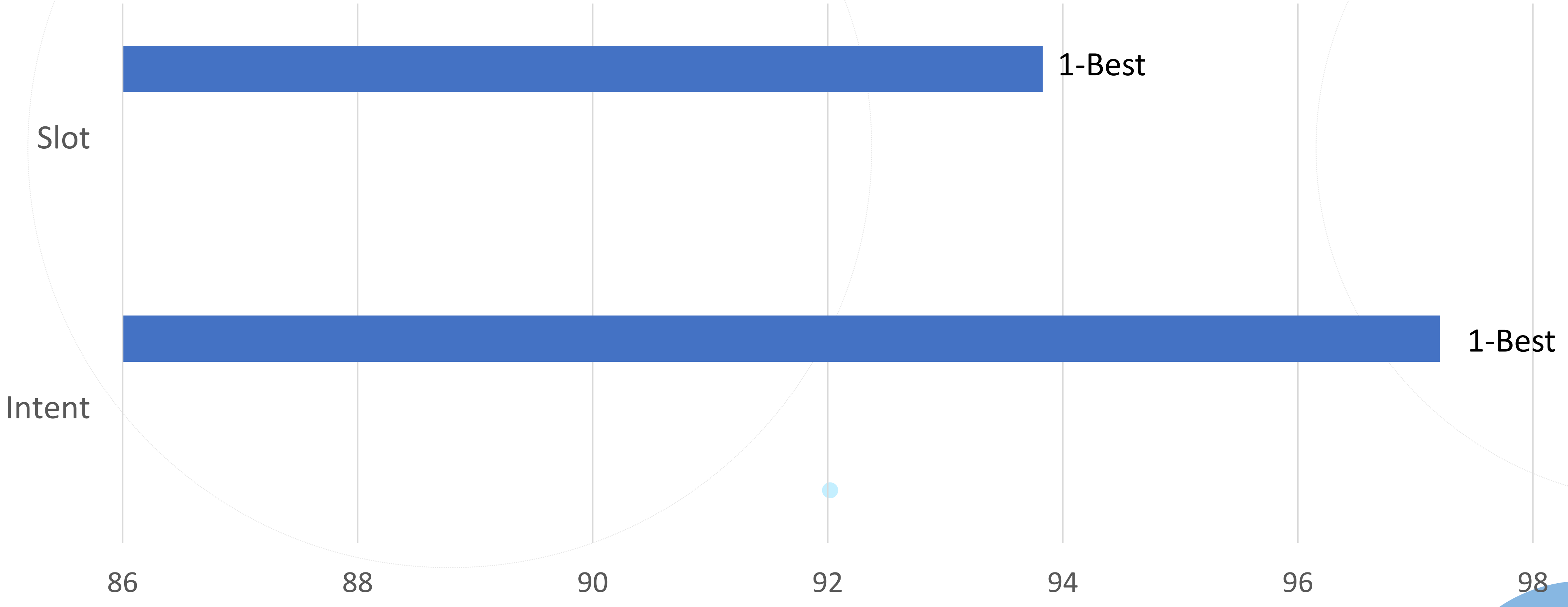
$$M_{ij}^{bin} = \begin{cases} 0 & v_j \in \mathrm{Pre}(v_i) \text{ or } v_i = v_j, \\ -\infty & \text{otherwise.} \end{cases}$$

- Probabilistic masks

$$M_{ij}^{prob} = \begin{cases} \log P(v_j \in \mathrm{Pre}(v_i) \mid v_i) & v_j \in \mathrm{Pre}(v_i), \\ 0 & v_i = v_j, \\ -\infty & \text{otherwise.} \end{cases}$$

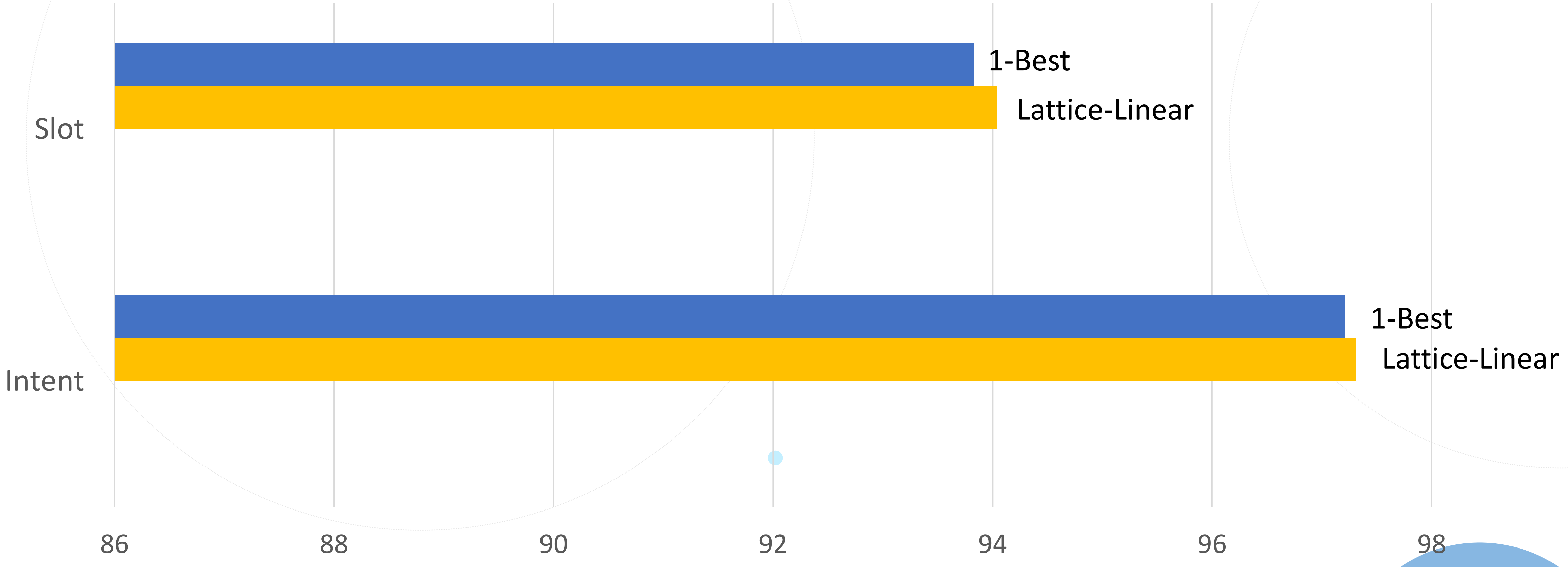
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%



