Slot Induction via Pre-trained Language Model Probing and Multi-level Contrastive Learning

Hoang Nguyen[1], **Chenwei Zhang**[2], **Ye Liu**[3], **Philip S. Yu**[1]

[1] University of Illinois at Chicago
[2] Amazon
[3] Salesforce Research

SIGDIAL 2023

Overview

- Background
- Motivation
- Framework
- Evaluation
- Conclusion

BACKGROUND

Background

Slot Filling [1]
(Closed world)

Utterance

Play me a top-ten song by phil ochs on groove shark

Knowledge
Slot Labels: PERSON, PLAYLIST_NAME

Slot

Phil ochs: PERSON,
groove shark:
PLAYLIST_NAME

[1] Goo et al., Slot-gated modeling for joint slot filling and intent prediction. ACL 2018

Background

Multi-task Learning[1] (Closed world)

Utterance

Play me a top-ten song by phil ochs on groove shark

Knowledge
Intent Labels: Play Music
Slot Labels: PERSON, PLAYLIST_NAME

Slot

Phil ochs: PERSON,
groove shark: PLAYLIST_NAME

Intent

Play Music

[1] Haihong et al., A novel bi-directional interrelated model for joint intent detection and slot fillin. ACL 2019.

Background

Zero-shot Learning[1]
(Closed world)

Utterance

Make me a reservation in south carolina

Knowledge
Intent Labels: Play Music
Slot Labels: PERSON, PLAYLIST_NAME

Slot

??

[1] Glass et al., Robust retrieval augmented generation for zero-shot slot filling. EMNLP 2021.

Background

Slot Induction (Open world)

Utterance

Make me a reservation in south carolina

Semantic Knowledge
Task-specific Knowledge

Make me a reservation in south carolina

Background

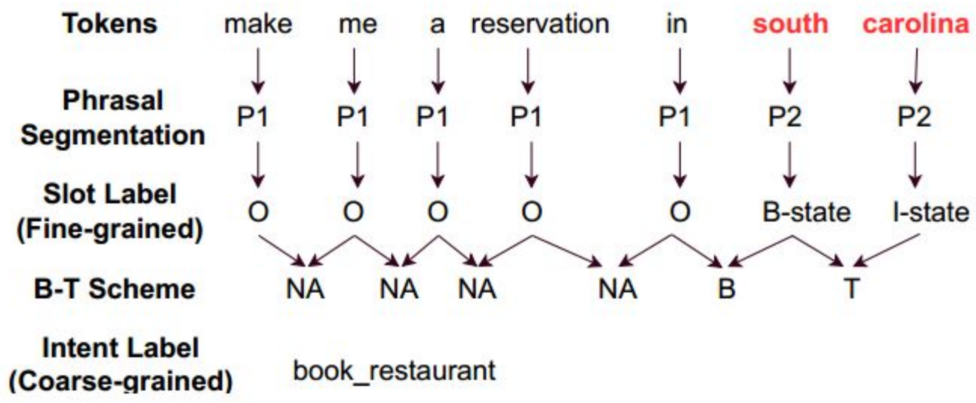
Slot Induction

Definition: Identifying phrases containing token-level slot labels
Assumption:

- Non-existence of token-level slot labels during training and testing

Comparison with Phrasal Segmentation:

- Slot phrases can be complex and not restricted to noun phrases
- **Utterances** and **intents** are the only sources of information



Sample Slot Phrases

Find movie times for **close by** movies
what are the most expensive first class tickets between atlanta and dallas

Background

Slot Induction

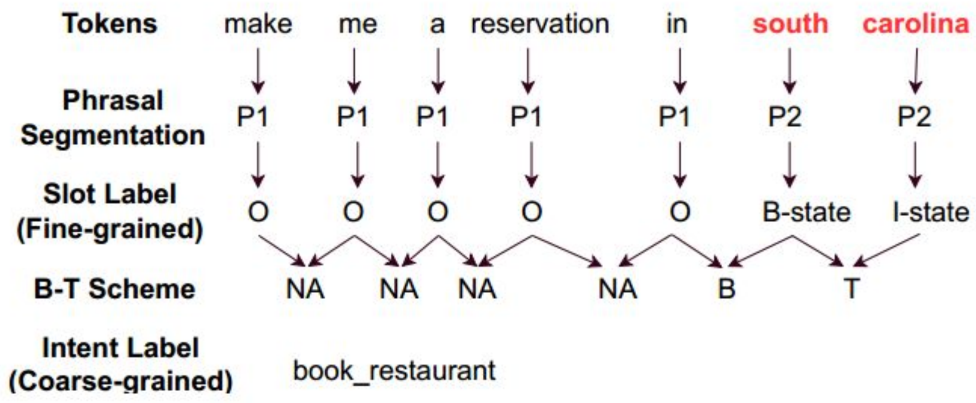
<u>Definition:</u> Identifying phrases containing token-level slot labels

<u>Assumption:</u>

- Non-existence of token-level slot labels during training and testing

<u>Comparison with Phrasal Segmentation:</u>

- Slot phrases can be complex and not restricted to noun phrases
- **Utterances** and **intents** are the only sources of information



