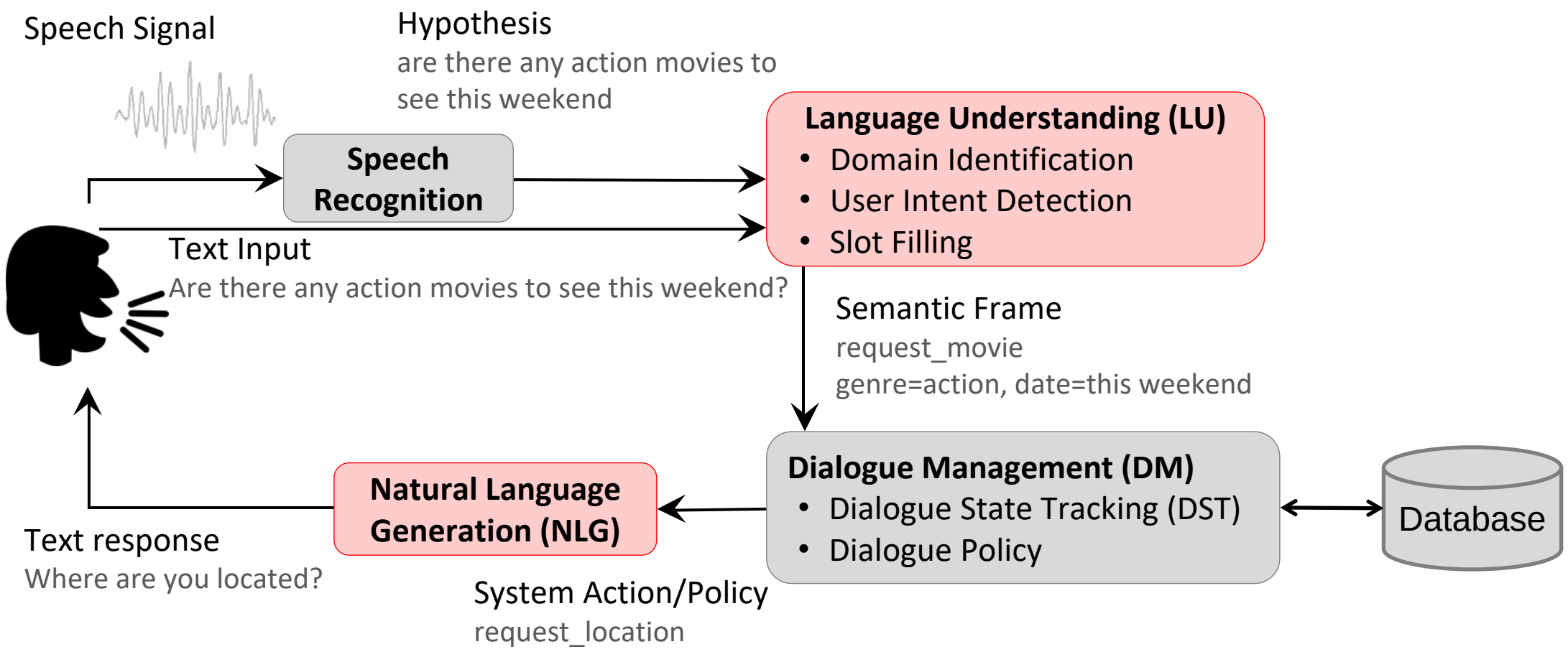
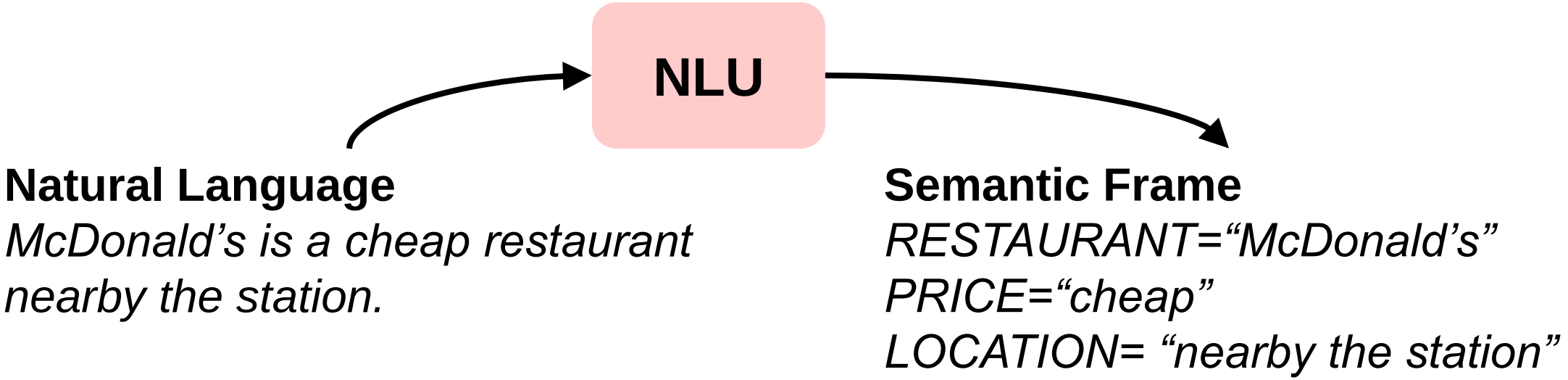
Task-Oriented Dialogue Systems (Young, 2000)



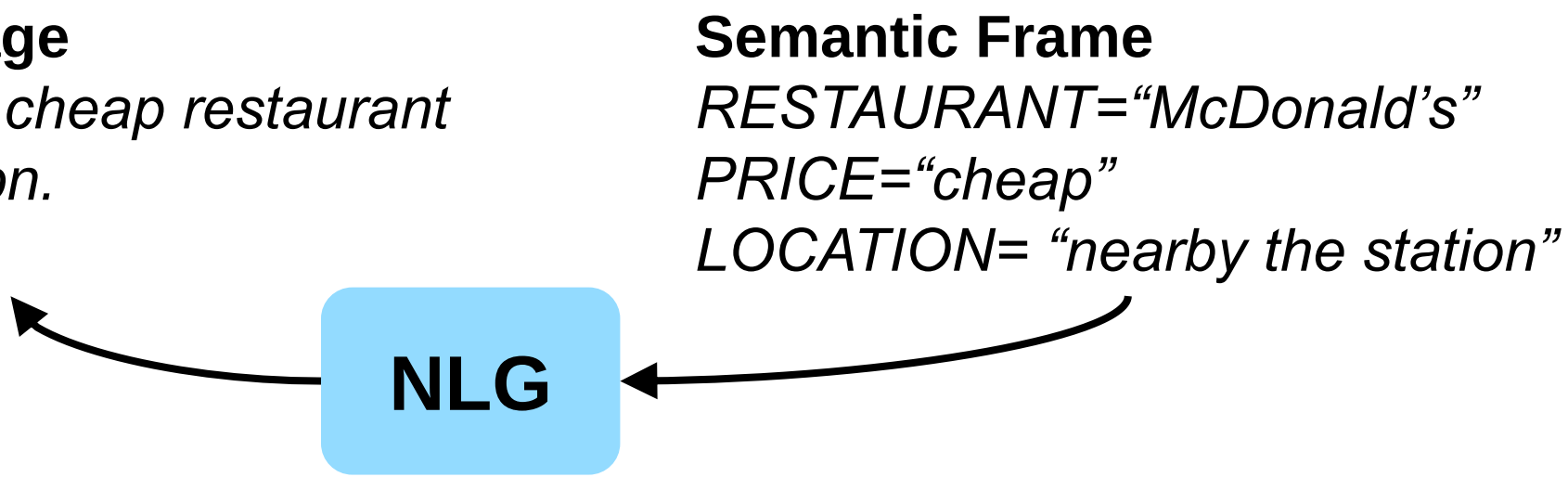
Natural Language Understanding (NLU)

