Exploiting the Duality between Language Understanding and Generation and Beyond

Shang-Yu Su 蘇上育

Advisor: Yun-Nung (Vivian) Chen 陳縕儂

MIULAB

National Taiwan University 國立臺灣大學

Self-Introduction

- B.S. NTUEE, 2017
- Dialogue Policy Learning
  - IJCNLP (2017), ACL (2018), EMNLP (2018)
- Natural Language Understanding (NLU)
  - ASRU (2017), NAACL-HLT (2018), ICASSP (2019)
- Natural Language Generation (NLG)
  - NAACL (2018), SLT (2018)
- **Duality between NLU and NLG**
  - ACL (2019, 2020), EMNLP (2020)

Outline

- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning
- Summary
- Related work

Training Stage

Inference Stage

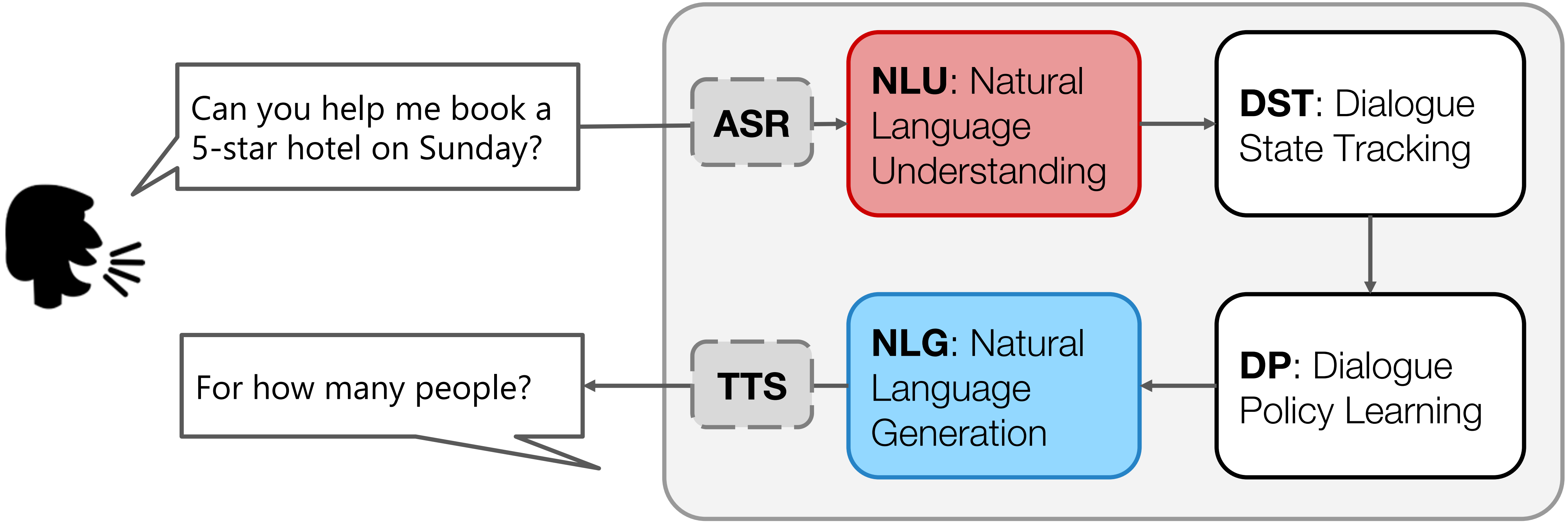
Finetuning Stage

Outline

- **Background**
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning
- Summary
- Related work

Background

- Natural language understanding (NLU) and natural language generation (NLG) are both critical research topics in the NLP and dialogue fields.



Natural Language Understanding (NLU)

- Parse **natural language** into **structured semantics**
- Many-to-one

Natural Language

1. Alimentum city centre is family-friendly.

2. Alimentum is a family-friendly city centre.

Semantic Frame

NAME="Alimentum"
familyFriendly ="yes"
area = "city centre"

Natural Language Generation (NLG)

- Construct **natural language** based on **structured semantics**
- One-to-many

Natural Language

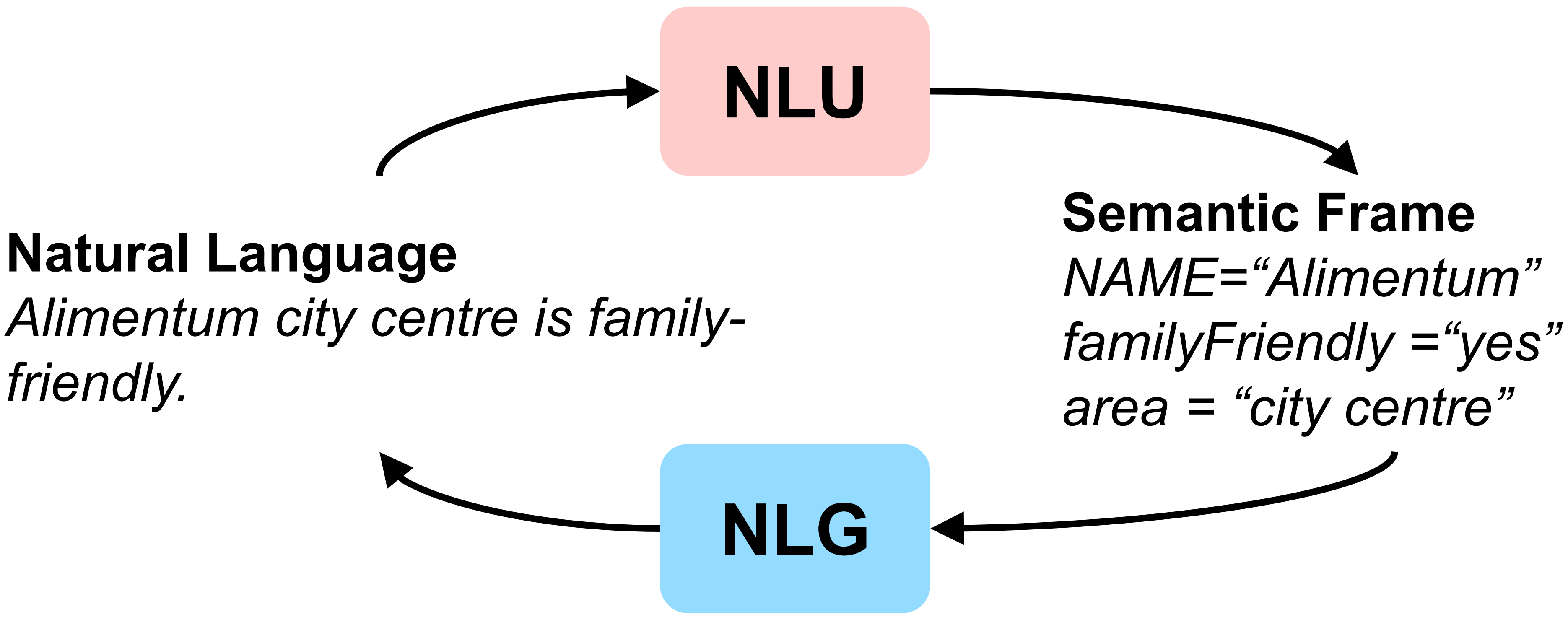
1. Alimentum city centre is family-friendly.

2. Alimentum is a family-friendly city centre.

Semantic Frame

NAME="Alimentum"
familyFriendly ="yes"
area = "city centre"

Duality between NLU and NLG



💡 NLU and NLG are a dual problem pair.

Dual Problems

- Machine Translation
  - English to Chinese, Chinese to English
  - Text vs Text
- Text-to-Speech and Speech Recognition
  - Speech vs Text
- **NLU and NLG**
  - Semantics vs Text
  - NLU is a huge family of tasks
  - Semantic frames/meaning representations are abstract

Problem Formulation

- Given n data pairs $\{(\underline{x_i}, \underline{y_i})\}_{i=1}^{n}$

semantics natural language

$$P(y \mid x; \theta_{x \to y})$$

NLG

$$P(x \mid y; \theta_{y \to x})$$

NLU

Independent Training

$$\min_{\theta_{x \to y}} (\mathbb{E}[l_1(f(x; \theta_{x \to y}), y)])$$
$$\min_{\theta_{y \to x}} (\mathbb{E}[l_2(g(y; \theta_{y \to x}), x)])$$

Outline

- Background
- Duality Exploitation
  - **Dual Supervised Learning**
    - Su et al., ACL 2019
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning (ongoing)
- Summary
- Related work

Training Stage

Probabilistic Duality

- 💡 Idea: bridge the bi-directional relationship from a probabilistic perspective.
- If two models are optimal, we have *probabilistic duality*:

$$
\begin{aligned}
P(x)P(y \mid x; \theta_{x \to y}) &= P(y)P(x \mid y; \theta_{y \to x}) \\
&= P(x, y) \ \forall x, y
\end{aligned}
$$

Objective

- Extended to a multi-objective optimization problem:

$$\begin{cases} \min_{\theta_{x\rightarrow y}} (\mathbb{E}[l_1(f(x;\theta_{x\rightarrow y}), y)]) \\ \min_{\theta_{y\rightarrow x}} (\mathbb{E}[l_2(g(y;\theta_{y\rightarrow x}), x)]) \\ \text{s.t. } P(x)P(y \mid x;\theta_{x\rightarrow y}) = P(y)P(x \mid y;\theta_{y\rightarrow x}) \end{cases}$$

Dual Supervised Learning (Xia et al., 2017)

- The standard supervised learning with an additional regularization term considering the duality between tasks.

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

$$\begin{aligned} l_{duality} = & (\log \hat{P}(x) + \log P(y \mid x;\theta_{x\to y}) \\ & - \log \hat{P}(y) - \log P(x \mid y;\theta_{y\to x}))^2. \end{aligned}$$

Dual Supervised Learning

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

Marginal Distribution

Conditional Distribution

? How to estimate the marginals?

Distribution Estimation as Autoregression

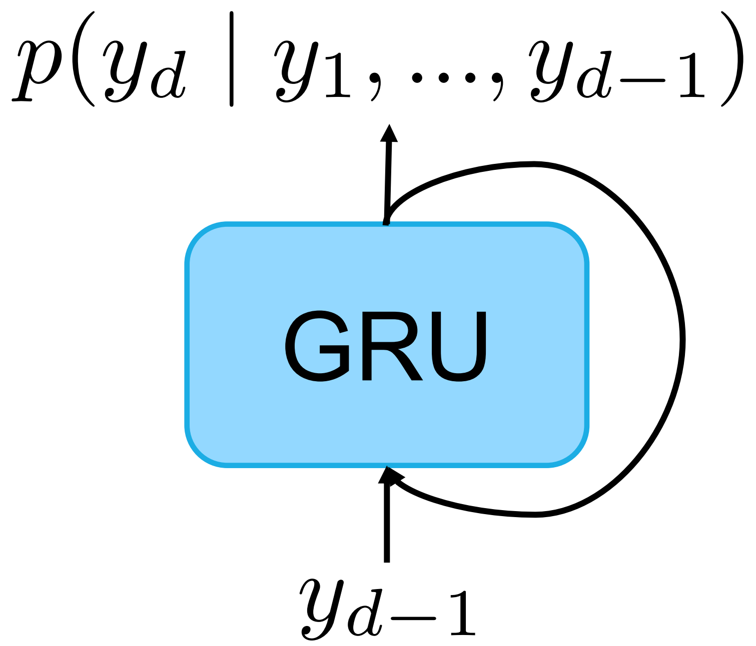
- Decompose any data distribution p(x) into the product of its nested conditional probability:

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

Natural Language

- Language has an intrinsic *sequential* nature
- Language modeling leverages the autoregressive property

$$\hat{P}(y) = \prod_{d}^{D} p(y_d \mid y_1, \ldots, y_{d-1})$$



Semantic Frames

Language

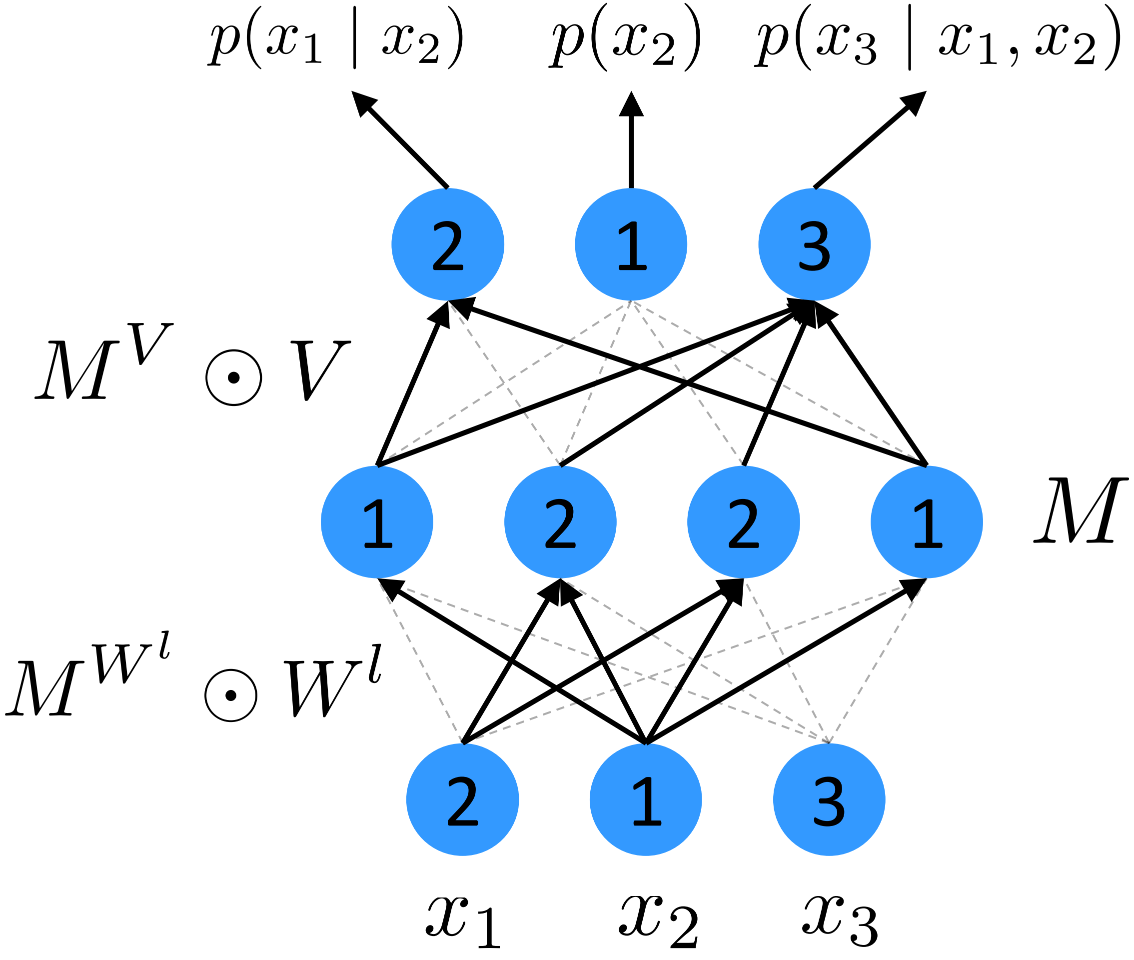
Bibimbap House is a moderately priced restaurant who's main cuisine is English food.
You will find this local gem near Clare Hall in the Riverside area.

Semantics

name[Bibimbap House], food[English], priceRange[moderate], area [riverside], near[Clare Hall]

💡 no uni-directional sequential relationship.