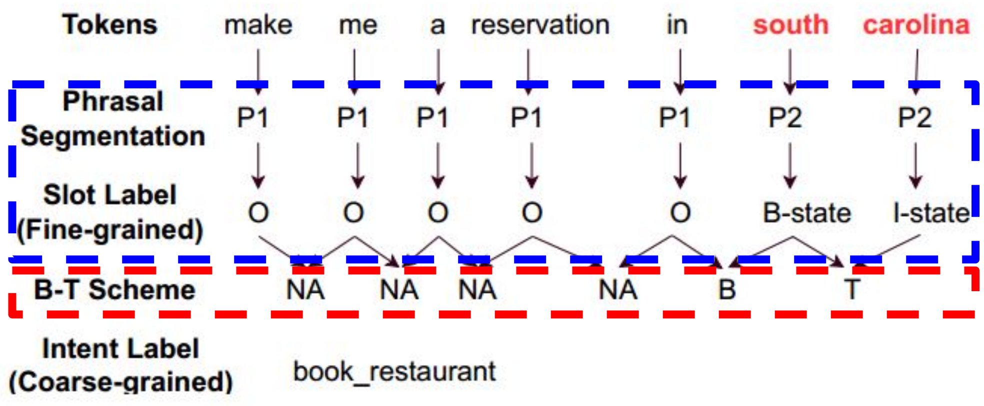
Requiring balance of **semantic knowledge** and **task-specific knowledge** for inference

Background

Slot Induction Evaluation

Assumption:
- Adoption of Break-Tie Mechanisms
- Allowing for direct comparison between Phrasal Segmentation and Slot Filling methods

Golden Metric: H-Mean
- Balance of correct Break, Tie predictions



MOTIVATION

MOTIVATION

Slot Induction

Utterance

Make me a reservation in south carolina

Semantic Knowledge

Pre-trained Language Model (PLM)

make me a reservation in south carolina

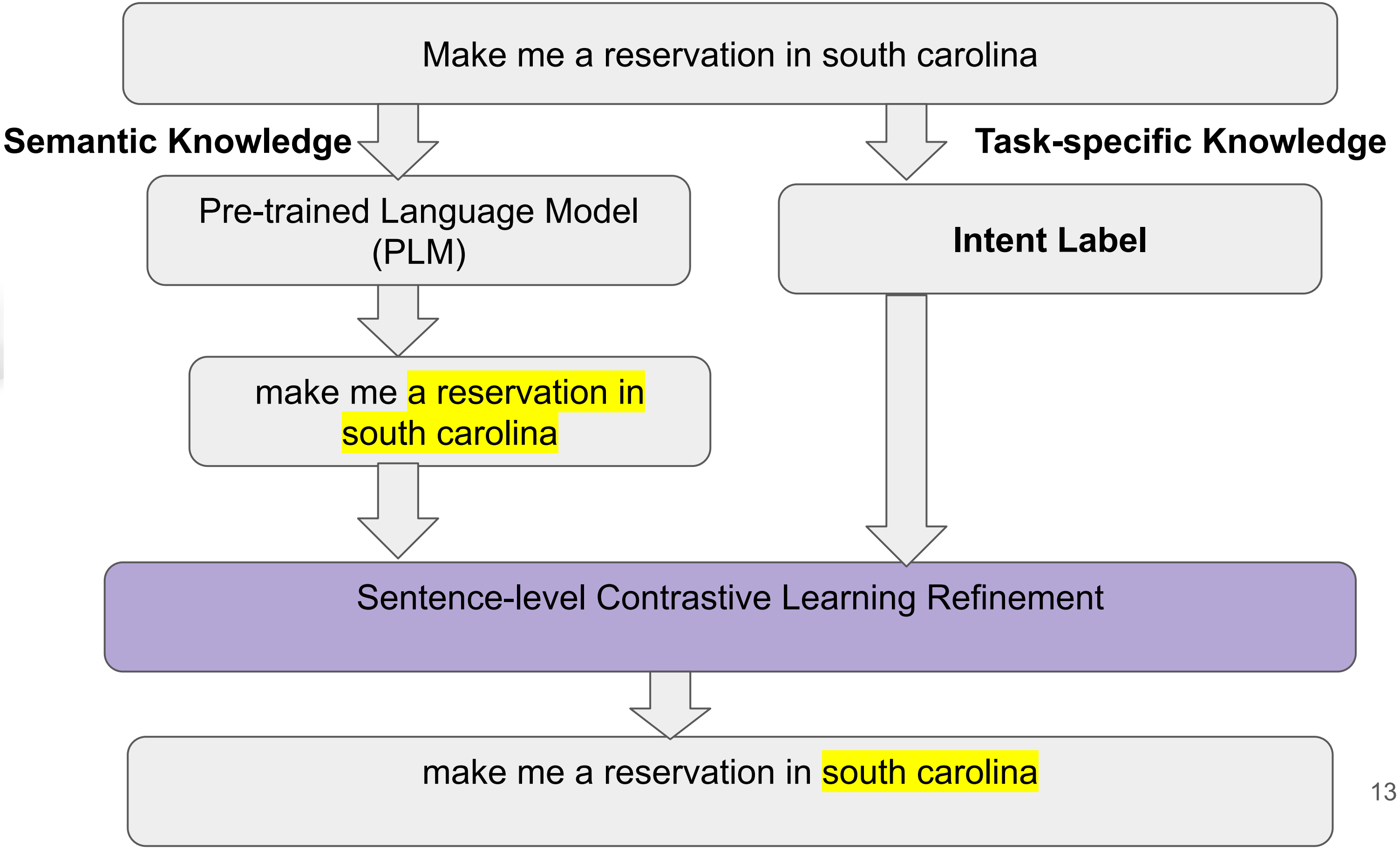
Segment-level Contrastive Learning Refinement