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

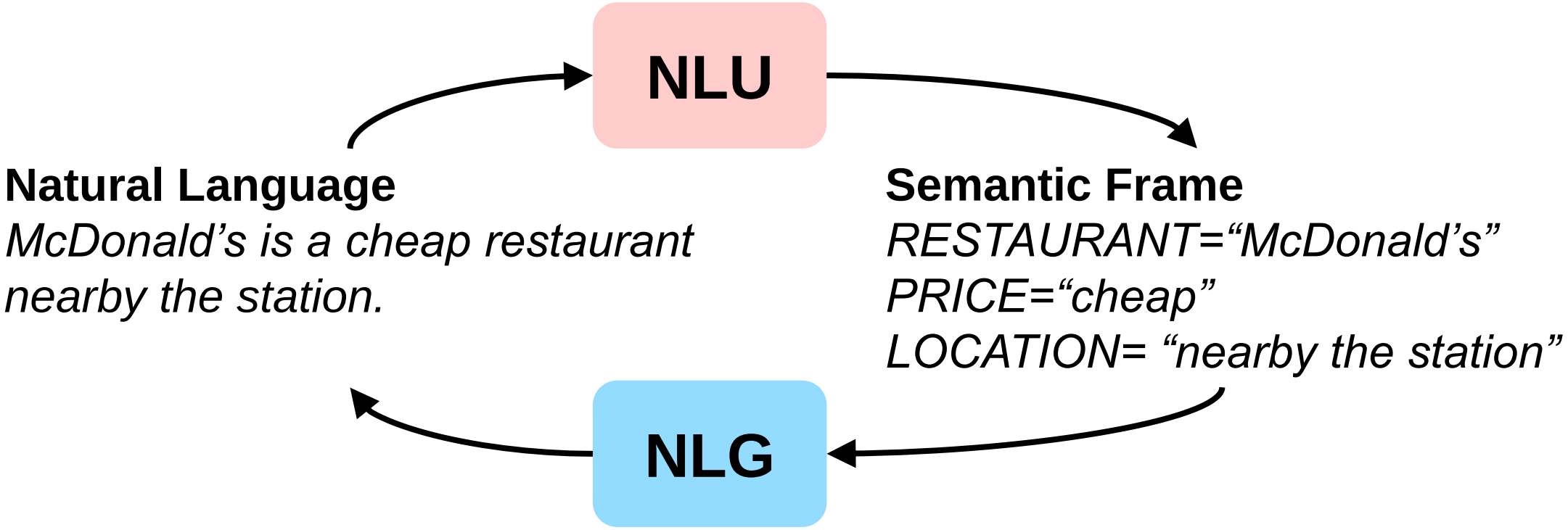


Natural Language Generation (NLG)

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



Duality between NLU and NLG



How can we leverage this dual relationship?

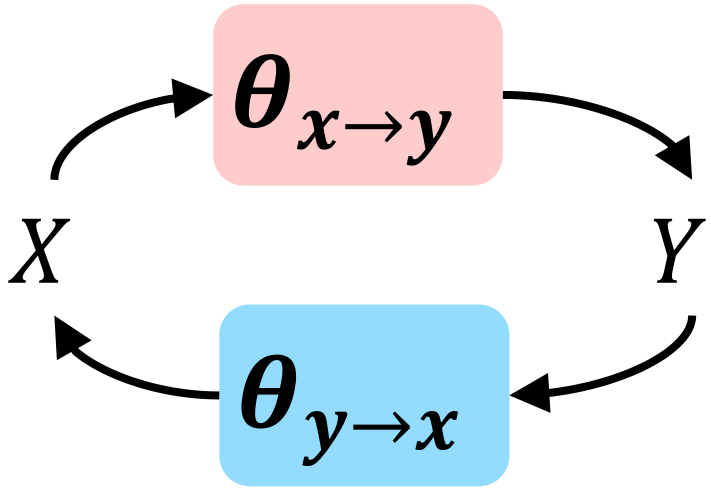
Solution:
Dual Supervised Learning for NLU & NLG

(Su et al., 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 ICML*, 2017.

Dual Supervised Learning

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

Objective function

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

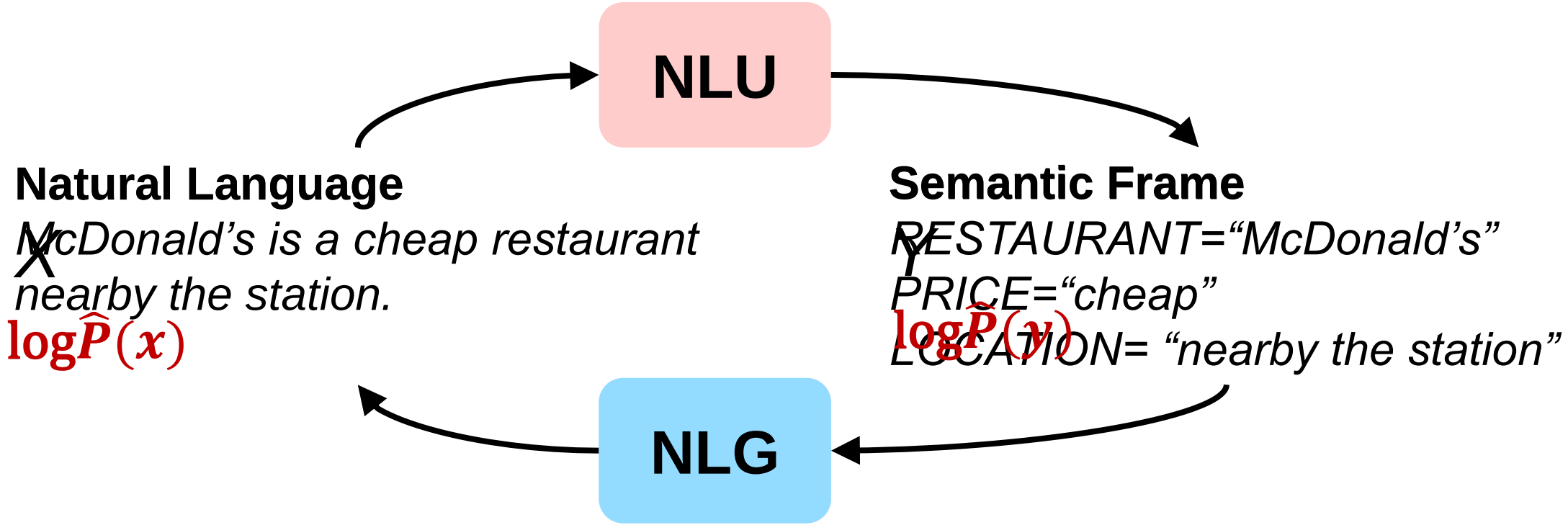
$$l_{duality} = \left(\log \hat{P}(x) + \log P(y \mid x; \theta_{x \to y}) - \log \hat{P}(y) - \log P(x \mid y; \theta_{y \to x})\right)^2$$

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 ICML*, 2017.