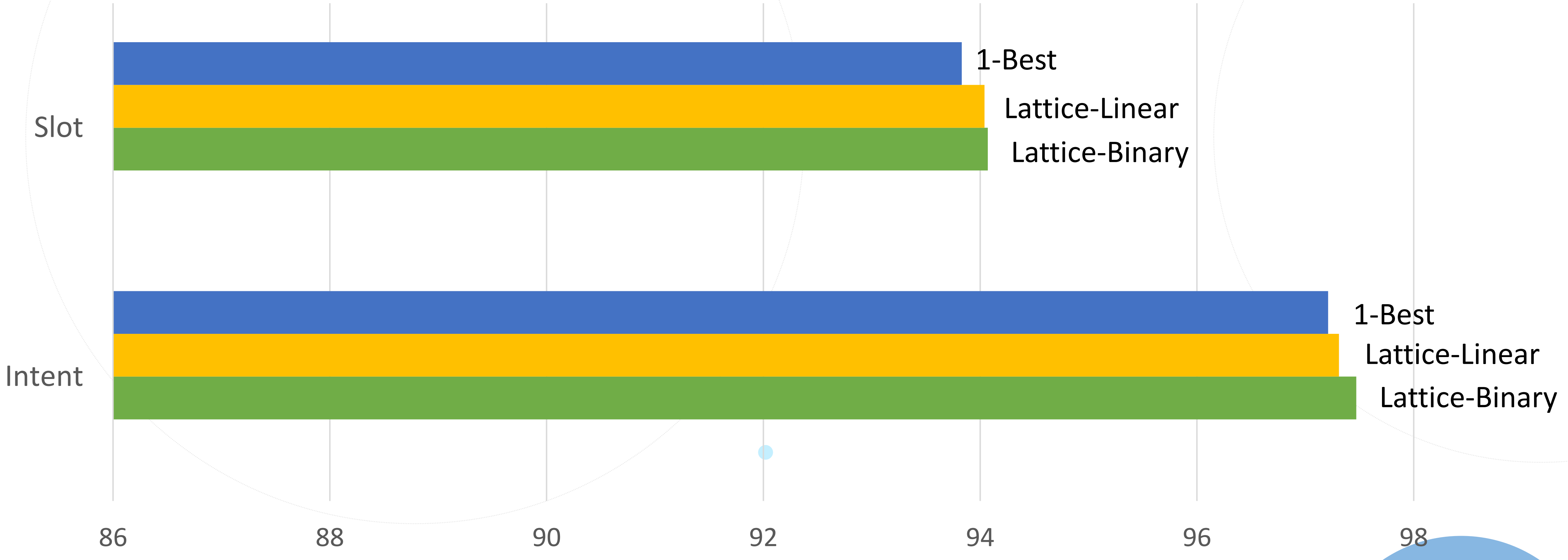
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%



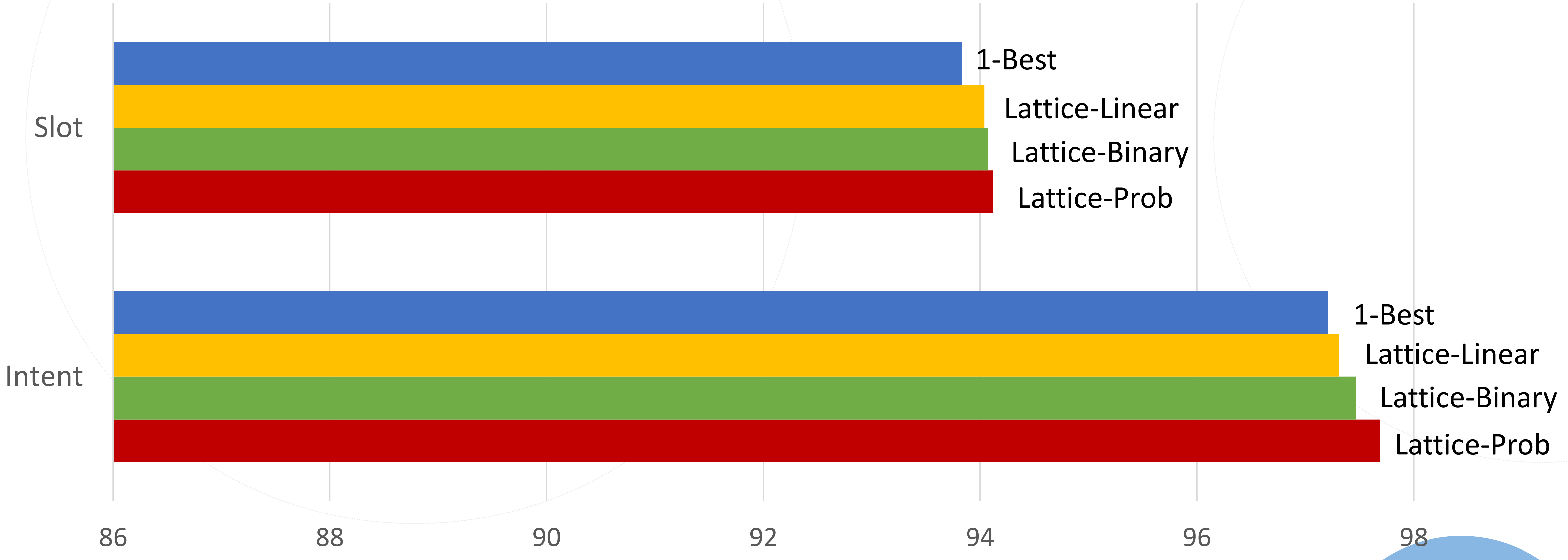
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%



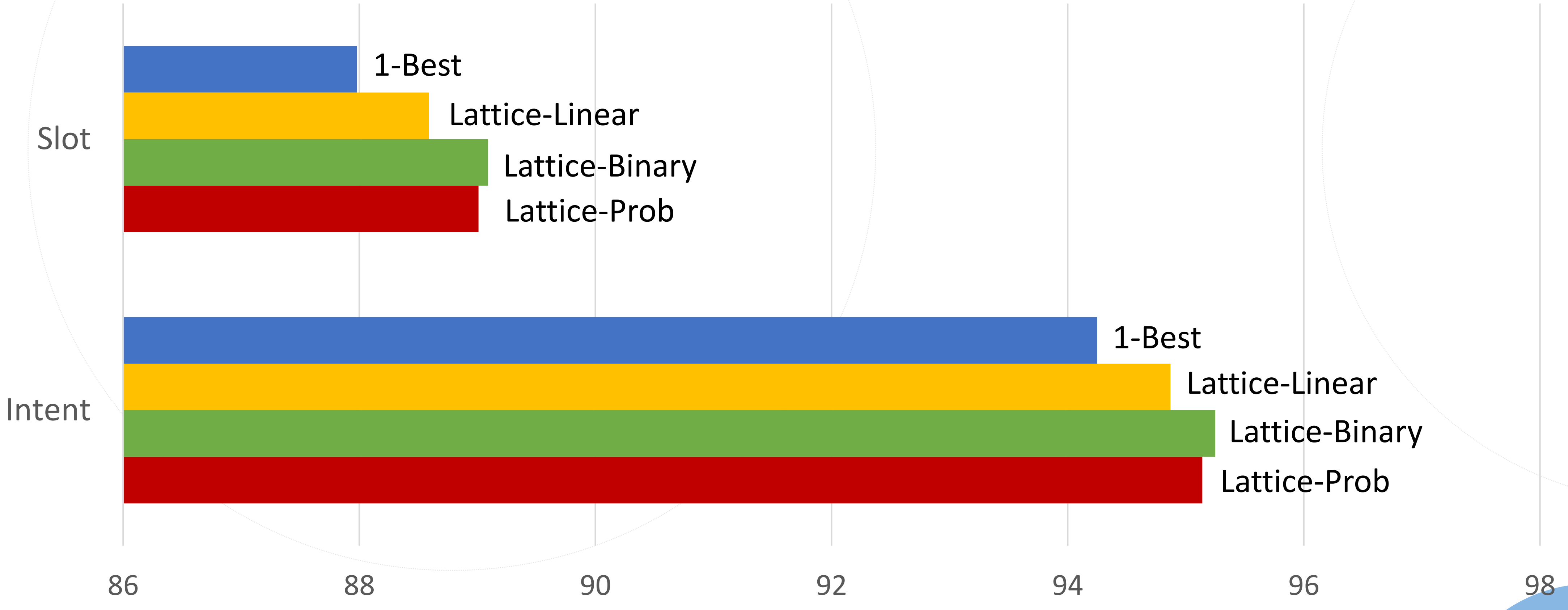
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%



Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
    - Word Error Rate: 26.3%



NTU MIULAB

What if we do not have ASR lattices?