make me a reservation in south carolina

MOTIVATION

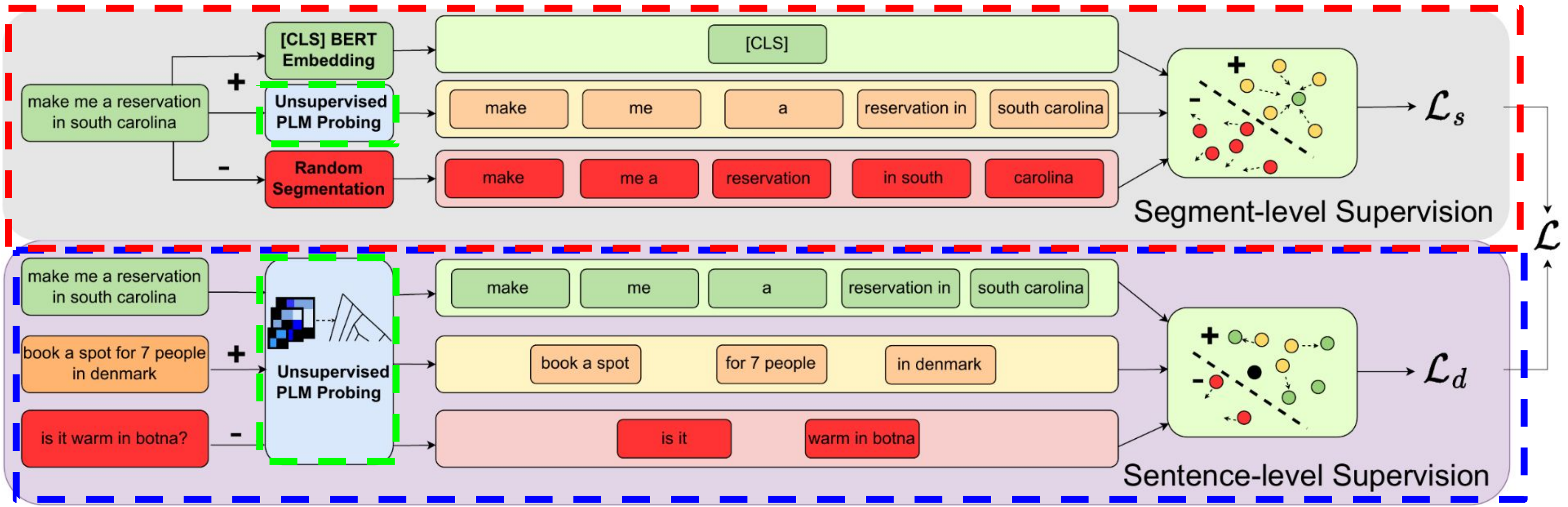
Utterance

Slot Induction



FRAMEWORK

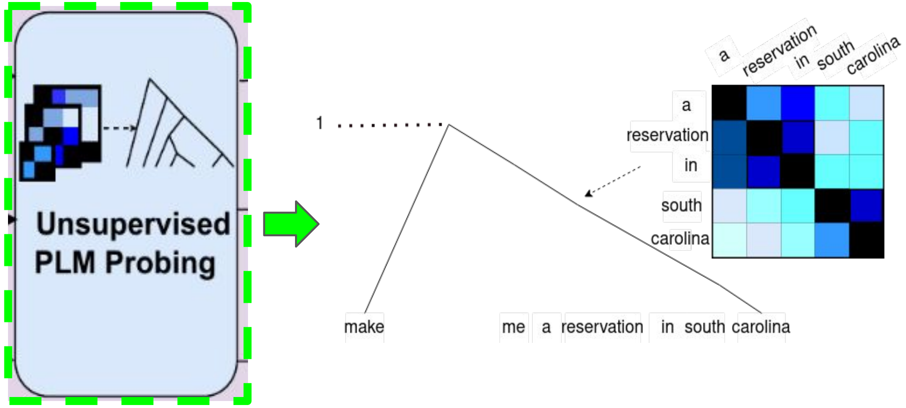
Framework



- **Segment-level Contrastive Learning (SegCL):**
  - **Refining semantic segments obtained from UPL via overall sentence semantic representation**
- **Sentence-level Contrastive Learning (SentCL):**
  - **Refining semantic segments obtained from UPL by exploiting samples with similar intent**