Dual Supervised Learning

- Let's go back to NLU and NLG

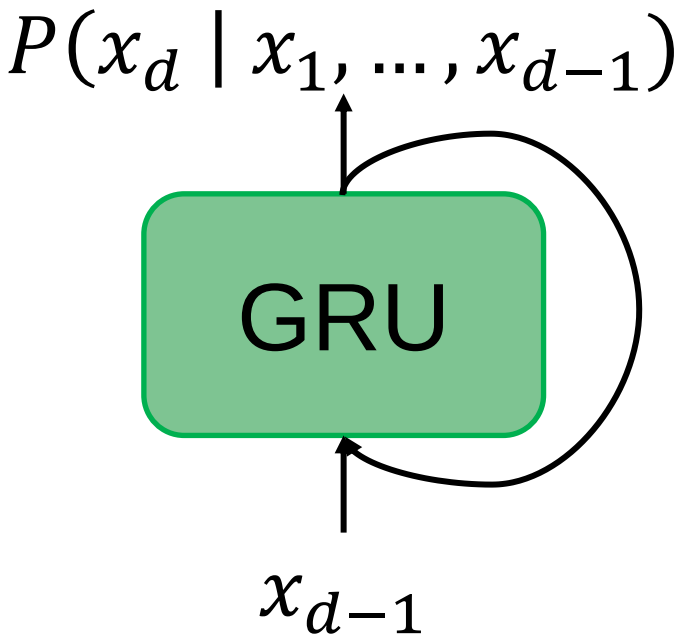


Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, "Dual Supervised Learning for Natural Language Understanding and Generation," in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

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

- Language modeling

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

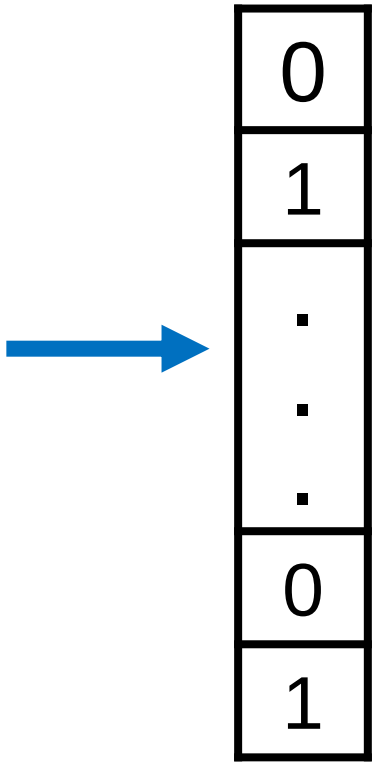


Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, “Dual Supervised Learning for Natural Language Understanding and Generation,” in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

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"



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

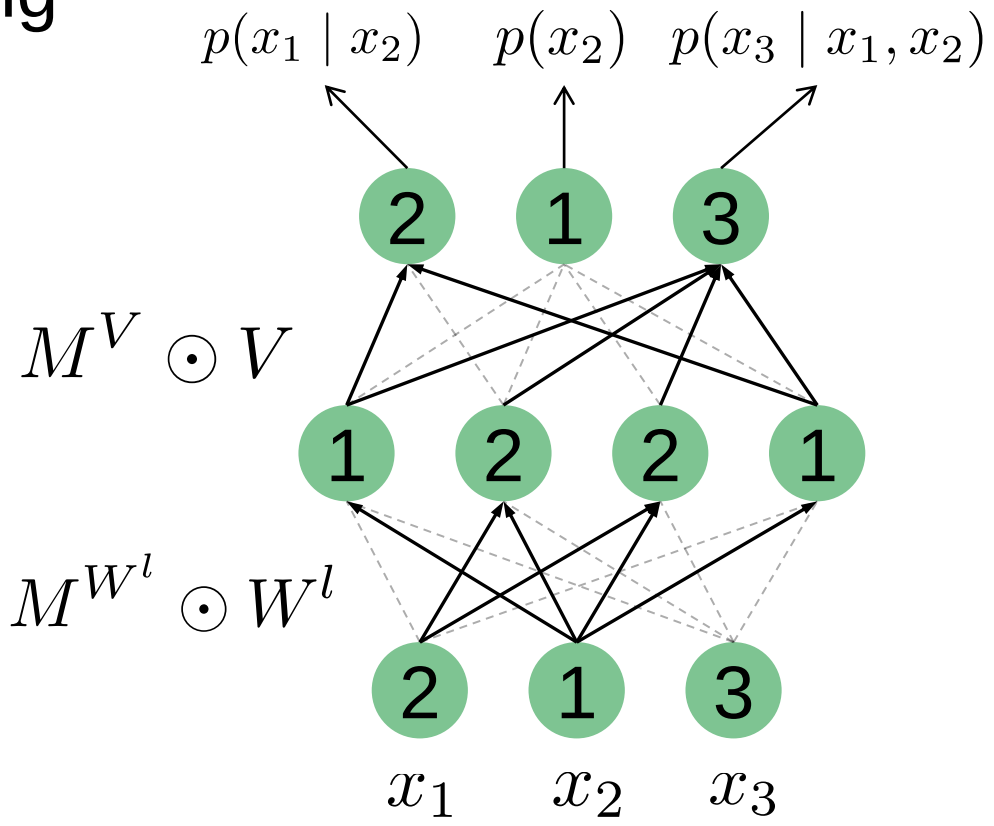
- Masked autoencoder for distribution estimation (MADE)

