

National
Taiwan
University
國立臺灣大學

陳縉儂 Yun-Nung (Vivian) Chen

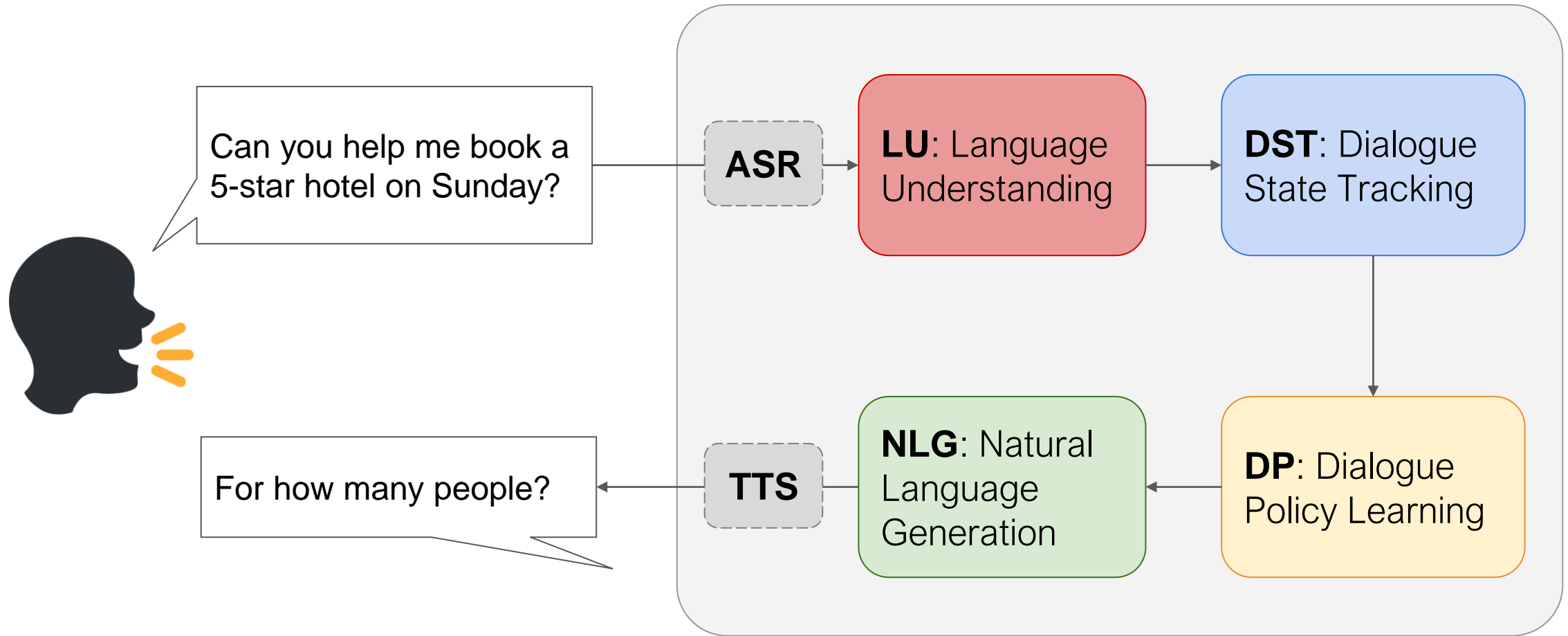
October 20th, 2023

Towards Human-Like Conversational AI

Advancing Understanding & Interaction



Task-Oriented Dialogue Systems ([Young, 2000](#))



3

Language Understanding (LU)



Can you help me book a
5-star hotel on Sunday?

LU: Language
Understanding

agoda

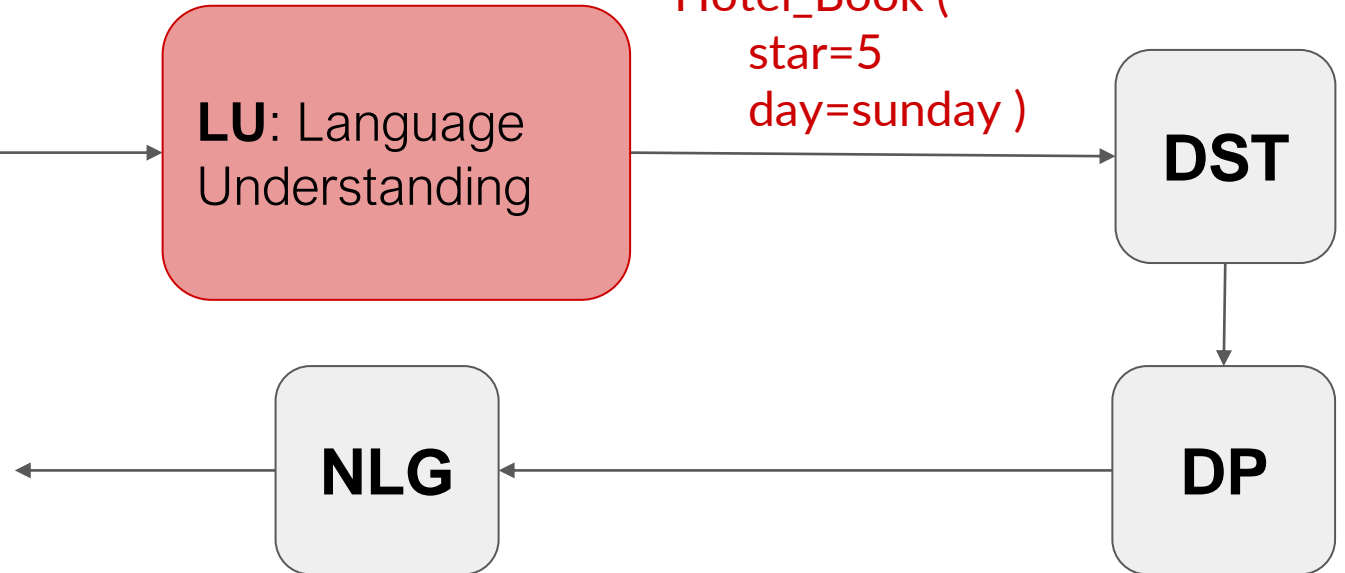


Hotel_Book (
star=5
day=sunday)