Framework

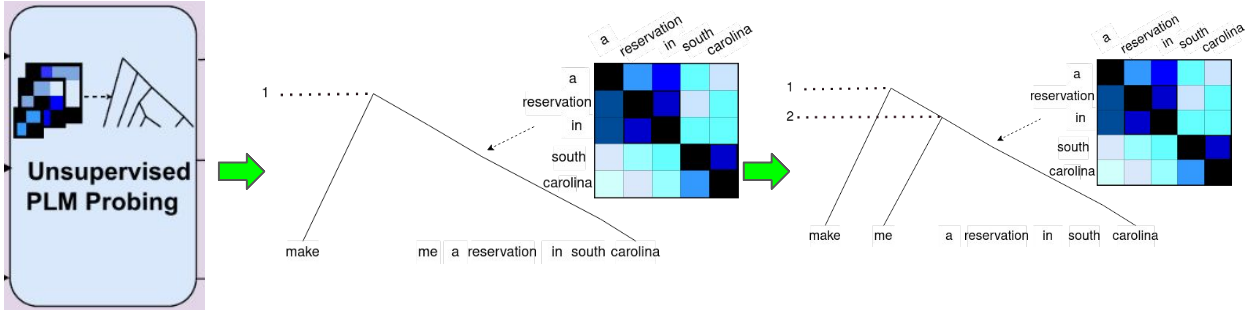
- **Unsupervised Pre-trained Language Model Probing (UPL)**
  - Extracting coherent semantic segments captured by PLM
  - **Perturbed Masking**[1]: Iteratively deciding the **split positions** of utterances via Impact Matrix **until token level is reached**
  - **Impact Matrix:** measuring impact score of every possible token pairs of utterances.



[1] Wu et al., Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT ACL 2020

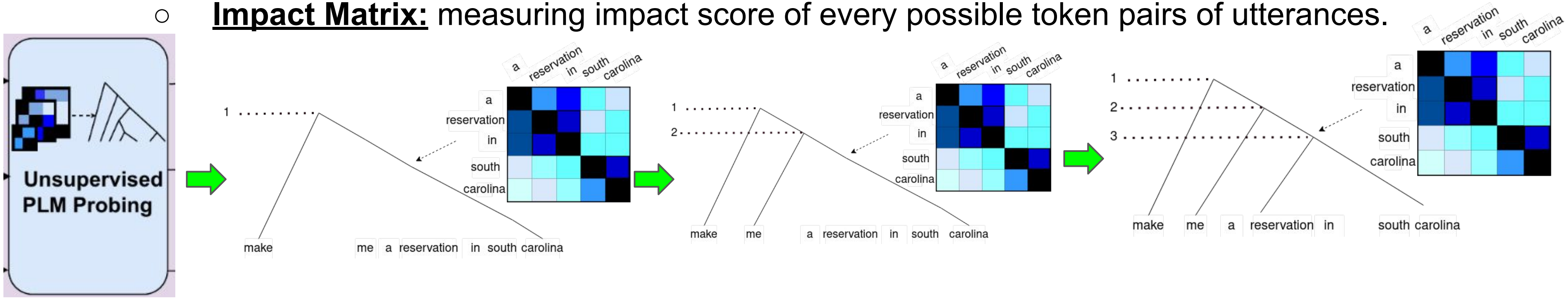
Framework

- **Unsupervised Pre-trained Language Model Probing (UPL)**
  - Extracting coherent semantic segments captured by PLM
  - **Perturbed Masking**: Iteratively deciding the **split positions** of utterances via Impact Matrix **until token level is reached**
  - **Impact Matrix:** measuring impact score of every possible token pairs of utterances.