Solution 2: Learning ASR-Robust Embeddings

ASR-Robust Contextualized Embeddings

- Confusion-Aware Fine-Tuning
  - Supervised

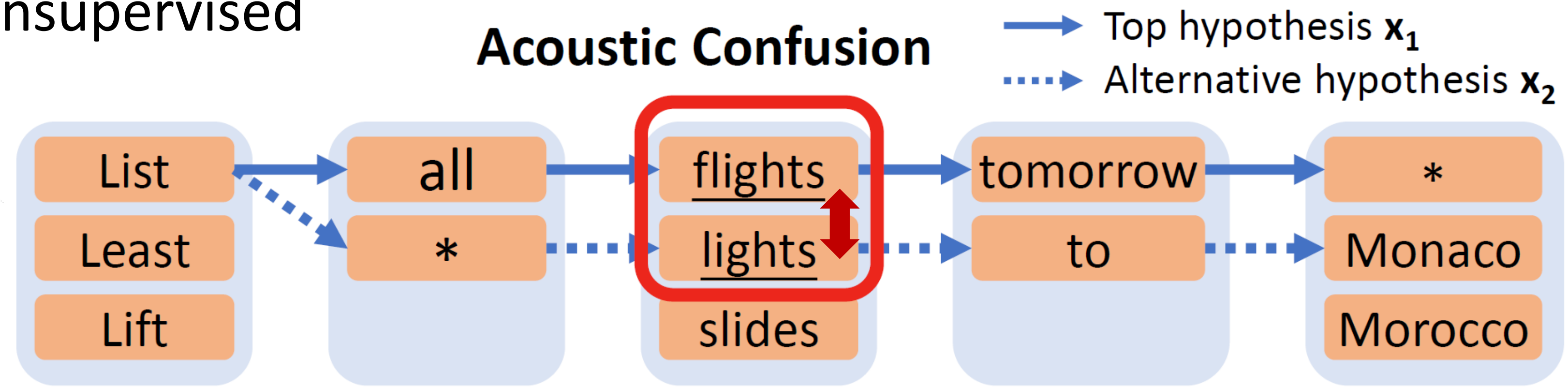
Acoustic Confusion $C = \{w_3^{x_{trs}}, w_2^{x_{asr}}\}$