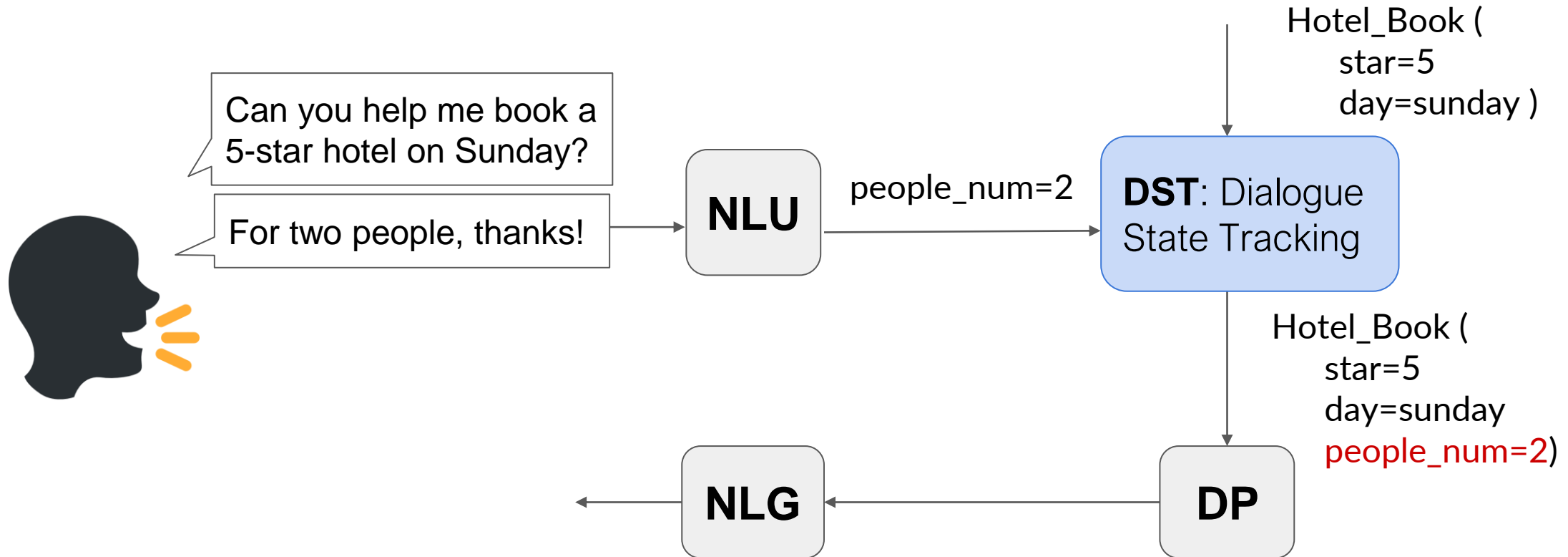
DST

DP

NLG

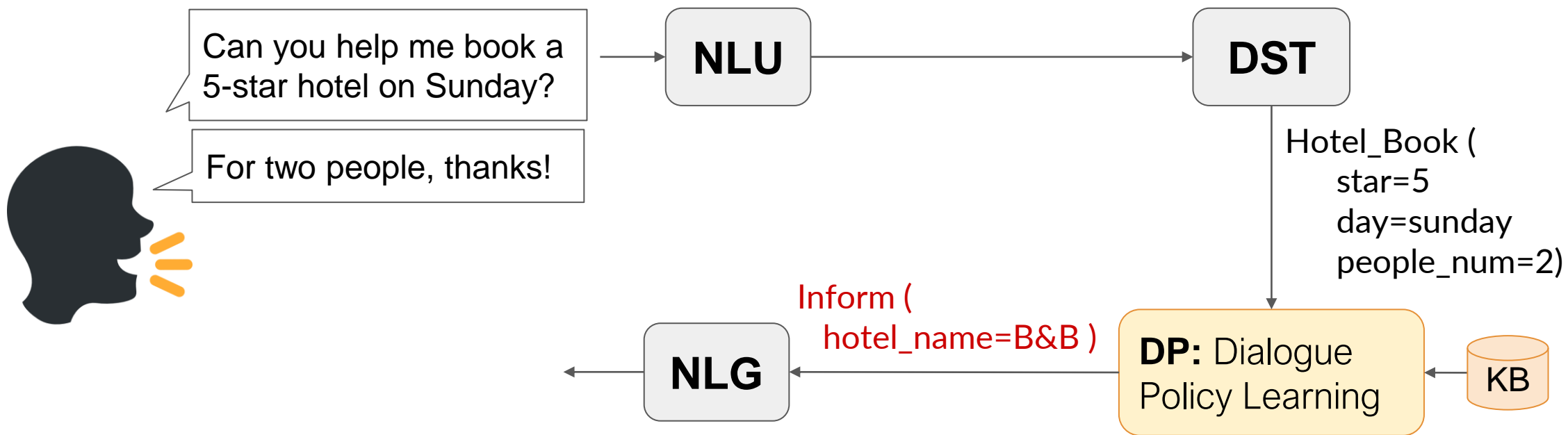


Dialogue State Tracking

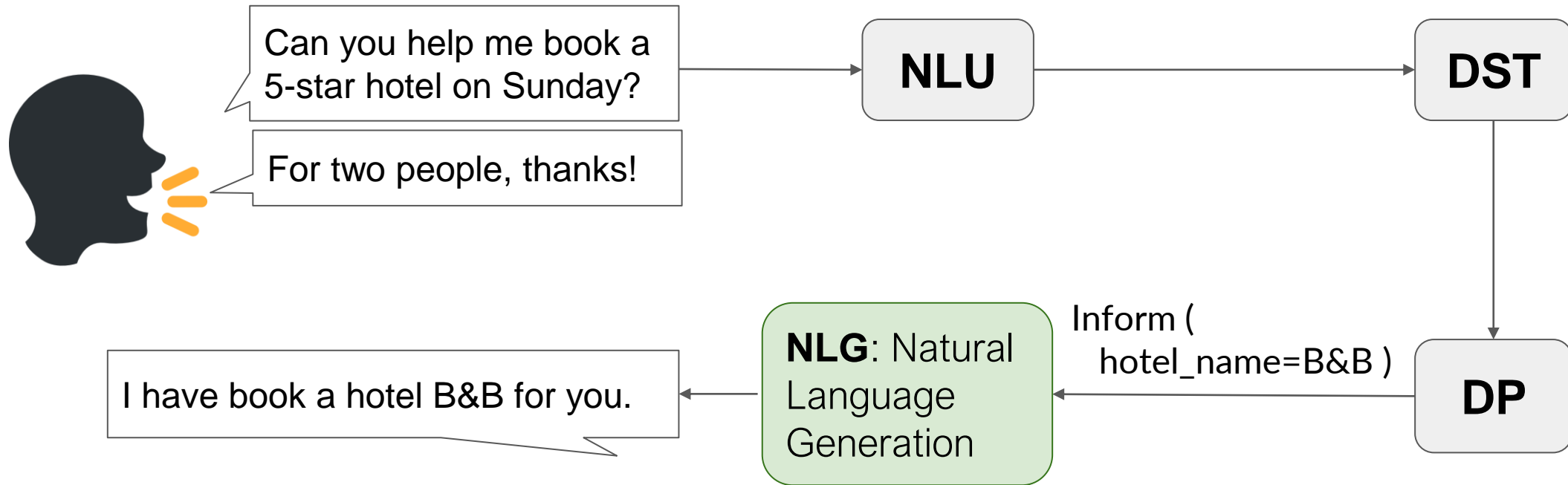


5

Dialogue Policy Learning



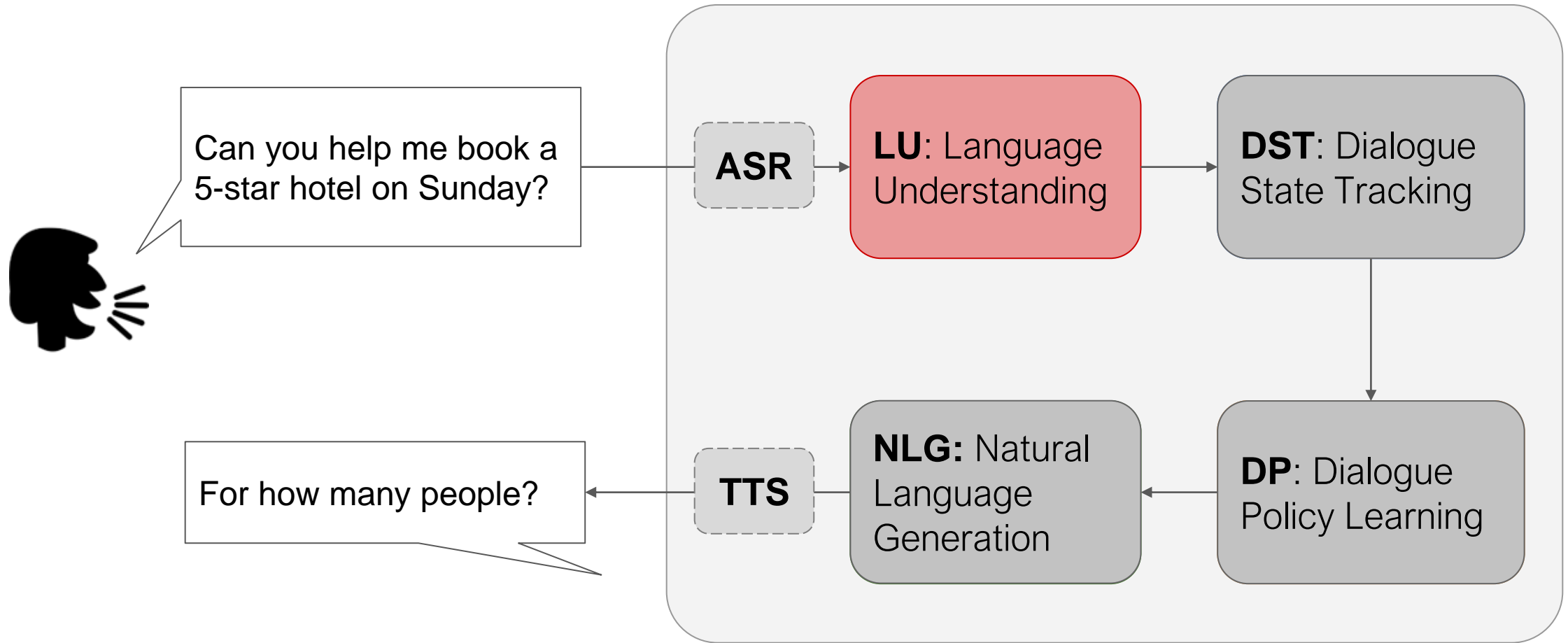
Natural Language Generation



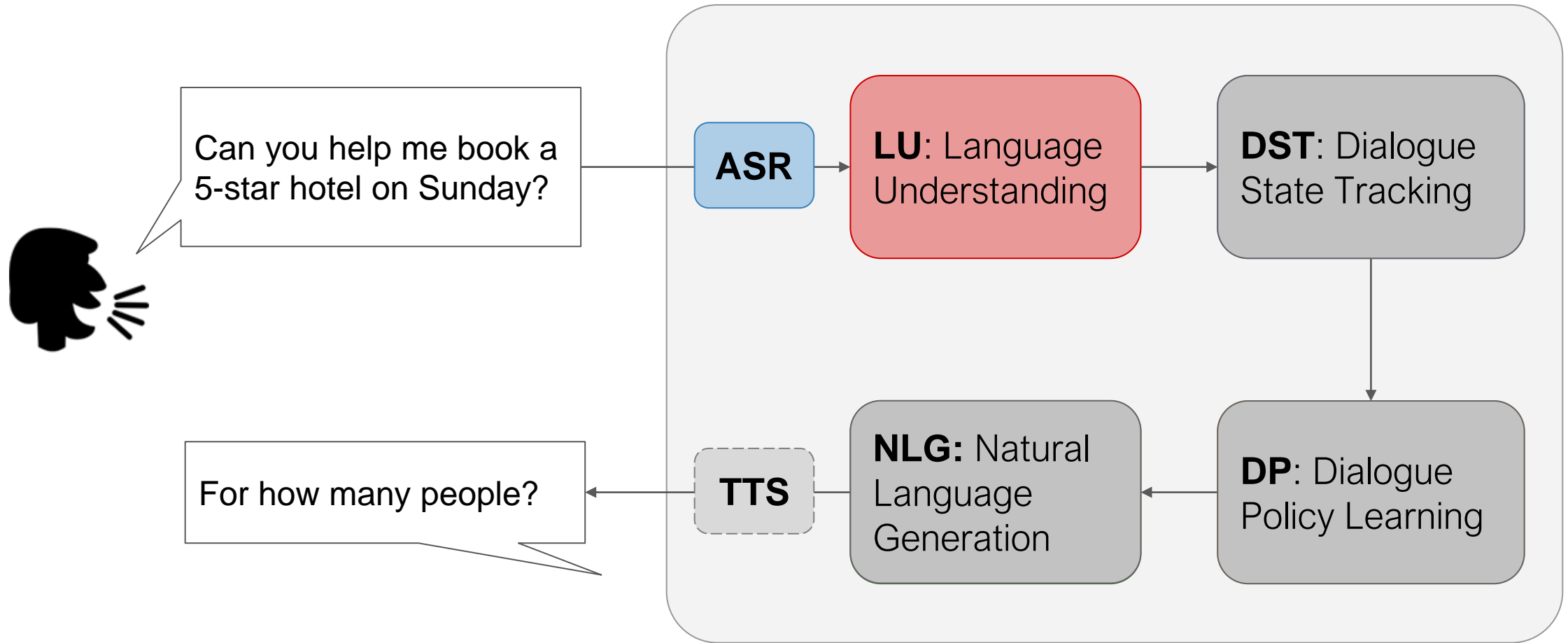
E R R O R

Understanding

Task-Oriented Dialogue Systems ([Young, 2000](#))

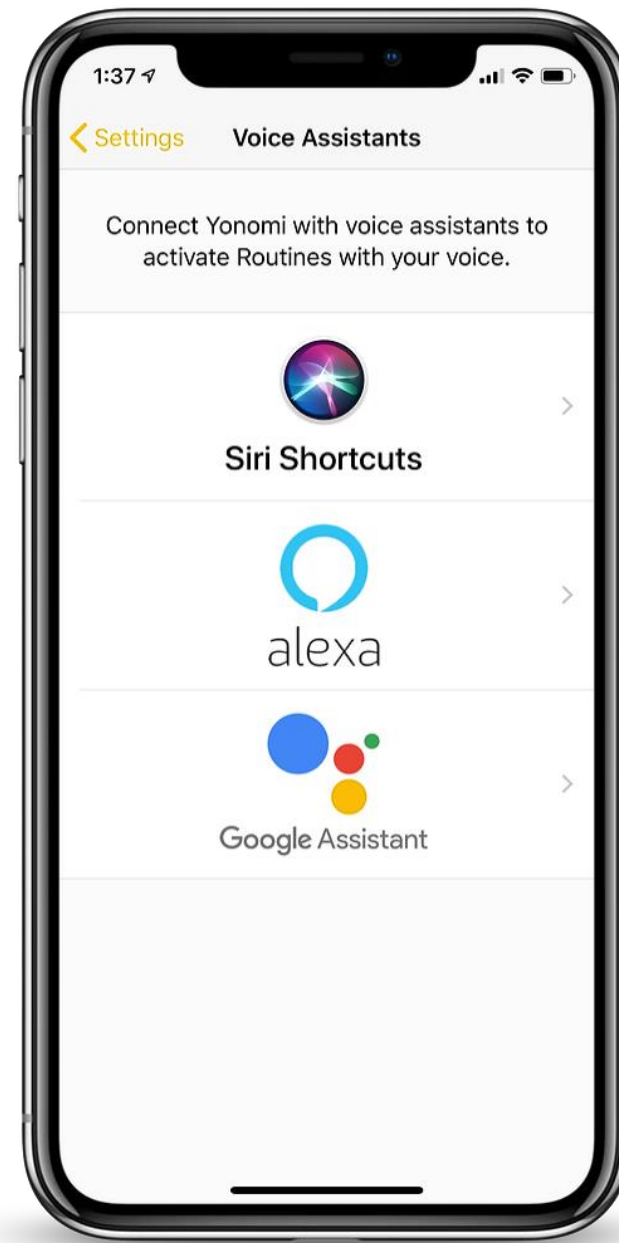


Task-Oriented Dialogue Systems ([Young, 2000](#))



Recent Advances in NLP

- Pre-trained models on text
 - ELMo, BERT, RoBERTa, XLM, GPT, etc.



Lift all lights ~~X~~ to Morocco
List all flights tomorrow



Mismatch between Written and Spoken Languages

Training

- Written language



Testing

- Spoken language
- Include recognition errors



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



Solution: LatticeLM

(Huang & Chen, ACL 2020)

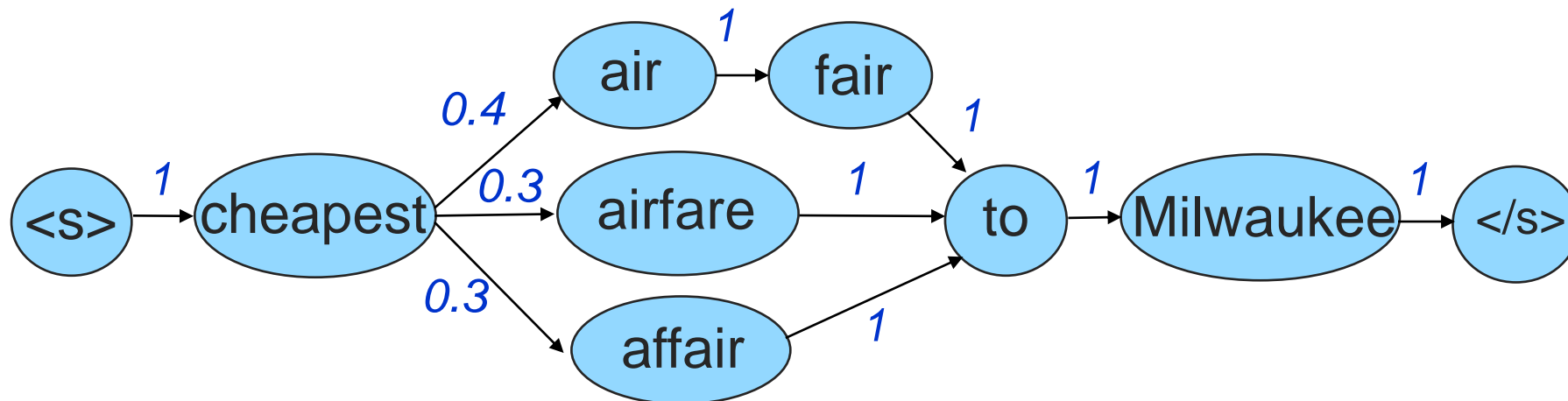
<https://github.com/MiuLab/LatticeLM>

13

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

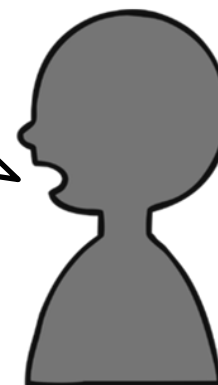


LatticeRNN
helps

(Ladhak, et al., 2016)