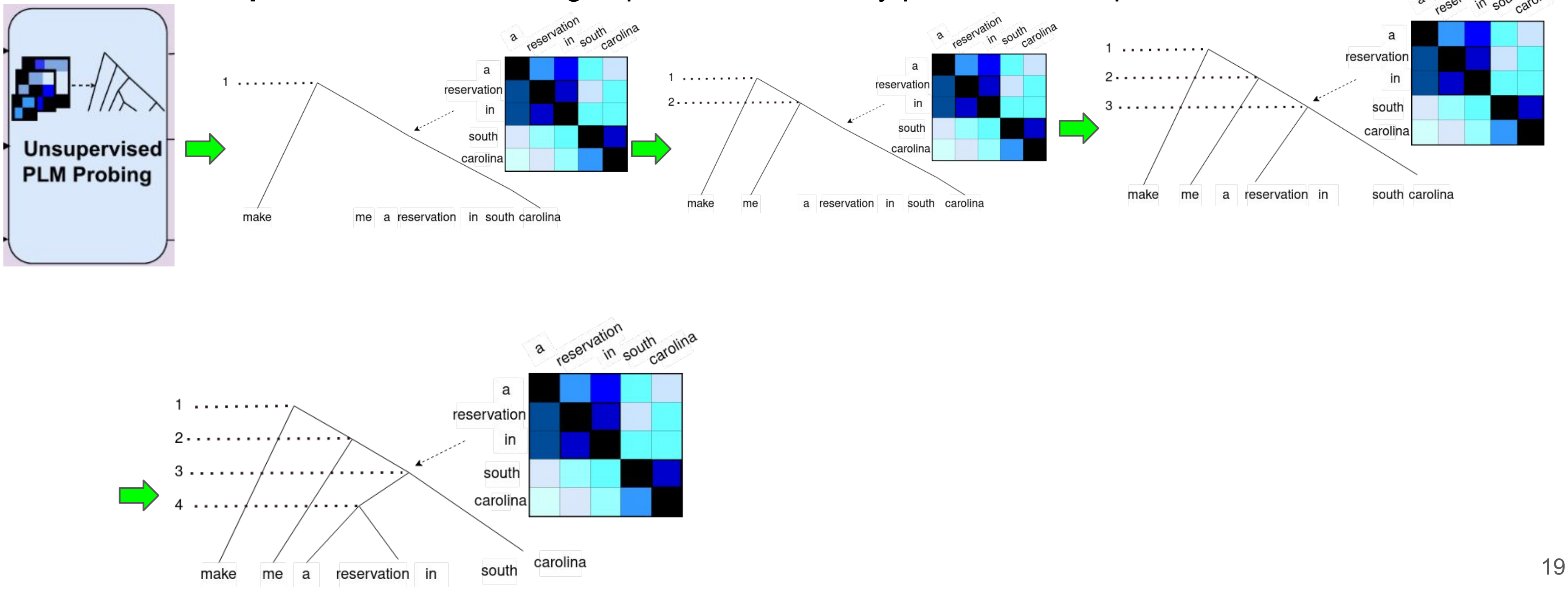
Framework

- **Unsupervised Pre-trained Language Model Probing (UPL)**
  - Extracting coherent semantic segments captured by PLM
  - **Perturbed Masking**: Iteratively deciding the **split positions** of utterances via Impact Matrix **until token level is reached**
  - **Impact Matrix:** measuring impact score of every possible token pairs of utterances.



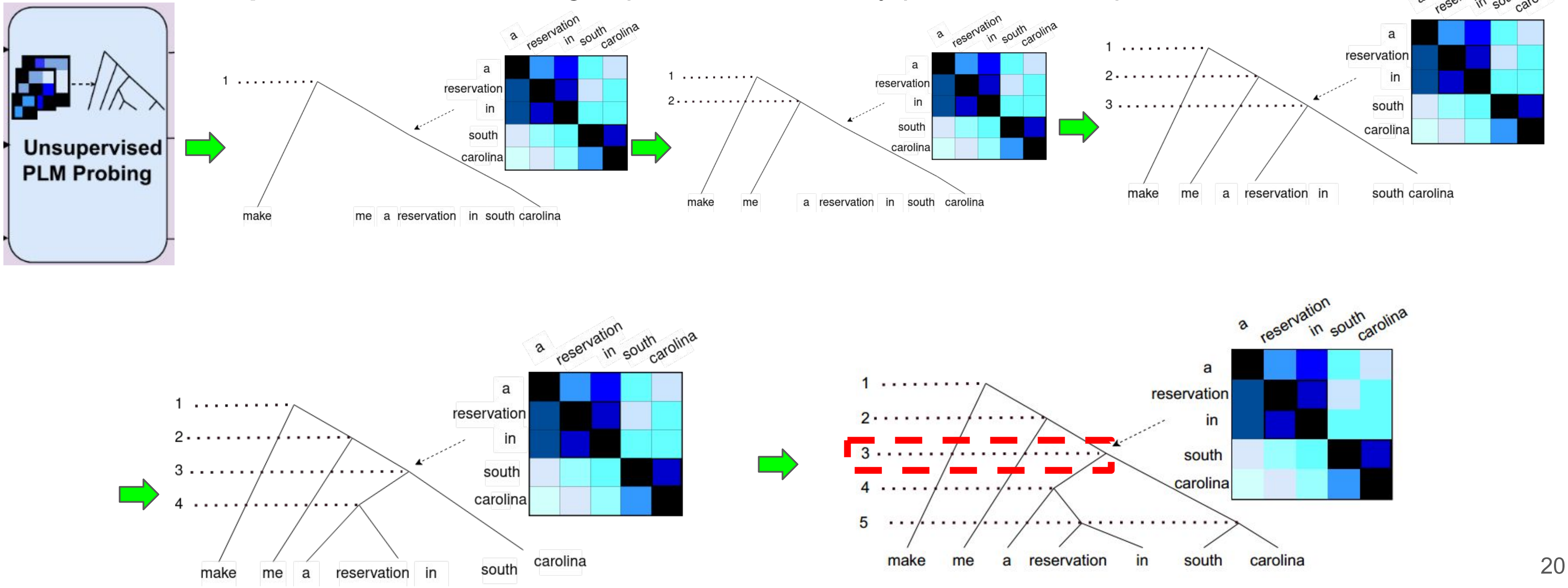
Framework

- **Unsupervised Pre-trained Language Model Probing (UPL)**
  - Extracting coherent semantic segments captured by PLM
  - <u>**Perturbed Masking**</u>: Iteratively deciding the **split positions** of utterances via Impact Matrix **until token level is reached**
  - <u>**Impact Matrix:**</u> measuring impact score of every possible token pairs of utterances.