Masked Autoencoder (Germain et al., 2015)



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

Masked Autoencoder

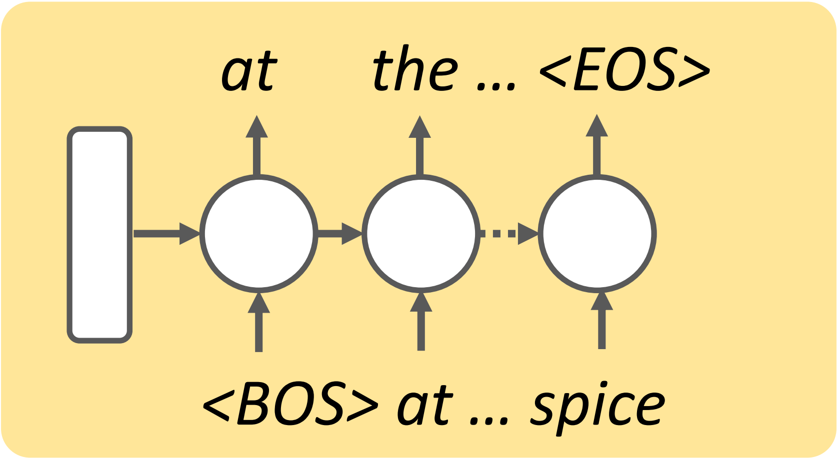
- Marginal distribution by product rule:

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

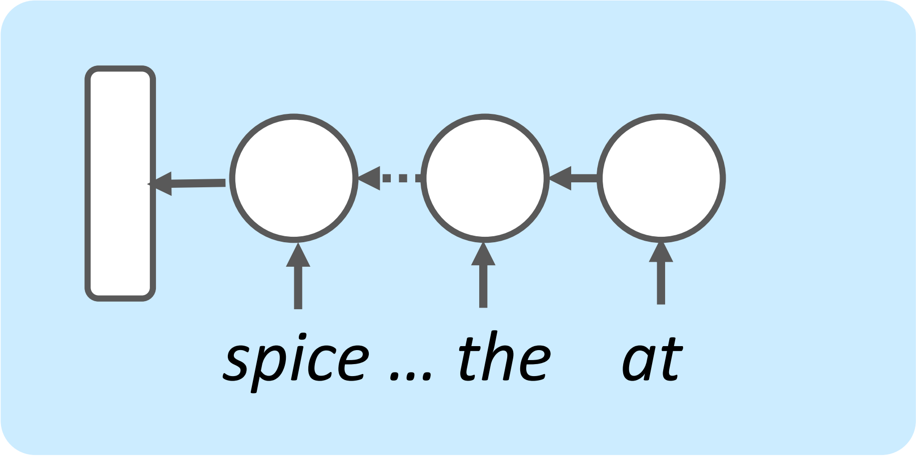
  - Note: no explicit rule specifying the exact dependencies between slot-value pairs in our data, we consider various dependencies via ensemble of multiple decomposition

Experiments

- Dataset: **E2E NLG** (restaurant domain)
- Model: GRU with identical fully-connected layers at two ends



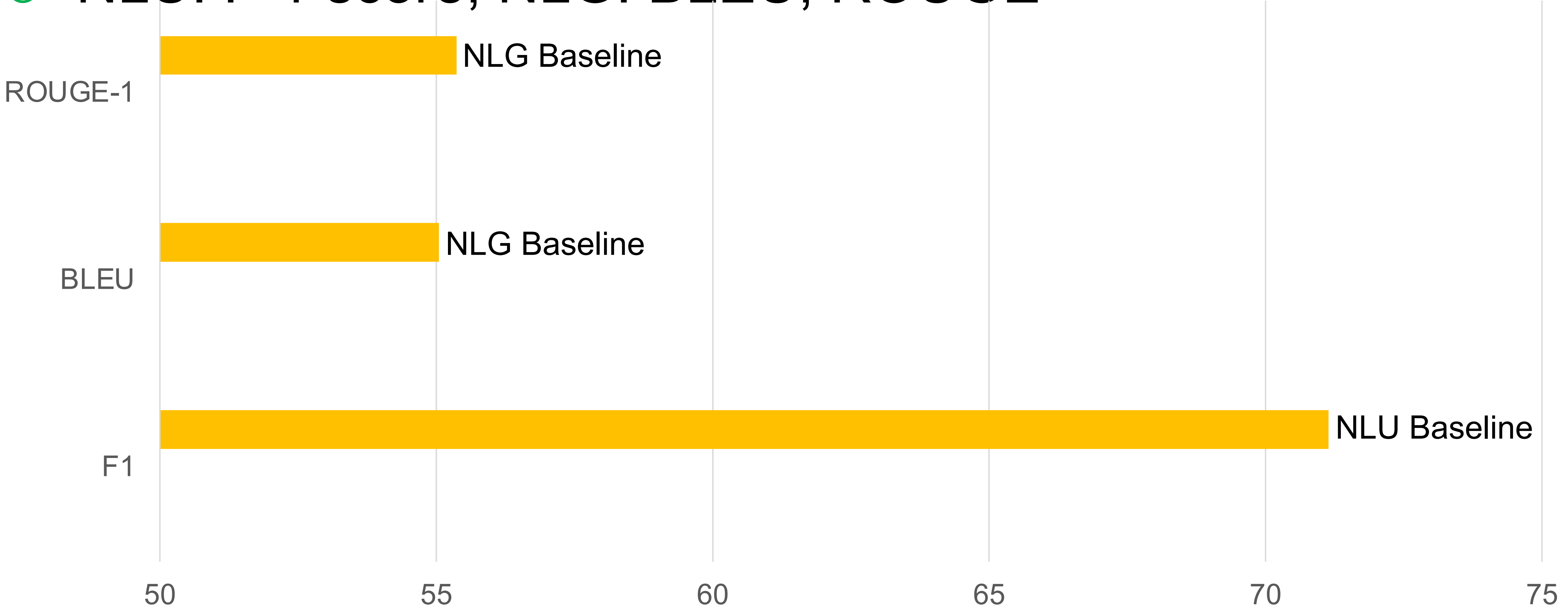
NLG



NLU

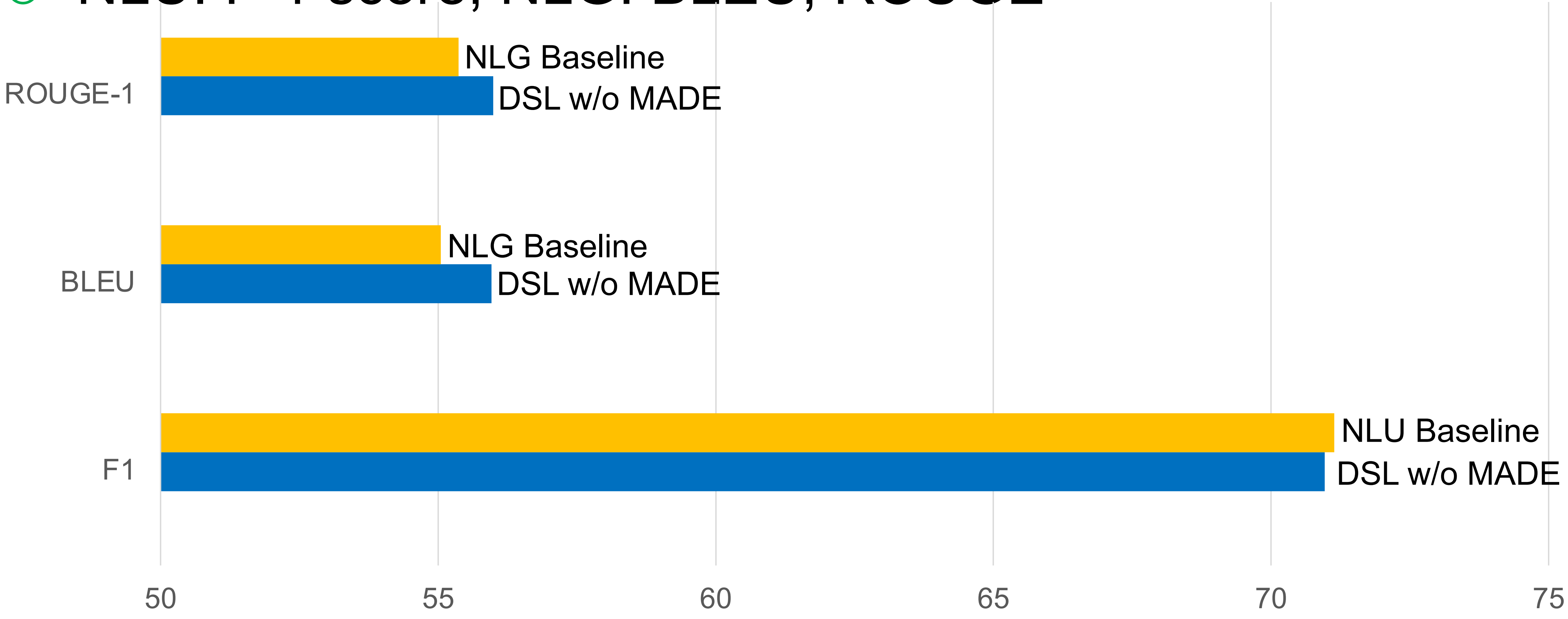
NLU/NLG Results

- NLU: F-1 score; NLG: BLEU, ROUGE



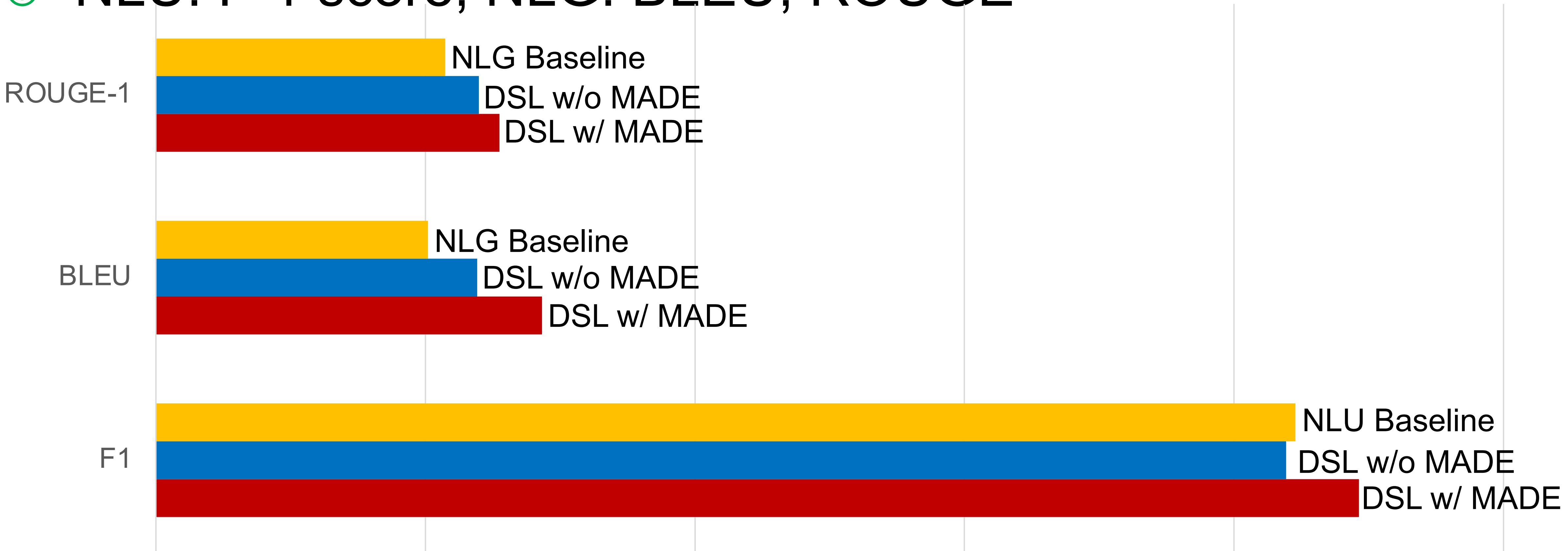
NLU/NLG Results

- NLU: F-1 score; NLG: BLEU, ROUGE



NLU/NLG Results

- NLU: F-1 score; NLG: BLEU, ROUGE



- ✓ Introducing a duality loss as the regularization term is useful
- ✓ Domain knowledge is introduced for estimating data distribution

Outline

- Background
- Duality Exploitation
  - Dual Supervised Learning
  - **Joint Dual Learning**
    - Su et al., ACL 2020
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning
- Summary
- Related work

Training Stage

A Step Forward

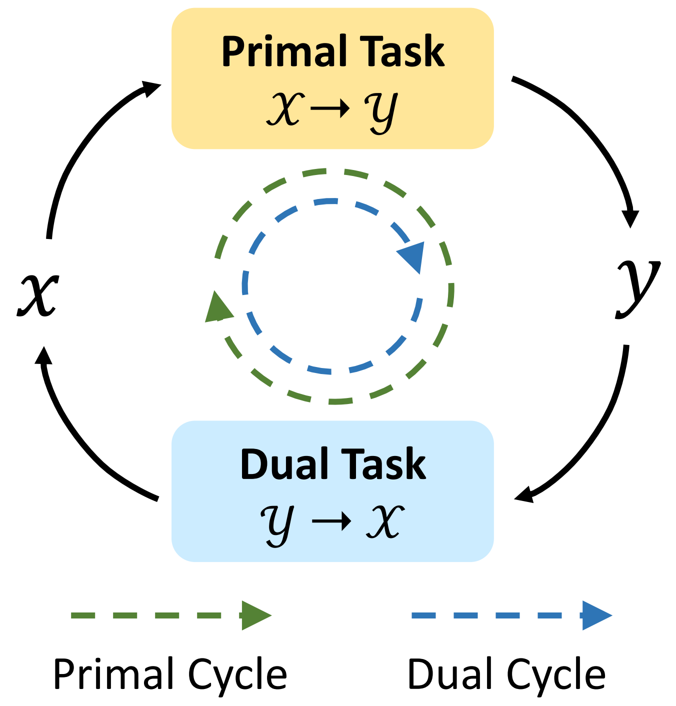
- Prior work learned both models in a *supervised manner.*
- Idea: design a more flexible and general learning framework

💡 Towards *semi-supervised* and *unsupervised* learning

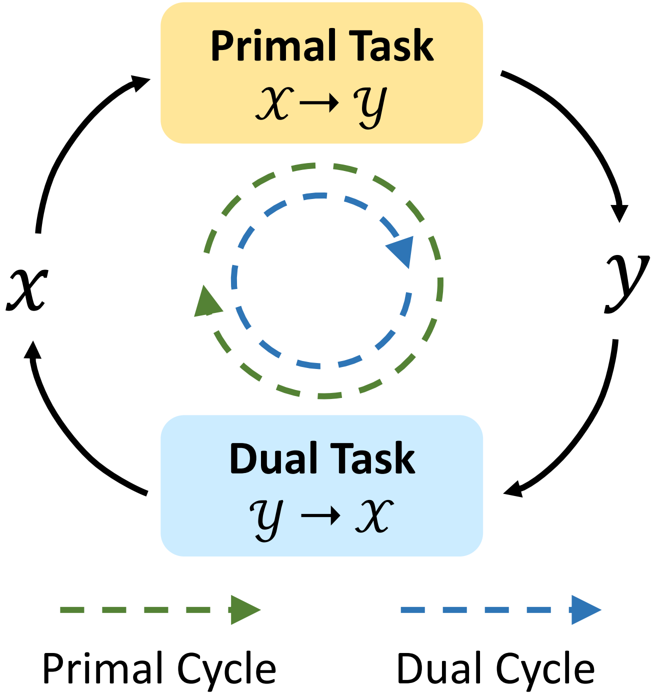
Joint Dual Learning

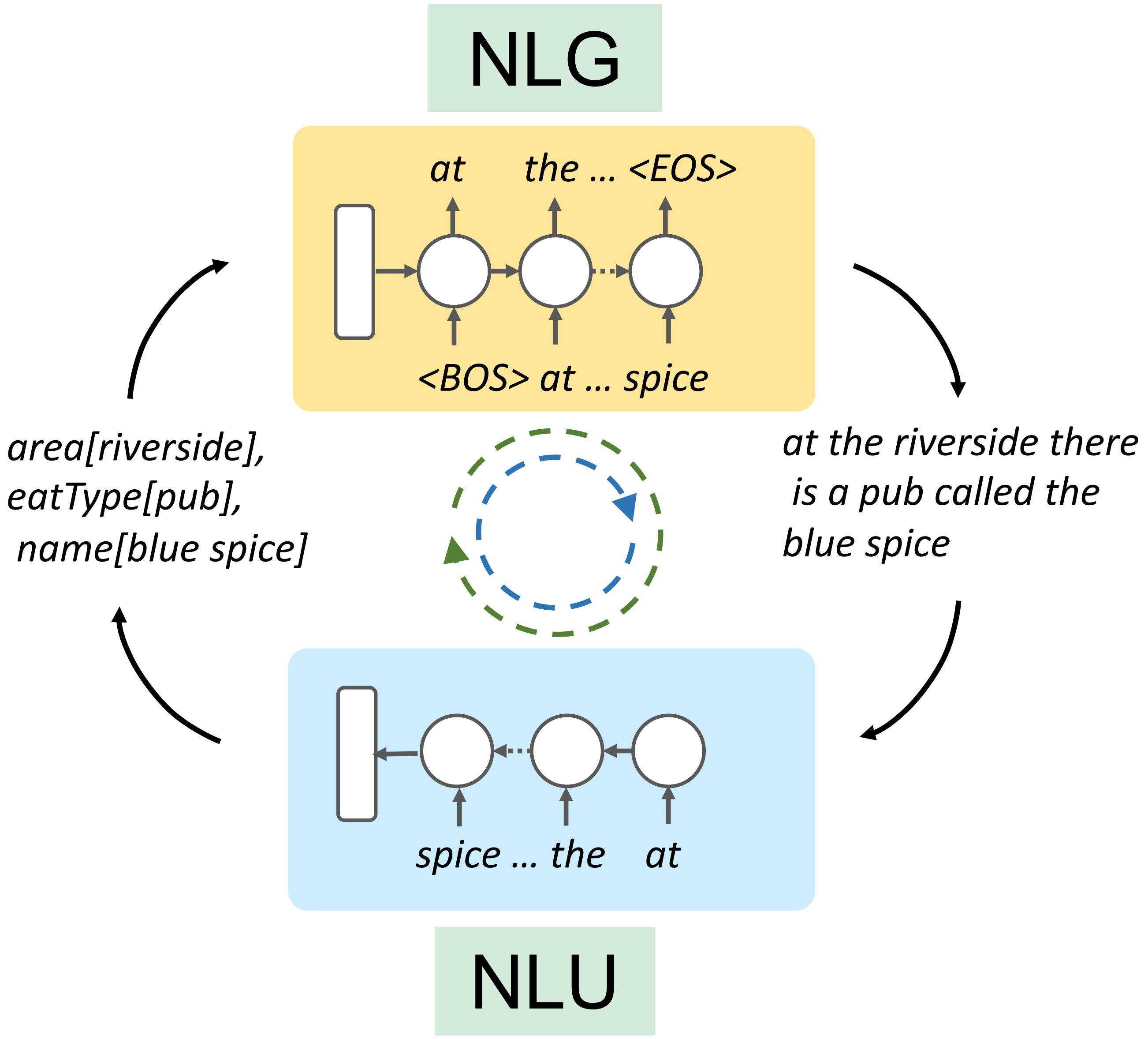
$$f(x) = \arg\max P(y \mid x; \theta_{x \to y})$$
$$g(y) = \arg\max P(x \mid y; \theta_{y \to x})$$