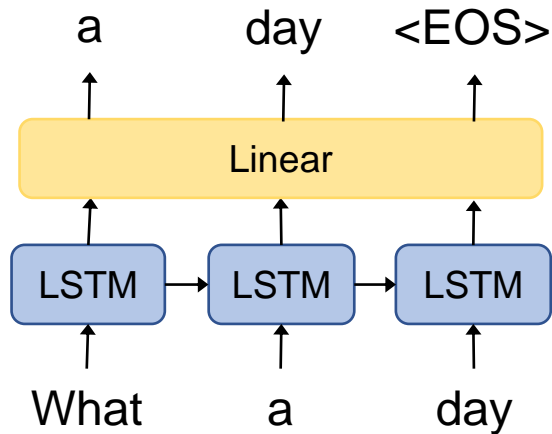


LM pre-training
helps

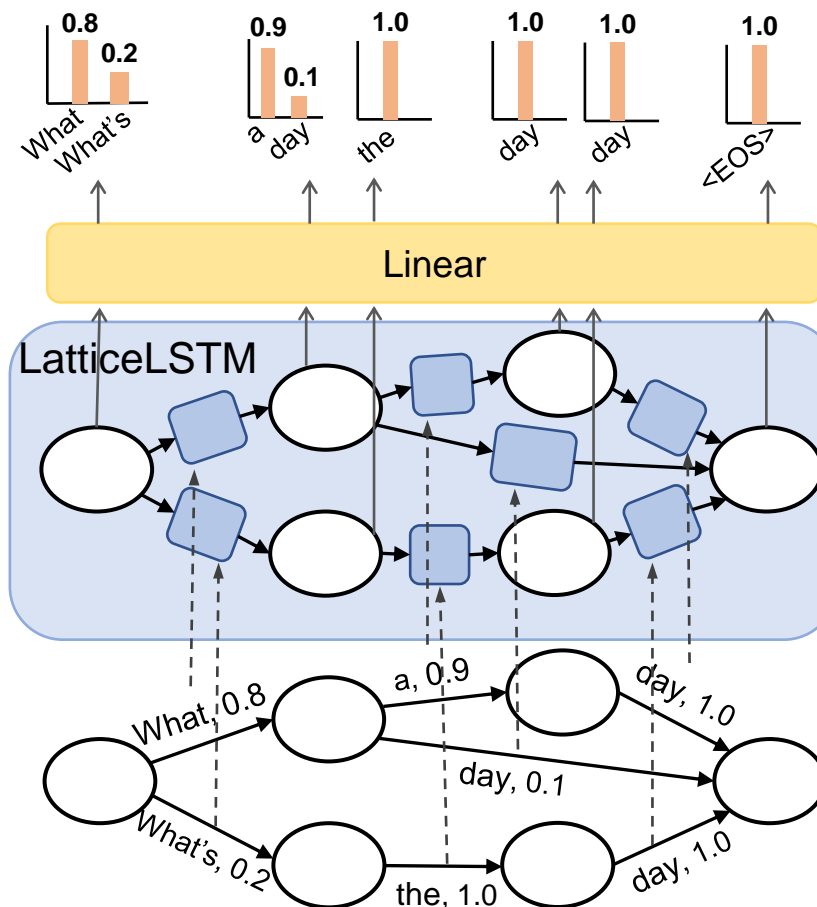


LatticeLM: Efficient Two-Stage Pre-Training

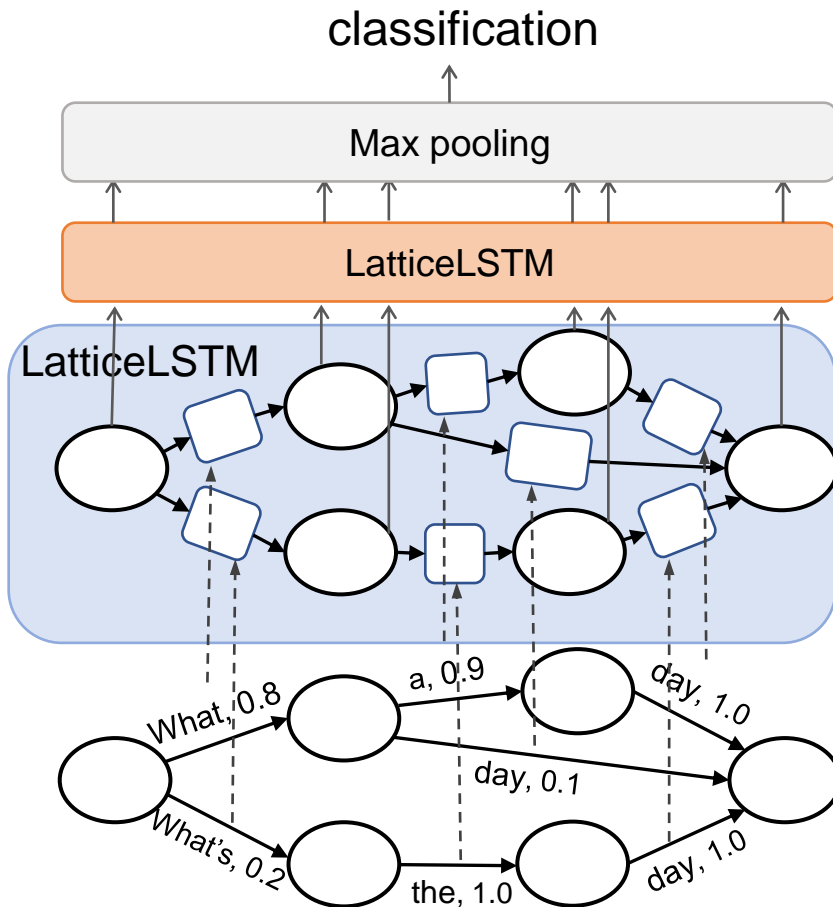
Stage 1: Pre-Training on Sequential Texts



Stage 2: Pre-Training on Lattices



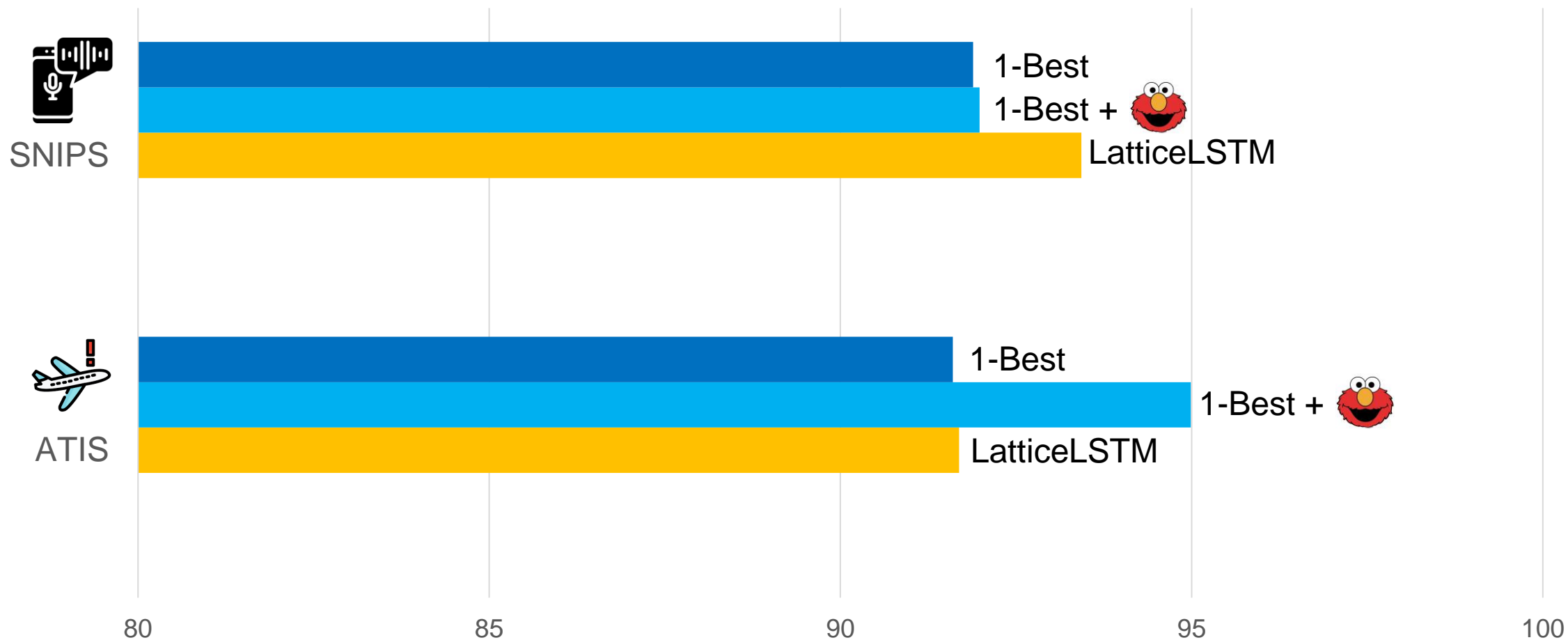
Fine-Tuning



Spoken Language Understanding Results

Intent Prediction

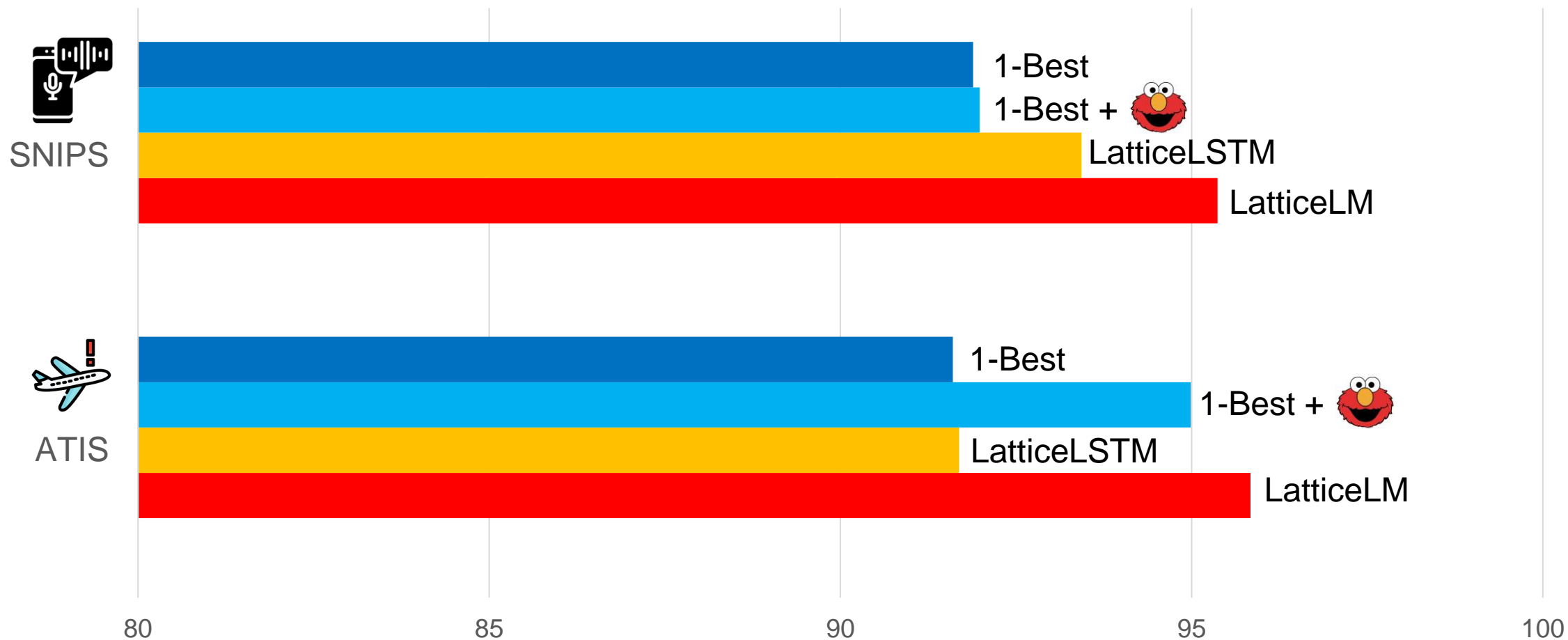
- Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