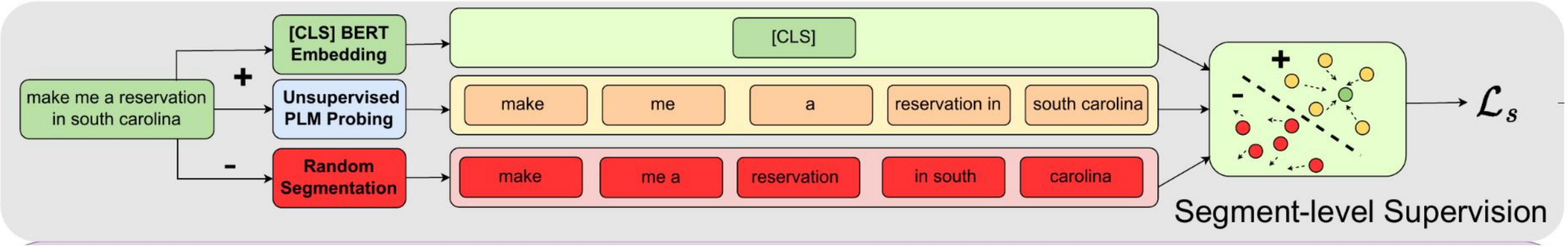


Framework

- **Unsupervised Pre-trained Language Model Probing (UPL)**
  - Extracting coherent semantic segments captured by PLM
  - **<u>Perturbed Masking</u>**: Iteratively deciding the **split positions** of utterances via Impact Matrix **until token level is reached**
  - **<u>Impact Matrix:</u>** measuring impact score of every possible token pairs of utterances.

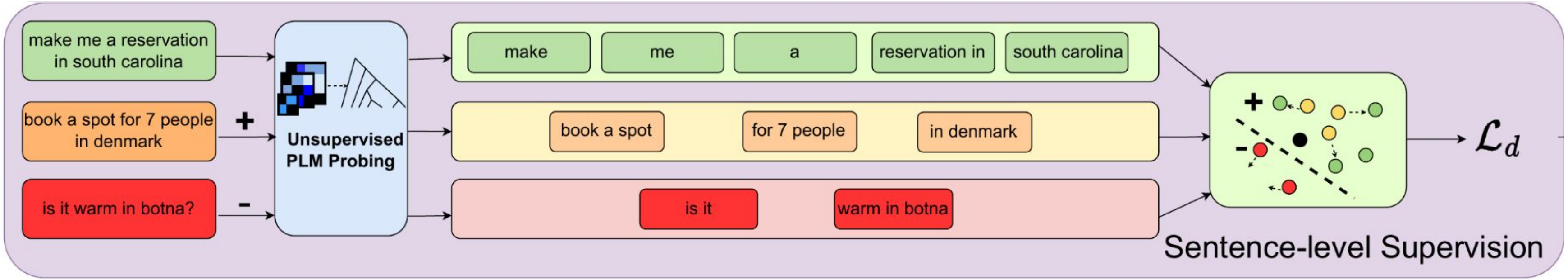


Framework



- **Segment-level Contrastive Learning (SegCL):**
  - Enhancing quality of PLM segments in an unsupervised way
  - Incorporating overall semantic representation from special [CLS] token as guidance
    - **Anchor:** Overall semantic representation
    - **Positive:** PLM's Segment representation of input utterance
    - **Negative:** Random segment representation of input utterance

Framework



- **Sentence-level Contrastive Learning (SentCL):**
  - Enhancing quality of semantic segments with available sentence-level intents
  - Encouraging semantic alignment between samples of similar intents
    - Anchor: PLM Segment representation of input utterance
    - Positive: PLM Segment representation of samples from **similar intents**
    - Negative: PLM Segment representation of samples from **different intents**

EVALUATION

DATASET

Table 1: Details of SNIPS and ATIS datasets.

	SNIPS_P1	SNIPS_P2	ATIS_P1	ATIS_P2
# Intents	5	2	14	7
# Slots	31	16	68	63
# Train Samples	9356	–	3811	–
# Validation Samples	500	–	414	–
# Test Samples	501	4127	750	895
Avg Train Sent Length	8.65	–	11.67	–
Avg Valid Sent Length	8.72	–	11.82	–
Avg Test Sent Length	8.71	9.87	10.68	8.92

- **Evaluation Task 1: Slot Induction (P1)**
  - **Objective:** Slot Induction Capability
  - **Metric:** H-Mean of Break-Tie mechanism

- **Evaluation Task 2: Generalization towards Emerging Intents (P2)**
  - **Objective:** Generalization capability of SI refinement method
  - **Metric:** Slot Filling metrics (Precision, Recall, F1)

EVALUATION TASK 1

Table 2: Experimental performance result on SNIPS dataset over 3 runs (**H-Mean** is considered the golden criteria for SI (Section 3)). ¶ denotes models that do not require random initializations.