💡 We want cycle-consistency: $g(f(x)) \approx x$



Joint Dual Learning





Primal Cycle

Start from data x, transform x by function f:

$$\hat{y} = f(x; \theta_{x \to y});$$

Compute the loss by $l_1(.)$;

Transform the output of the primal task by function g:

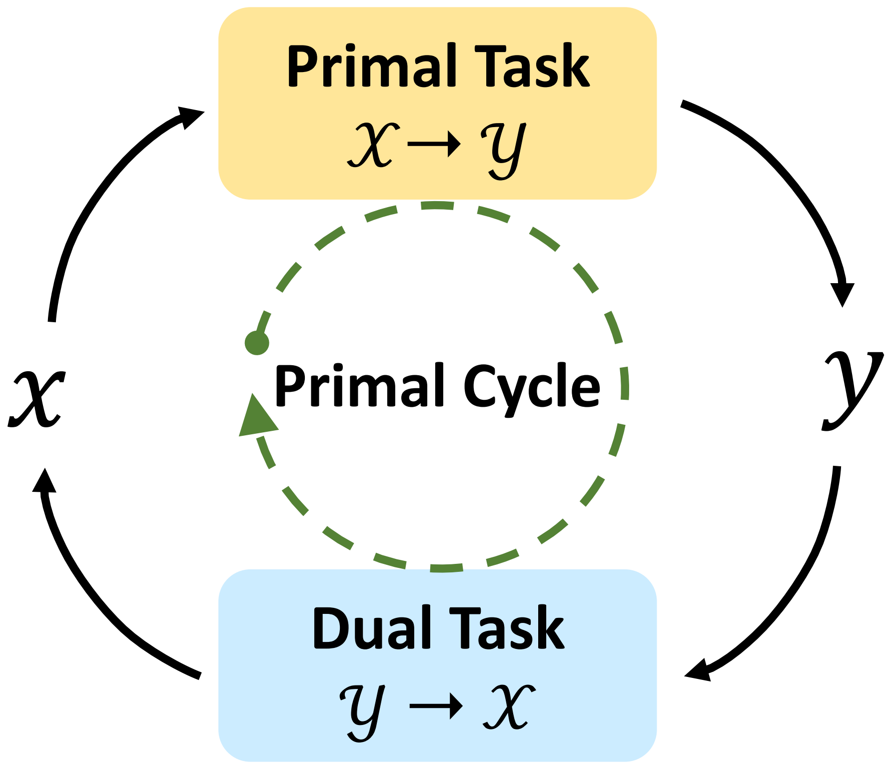
$$\hat{x} = g(\hat{y}; \theta_{y \to x});$$

Compute the loss by $l_2(.)$;

Update model parameters:

$$\theta_{x \to y} \leftarrow \theta_{x \to y} - \gamma_1 \nabla_{\theta_{x \to y}}([l_1(\hat{y}) + l_2(\hat{x})]);$$

$$\theta_{y \to x} \leftarrow \theta_{y \to x} - \gamma_2 \nabla_{\theta_{y \to x}}([l_2(\hat{x})]);$$



Dual Cycle

Start from data y, transform y by function g:

$$\hat{x} = g(y; \theta_{y\to x});$$

Compute the loss by $l_2(.)$;

Transform the output of the dual task by function f:

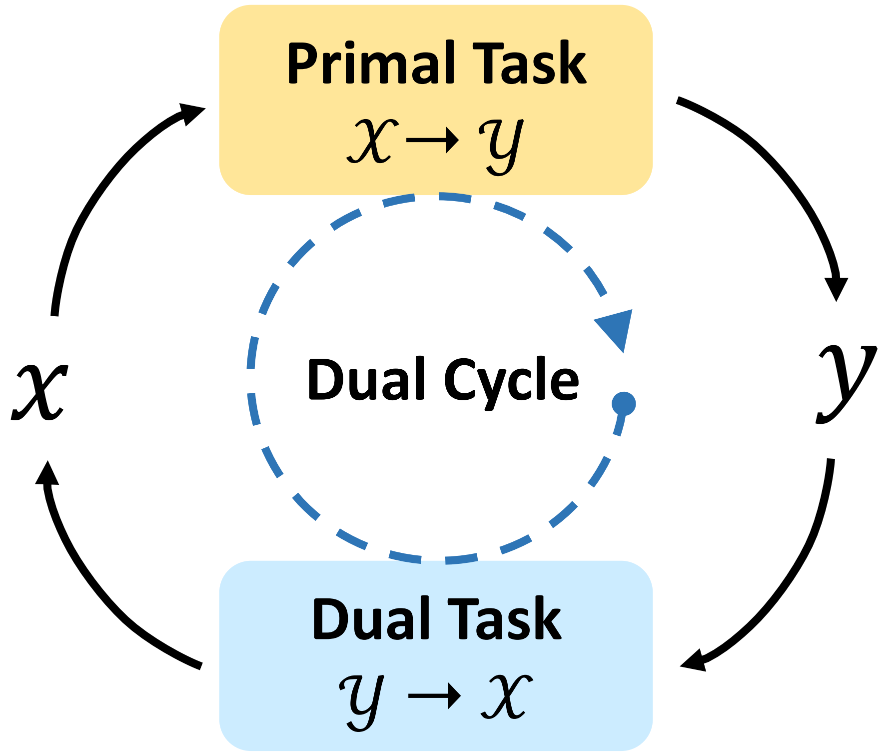
$$\hat{y} = f(\hat{x}; \theta_{x\to y});$$

Compute the loss by $l_1(.)$;

Update model parameters:

$$\theta_{y\to x} \leftarrow \theta_{y\to x} - \gamma_2 \nabla_{\theta_{y\to x}}([l_2(\hat{x}) + l_1(\hat{y})]);$$

$$\theta_{x\to y} \leftarrow \theta_{x\to y} - \gamma_1 \nabla_{\theta_{x\to y}}([l_1(\hat{y})]);$$



Learning Objective

- Loss function: cross entropy, policy gradient (REINFORCE), or their combination

$$\nabla \mathbb{E}[r] = \mathbb{E}[r(y) \nabla \log p(y \mid x)]$$ (Policy Gradient)

- Reward functions
  - Explicit reward
  - Implicit feedback

Explicit Reward

- Reconstruction Likelihood

$$\begin{cases} \log p(x \mid f(x_i; \theta_{x \to y}); \theta_{y \to x}) & \textbf{Primal} \\ \log p(y \mid g(y_i; \theta_{y \to x}); \theta_{x \to y}) & \textbf{Dual} \end{cases}$$

- Automatic Evaluation Score
  - BLEU and ROUGE for language (NLG)
  - F-score for semantic (NLU)

Implicit Reward

- Model-based methods estimating data distribution
  - Language Modeling (LM) for language
  - Masked Autoencoder (MADE) for semantics

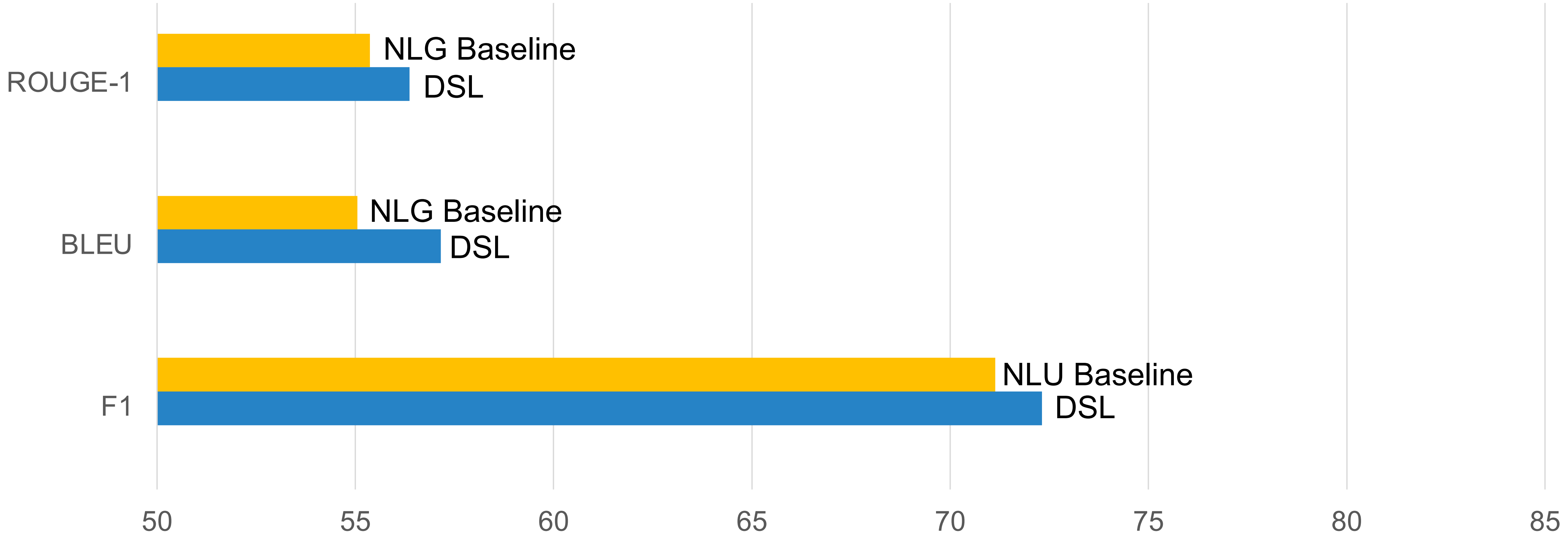
Joint Learning

- Proposed methods to enable gradient propagation over discrete prediction:
  - Straight-Through Estimator
  - Distribution as Input

- Flexibility:
  - **Hybrid objective:** could apply multiple objective functions (including supervised and unsupervised ones)
  - **Towards unsupervised learning:** the models could be potentially trained with unpaired data by full cycles

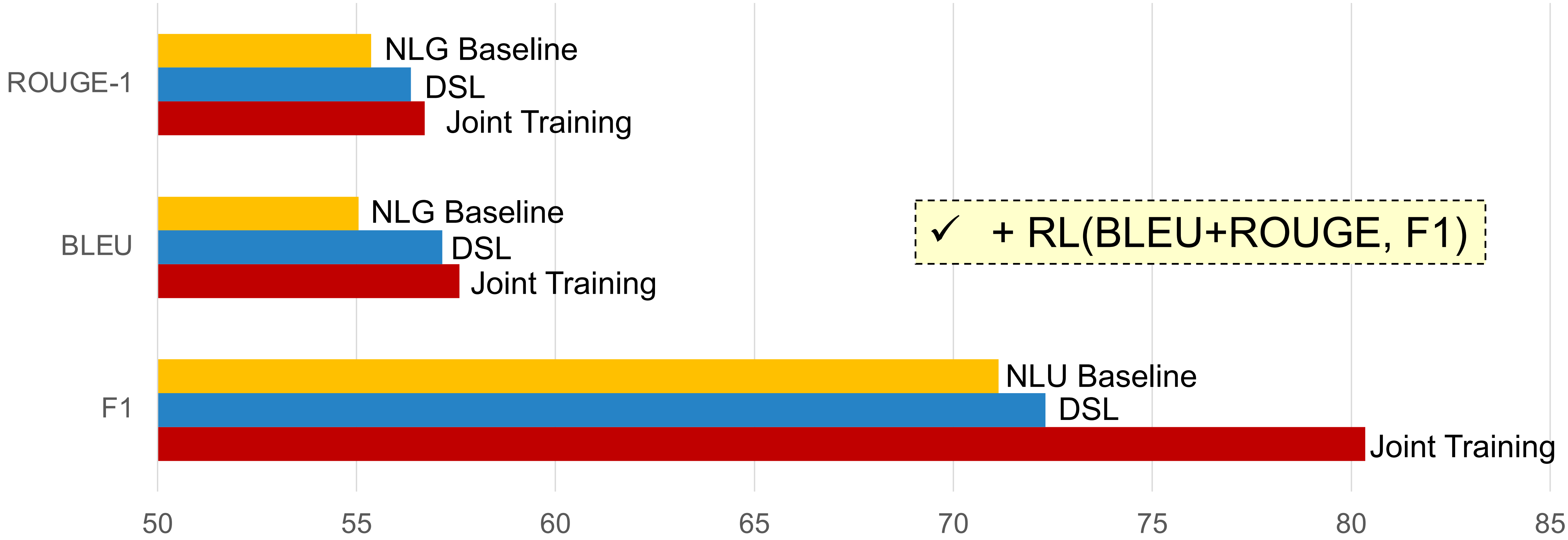
NLU/NLG Results

- NLU: F-1 score; NLG: BLEU, ROUGE



NLU/NLG Results

- NLU: F-1 score; NLG: BLEU, ROUGE



✓ A joint learning framework provides the flexibility of incorporating supervised and unsupervised learning algorithms to jointly train two models.

Generated Examples

familyFriendly[yes], area[city centre], eatType[pub], food[chinese], name[blue spice], near[rainbow vegetarian cafe]

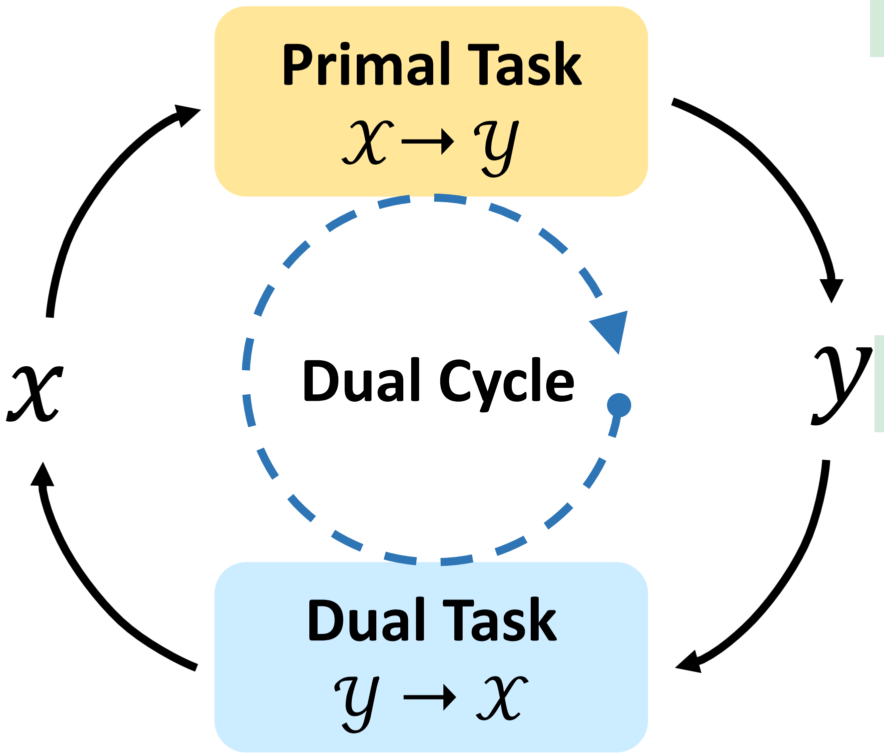
blue spice is a family friendly pub located in the city centre it serves chinese food and is near the rainbow vegetarian cafe

Baseline

familyFriendly[yes], food:[chinese]

Proposed

familyFriendly[yes], area[city centre], eatType[pub], priceRange[moderate], food[chinese], name[blue spice]



Baseline

the chinese restaurant the twenty two is a family friendly restaurant

Proposed

the chinese restaurant the blue spice is located in the city centre it is moderately priced and kid friendly

Outline

- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - **Dual Mutual Information Maximization**
    - Unpublished
  - Dual Inference
  - Dual Finetuning
- Summary
- Related work

Training Stage

Motivation

- Challenges might come from the nature of data

Natural Language

1. Alimentum city centre is family-friendly.

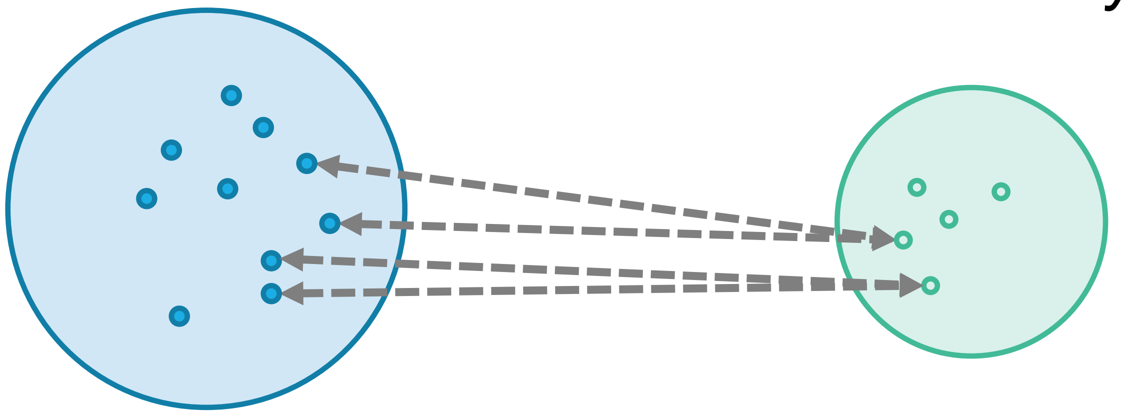
2. Alimentum is a family-friendly city centre.