Spoken Language Understanding Results

Intent Prediction

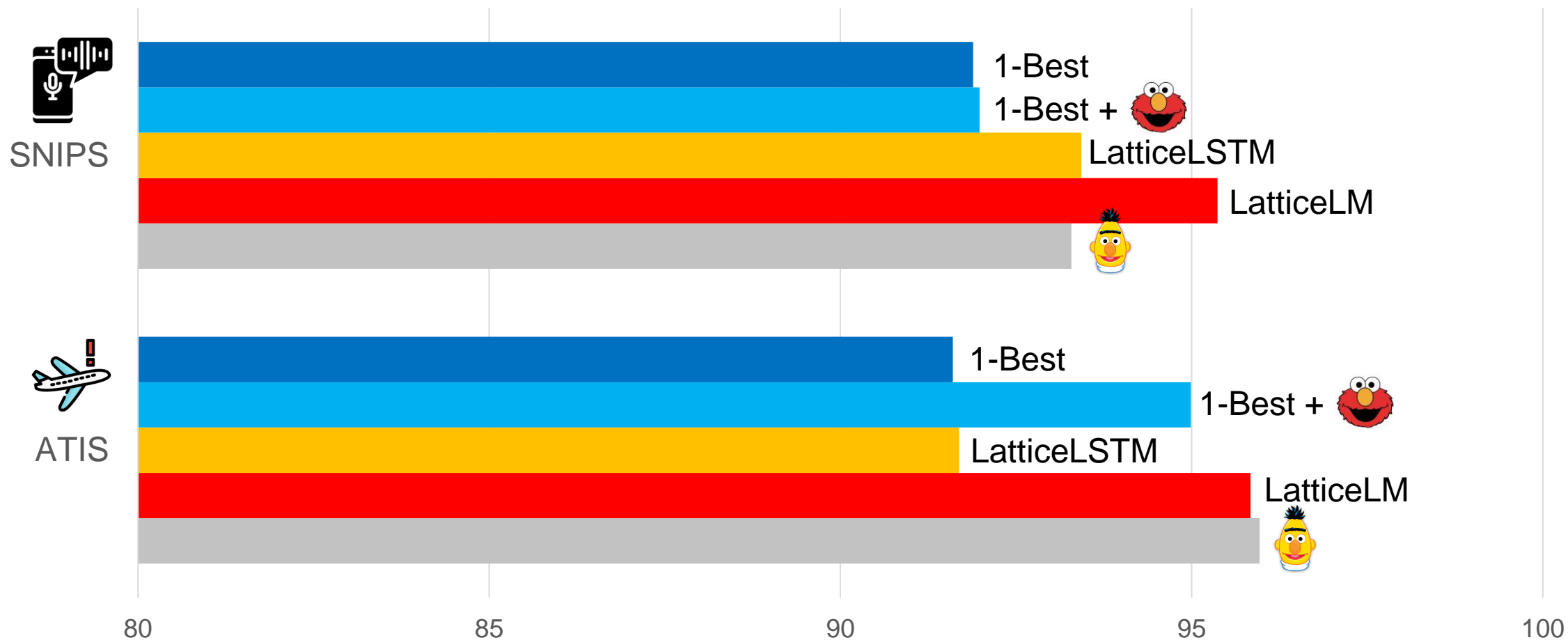
- Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



Spoken Language Understanding Results

Intent Prediction

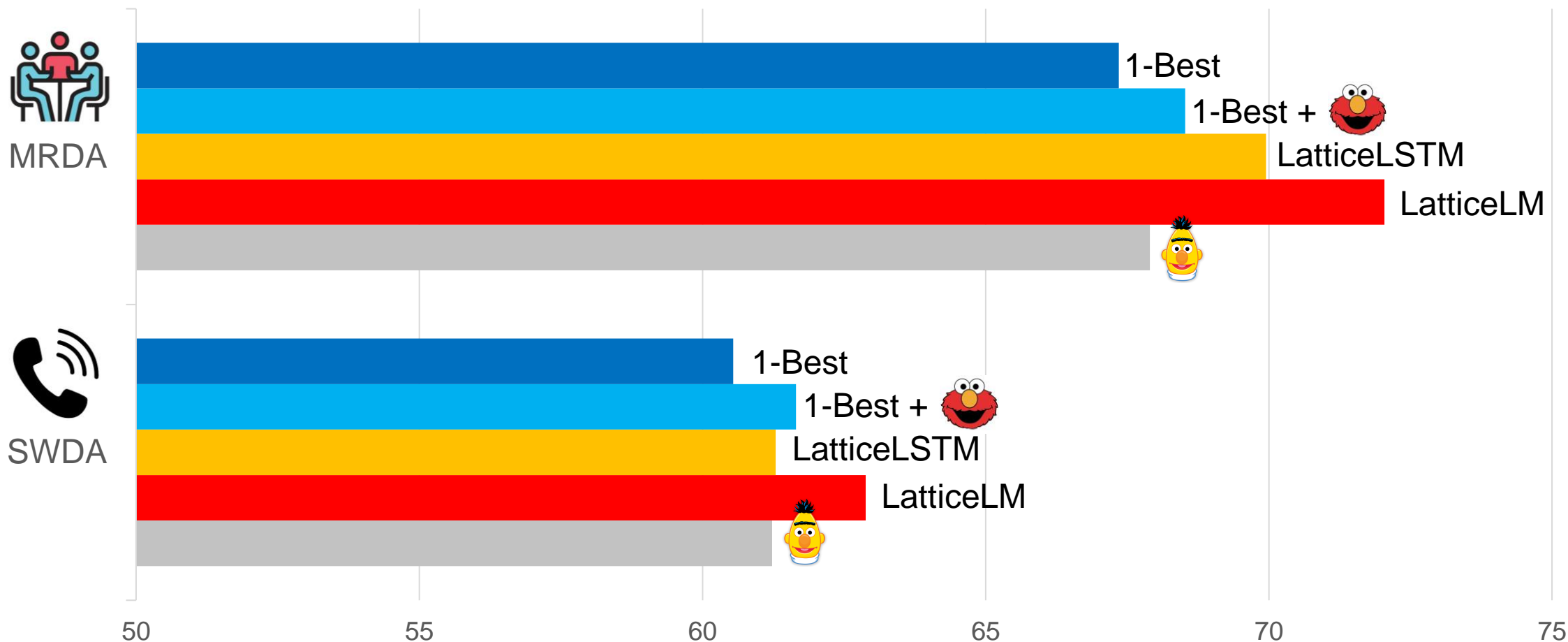
- Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)



Spoken Language Understanding Results

Dialogue Act Prediction

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



What if we only have texts from ASR?



Solution:

Contrastive Learning for

ASR-Robust Embeddings

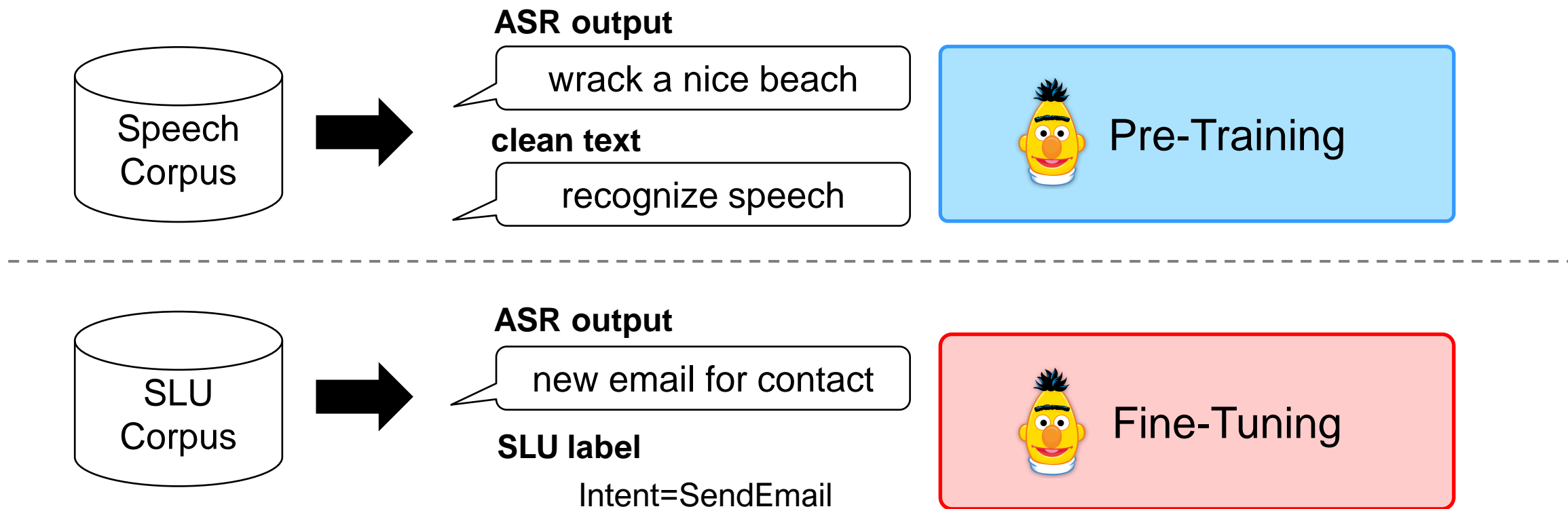
(Chang & Chen, INTERSPEECH 2022)