	Model	Prior Knowledge	Break			Tie			H-Mean
			B-P	B-R	B-F1	T-P	T-R	T-F1	
Upper Bound	Joint BERT FT	Slot + Intent	96.91 ± 0.17	96.62 ± 0.69	96.76 ± 0.26	73.55 ± 0.38	73.39 ± 1.03	73.47 ± 0.38	83.52 ± 0.16
	FlairNLP ¶	POS & NER	80.04	62.81	70.38	48.25	63.31	54.77	61.60
	SpaCy ¶	POS	75.73	50.29	60.45	41.71	62.97	50.18	54.84
Comparable	DP-LB ¶	–	59.68	34.27	43.54	21.69	38.53	27.76	33.90
	DP-RB ¶	–	66.53	52.56	58.73	33.97	52.24	41.17	48.40
	AutoPhrase	External KB	65.51 ± 0.23	57.16 ± 2.59	61.05 ± 1.15	33.39 ± 0.74	36.62 ± 1.67	34.93 ± 1.50	44.43 ± 1.64
	UCPhrase	PLM	42.25 ± 4.90	20.26 ± 2.71	27.39 ± 1.95	36.06 ± 2.42	**73.53 ± 3.33**	**48.39 ± 2.91**	34.98 ± 2.35
	USSI ¶	PLM	**83.21**	62.12	71.14	33.96	49.93	40.42	51.55
	Ours (w/o CL) ¶	PLM	75.36	66.70	70.76	38.51	45.81	41.84	52.59
	Ours (w/o SentCL)	PLM	76.09 ± 0.73	66.43 ± 0.29	70.94 ± 0.49	39.15 ± 0.60	47.9 ± 0.91	43.09 ± 0.73	53.61 ± 0.71
	Ours (full)	**PLM + Intent**	76.87 ± 0.25	**67.77 ± 0.26**	**72.00 ± 0.24**	**40.39 ± 0.16**	48.49 ± 0.19	44.07 ± 0.04	**54.68 ± 0.08**

- **Upper Bound:** requiring token-level annotations during training/ pre-training
- **Comparable Method:** no token-level annotations are involved during training.

- **Ours** bridges the gap with Upper Bound Methods in terms of **H-Mean**
- **Ours** exceeds the Comparable Methods in terms of **H-Mean**

EVALUATION TASK 1

Table 3: Experimental performance result on ATIS dataset over 3 runs (**H-Mean** is considered the golden criteria for SI (Section 3)). ¶ denotes models that do not require random initializations.

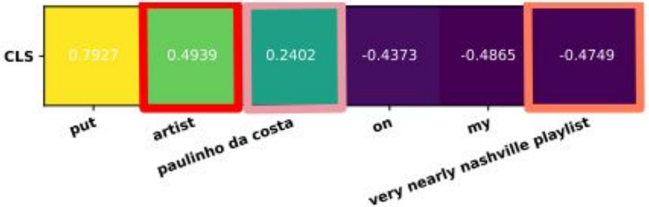
	Model	Prior Knowledge	Break			Tie			H-Mean
			B-P	B-R	B-F1	T-P	T-R	T-F1	
Upper Bound	Joint BERT FT	Slot + Intent	98.49 ± 0.24	99.33 ± 0.08	98.91 ± 0.09	59.07 ± 0.36	58.27 ± 0.89	58.67 ± 0.63	73.65 ± 0.54
	FlairNLP ¶	POS & NER	95.44	77.90	85.78	41.34	61.91	49.58	62.84
	SpaCy ¶	POS	94.45	69.64	80.17	35.33	61.17	44.79	57.47
Comparable	DP-LB ¶	–	80.80	36.38	50.17	12.32	38.51	18.67	27.21
	DP-RB ¶	–	84.24	**66.84**	**74.54**	14.81	30.52	19.94	31.46
	AutoPhrase	External KB	75.96 ± 0.04	40.06 ± 0.28	52.46 ± 0.18	19.75 ± 0.21	49.33 ± 0.38	**28.20 ± 0.28**	36.68 ± 0.21
	UCPhrase	PLM	47.25 ± 0.04	17.27 ± 0.72	25.29 ± 0.78	17.36 ± 0.16	**58.21 ± 0.68**	26.75 ± 0.11	26.00 ± 0.47
	USSI ¶	PLM	**95.06**	56.36	70.77	14.78	45.22	22.28	33.89
	Ours (w/o CL) ¶	PLM	86.40	61.53	71.87	18.23	35.27	24.04	36.03
	Ours (w/o SentCL)	PLM	87.29 ± 0.15	64.21 ± 0.27	73.99 ± 0.13	20.09 ± 0.08	35.86 ± 0.35	25.75 ± 0.08	38.20 ± 0.08
	Ours (full)	**PLM + Intent**	87.80 ± 0.27	63.27 ± 0.67	73.54 ± 0.36	**20.53 ± 0.14**	37.89 ± 0.99	26.63 ± 0.26	**39.10 ± 0.24**

- **Ours** remains competitive among Comparable methods
- The gap between **Comparable** and **Upper Bound** methods are more significant

ABLATION STUDY

Contribution of Multi-level CL

	SNIPS	ATIS
Ours (w/o CL)	52.59	36.03
+ SegCL	53.61 ± 0.71	38.20 ± 0.08
+ SentCL (w/o aug)	53.44 ± 0.22	37.59 ± 0.81
+ SentCL (w aug)	54.23 ± 0.10	38.12 ± 0.36
Ours (full)	$\mathbf{54.68 \pm 0.08}$	$\mathbf{39.10 \pm 0.24}$