Semantic Frame

NAME="Alimentum"

familyFriendly ="yes"

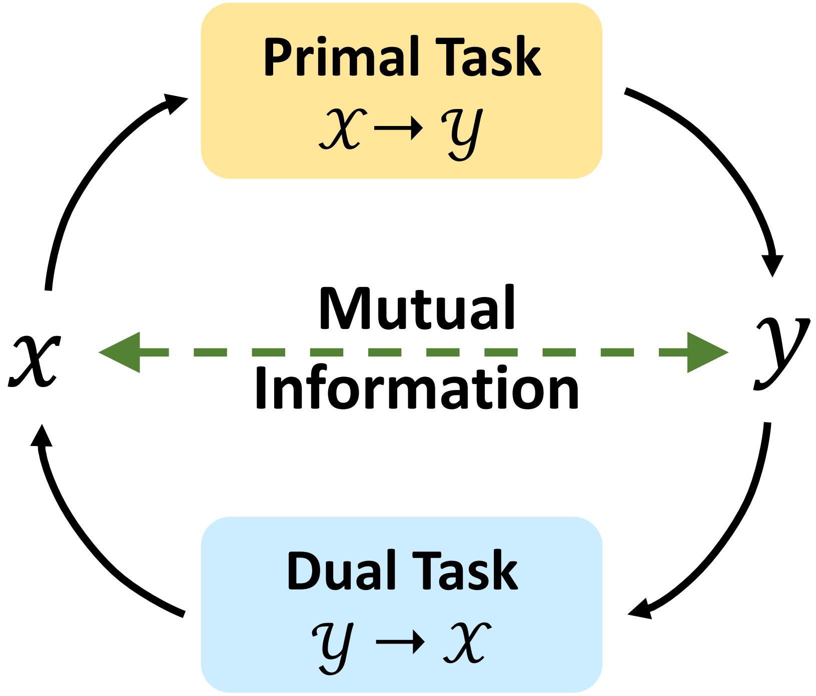
area = "city centre"



💡 MMI between the representation of language and semantics.

Mutual Information Maximization

- We aim to enhance the joint learning framework by *maximizing mutual information* between the representation of language and semantics.



Mutual Information Estimation

- MI cannot be directly used as a training objective due to intractability.
- Deep Infomax (DIM) (Hjelm et al., 2018) enables estimating MI by back-propagation in neural networks.

Deep Infomax (DIM) (Hjelm et al., 2018)

- A discriminator distinguishes between positive samples from the joint distribution and negative samples from the product of marginals.
  - Use Jensen-Shannon divergence via BCE loss (Yeh et al., 2019)

$$
\begin{aligned}
MI(X;Y) \geq & \mathbb{E}_{\mathbb{P}}[\log(d(x,y))] + \\
& \frac{1}{2}\mathbb{E}_{\mathbb{N}}[\log(1-d(x,\bar{y}))] + \\
& \frac{1}{2}\mathbb{E}_{\mathbb{N}}[\log(1-d(\bar{x},y))]
\end{aligned}
$$

Primal Cycle

Start from data x, transform x by function f:

$$\hat{y} = f(x; \theta_{x \to y});$$

Compute the loss by $\mathcal{L}_f(\hat{y}, y)$;

Random shuffle B and map the data pairs to original order to have negative samples $(\hat{x}, \bar{y})$ and $(\bar{x}, y)$;

Compute MI regularization:

$$\mathcal{L}_{MI} = \frac{1}{n} \sum \log(d(\hat{x}, y)) + \log(1 - d(\bar{x}, y)) + \log(1 - d(\hat{x}, \bar{y}));$$

Transform the output of the primal task by function g:

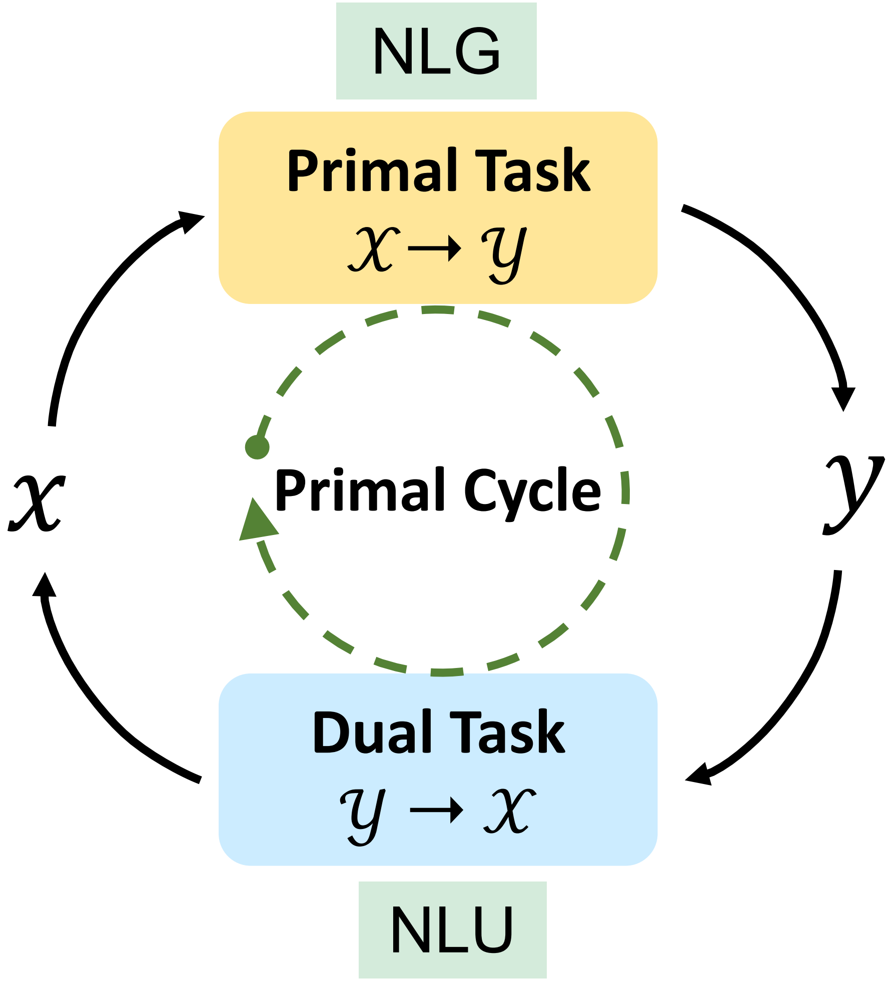
$$\hat{x} = g(\hat{y}; \theta_{y \to x});$$

Compute the loss by $\mathcal{L}_g(\hat{x}, x)$;

Update model parameters:

$$\theta_{x \to y} \leftarrow \theta_{x \to y} - \gamma \nabla_{\theta_{x \to y}} (\mathcal{L}_f(.) + \mathcal{L}_g(.) - \lambda \mathcal{L}_{MI}(.));$$

$$\theta_{y \to x} \leftarrow \theta_{y \to x} - \gamma \nabla_{\theta_{y \to x}} (\mathcal{L}_f(.) + \mathcal{L}_g(.) - \lambda \mathcal{L}_{MI}(.));$$



Dual Cycle

Start from word representations y, transform y by function g:

$$\hat{x} = g(y; \theta_{y \to x});$$

Compute the loss $\mathcal{L}_g(\hat{x}, x)$;

Random shuffle B and map the data pairs to original order to have negative samples $(\hat{x}, \bar{y})$ and $(\bar{x}, y)$;

Compute MI regularization:

$$\mathcal{L}_{MI} = \frac{1}{n} \sum \log(d(\hat{x}, y)) + \log(1 - d(\bar{x}, y)) + \log(1 - d(\hat{x}, \bar{y}));$$

Transform the output of the dual task by function f:

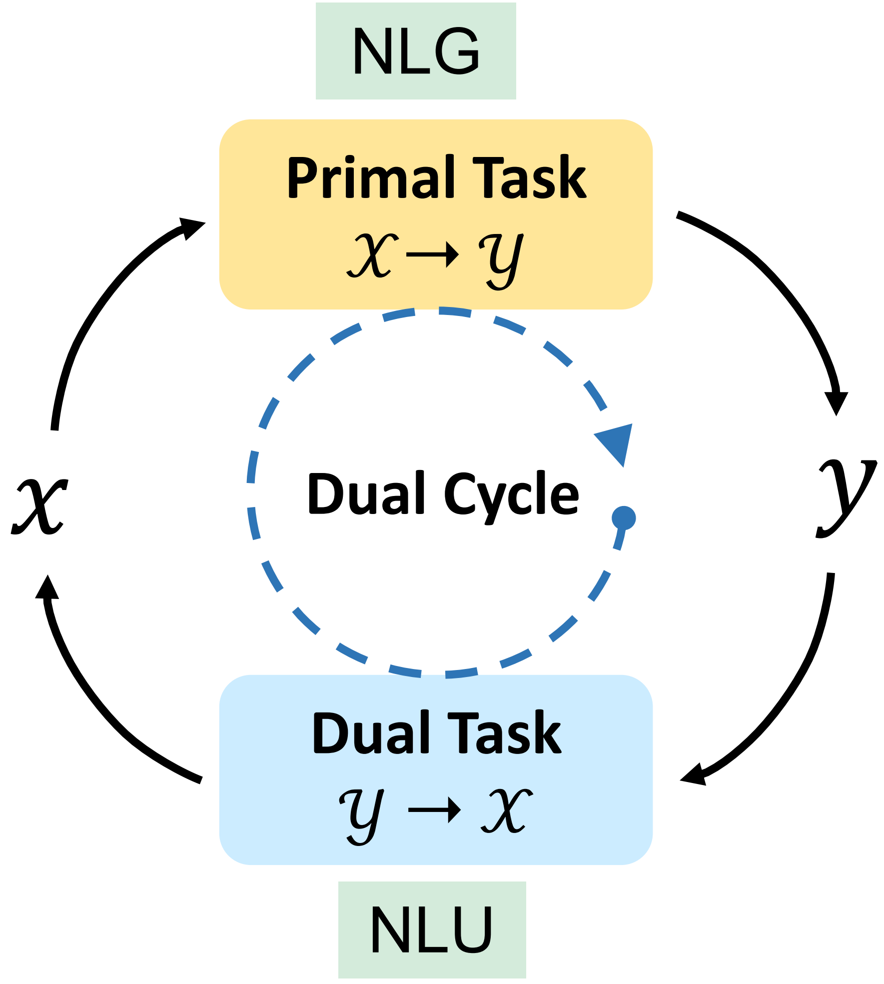
$$\hat{y} = f(\hat{x}; \theta_{x \to y});$$

Compute the loss by $\mathcal{L}_f(\hat{y}, y)$;

Update model parameters:

$$\theta_{x \to y} \leftarrow \theta_{x \to y} - \gamma \nabla_{\theta_{x \to y}} (\mathcal{L}_f(.) + \mathcal{L}_g(.) - \lambda \mathcal{L}_{MI}(.));$$

$$\theta_{y \to x} \leftarrow \theta_{y \to x} - \gamma \nabla_{\theta_{y \to x}} (\mathcal{L}_f(.) + \mathcal{L}_g(.) - \lambda \mathcal{L}_{MI}(.));$$

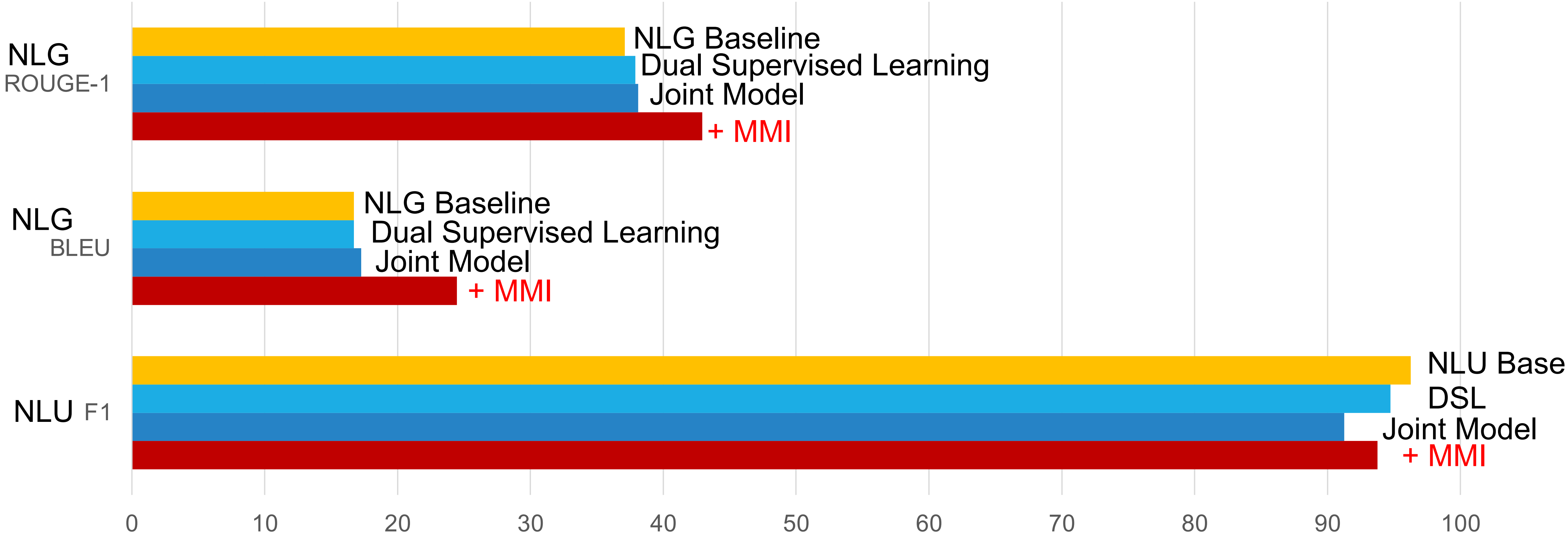


Datasets

- **ATIS:** flight reservations 
  - Sentence-level intents and word-level slot tags
- **SNIPS:** voice assistants for multiple domains
  - Sentence-level intents and word-level slot tags
- **E2E NLG:** restaurant domain
  - Each meaning representation has up to 5 references in natural language and no intent labels