<https://github.com/MiuLab/SpokenCSE>

21

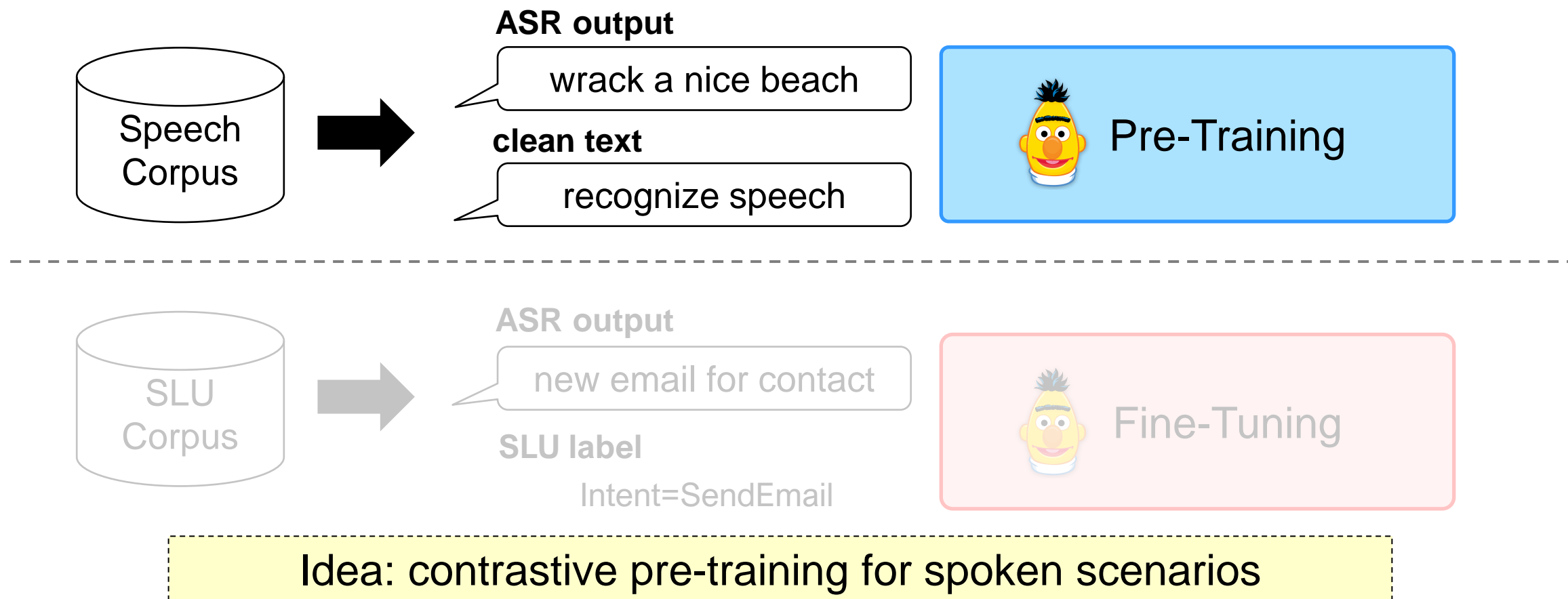
Improving ASR Robustness of Embeddings

- Idea: adapt embeddings robust to errors with only textual information



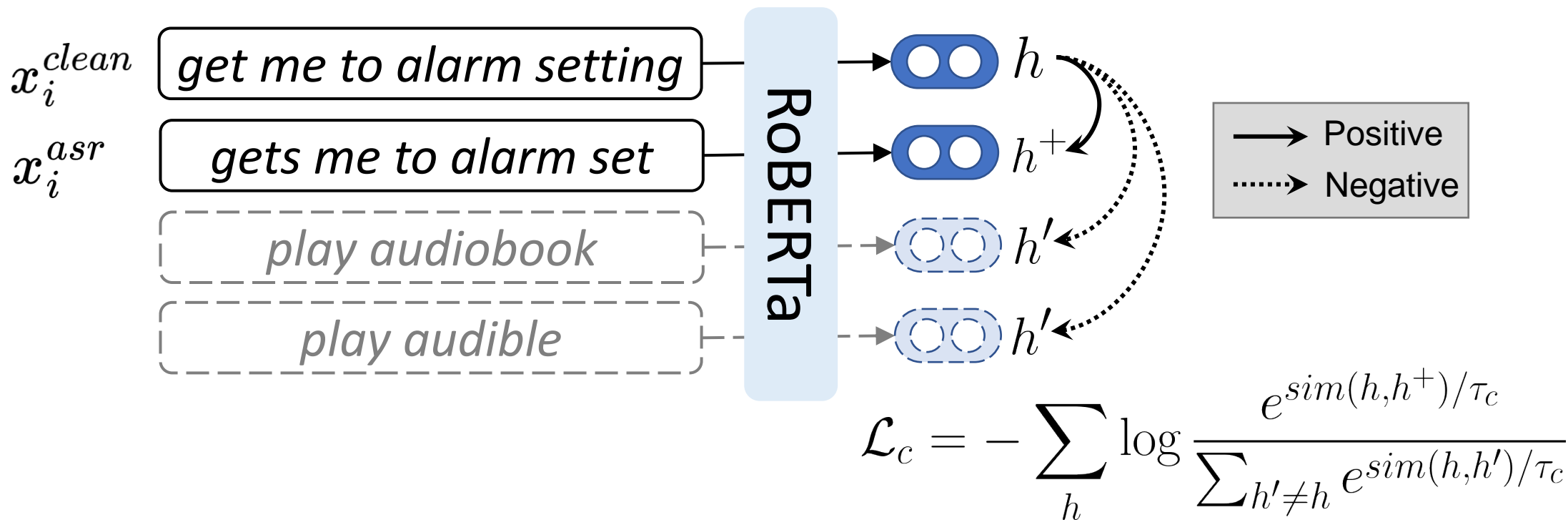
Improving ASR Robustness of Embeddings

- Idea: adapt embeddings robust to errors with only textual information



Contrastive Pre-Training

- Idea: ASR outputs have similar embeddings as their clean texts



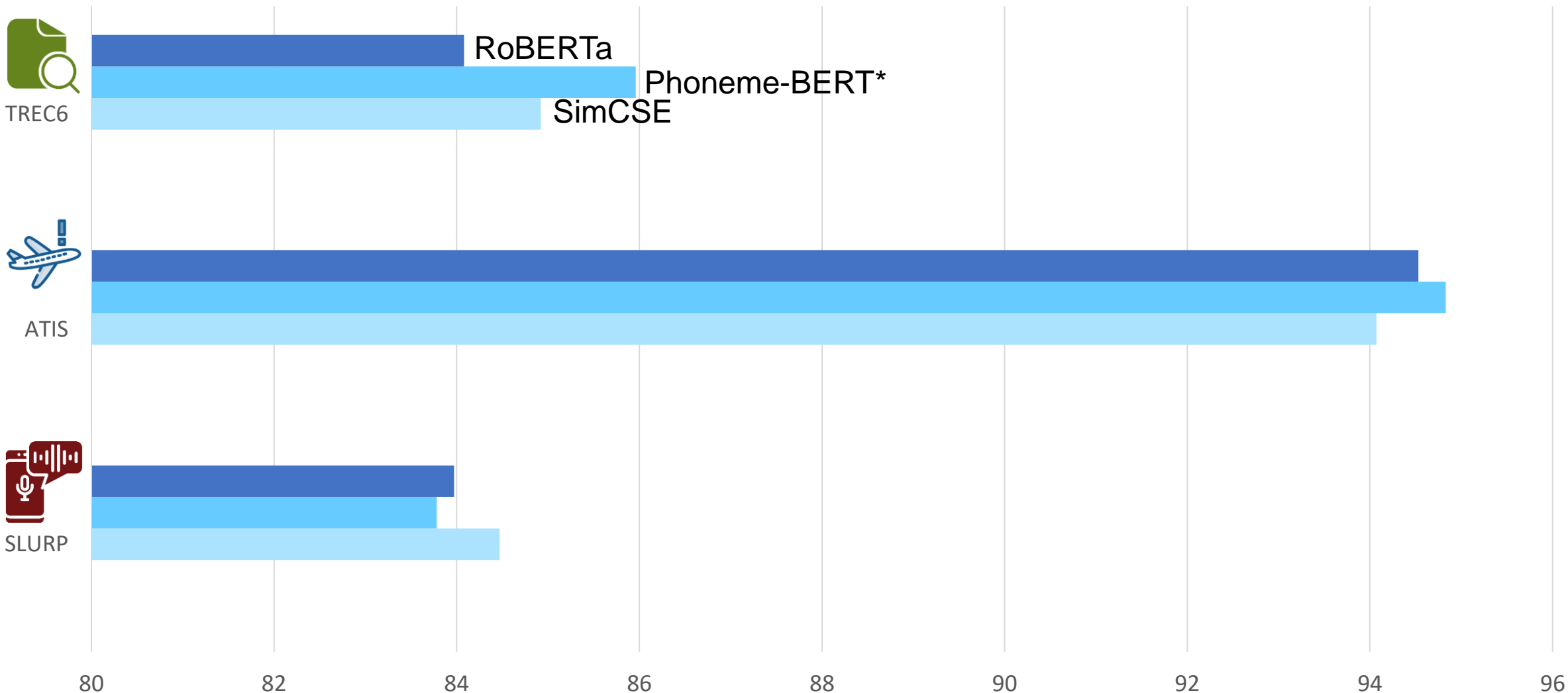
- Pre-training objective: $\mathcal{L}_{pt} = \mathcal{L}_c + \lambda_{mlm} \cdot \mathcal{L}_{mlm}$
prevent catastrophic forgetting

Spoken Language Understanding Results

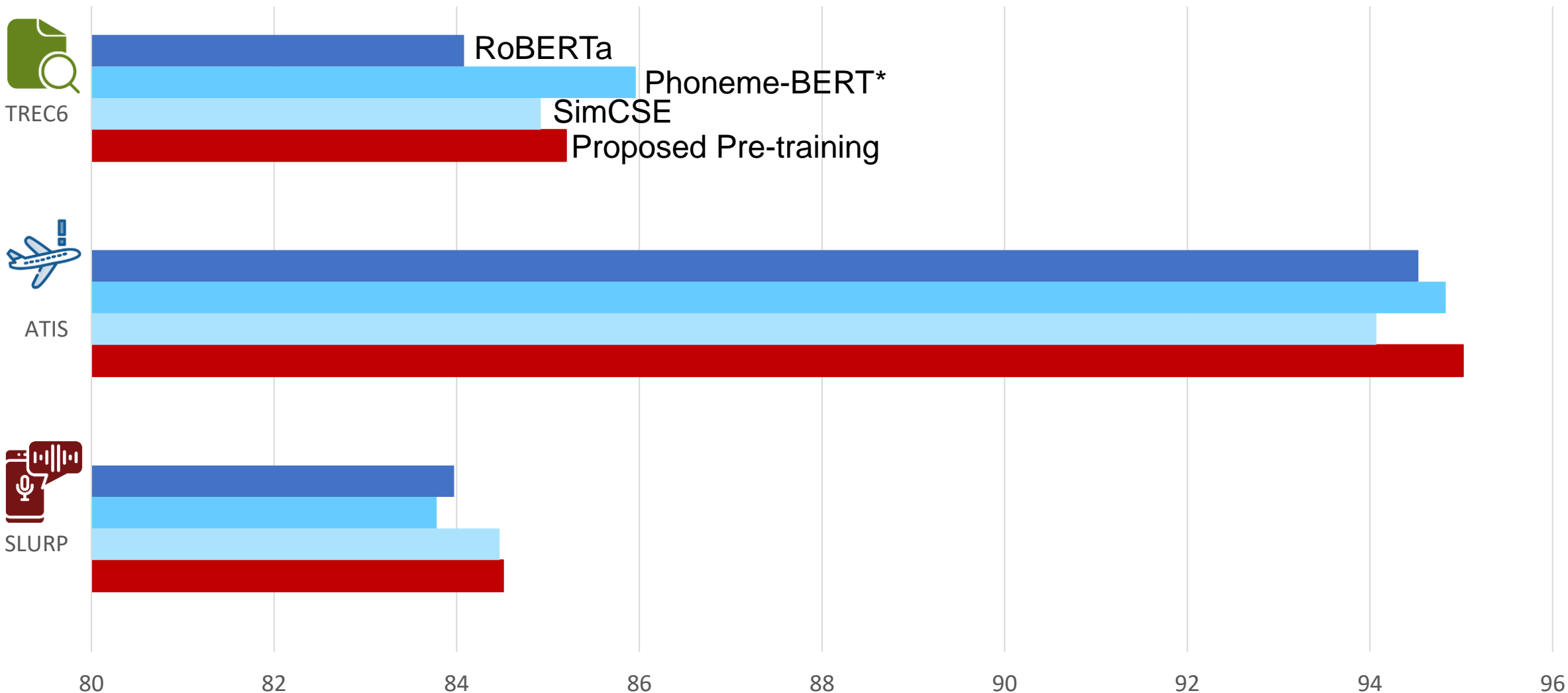
- SLU data
 - Synthesized TREC6 (WER=29%) & ATIS (WER=32%)
 - SLURP: Spoken Language Understanding Resources Package (WER=25%)

Dataset	#Class	Avg. Length	Train	Test
TREC6	6	8.89	5,452	500
ATIS	22	11.14	4,978	893
SLURP	18 * 46	8.89	50,628	10,992

Spoken Language Understanding Results



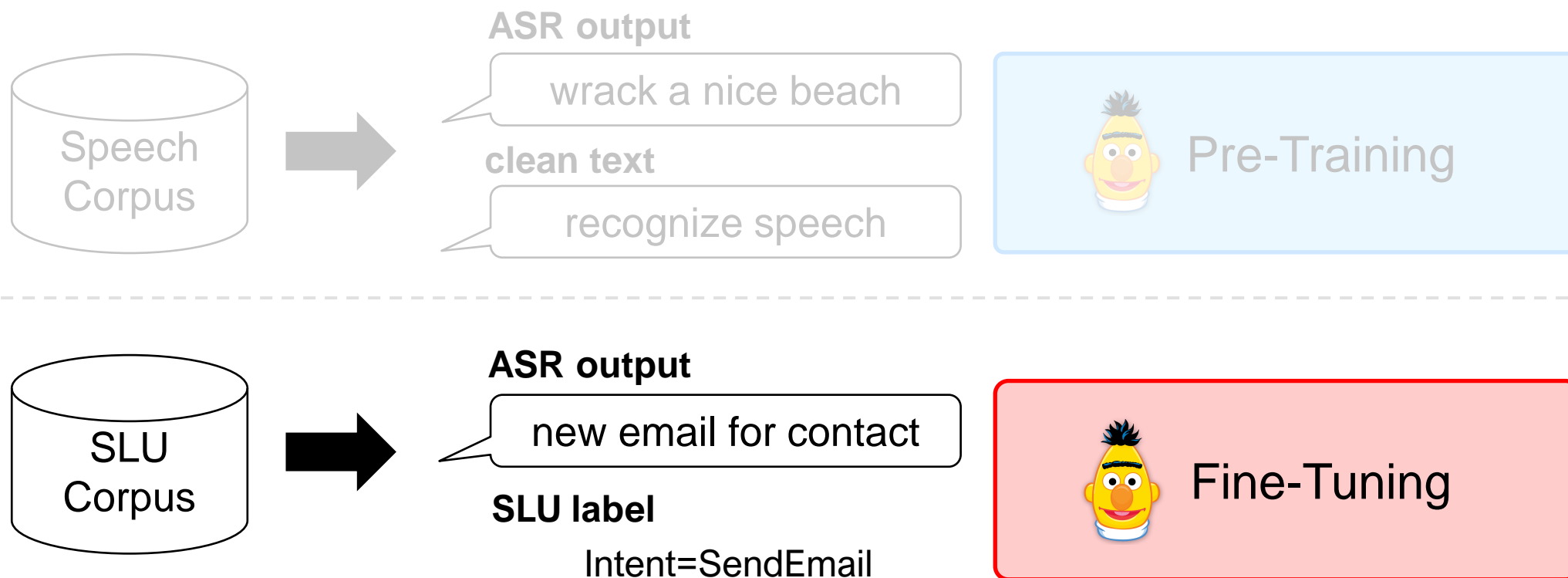
Spoken Language Understanding Results



Contrastive pre-training improves robustness of spoken embeddings with only texts