x_{trs} : Show me the <u>fares</u> from Dallas to Boston

x_{asr} : Show me * <u>affairs</u> from Dallas to Boston

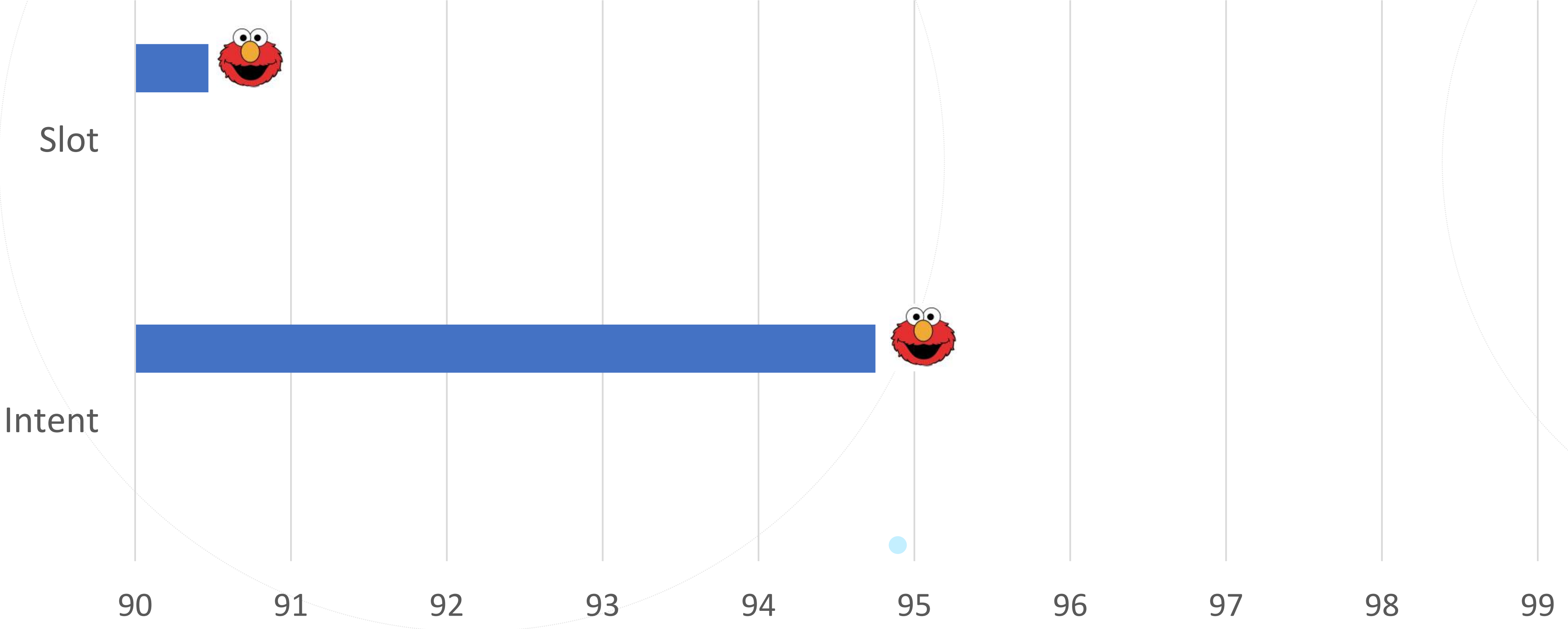
  - Unsupervised

Acoustic Confusion



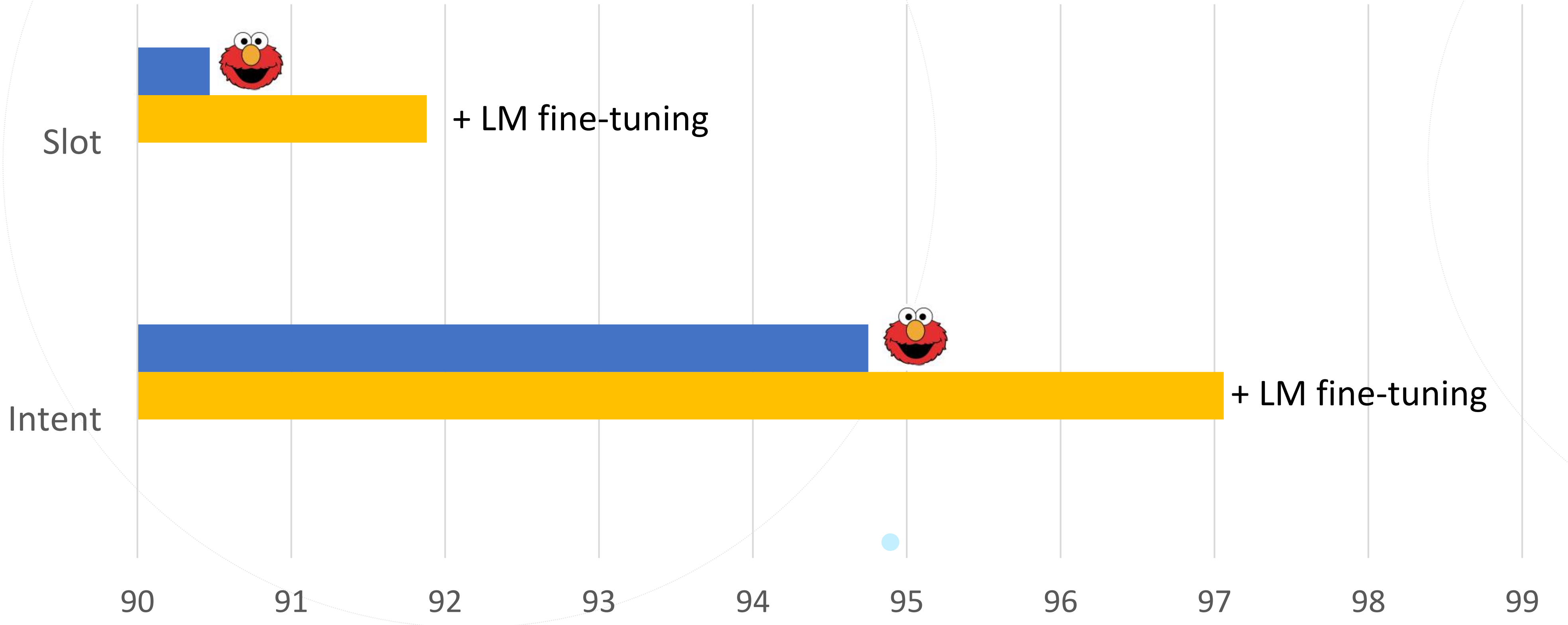
Spoken Language Understanding Results

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



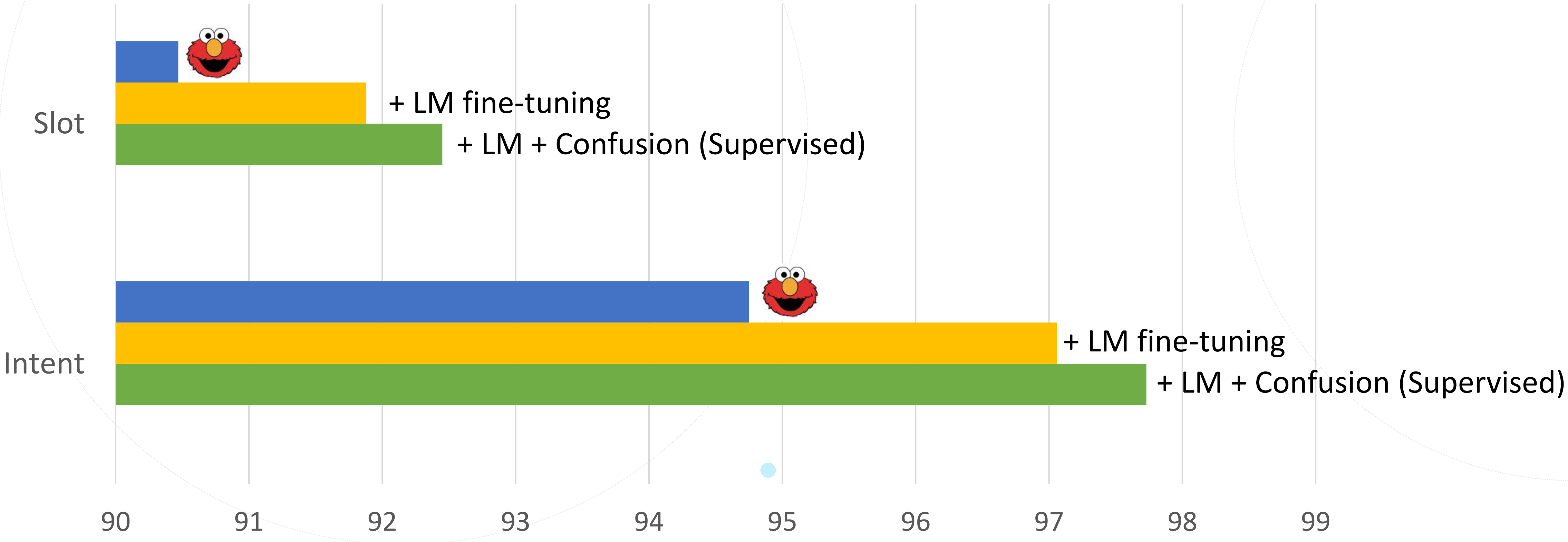
Spoken Language Understanding Results

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



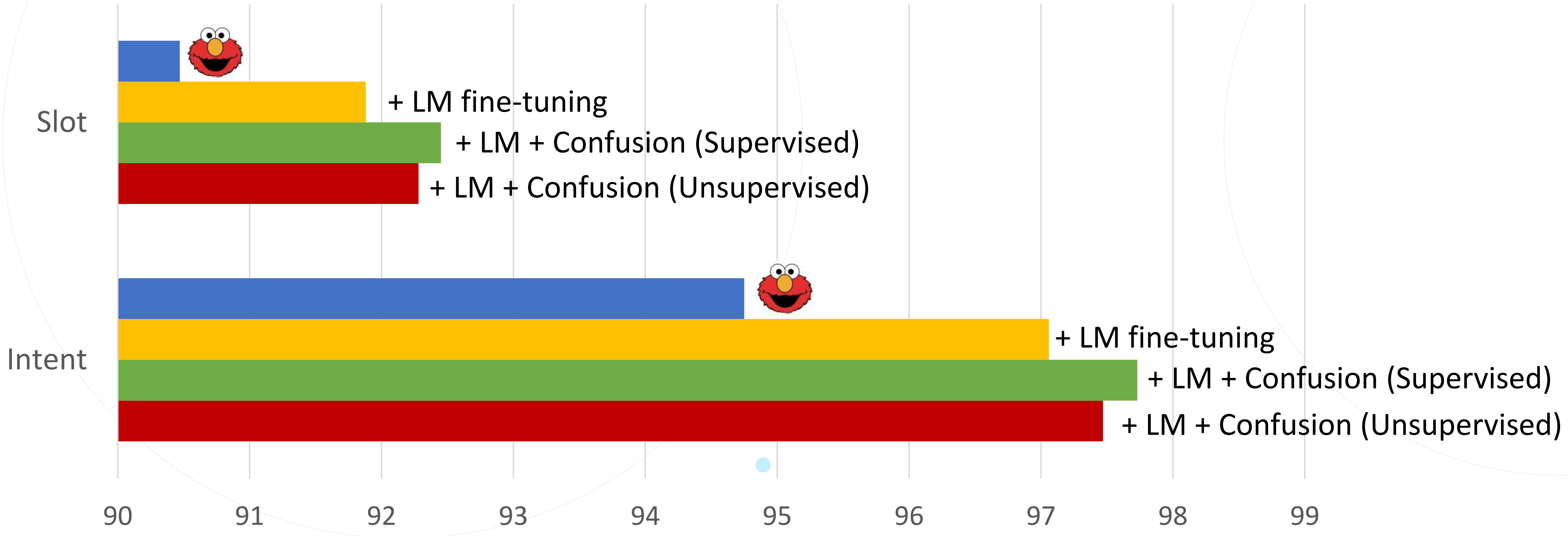