  Introduce sequential dependency among labels by masking certain connections

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

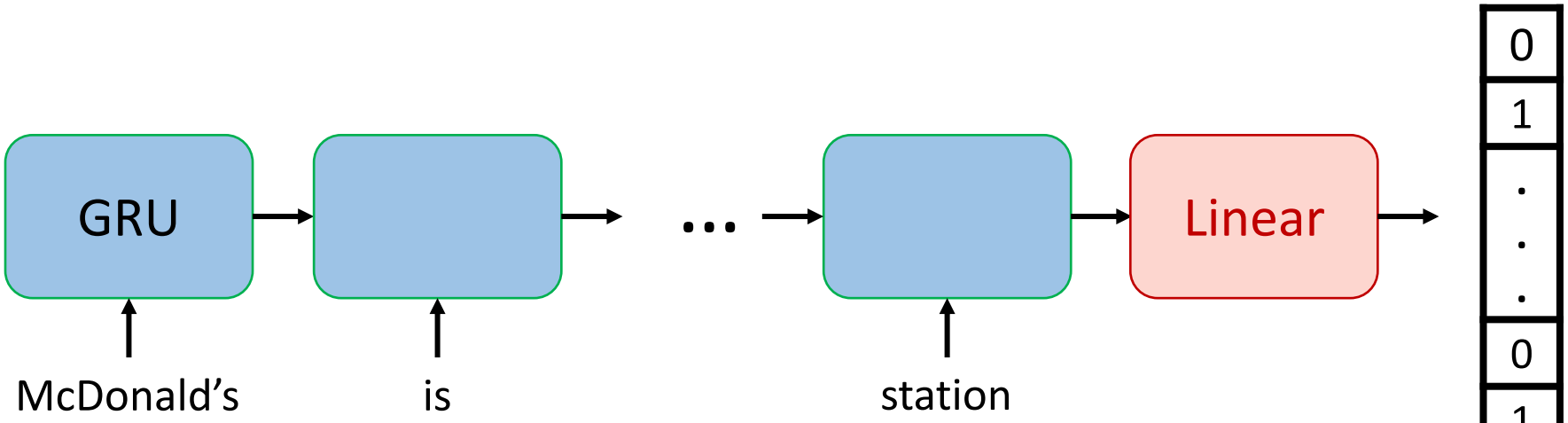
$$p(x) = \prod_d^D p(x_d \mid S_d)$$

→ marginal distribution of y

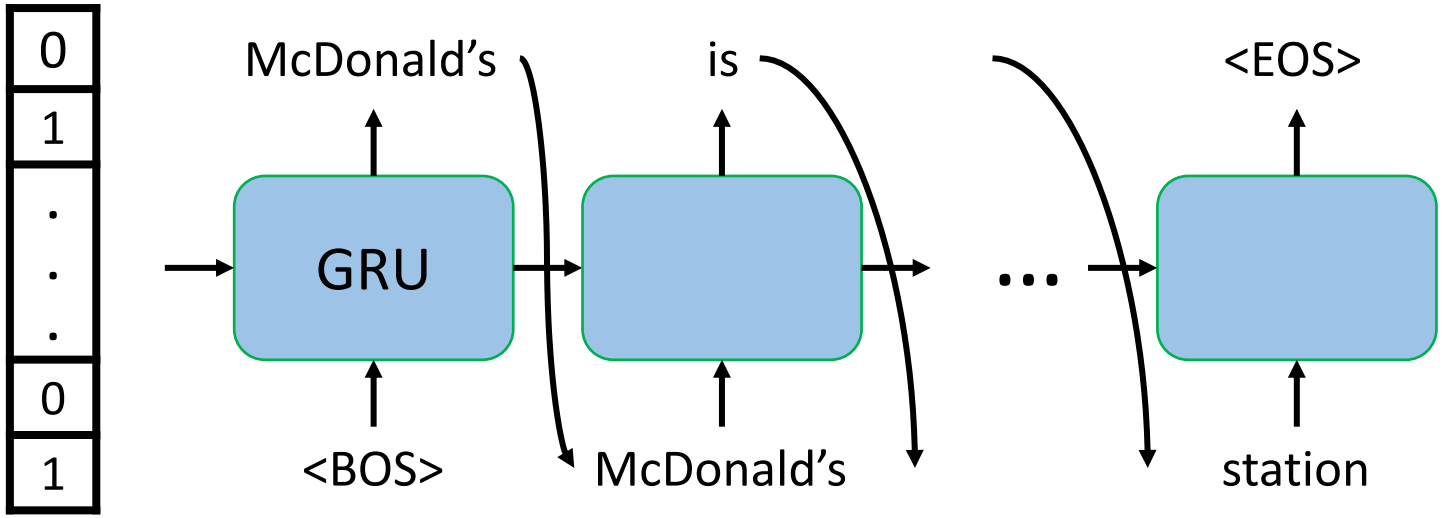


Germain, M., Gregor, K., Murray, I., & Larochelle, H., "MADE: Masked autoencoder for distribution estimation," in *Proceedings of International Conference on Machine Learning*, 2015.

NLU



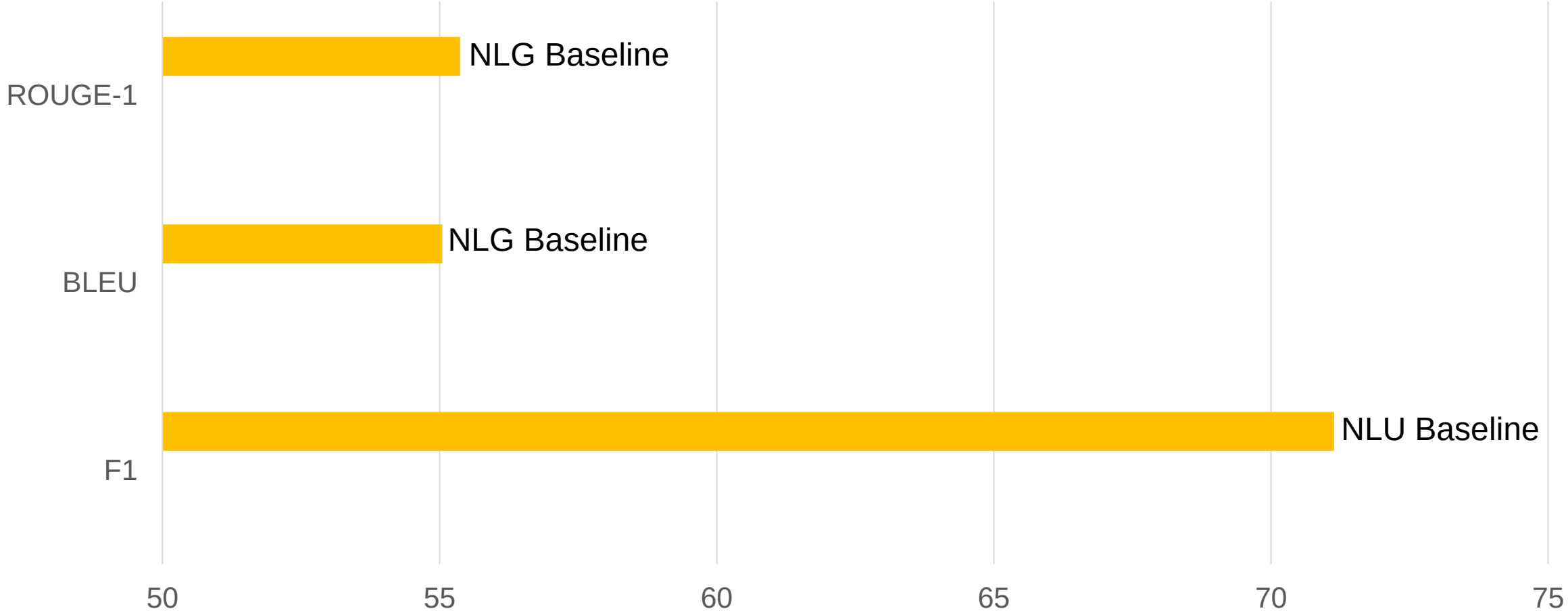
NLG



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.