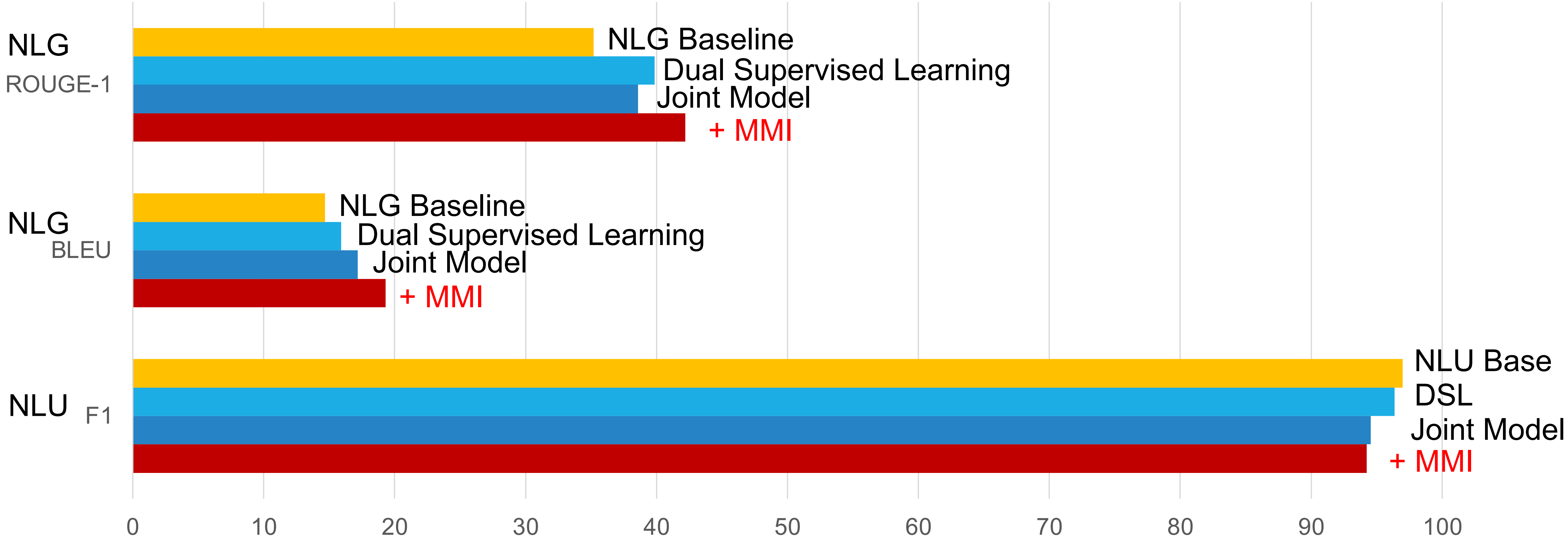
NLU/NLG Results

- **ATIS** data: 5k examples in the flight booking domain



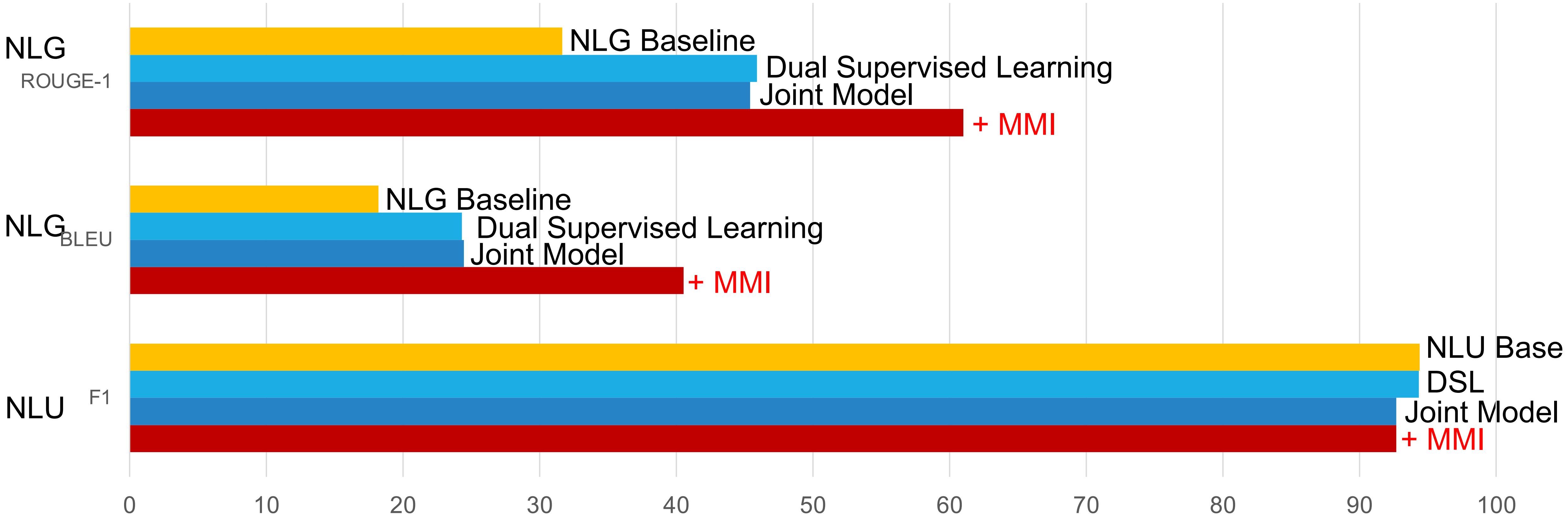
NLU/NLG Results

- **SNIPS** data: voice assistants for multiple domains



NLU/NLG Results

- **E2E NLG** data: 50k examples in the restaurant domain



✓ Connecting models via MMI between semantics and language is useful

Outline

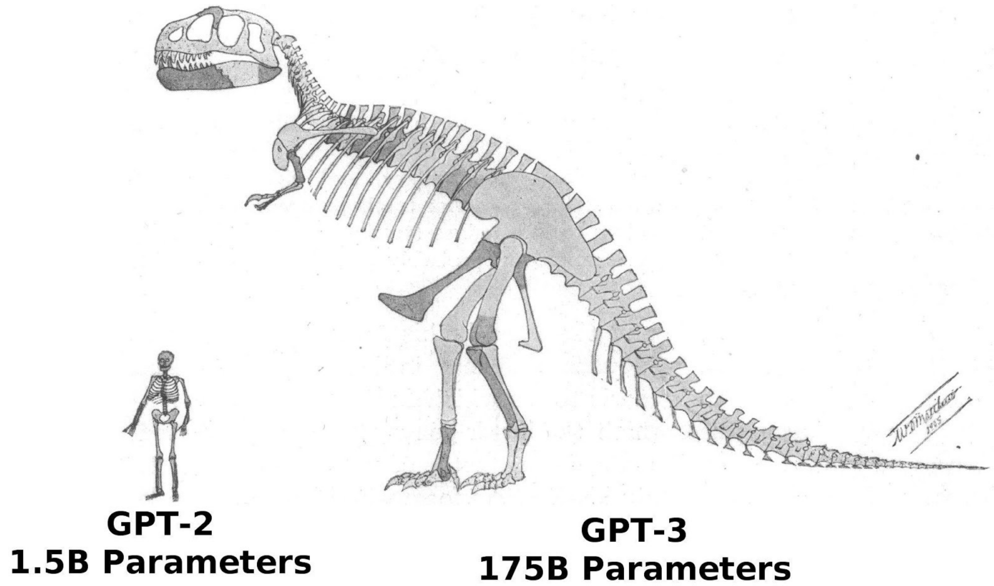
- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - **Dual Inference**
    - Su et al., Findings in EMNLP 2020
  - Dual Finetuning (ongoing)
- Summary
- Related work

Inference Stage

Motivation

- Prior work utilized the duality in the *training* stage
- Due to current large-scaled NLP models, it is difficult/impractical to *re-train* models.

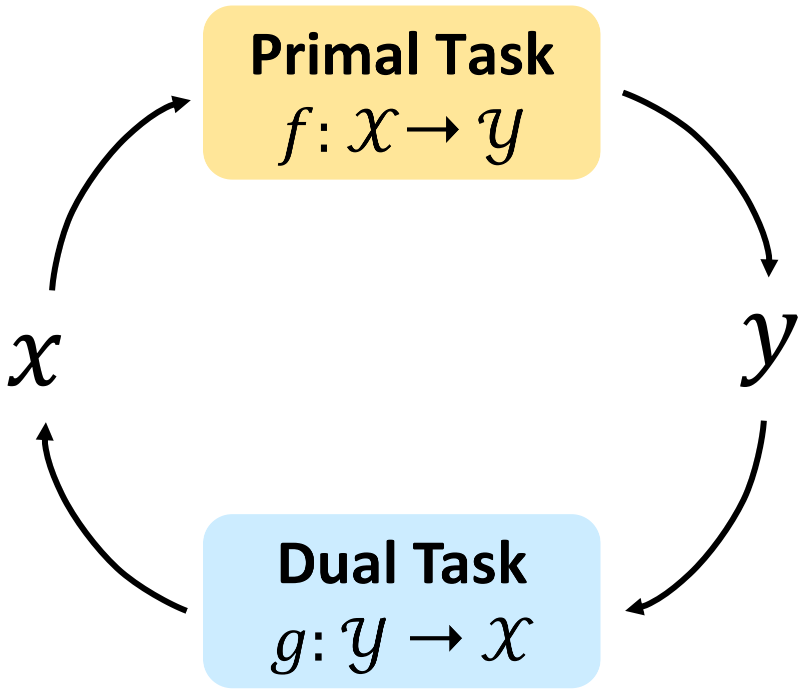
💡 Exploiting the duality in the inference stage



Dual Inference for NLU / NLG

- Normal inference process

$$f(x) = \arg\max \{\log P(y \mid x; \theta_{x \to y})\}$$
$$g(y) = \arg\max \{\log P(x \mid y; \theta_{y \to x})\}$$



Dual Inference for NLU / NLG

- Inference with duality (Xia et al., 2017)

$$f(x) = \arg\max \{\log P(y \mid x; \theta_{x \to y})\}$$
$$\simeq \arg\max \{\alpha \log P(y \mid x; \theta_{x \to y}) + (1-\alpha) \log P(y \mid x; \theta_{y \to x})\}$$

Estimated by forward model

Estimated by backward model

$$\log P(y \mid x; \theta_{y \to x})$$
$$= \log\left(\frac{P(x \mid y; \theta_{y \to x}) P(y; \theta_y)}{P(x; \theta_x)}\right)$$
$$= \log P(x \mid y; \theta_{y \to x}) + \log P(y; \theta_y) - \log P(x; \theta_x)$$

Dual Inference for NLU / NLG

$$f(x) \simeq \arg\max\{\alpha \log P(y \mid x; \theta_{x\to y}) + (1-\alpha)(\log P(x \mid y; \theta_{y\to x}) + \log P(y; \theta_y) - \log P(x; \theta_x))\}$$

forward model

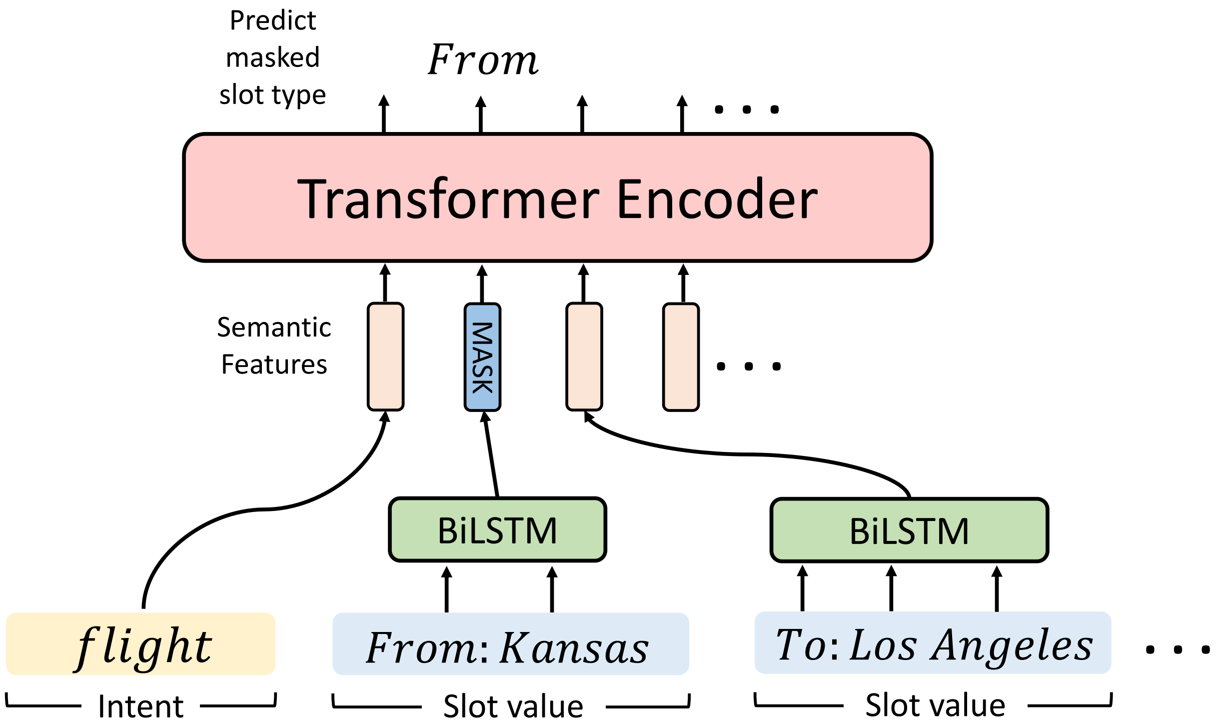
backward model

Marginal of y

Marginal of x

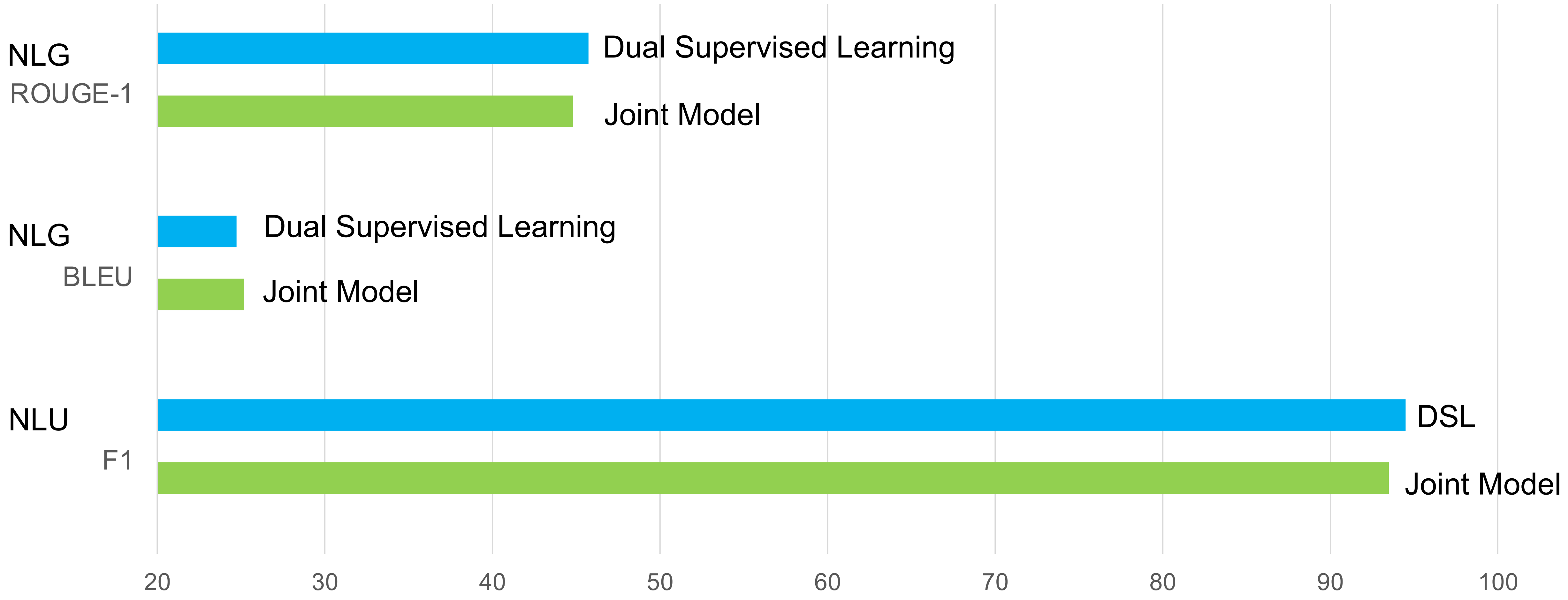
Marginal Distribution Estimation

- Prior work uses MADE, treating semantics as a finite number of labels.
- Considering scalability, we propose a non-autoregressive masked-model.