Improving ASR Robustness of Embeddings

- Idea: adapting embeddings robust to misrecognitions

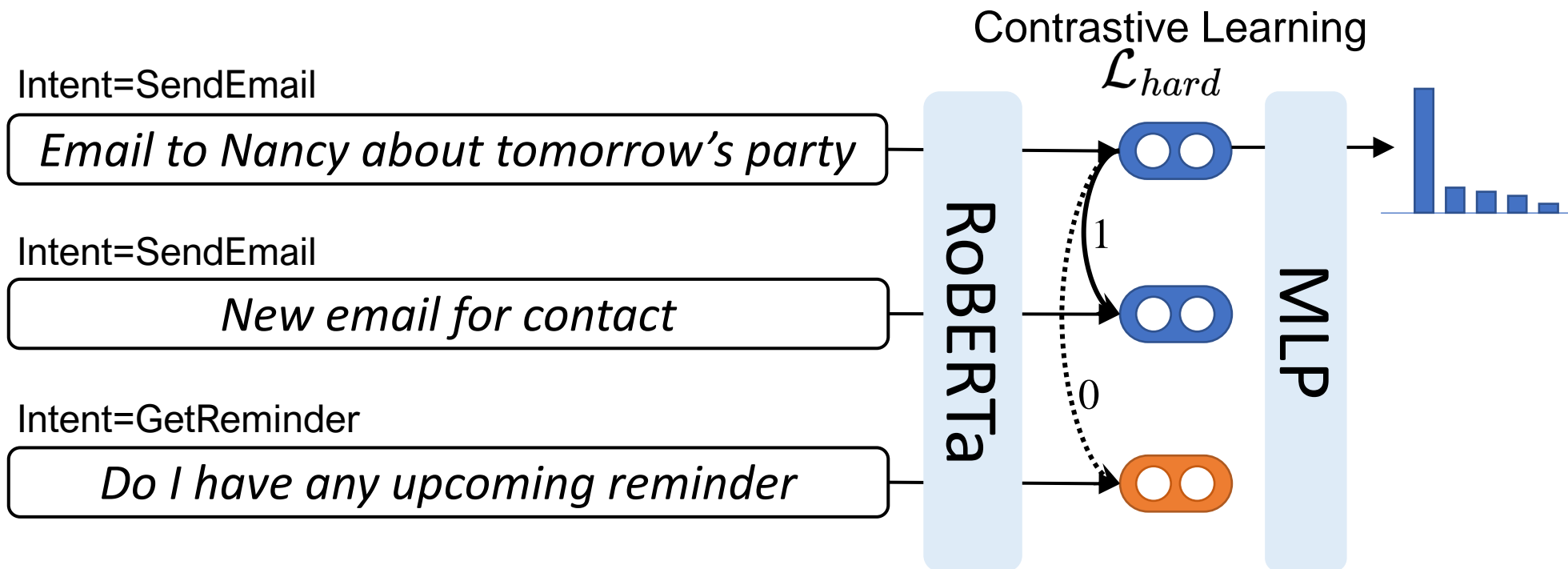


Idea: supervised contrastive learning with self-distillation

Supervised Contrastive Learning

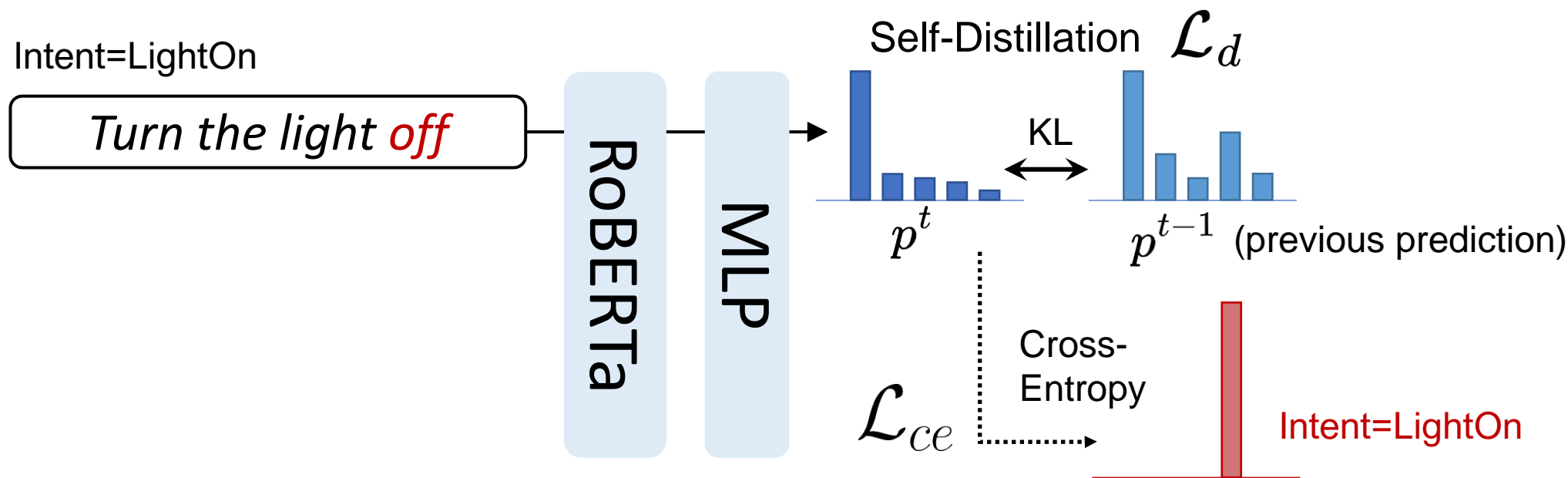
- Idea: data with the same label should be close to each other

- Objective:
$$\mathcal{L}_{hard} = - \sum_i \sum_{j \neq i} 1_{y_i=y_j} \log \frac{e^{sim(h_i, h_j)/\tau_{sc}}}{\sum_{k \neq i} e^{sim(h_i, h_k)/\tau_{sc}}}$$



Self-Distillation

- Issue: misrecognitions may lead to wrong or vague intents
- Objective: $\mathcal{L}_d = \sum_i KL(p_i^{t-1} \| p_i^t)$

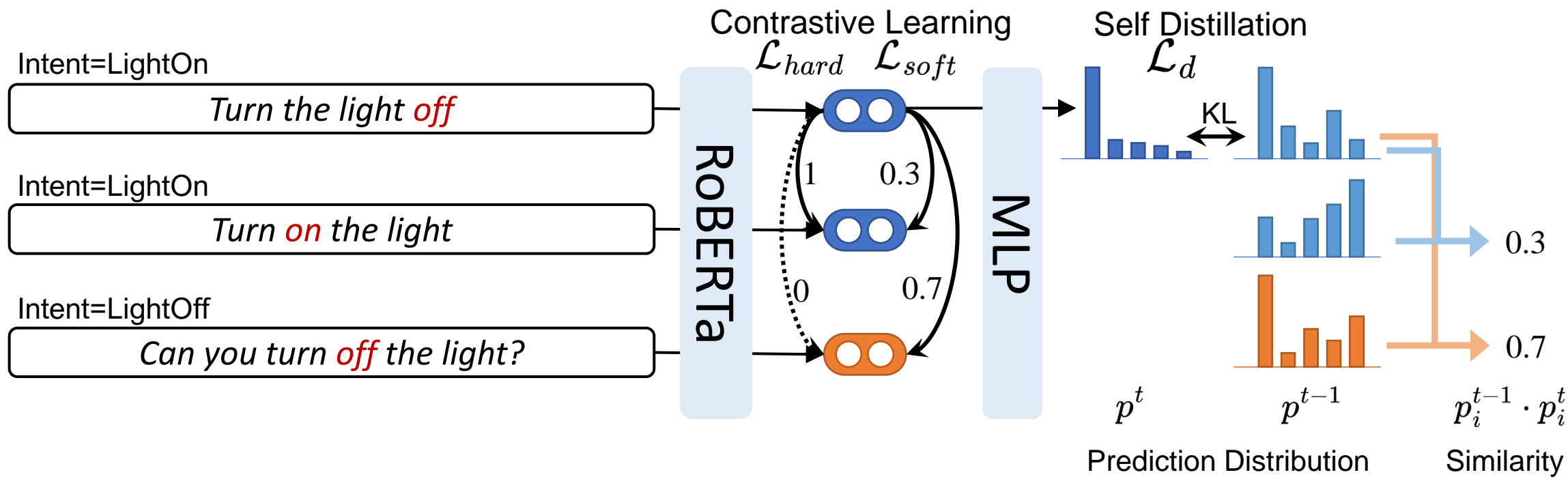


Supervised Contrastive with Self-Distillation

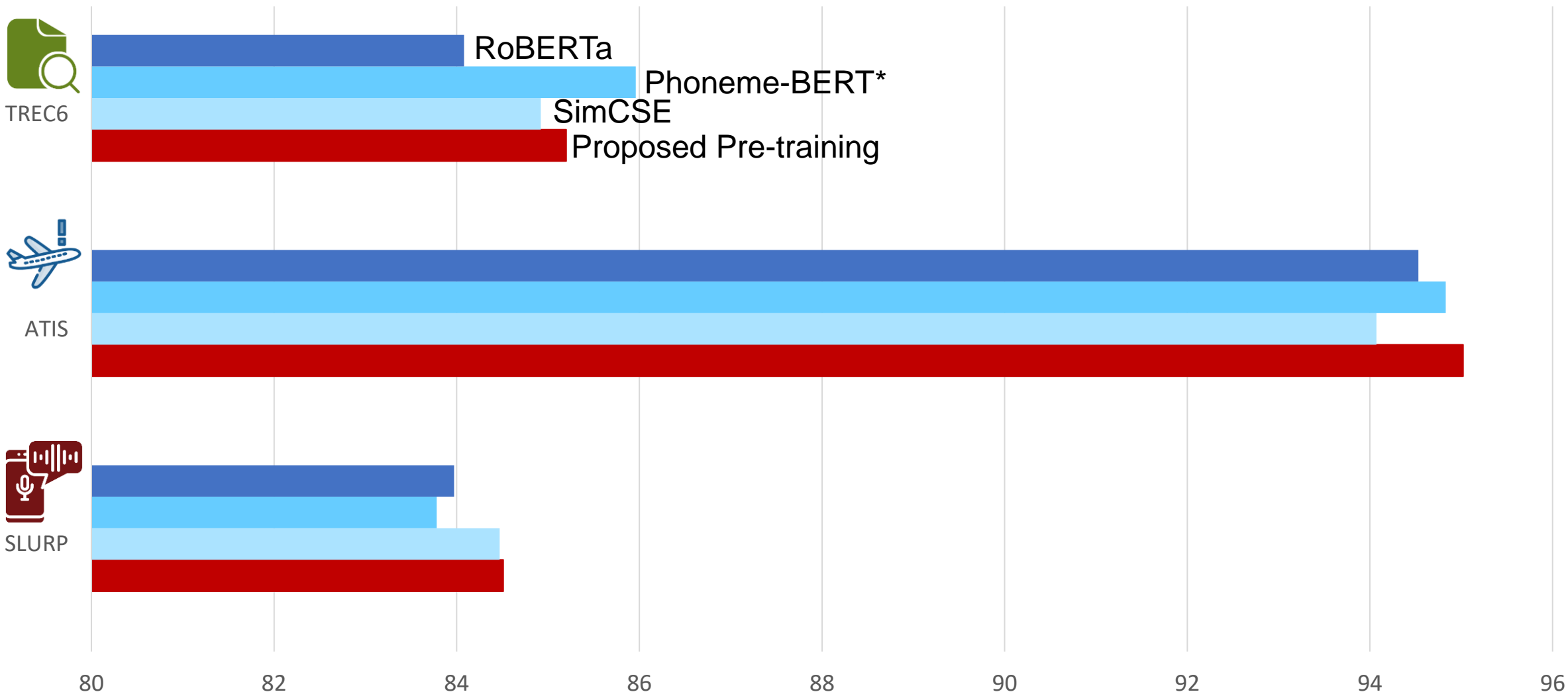
- Issue: noisy labels also affect \mathcal{L}_{hard}

$$\mathcal{L}_{soft} = - \sum_i \sum_{j \neq i} \underbrace{(p_i^{t-1} \cdot p_j^{t-1})}_{(1_{y_i=y_j})} \log \frac{e^{sim(h_i, h_j)/\tau_{sc}}}{\sum_{k \neq i} e^{sim(h_i, h_k)/\tau_{sc}}}$$