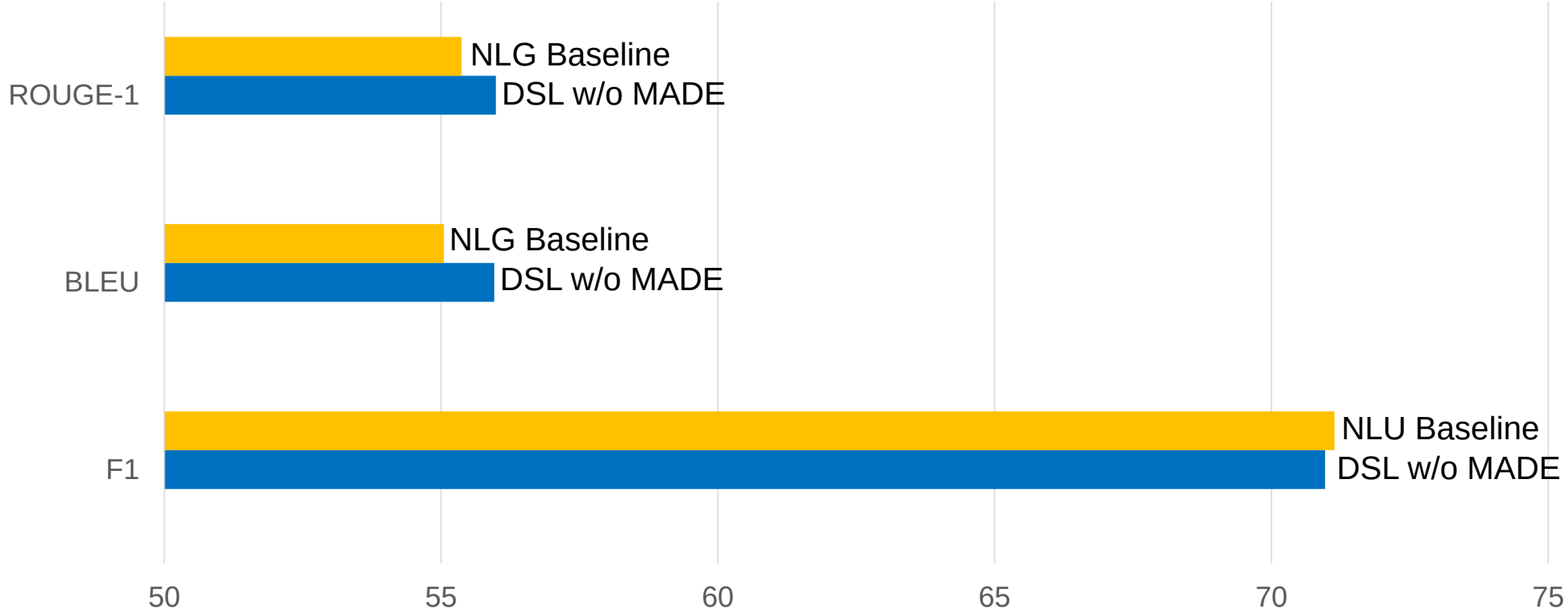
NLU/NLG Results

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



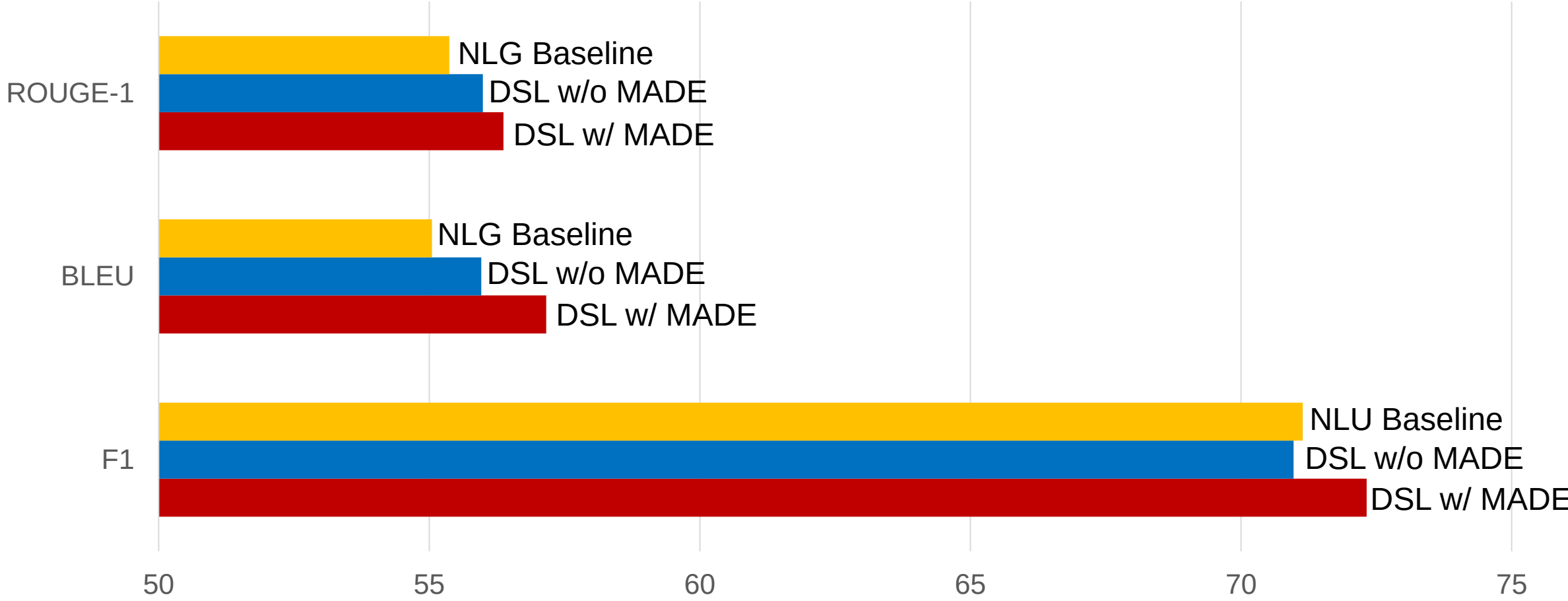
NLU/NLG Results

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

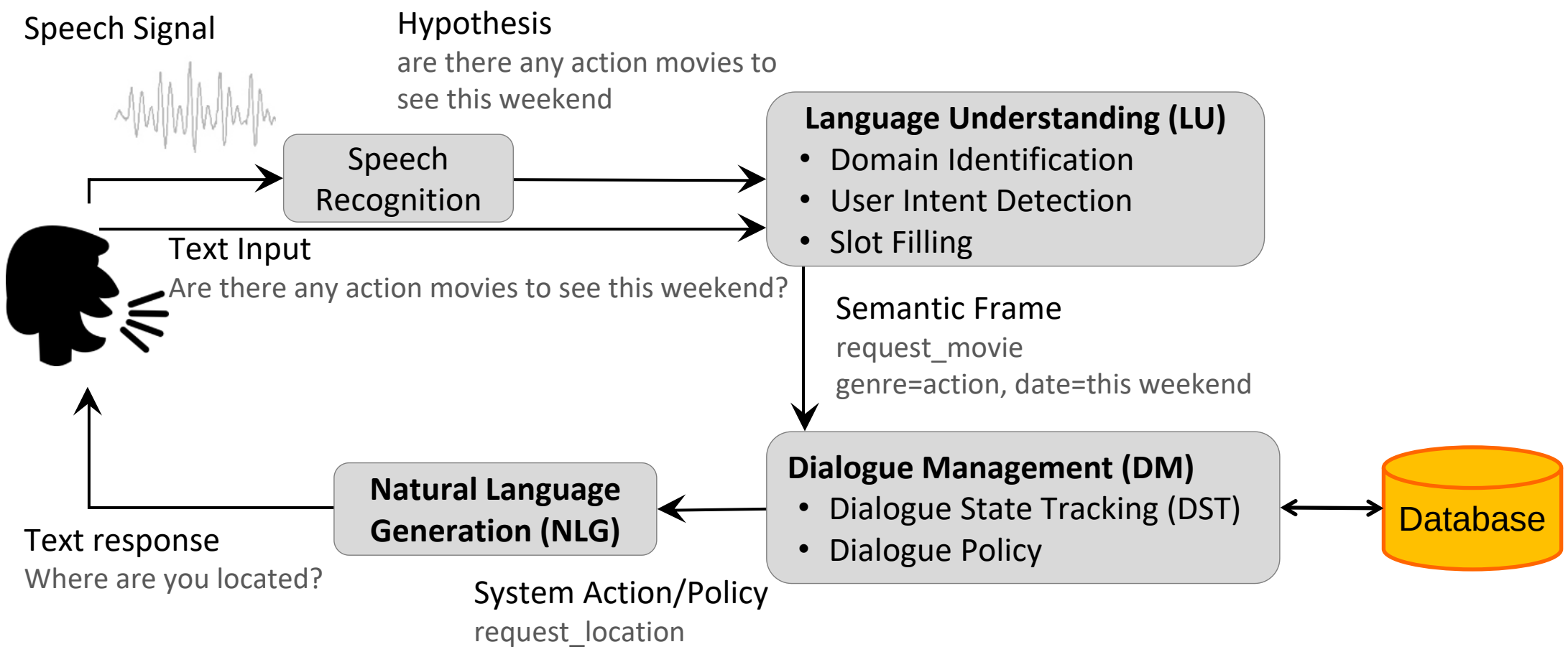


NLU/NLG Results

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

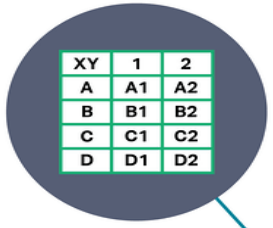


Task-Oriented Dialogue Systems (Young, 2000)

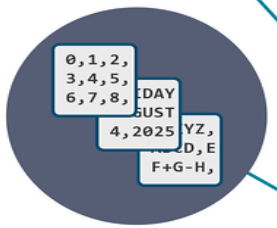


Structured Data vs Unstructured Data

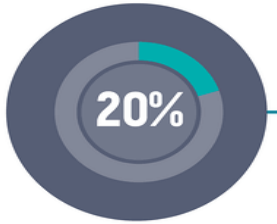
Can be displayed in rows, columns and relational databases