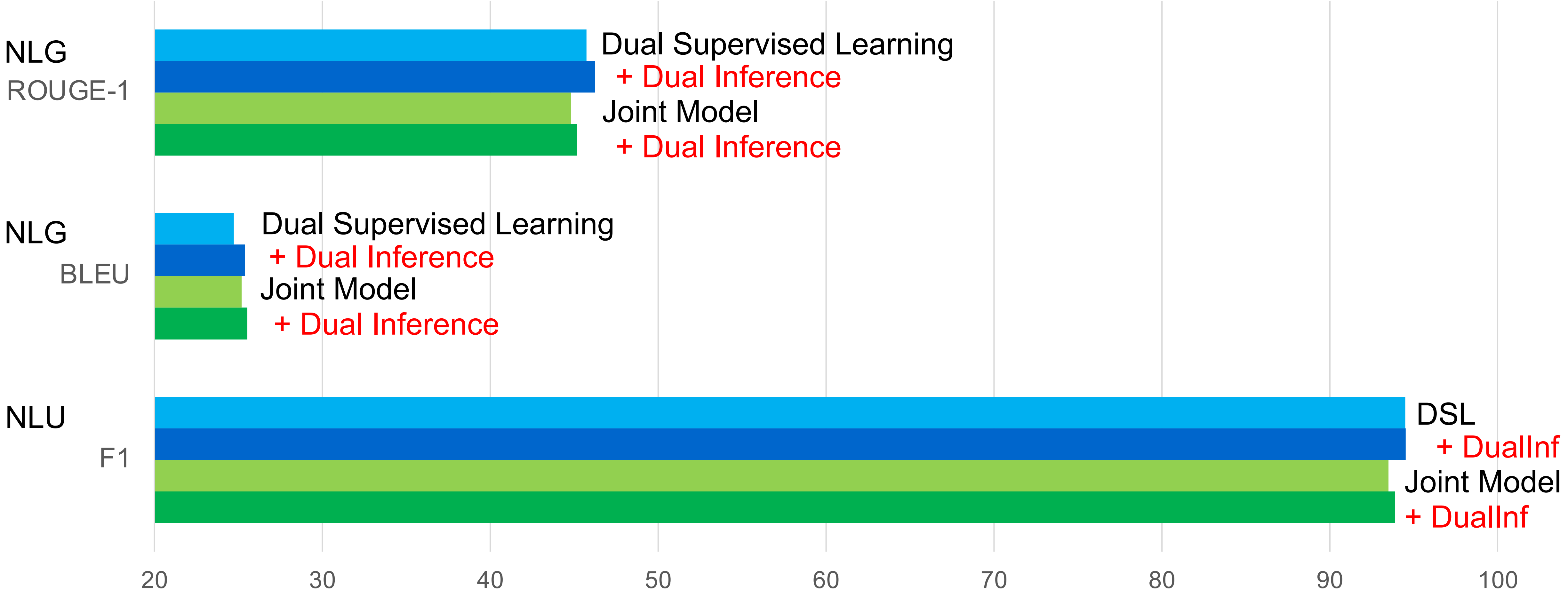
NLU/NLG Results

- **E2E NLG** data: 50k examples in the restaurant domain



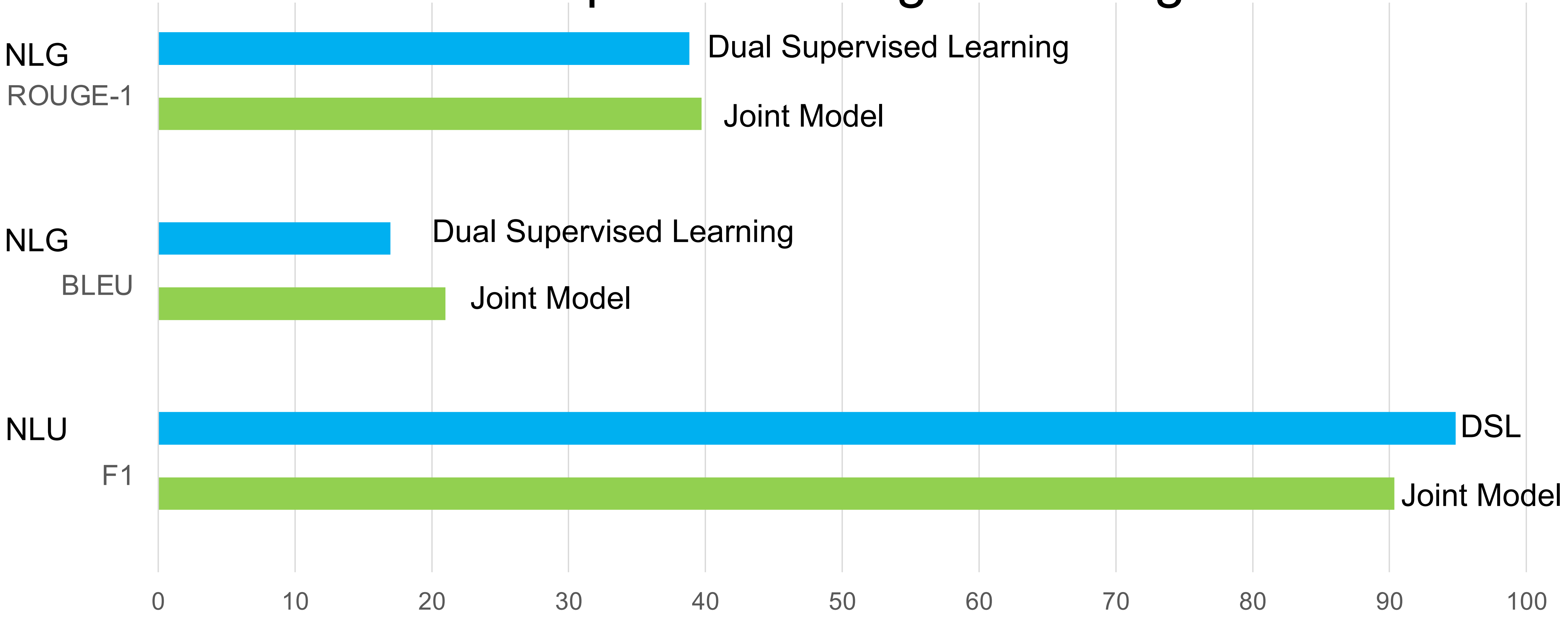
NLU/NLG Results

- **E2E NLG** data: 50k examples in the restaurant domain



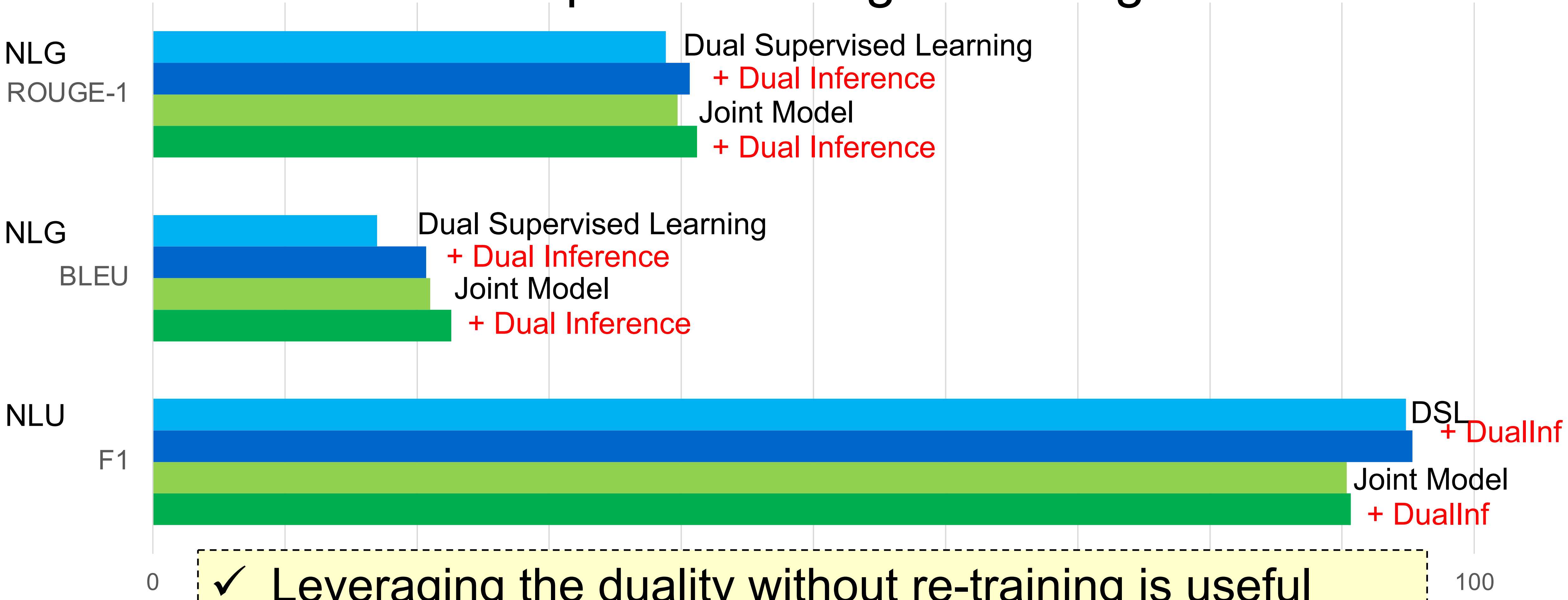
NLU/NLG Results

- **ATIS** data: 5k examples in the flight booking domain



NLU/NLG Results

- **ATIS** data: 5k examples in the flight booking domain



- ✓ Leveraging the duality without re-training is useful
- ✓ Consistent improvement for multiple datasets

Outline

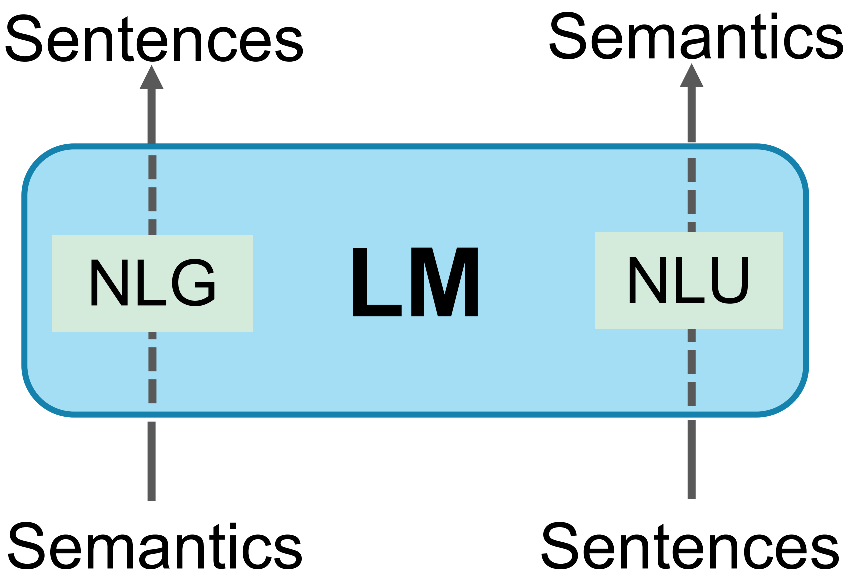
- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - **Dual Finetuning**

Finetuning Stage

- Summary
- Related work

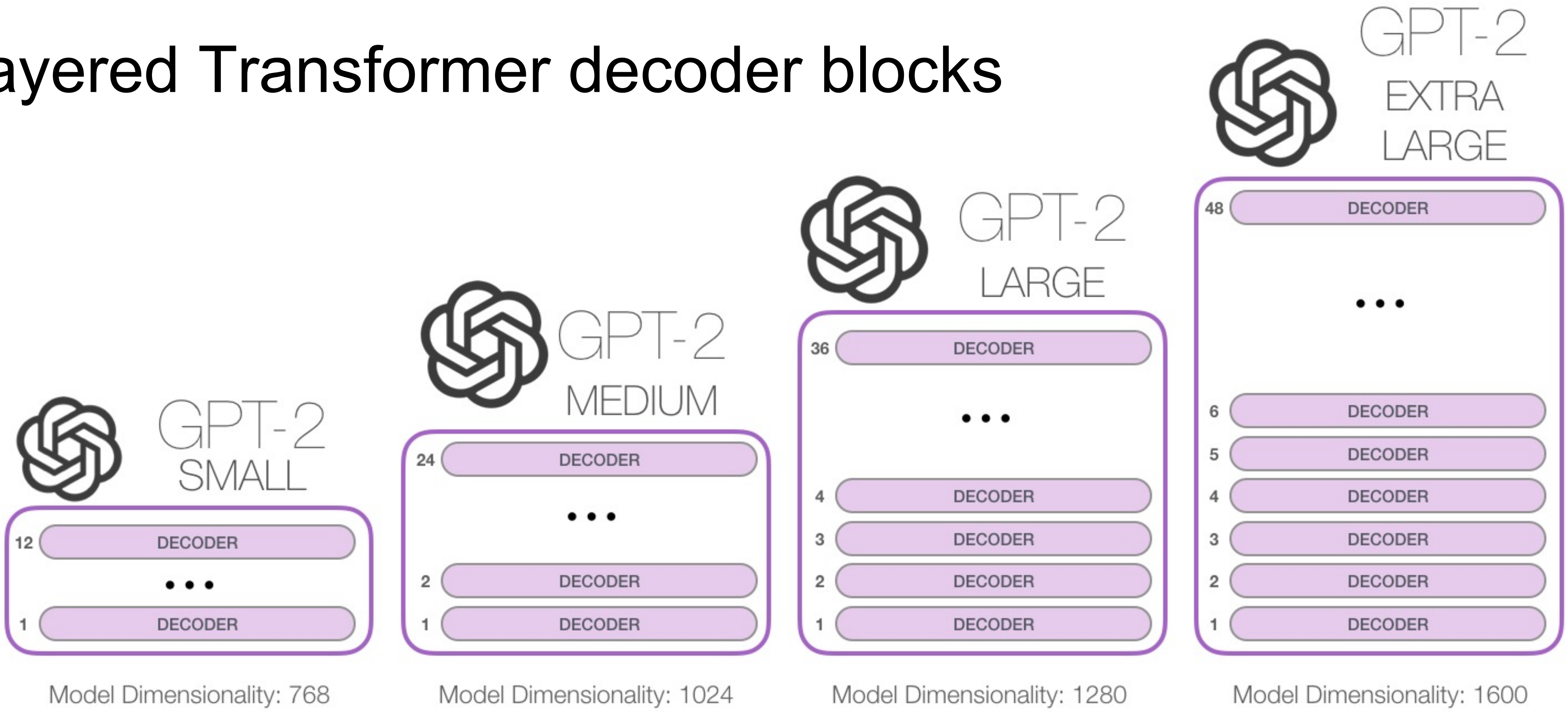
Motivation

- Nowadays, finetuning pre-trained language models is often the first choice for a NLP problem.
- One model for two dual tasks.



GPT-2 (Radford et al., 2019)

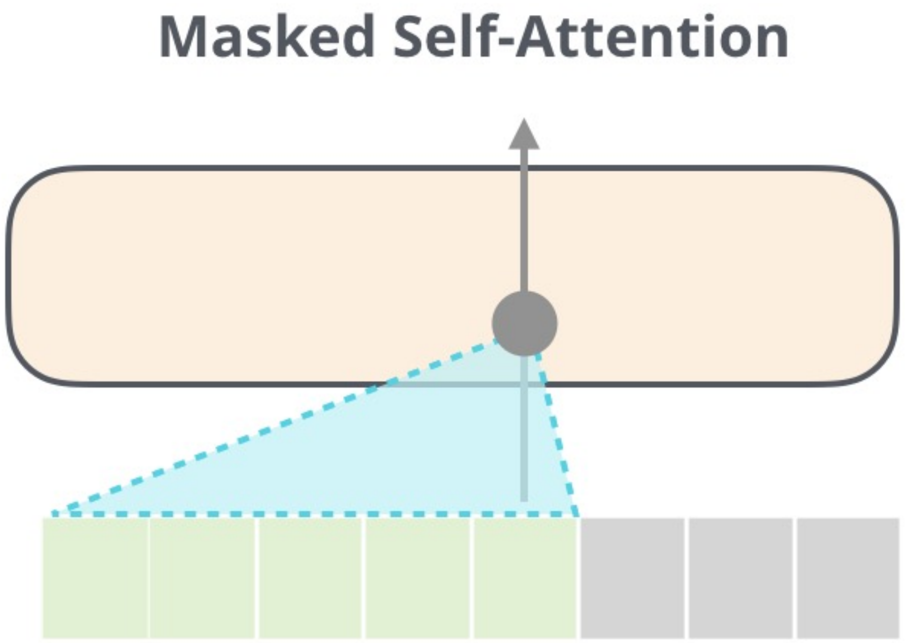
- **G**enerative **P**re-trained **T**ransformer **2**
- Layered Transformer decoder blocks



"Language models are unsupervised multitask learners." Radford, et al. 2019

GPT-2 (Radford et al., 2019)

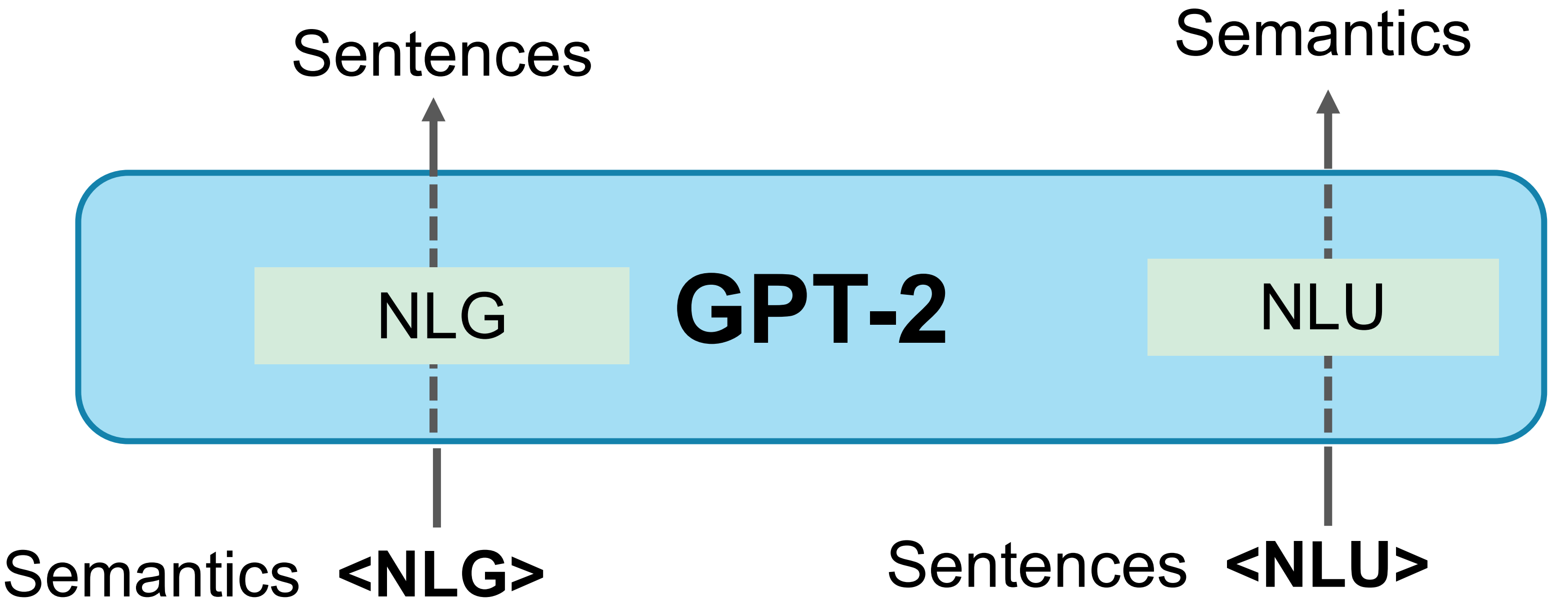
- Pretrained on WebText, which has over 8 million documents for a total of 40 GB of text
- Language Modeling
- Auto-regressive nature