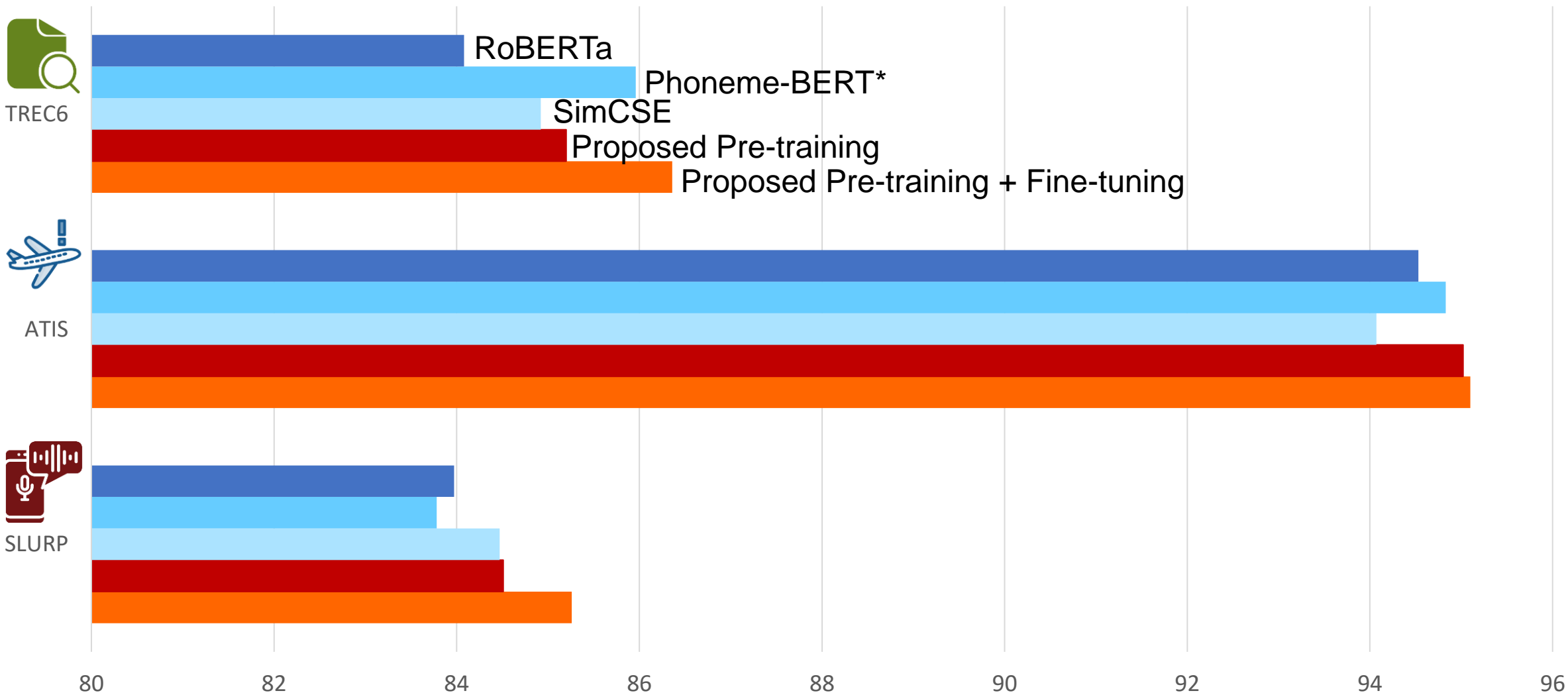
- Fine-tuning objective: $\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d \mathcal{L}_d + \lambda_{hard} \mathcal{L}_{hard} + \lambda_{soft} \mathcal{L}_{soft}$



Spoken Language Understanding Results



Spoken Language Understanding Results



Proposed methods improve robustness to ASR errors and label noises

Ablation Study

$$\mathcal{L}_{pt} = \mathcal{L}_c + \lambda_{mlm} \cdot \mathcal{L}_{mlm}$$

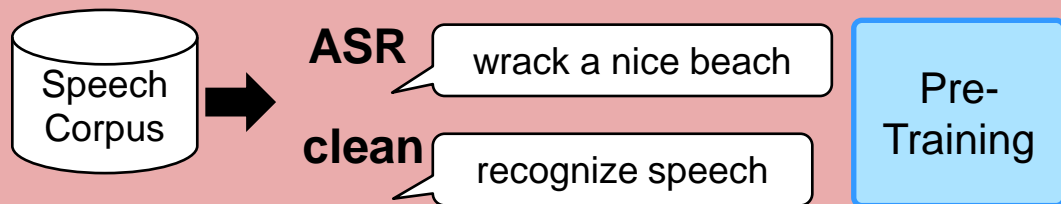
$$\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d \mathcal{L}_d + \lambda_{hard} \mathcal{L}_{hard} + \lambda_{soft} \mathcal{L}_{soft}$$

Pre-Training	Fine-Tuning	SLURP	ATIS	TREC6
Full	Full	85.26	95.10	86.36
No \mathcal{L}_{mlm}	Full	84.83	93.75	85.32
No \mathcal{L}_c	Full	85.15	95.00	85.53
Full	No $\mathcal{L}_{hard} + \mathcal{L}_{soft}$	85.14	94.83	86.08
Full	No $\mathcal{L}_d + \mathcal{L}_{soft}$	84.77	94.75	85.60
Full	No \mathcal{L}_{soft}	84.81	94.65	86.20

All parts in the proposed approach are necessary to achieve better SLU performance.

Robustness

- ✓ **LatticeLM** for preserving uncertainty
- ✓ **Contrastive learning** with only textual information



- **Contrastive Pre-training** learns error-invariant sentence embeddings



- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

Practicality

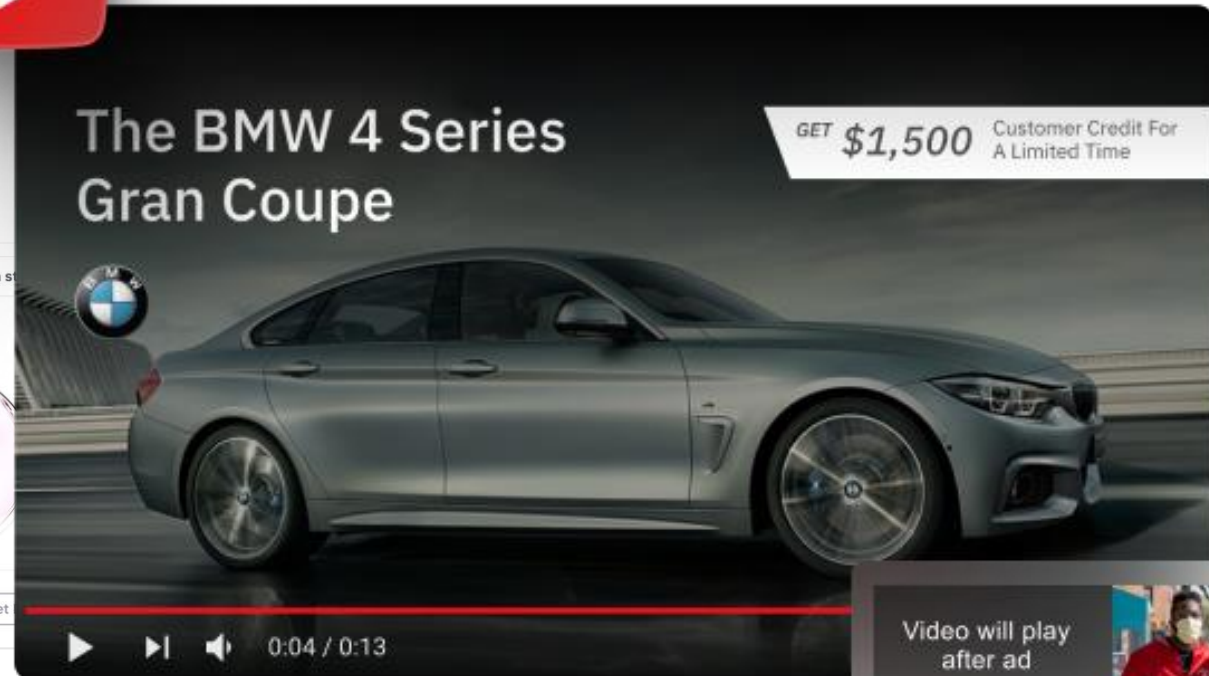
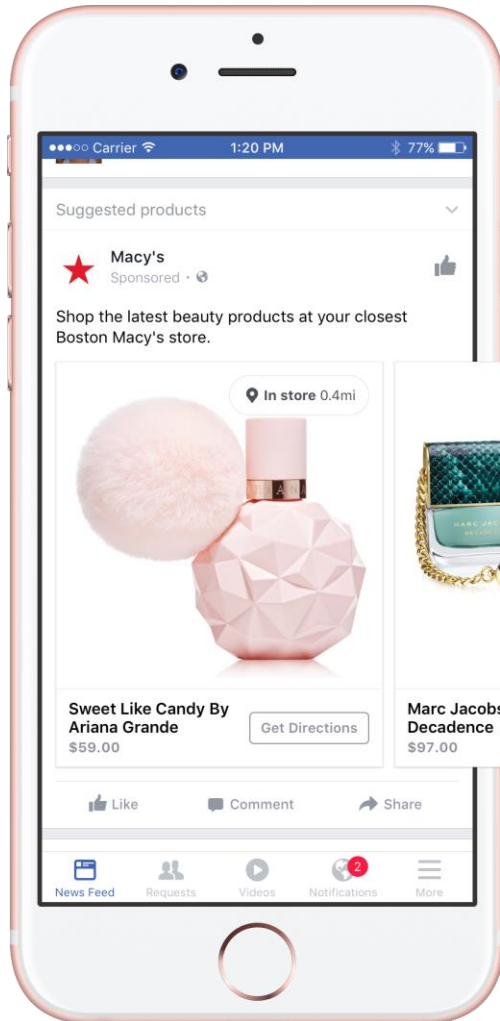
Scalability

The background is a collage of various business-related terms on torn paper scraps. The words are in different orientations and colors, including yellow and grey. Some visible words include 'EDUCATION', 'GROWTH', 'FINANCE', 'DEVELOPMENT', 'MARKETING', 'INSPIRATION', 'CUSTOMER', 'PROFIT', 'VISION', 'TEAMWORK', 'MARKET', 'EFFICIENCY', 'PROGRESS', 'CREATIVITY', 'SERVICE', 'PLANNING', 'SUCCESS', 'BUSINESS', and 'OPPORTUNITY'.