Numbers, dates and strings

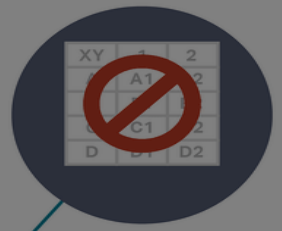


Estimated 20% of enterprise data *(Gartner)*



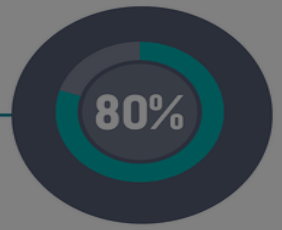
Requires less storage

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

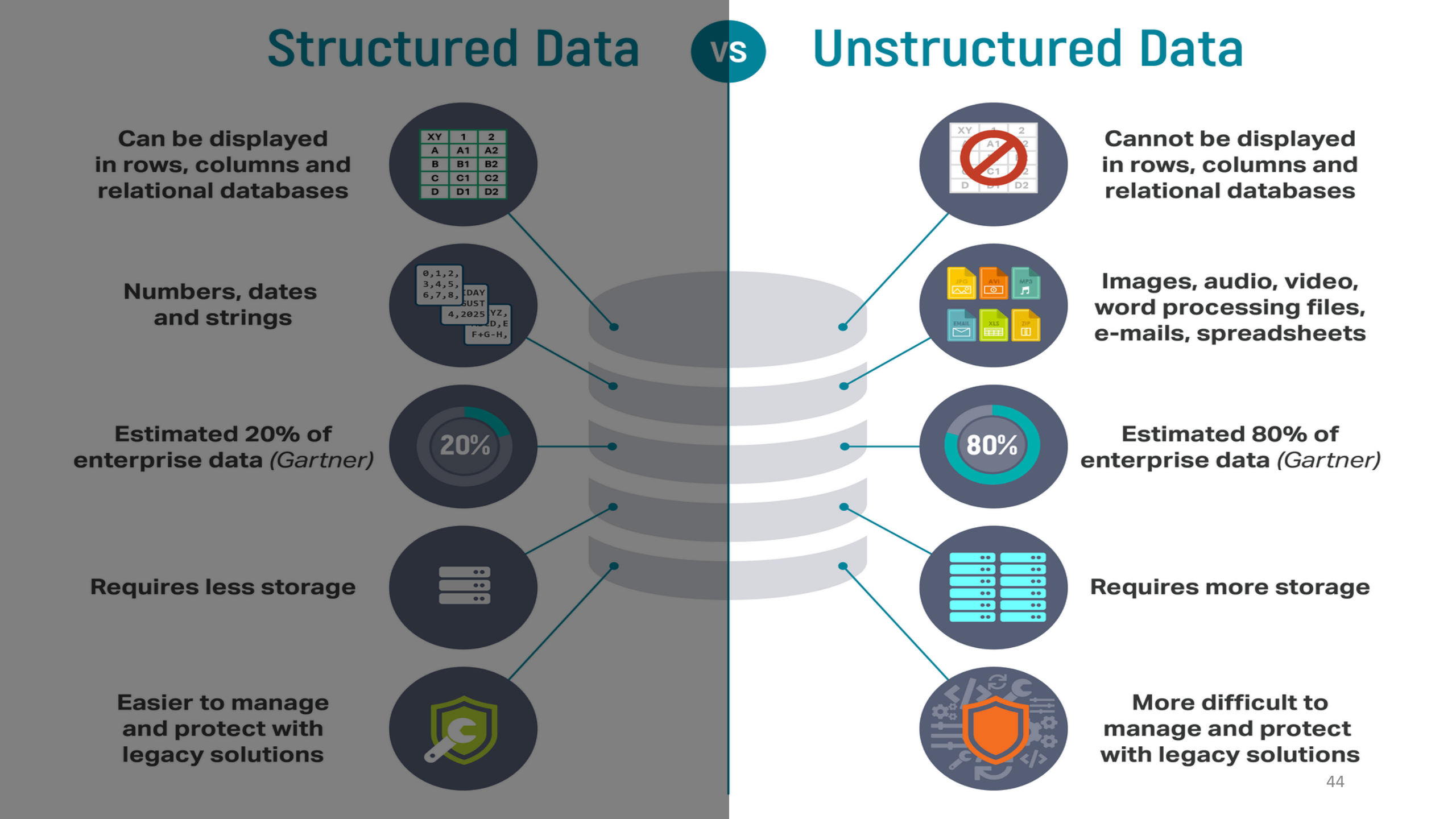


Estimated 80% of enterprise data *(Gartner)*

Requires more storage

More difficult to manage and protect with legacy solutions

Structured Data vs Unstructured Data



Unstructured Knowledge Access

- 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
. . .

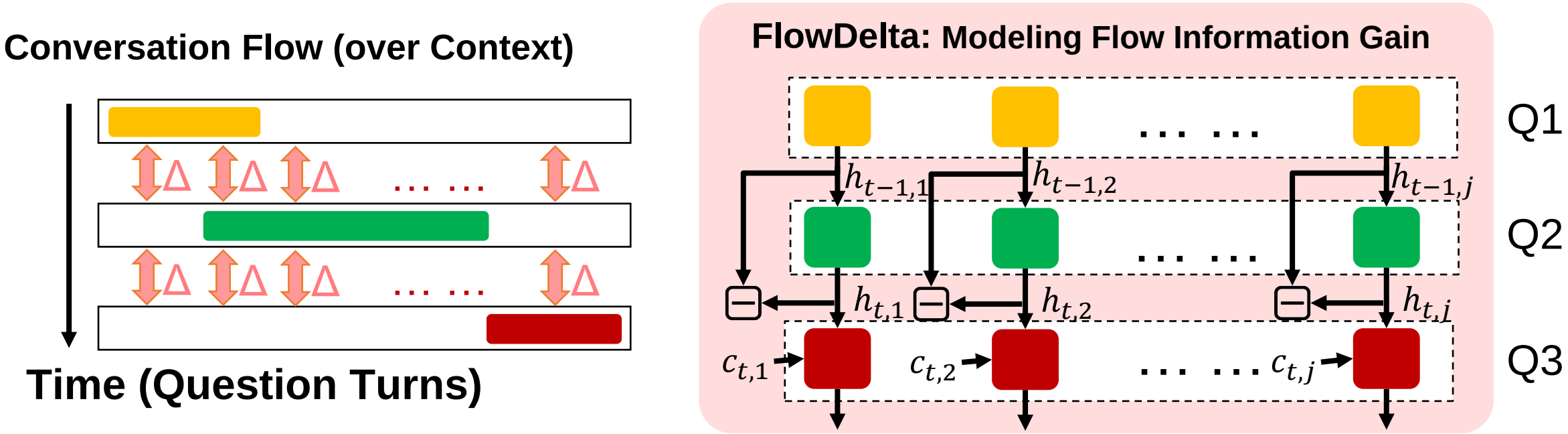
Solution: FlowDelta

(Yeh & Chen, 2019)

Yi-Ting Yeh and Yun-Nung Chen, "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension," in *Proceedings of Machine Reading for Question Answering Workshop at EMNLP (MRQA),* 2019.