(a) Segment-level Supervised Positive-Anchor Pair

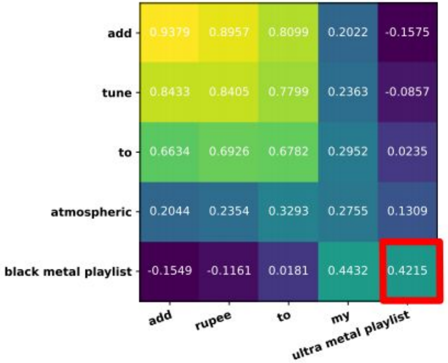
(b) Segment-level Supervised Negative-Anchor Pair

Segment-level Contrastive Learning is effective

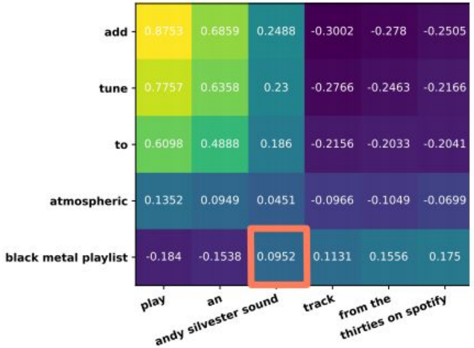
ABLATION STUDY

Contribution of Multi-level CL

	SNIPS	ATIS
Ours (w/o CL)	52.59	36.03
+ SegCL	53.61 ± 0.71	38.20 ± 0.08
+ SentCL (w/o aug)	53.44 ± 0.22	37.59 ± 0.81
+ SentCL (w aug)	54.23 ± 0.10	38.12 ± 0.36
Ours (full)	$\mathbf{54.68 \pm 0.08}$	$\mathbf{39.10 \pm 0.24}$



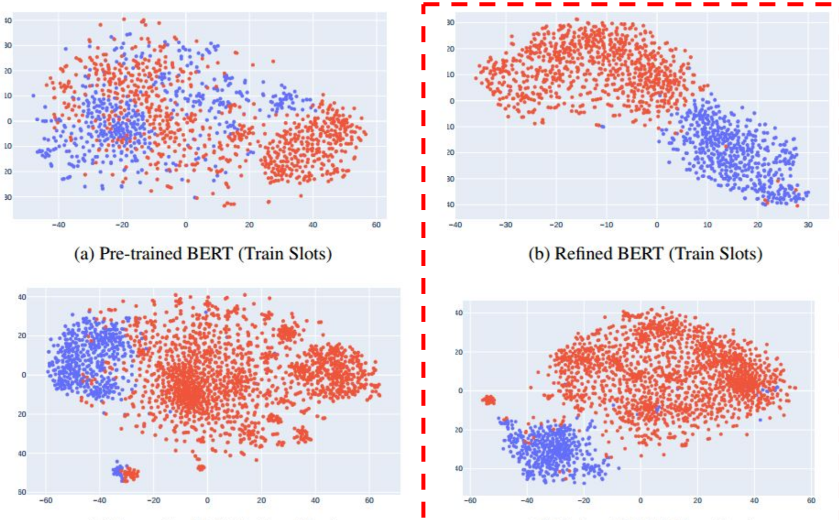
(c) Sentence-level Supervised Positive-Anchor Pair



(d) Sentence-level Supervised Negative-Anchor Pair

Sentence-level Contrastive Learning is effective

EVALUATION TASK 2



(a) Pre-trained BERT (Train Slots) (b) Refined BERT (Train Slots) (c) Pre-trained BERT (Test Slots) (d) Refined BERT (Test Slots)

Figure 6: Slot Value Representation Visualization of the raw original pre-trained BERT and raw Refined BERT via SI on sample slot types from training set SNIPS_P1 ((a), (b)) and testing set SNIPS_P2 ((c), (d)). Blue and Red denotes slot values from randomly sampled ground truth slot types.

Table 5: Evaluation of SF task over 3 runs on Emerging Intents in SNIPS_P2 and ATIS_P2 datasets.

	SNIPS_P2		
	S-P	S-R	S-F1
Original BERT	14.11 ± 0.47	17.78 ± 0.82	15.73 ± 0.62
Refined BERT	$\mathbf{15.08 \pm 0.48}$	$\mathbf{19.61 \pm 0.23}$	$\mathbf{17.05 \pm 0.38}$
	ATIS_P2		
Original BERT	66.67 ± 0.82	63.35 ± 1.35	64.96 ± 0.74
Refined BERT	$\mathbf{70.12 \pm 0.85}$	$\mathbf{63.64 \pm 0.48}$	$\mathbf{66.72 \pm 0.66}$

SI Refinement provides effective initializations for token-level slot when generalized towards emerging intents

CONCLUSION

CONCLUSION

- Token-level Slot Induction via
  - **Unsupervised Pre-trained Language Model Probing:** inherent semantic knowledge extraction from PLM
  - **Multi-level Contrastive Learning:** semantic segment refinement
- Capability of improved initialization for token-level slot label tasks when generalized towards emerging intents

Thank you for your attendance

Questions?

Code + Data: https://github.com/nhhoang96/MultiCL_Slot_Induction