💡 Model both NLU and NLG as text generation

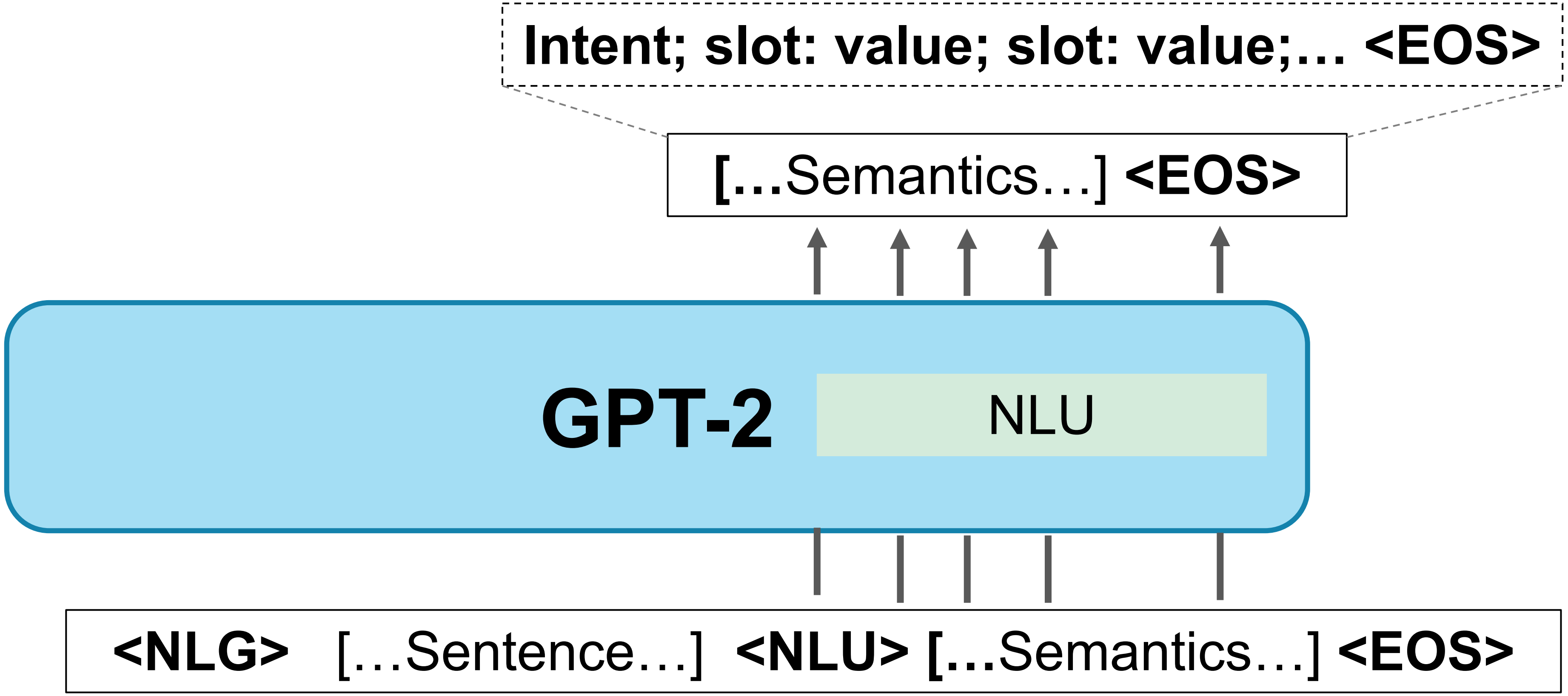
Objective Design

- **How to enforce the model to execute the target task?**
- Special task tokens



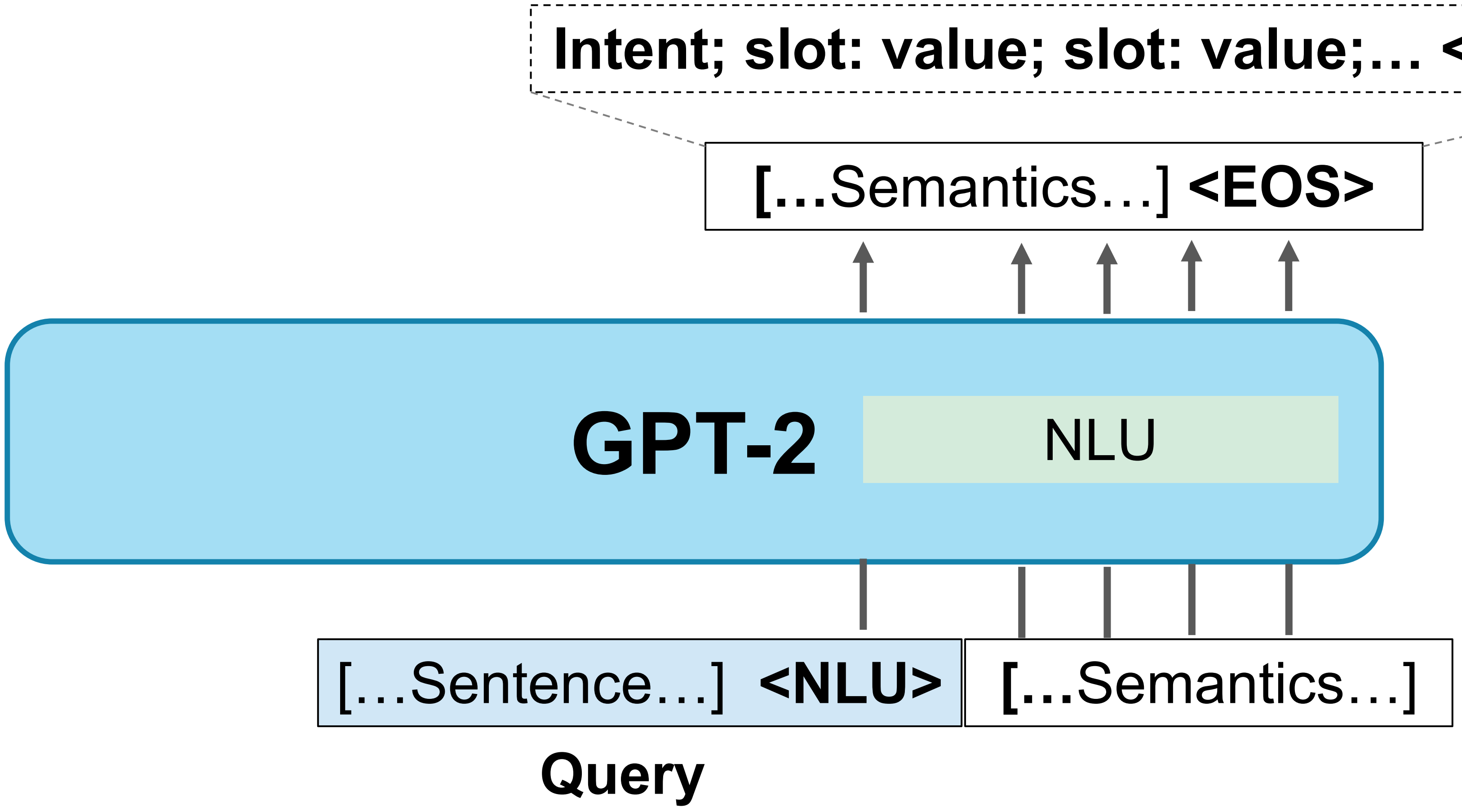
Objective Design

- Language Modeling training

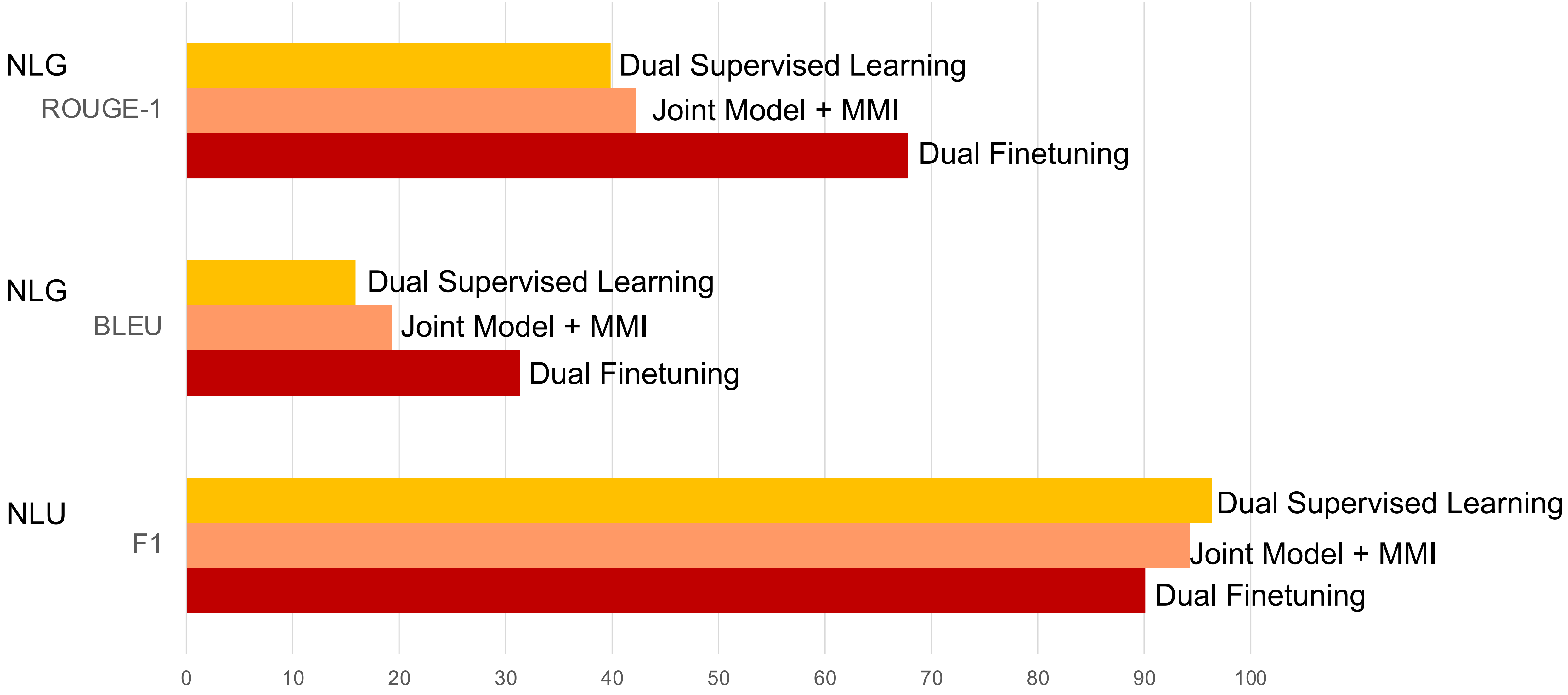


Inference

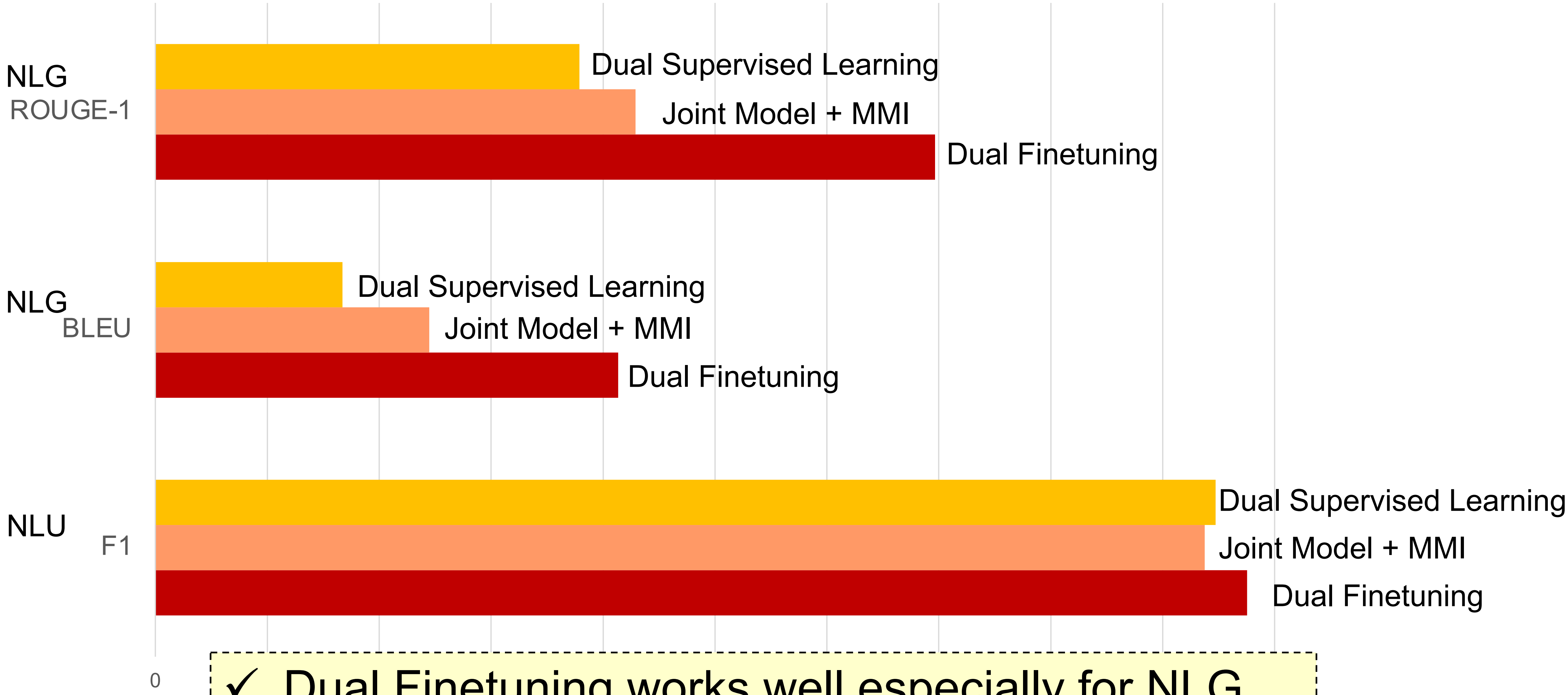
- Let the model generate sequences auto-regressively



NLU/NLG Results on SNIPS



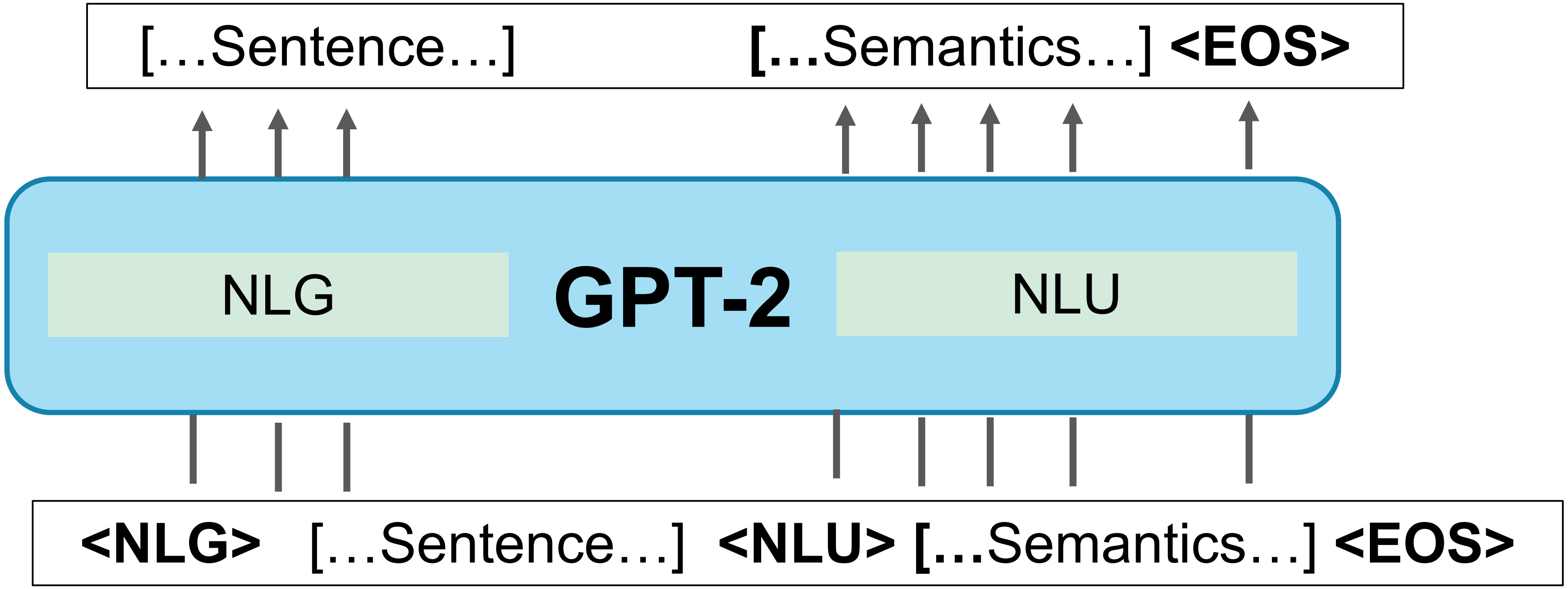
NLU/NLG Results on ATIS



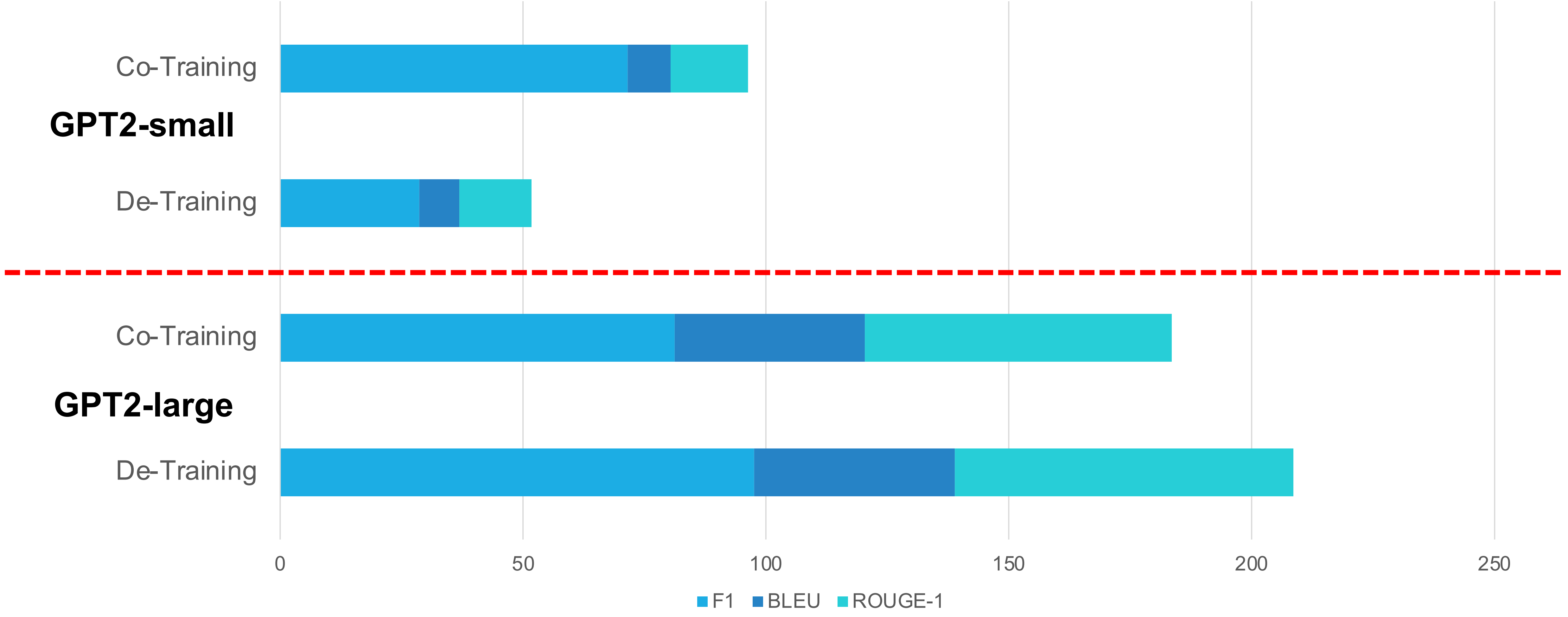
✓ Dual Finetuning works well especially for NLG.