Spoken Language Understanding Results

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

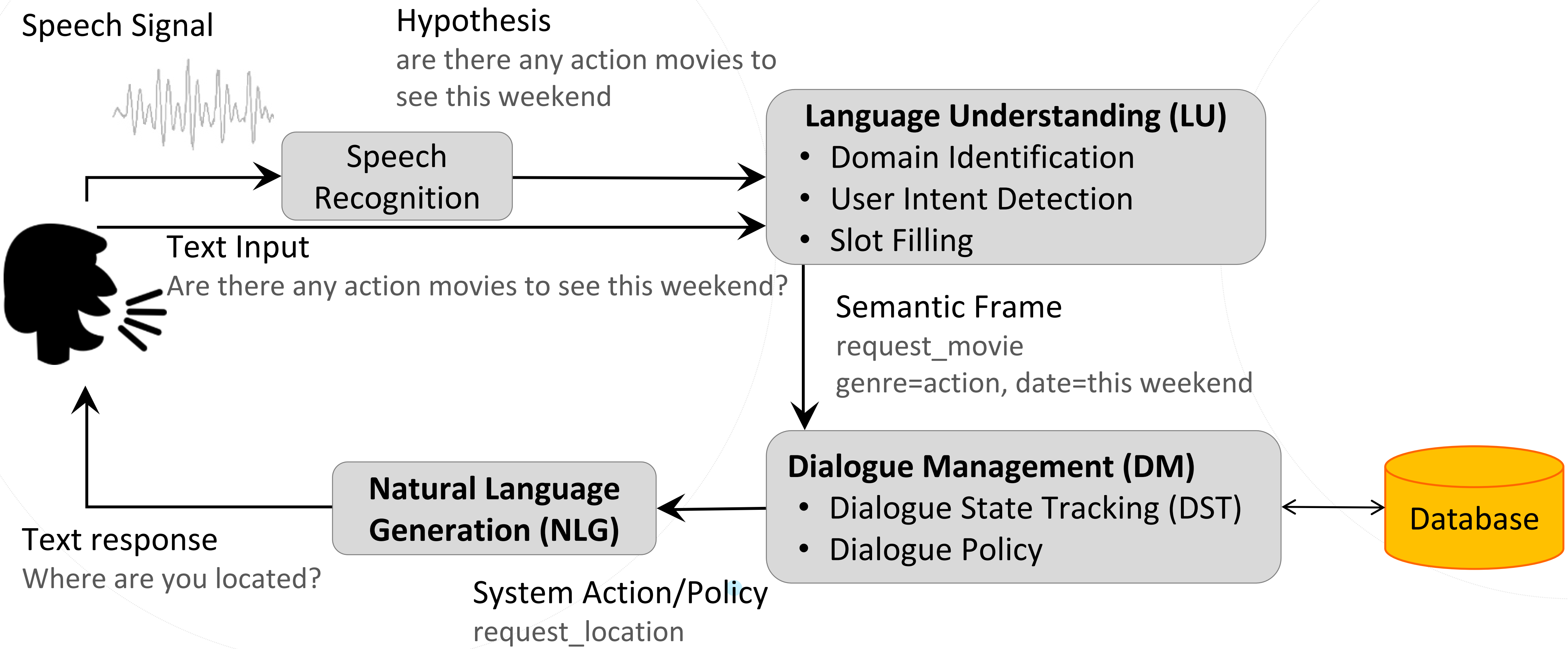


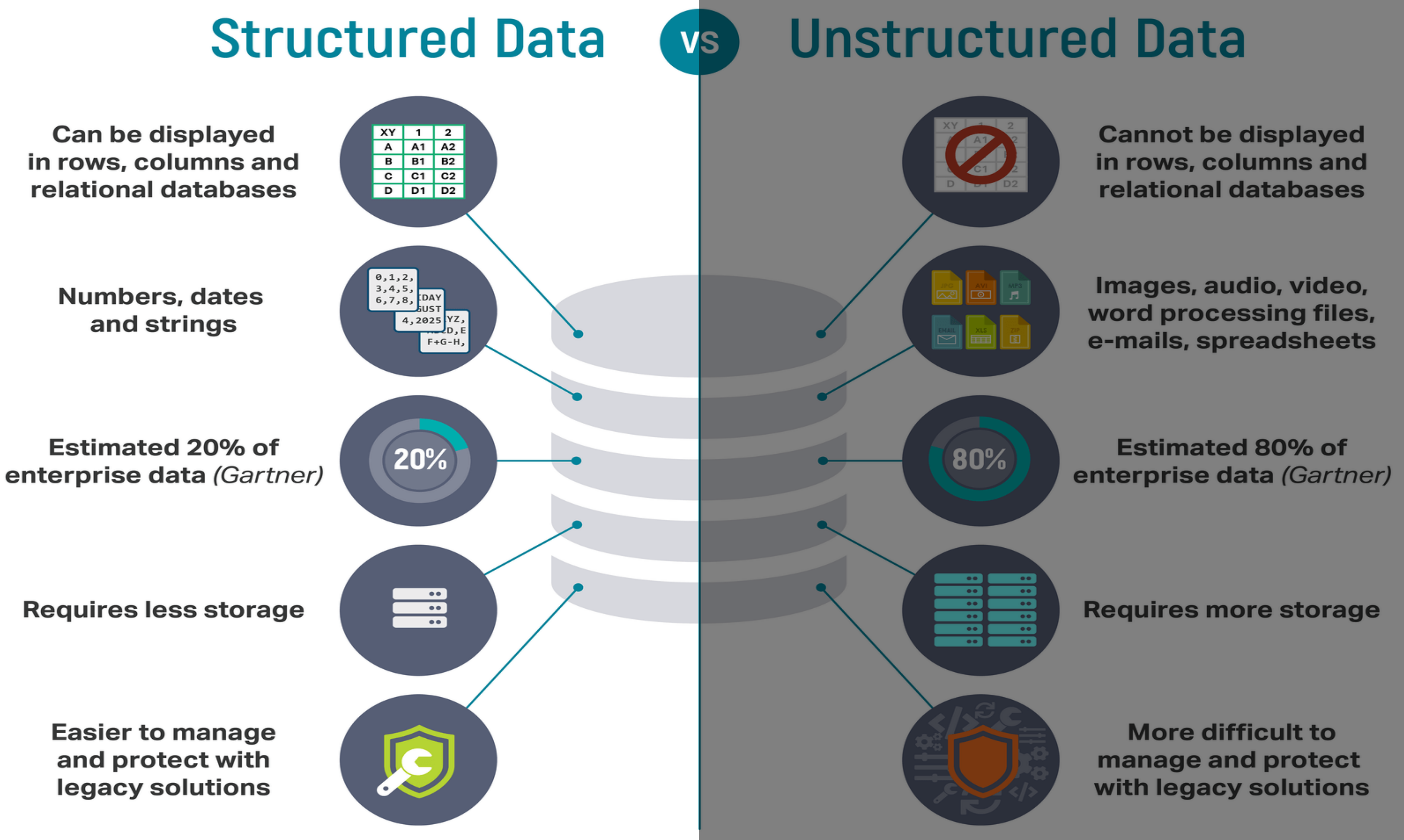
Spoken Language Understanding Results

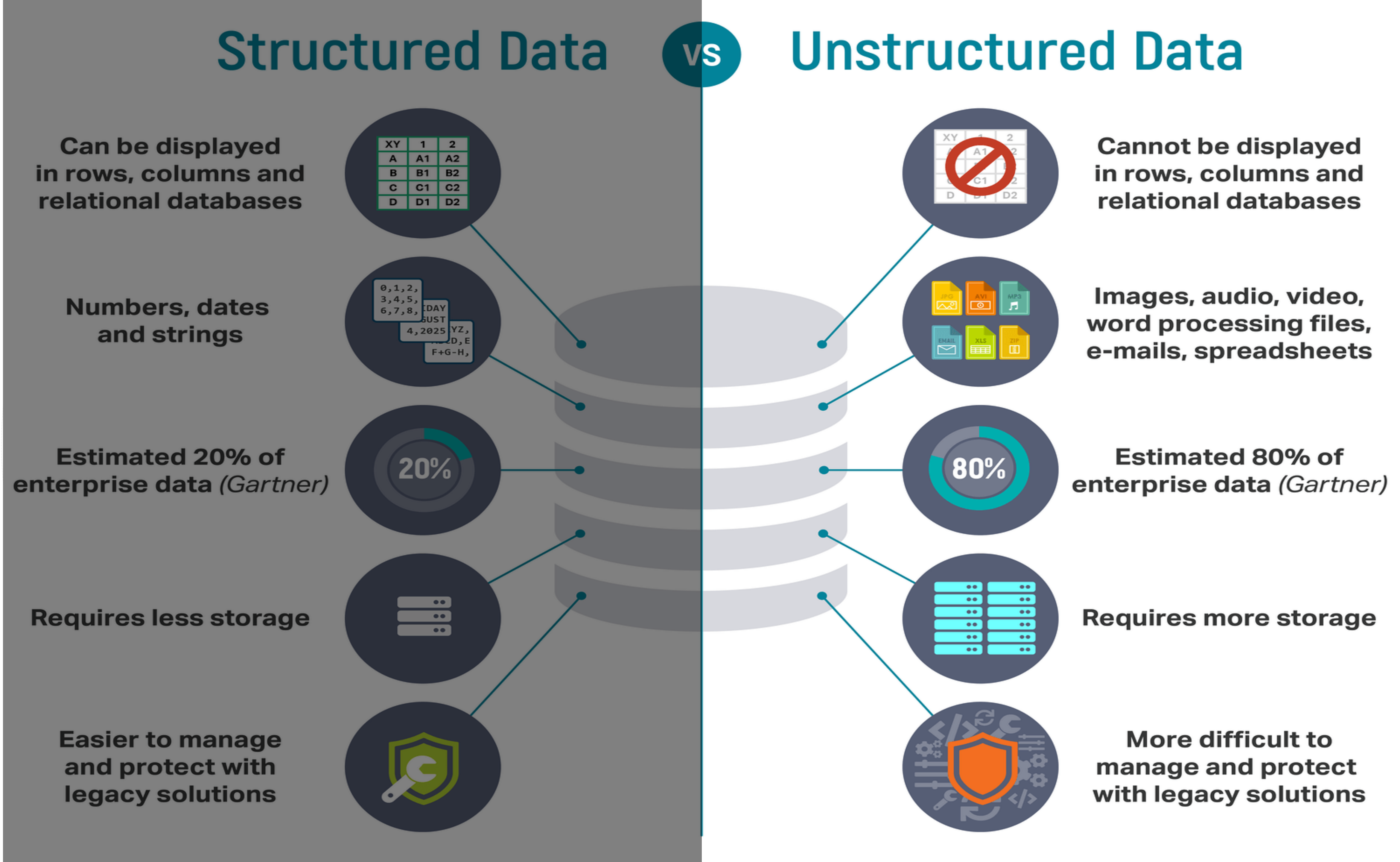
- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%



Task-Oriented Dialogue Systems (Young, 2000)




Structured Data
VS
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
Requires more storage