FlowDelta: Information Gain in Dialogue Flow

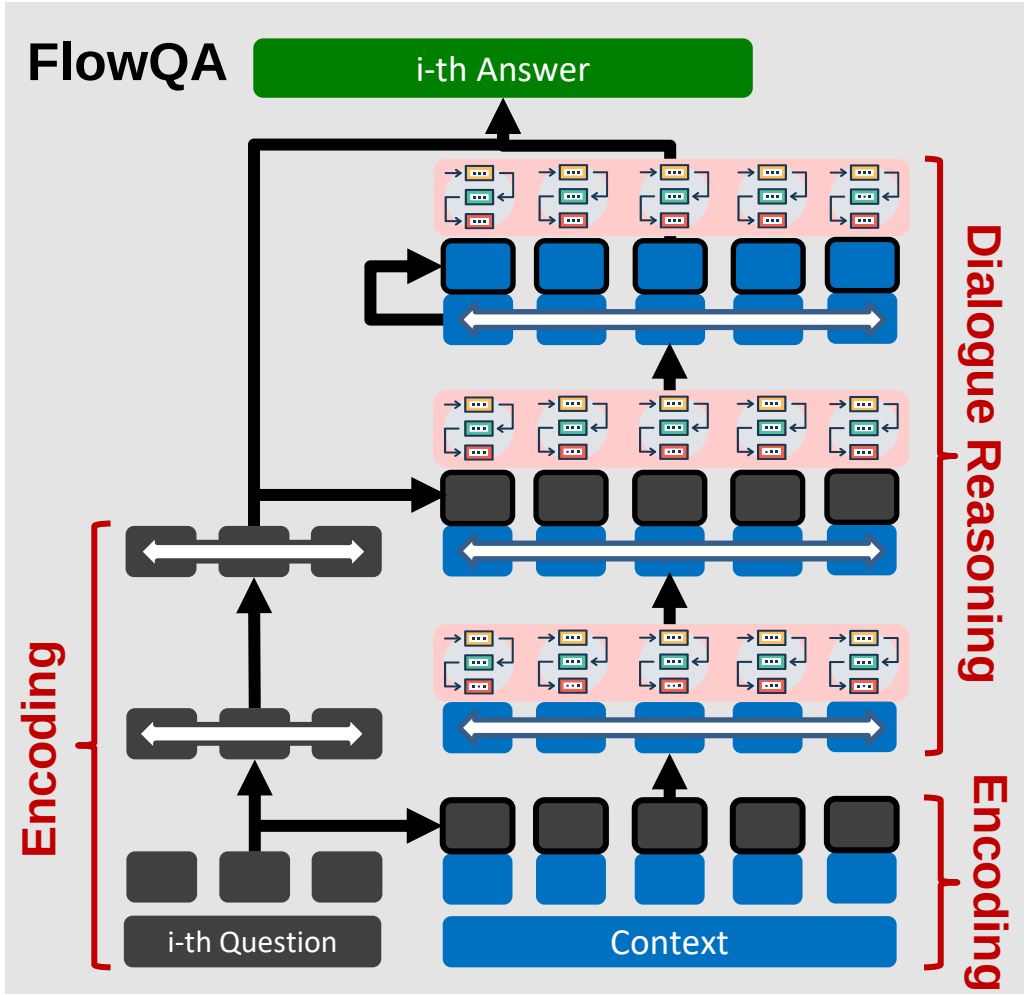
- Idea: model the *difference* of hidden states in multi-turn dialogues

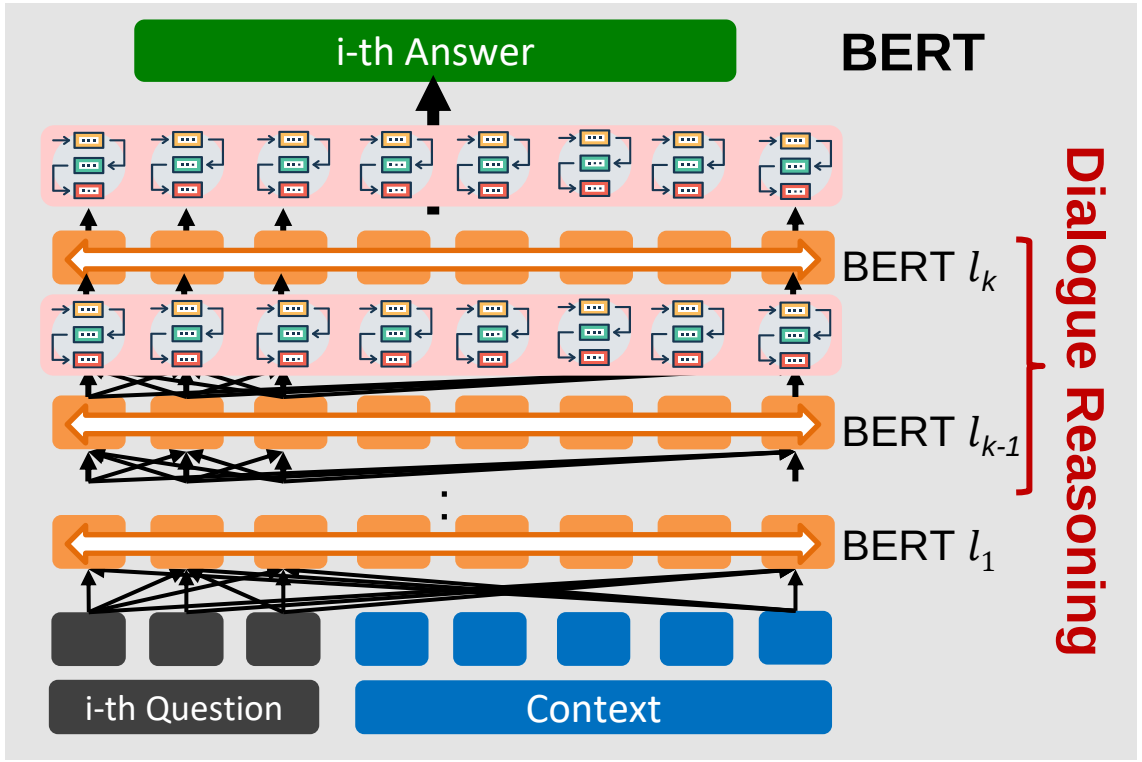


Yi-Ting Yeh and Yun-Nung Chen, "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension," in *Proceedings of Machine Reading for Question Answering Workshop at EMNLP (MRQA)*, 2019.

FlowDelta (Yeh & Chen, 2019)

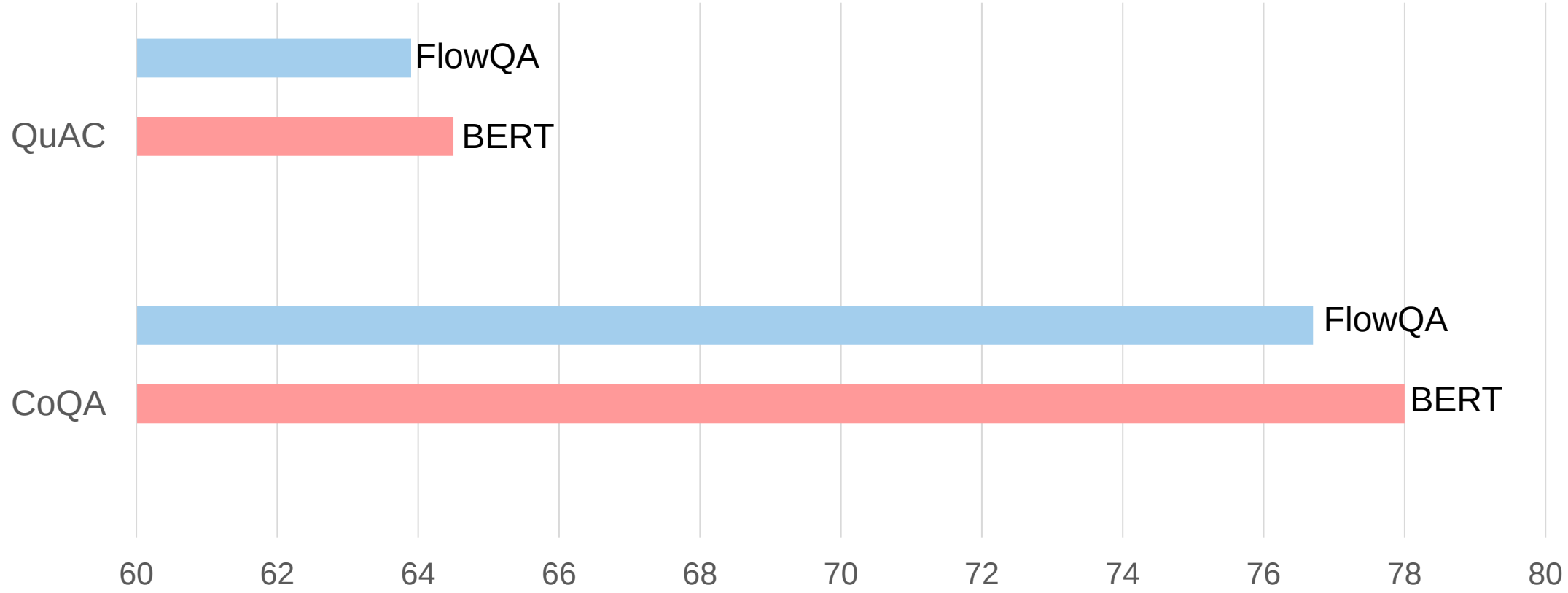
- Idea: model the *difference* of hidden states in multi-turn dialogues