Co-Training

- Train two tasks at one time
- Language modeling training



Performance Comparison (SNIPS)



✓ Co-training only works better with smaller models.

Outline

- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning
- **Summary**
- Related work

Summary

- Dual Supervised Learning
  - *Supervised Learning:* duality loss as regularization term
- Joint Dual Learning
  - *Semi-supervised Learning*: joint learning framework
- Dual Mutual Information Maximization
  - *Supervised Learning + MMI:* auxiliary MMI objective

Training Stage

- Dual Inference
  - *Inference:* enhanced inference process

Inference Stage

- Dual Finetuning
  - *Finetuning:* dual finetuning objectives

Finetuning Stage

Challenges

- Not every NLU data is suitable for augmenting into NLG data.
- NLU always requires human annotations, technically it is infeasible to perform “fully” unsupervised learning.
- Different relationships between tasks

Outline

- Background
- Duality Exploitation
  - Dual Supervised Learning
  - Joint Dual Learning
  - Dual Mutual Information Maximization
  - Dual Inference
  - Dual Finetuning
- Summary
- **Related work**

Semantic Parsing with Dual Learning

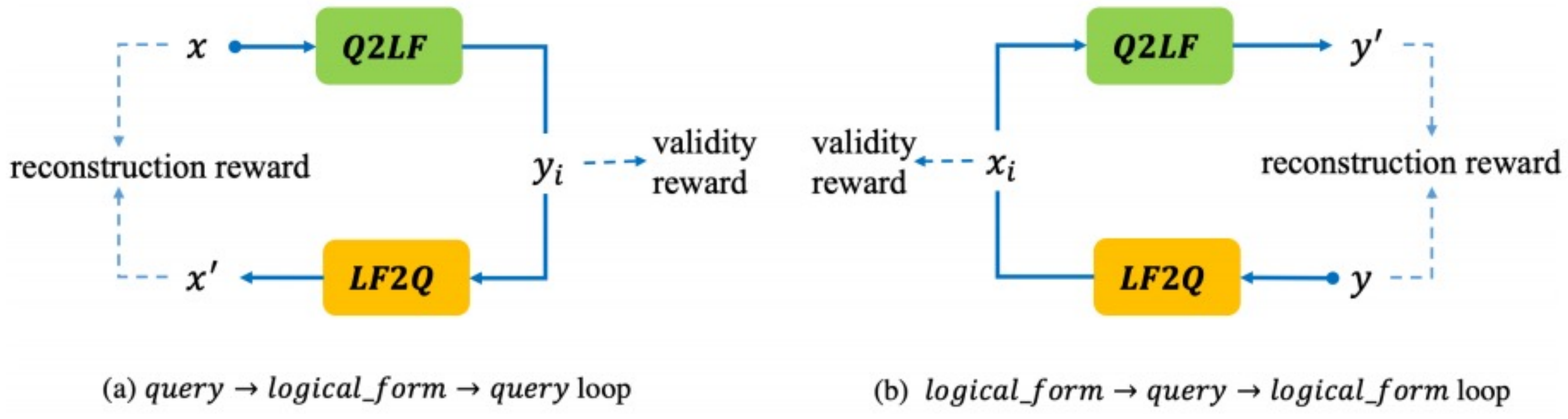
- Contemporaneous work focusing on semantic parsing
- Similar to our Joint Dual Learning

Query

"show flight from ci0 to ci1"

Logical Form

"(lambda $0 e (and (from $0 ci0) (to $0 ci1) (flight $0)))"

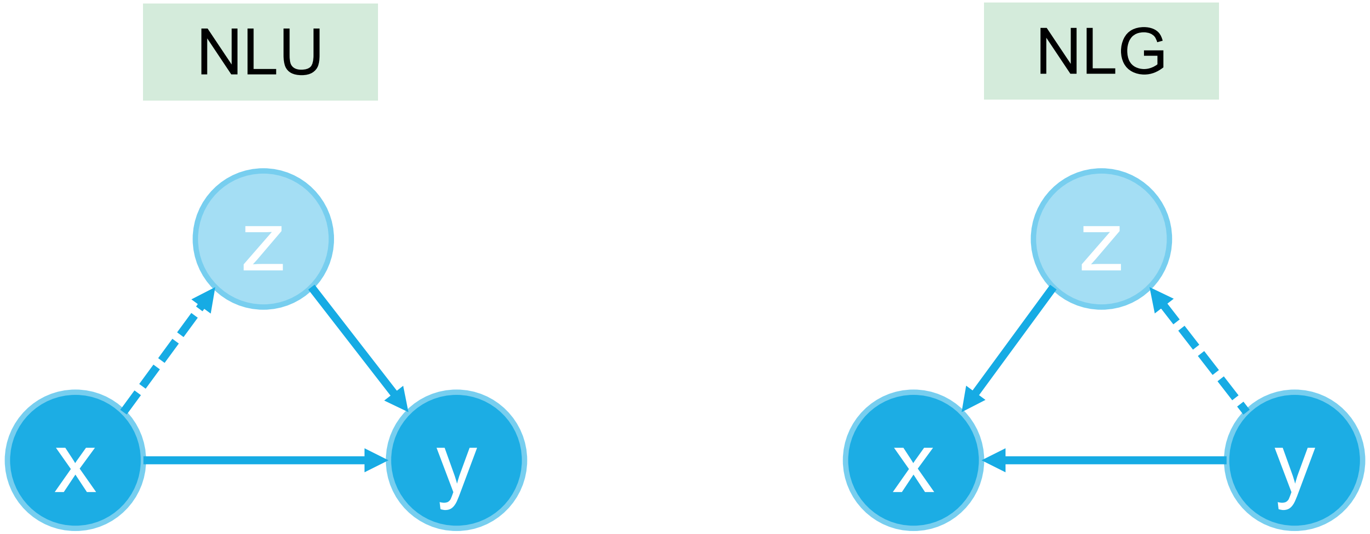


💡 Different NLU task

"Semantic Parsing with Dual Learning" Cao et al., NAACL 2019

Latent Variable Model (Tseng et al., 2020)

- Coupling NLU and NLG with a latent variable representing the shared intent between natural language and formal representations



💡 Advanced neural modeling

"A Generative Model for Joint Natural Language Understanding and Generation" Tseng et al., ACL 2020

Pragmatically Text Generation

- Computational pragmatics: Listener vs Speaker
- The listener model and the base speaker model together define a *pragmatic speaker*

$$S_1^R(o \mid i) = L^R(i \mid o)^\lambda \cdot S_0(o \mid i)^{1-\lambda}$$

NLU NLG

- Similar to our Dual Inference

💡 Traditional linguistic perspective

"Pragmatically Informative Text Generation" Shen et al., NAACL 2019

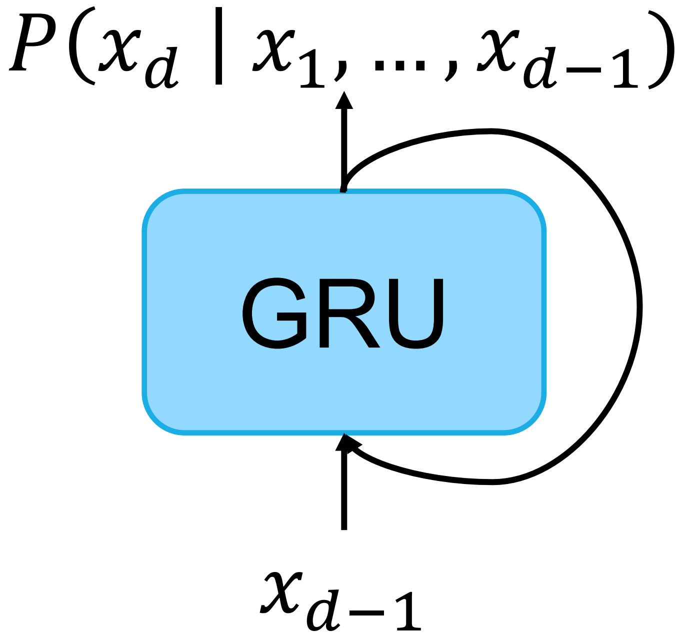
Thanks for your attention.

Appendix

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

- Language modeling

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

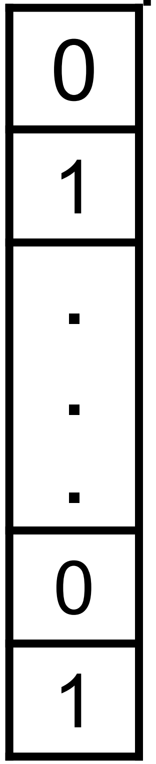


Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen, "Dual Supervised Learning for Natural Language Understanding and Generation," in *Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

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"

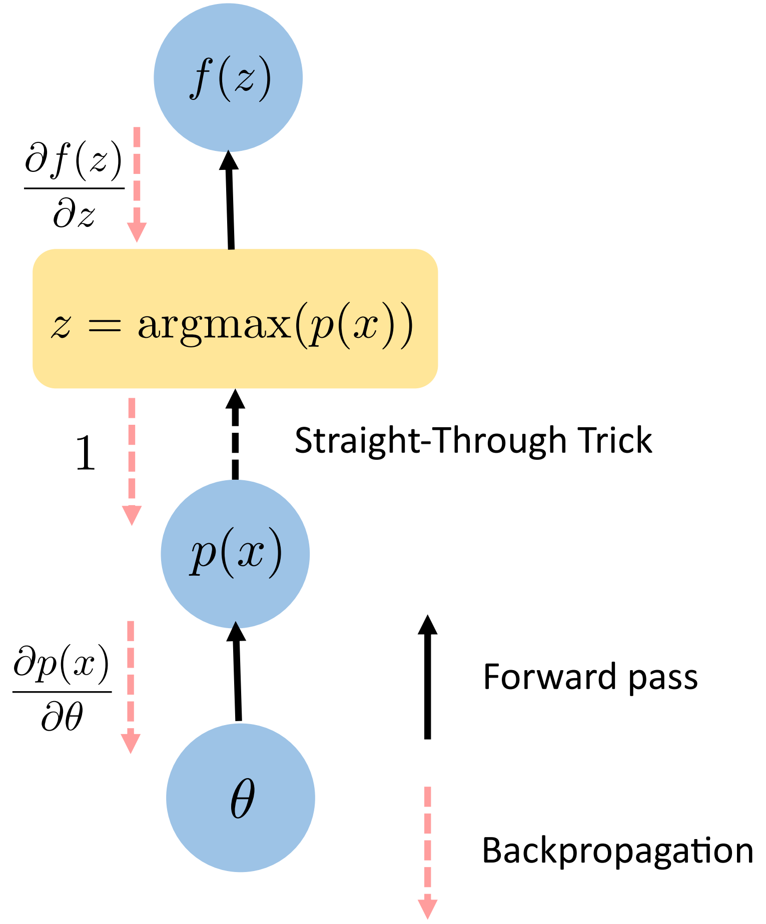


How to model the marginal distributions of y?

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.

Straight-Through Estimator

- Directly using the gradients of discrete samples as the gradients of the distribution parameters.



Distribution as Input

- For NLU, we use the predicted distribution over the vocabulary from NLG to perform the weighted-sum of word embeddings.
- For NLG, the probability distribution of slot-value pairs predicted by NLU can directly serve as the input vector.

Dual Supervised Learning Results

Model	NLU	NLG			
	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative training	71.14	55.05	55.37	27.95	39.90
Dual Supervised Learning with λ = 0.1	**72.32**	**57.16**	**56.37**	**29.19**	**40.44**
Dual Supervised Learning with λ = 0.01	72.08	55.07	55.56	28.42	40.04
Dual Supervised Learning with λ = 0.001	71.71	56.17	55.90	28.44	40.08
Dual Supervised Learning w/o MADE	70.97	55.96	55.99	28.74	39.98

Joint Dual Learning Results

Model	NLU	NLG			
	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative training	71.14	55.05	55.37	27.95	39.90
Dual Supervised Learning	72.32	57.16	56.37	29.19	40.44
Joint Training (Straight-Through)	71.73	55.19	55.16	27.45	39.33
Joint Training (Distribution as Input)	80.03	55.34	56.17	28.48	39.24
+ RL(BLEU+ROUGE, F1)	**80.35**	**57.59**	**56.71**	**29.06**	**40.28**
+ RL(LM, MADE)	79.52	55.61	55.97	28.57	39.97

MMI Results on ATIS

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	85.98	**96.28**	16.71	37.11	13.47	35.88
Dual Supervised Learning	83.02	94.73	16.72	37.89	14.60	36.52
Joint Baseline	80.61	91.26	17.26	38.10	14.69	36.73
+ MI(semantics, word)	88.15	93.75	**24.46**	**42.92**	**23.01**	**41.78**
+ MI(semantics, sentence)	**88.50**	93.85	19.28	39.55	16.88	38.19

MMI Results on SNIPS

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	**97.40**	**96.98**	14.69	35.20	13.27	34.19
Dual Supervised Learning	97.39	96.35	15.90	39.85	16.39	38.69
Joint Baseline	97.32	94.56	17.19	38.59	16.36	37.53
+ MI(semantics, word)	97.02	94.25	**19.30**	**42.20**	**19.66**	**40.83**
+ MI(semantics, sentence)	96.93	95.42	16.82	39.06	16.45	37.75

MMI Results on E2E NLG

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	-	**94.41**	18.21	31.66	12.47	27.39
Dual Supervised Learning	-	94.36	24.32	45.91	19.31	39.92
Joint Baseline	-	92.69	24.47	45.41	19.22	39.10
+ MI(semantics, word)	-	92.69	**40.53**	**61.00**	**36.14**	**52.60**
+ MI(semantics, sentence)	-	92.64	28.21	49.52	23.18	41.63

Dual Inference Results on ATIS

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	84.10	94.26	16.08	35.10	11.94	33.73
+ DualInf(α=0.5, β=0.5)	85.07	93.84	**17.38**	**36.40**	**13.33**	**35.09**
+ DualInf(α*, β*)	**85.57**	**94.63**	16.16	35.19	11.93	33.75
Dual Supervised Learning	82.98	94.85	16.98	38.83	15.56	37.50
+ DualInf(α=0.5, β=0.5)	83.68	94.89	**20.69**	**40.62**	**17.72**	**39.31**
+ DualInf(α*, β*)	**84.26**	**95.32**	17.05	38.82	15.57	37.42
Joint Baseline	81.44	90.37	21.00	39.70	18.91	38.48
+ DualInf(α=0.5, β=0.5)	81.21	88.42	**22.60**	**41.19**	**20.24**	**39.88**
+ DualInf(α*, β*)	**85.88**	**90.66**	20.67	39.41	18.68	38.16

Dual Inference Results on SNIPS

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	96.58	96.67	15.49	34.32	13.75	33.26
+ DualInf(α=0.5, β=0.5)	**97.07**	**96.70**	**16.90**	**35.43**	**15.18**	**34.41**
+ DualInf(α^*, β^*)	96.88	96.76	15.46	34.21	13.78	33.14
Dual Supervised Learning	96.83	96.71	15.96	36.69	15.39	35.73
+ DualInf(α=0.5, β=0.5)	**96.88**	**96.80**	**18.07**	**37.63**	**16.75**	**36.67**
+ DualInf(α^*, β^*)	95.34	96.68	16.08	36.97	15.62	36.04
Joint Baseline	97.18	94.57	17.15	36.32	15.68	35.36
+ DualInf(α=0.5, β=0.5)	**97.27**	95.59	**18.56**	37.87	17.25	36.90
+ DualInf(α^*, β^*)	95.54	**96.06**	18.26	**38.16**	**17.70**	**37.40**

Dual Inference Results on E2E NLG

Model	NLU		NLG			
	Accuracy	F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Iterative Baseline	-	94.25	24.98	44.60	19.40	37.99
+ DualInf(α=0.5, β=0.5)	-	94.29	25.34	44.82	19.73	38.23
+ DualInf(α^*, β^*)	-	94.55	25.35	44.87	19.74	38.30
Dual Supervised Learning	-	94.49	24.73	45.74	19.60	39.91
+ DualInf(α=0.5, β=0.5)	-	94.53	25.40	46.25	20.18	40.42
+ DualInf(α^*, β^*)	-	94.47	24.67	45.71	19.56	39.88
Joint Baseline	-	93.51	25.19	44.80	19.59	38.20
+ DualInf(α=0.5, β=0.5)	-	93.43	25.57	45.11	19.90	38.56
+ DualInf(α^*, β^*)	-	93.88	25.54	45.17	19.89	38.61

Transformer

- Multi-Head Attention