More difficult to manage and protect with legacy solutions
20%
80%



Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

NTU MIULAB

Conversational AI for Unstructured Knowledge

- A machine reads big text data
  - serves as a teacher
- A user can ask questions
  - serves as a student
  - in a conversational manner

→ Conversational QA

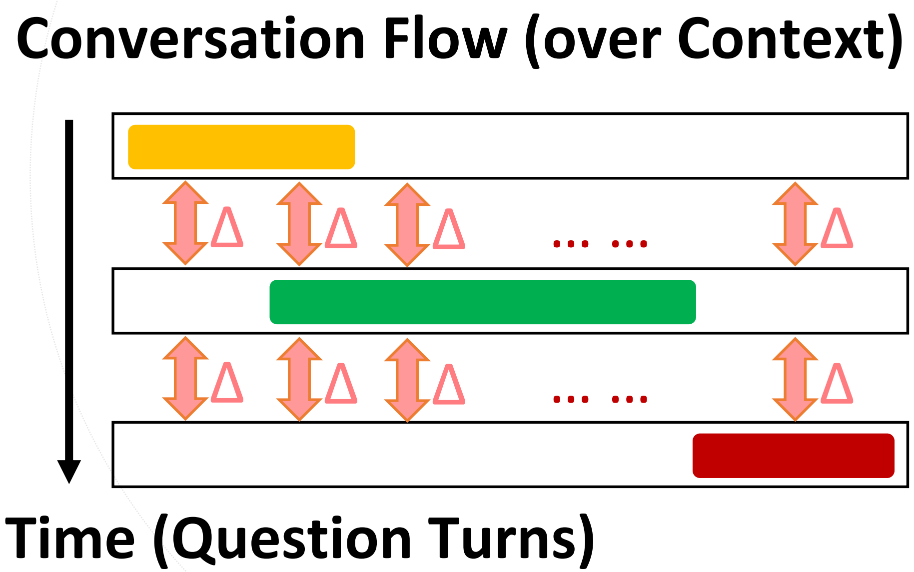
Section: Daffy Duck, Origin & History

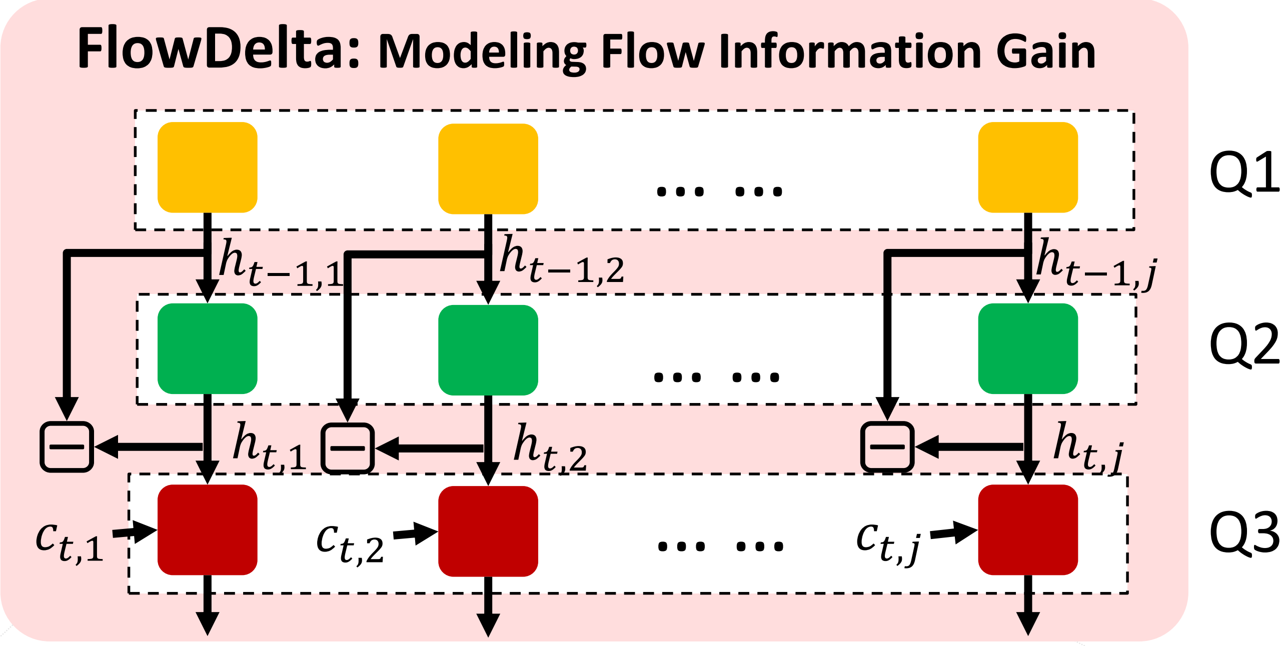
STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
...