Conversational QA Results

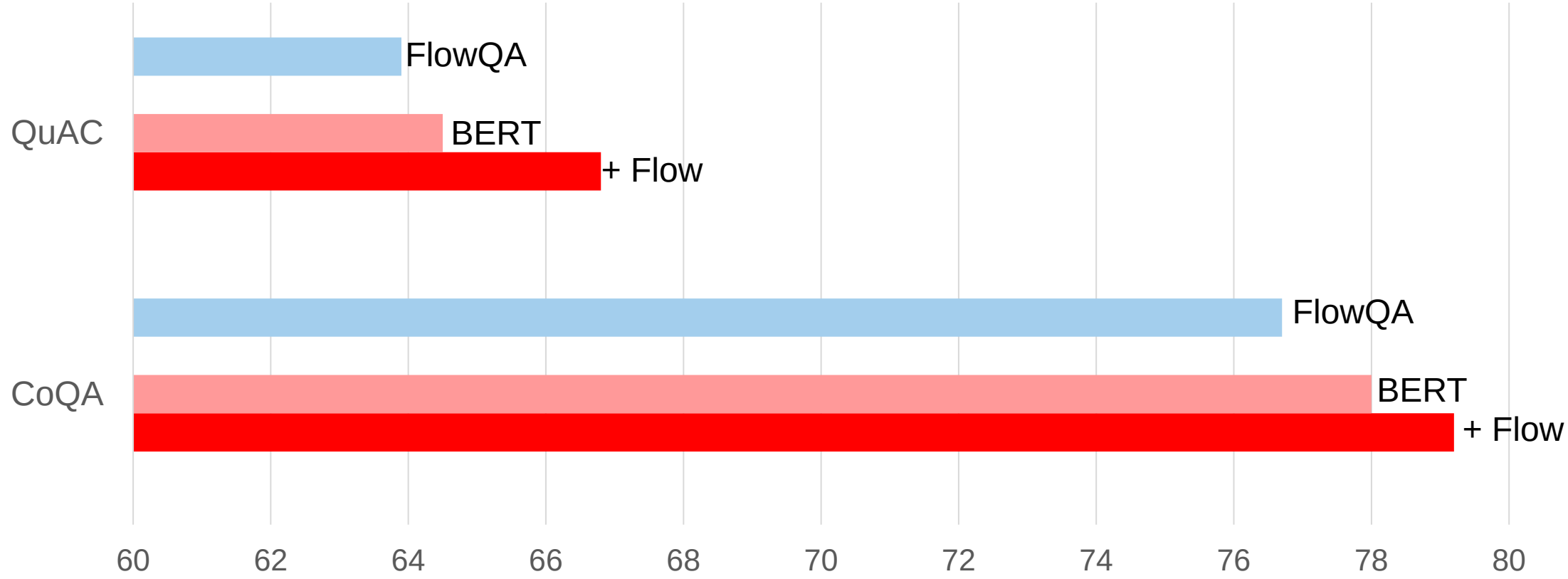
- Data: QuAC, CoQA



Yi-Ting Yeh and Yun-Nung Chen, "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension," in *Proceedings of Machine Reading for Question Answering Workshop at EMNLP (MRQA)*, 2019.

Conversational QA Results

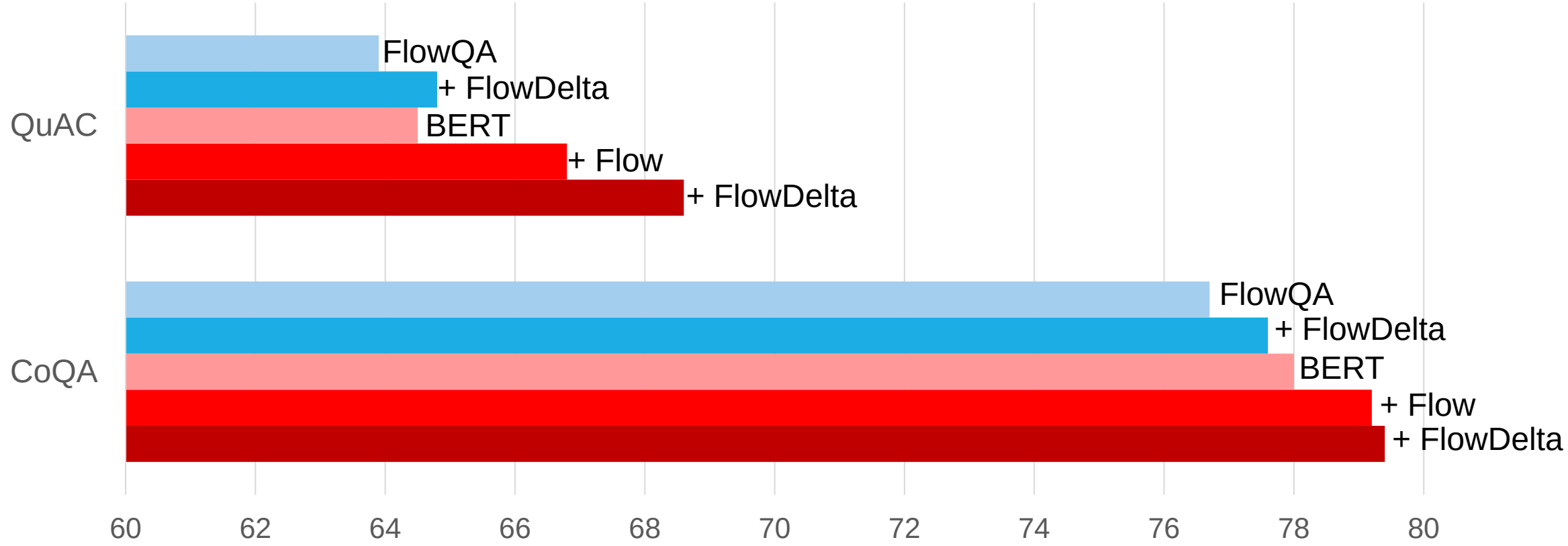
- Data: QuAC, CoQA



Yi-Ting Yeh and Yun-Nung Chen, "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension," in *Proceedings of Machine Reading for Question Answering Workshop at EMNLP (MRQA)*, 2019.

Conversational QA Results

- Data: QuAC, CoQA



Yi-Ting Yeh and Yun-Nung Chen, "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension," in *Proceedings of Machine Reading for Question Answering Workshop at EMNLP (MRQA)*, 2019.

QuAC Leaderboard

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
 Sep 15, 2019	History-Attentive-TransBERT (single model) *Alibaba AI Labs*	**72.9**	**69.7**	**13.6**
2 Aug 31, 2019	TransBERT (single model) *Anonymous*	71.4	68.1	10.0
3 Apr 24, 2019	Bert-FlowDelta (single model) *National Taiwan University, MiuLab* https://arxiv.org/abs/1908.05117	67.8	63.6	12.1
4 June 13, 2019	Context-Aware-BERT (single model) *Anonymous*	69.6	65.7	8.1
5 Aug 22, 2019	BertMT (single model) *WeChat AI*	68.9	65.2	8.9
6 Sep 9, 2019	BertInfoFlow (single model) *PINGAN Omni-Sinitic*	69.3	65.2	8.5

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

- Leveraging the duality of NLU and NLG improves the scalability
  - Apply dual supervised learning to leverage the duality
  - Data distribution property is important
  - Better performance and flexibility for diverse NLU/NLG models
- **Conversational QA** enables unstructured information access
  - **FlowDelta: information gain in dialogue flow** guides better understanding