NTU MIULAB

FlowDelta: Information Gain in Dialogue Flow

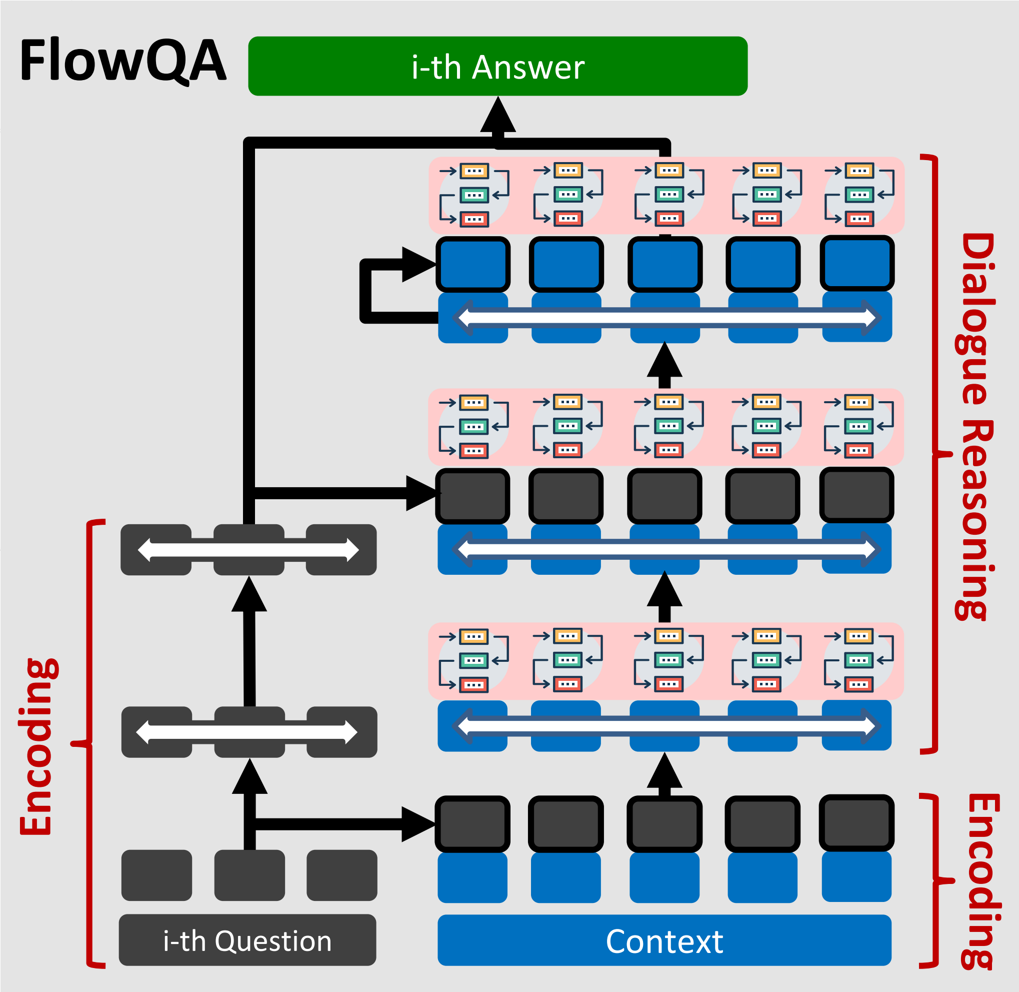
- Idea: model the *difference* of hidden states in multi-turn dialogues

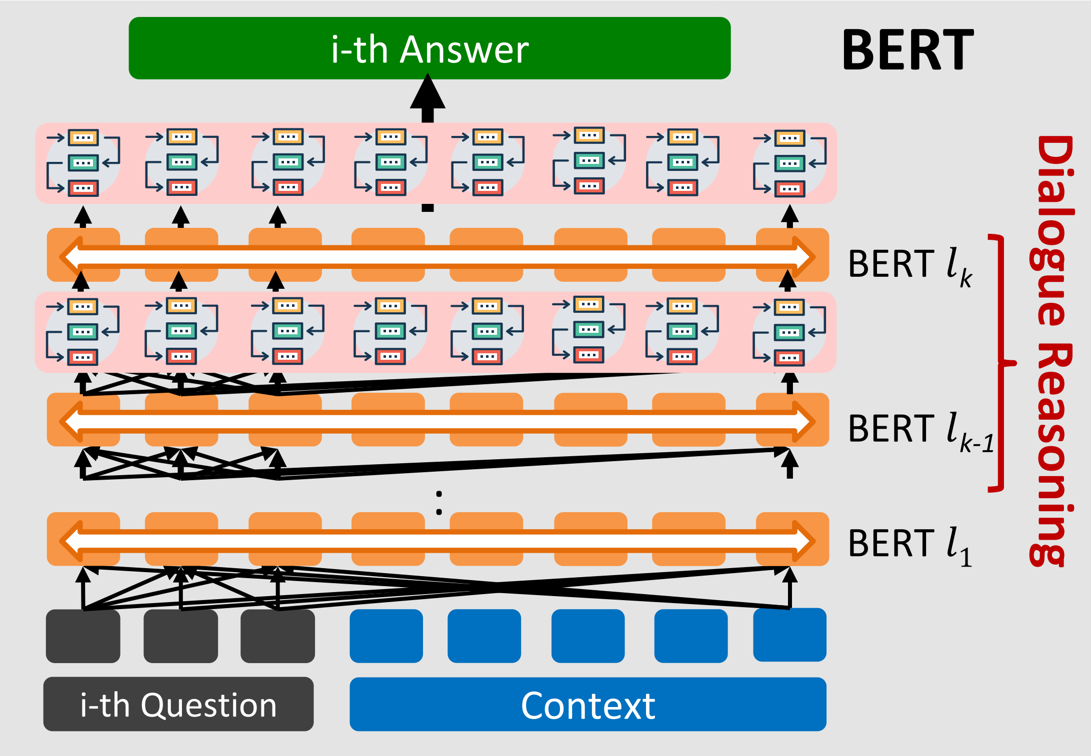




FlowDelta: Information Gain in Dialogue Flow

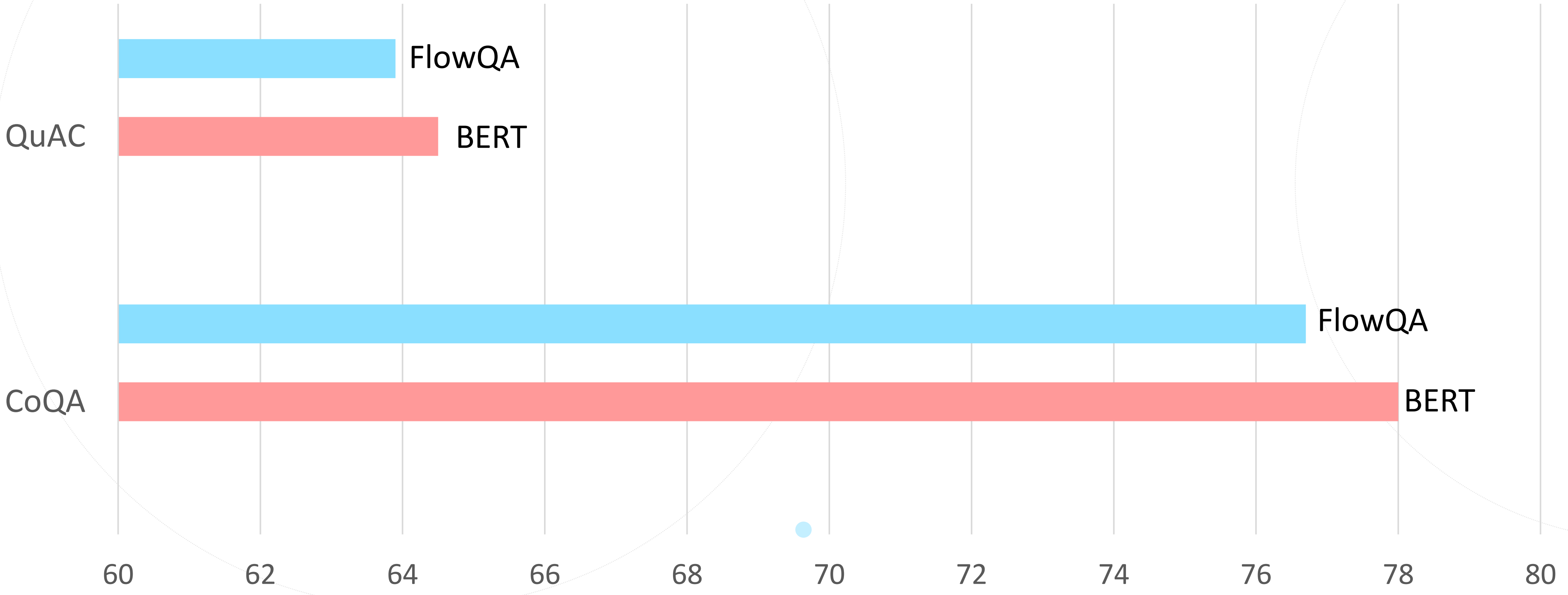
- Idea: model the *difference* of hidden states in multi-turn dialogues





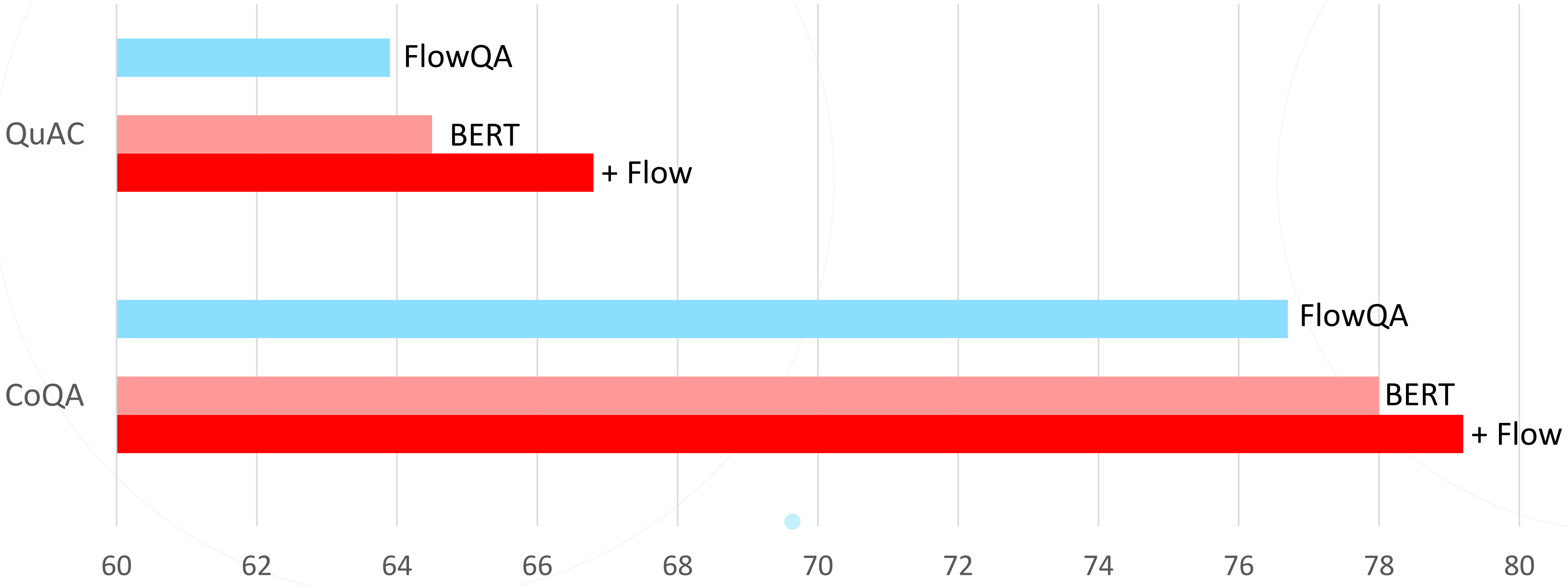
Conversational QA Results

- Data: QuAC, CoQA



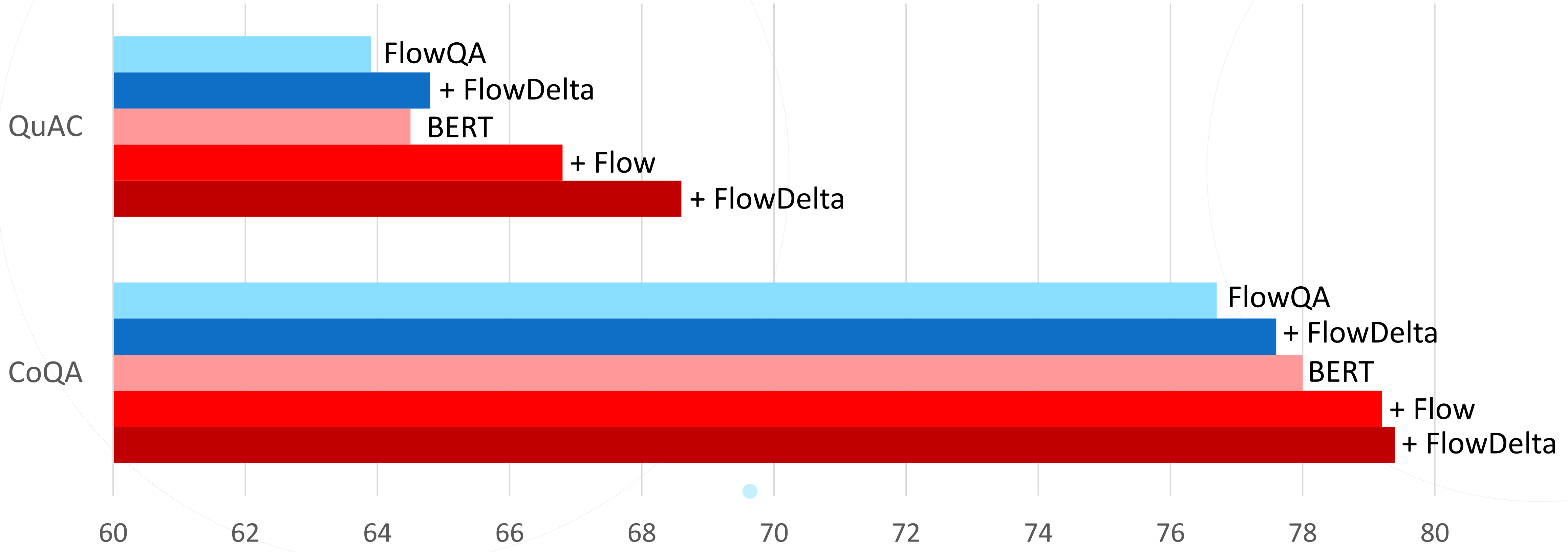
Conversational QA Results

- Data: QuAC, CoQA



Conversational QA Results

- Data: QuAC, CoQA



QuAC Leaderboard

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
 Jun 23, 2019	TransBERT (single model) *Anonymous*	69.4	65.4	9.3
2 Apr 24, 2019	Bert-FlowDelta (single model) *Anonymous*	67.8	63.6	**12.1**
3 June 13, 2019	Context-Aware-BERT (single model) *Anonymous*	**69.6**	**65.7**	8.1

Summary

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

1) Adapting Transformer to ASR lattices
2) Adapting contextualized embeddings robust to misrecognition

- **Conversational QA** enables unstructured information access
    - **FlowDelta: information gain in dialogue flow** guides better understanding

Yun-Nung (Vivian) Chen

Assistant Professor, National Taiwan University

y.v.chen@ieee.org / http://vivianchen.idv.tw

NTU MIULAB