Interaction

**BUSINESS
OPPORTUNITY**

Current Product Advertisements



Machine-based marketing sometimes interrupts user experience

Salesperson: Discover Business Potential

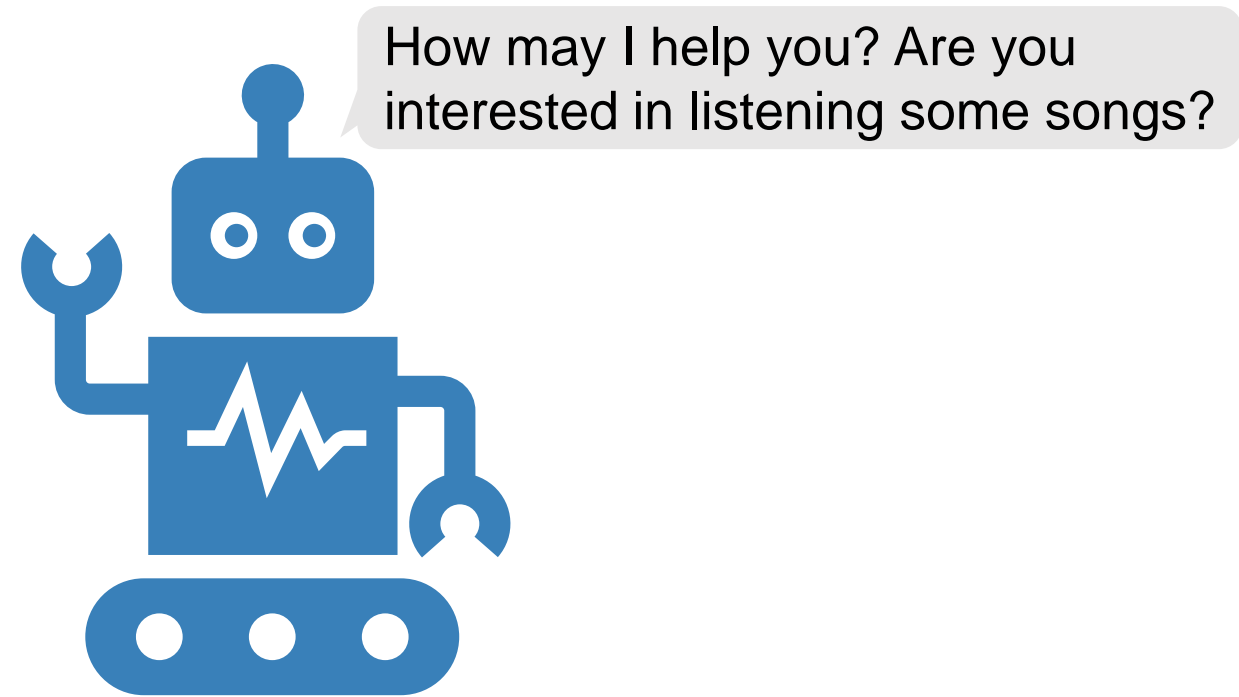
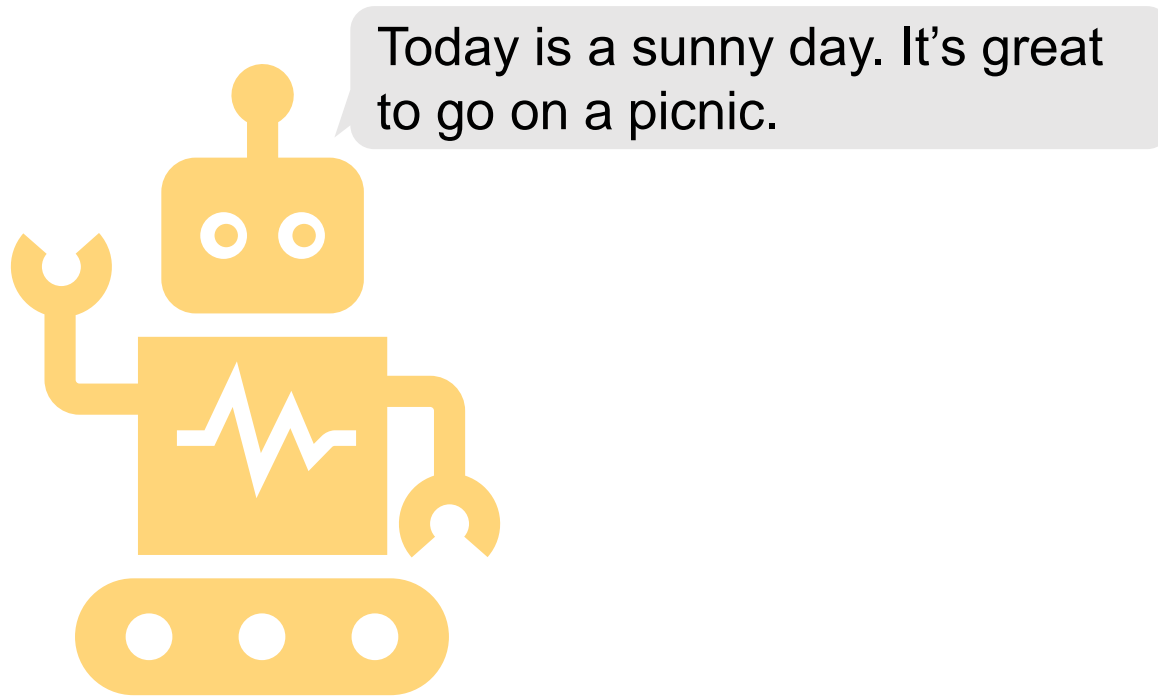
- Human salesperson is capable of
 - mining business-related intentions* from customers *via chatting*
 - smoothly promoting* the related product *through conversations*



39 Two Types of Dialogue Systems

● Open-Domain Chatting

● Task-Oriented



Issue: most prior work focused on **only one** type of dialogue systems



Solution:

SalesBot: Transitioning from Open-Domain to Task-Oriented Dialogues

(Chiu et al., ACL 2022)

<https://github.com/MiuLab/SalesBot>

40

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen, "SalesBot: Transitioning from Chit-Chat to Task-Oriented Dialogues," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

SalesBot: Social to Task-Oriented

Simulated User

My son is 2 and a half. He is handful, but I love him so much.

That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?

No, I don't. I work too much to have time for one. How about you?

I work a lot too. I'm a stay at home mom. What do you do for work?

I work at a library. It's a lot of fun, and I get to meet lots of interesting people.

That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?

I like to read a lot. I also like to go to the movies. What about yourself?

Are you interested in watching any movie?

I'm looking for a movie to watch. A regular showing would be fine.

Where are you located?

I'm in San Francisco, please look for movies there.

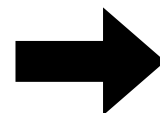
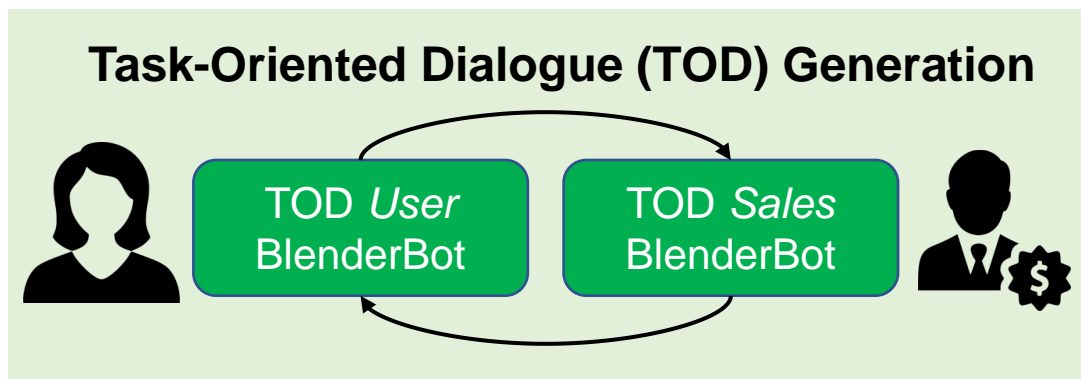
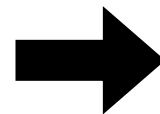
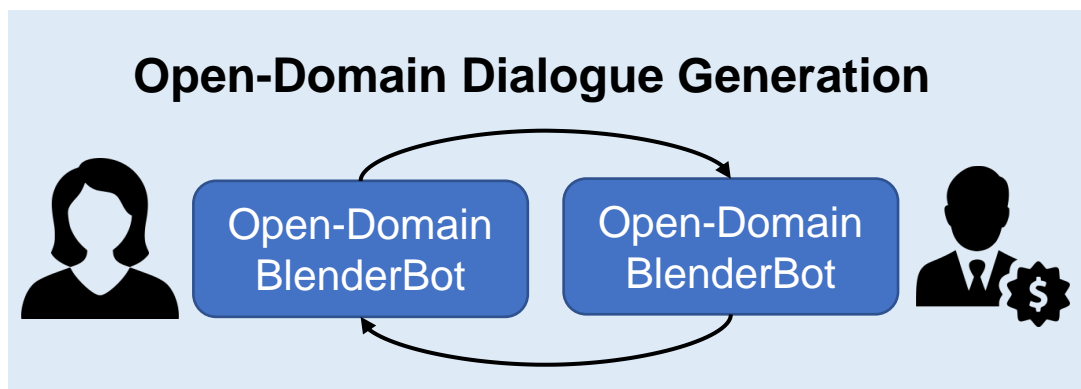
There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

Open-Domain
Task-Oriented

Such data can allow us to train a conversational agent with a salesperson's capability

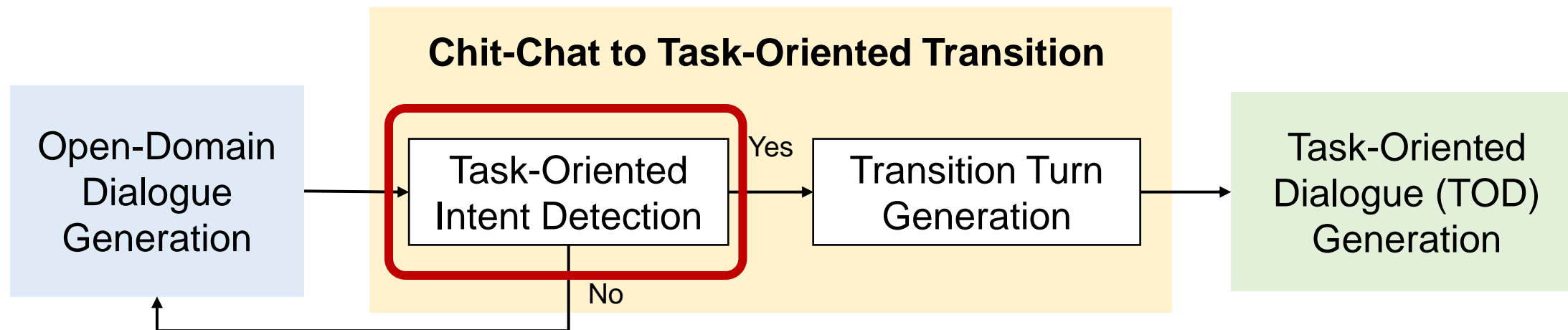
SalesBot: Social to Task-Oriented

- Motivation: no existing data with the property
- Approach: simulate the scenarios to generate unlimited data



How to bridge two types of dialogues?

SalesBot: Social to Task-Oriented

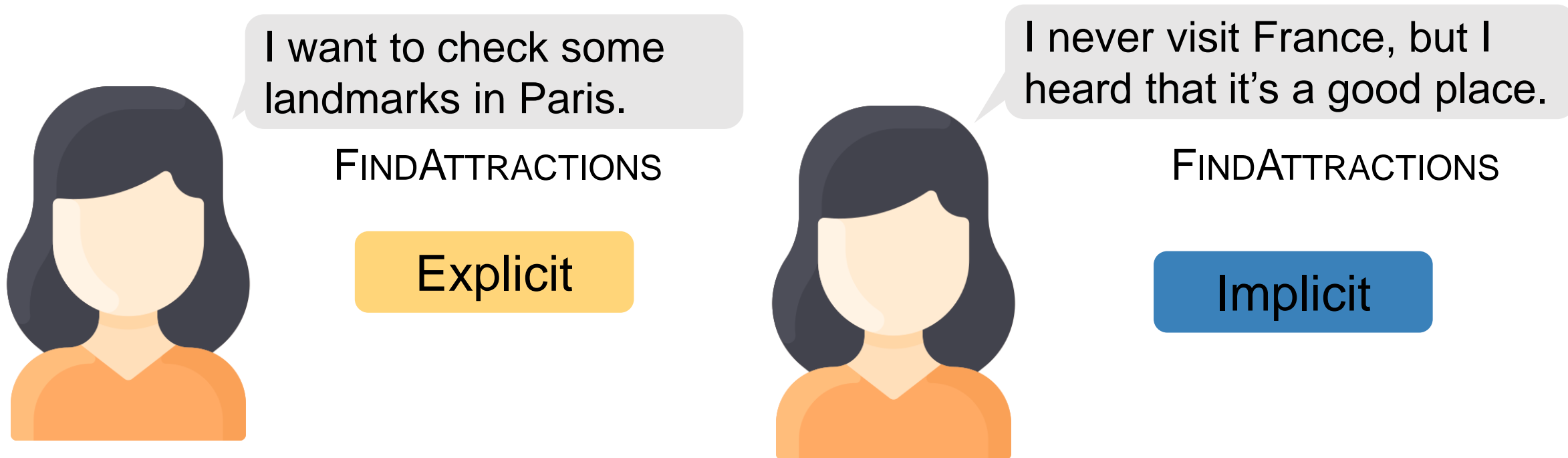


Challenges

- ① **When** to switch to the task-oriented dialogue system?
→ **Task-Oriented (Implicit) Intent Detection**
- ② **How** to smoothly switch from chit-chat to task-oriented dialogues?
→ **Transition Turn Generation**

Task-Oriented (Implicit) Intent Detector

- Goal: identify if the user is likely to have task-related intents



Issue: no data with annotated implicit intents

Zero-Shot Intent Detector

- Idea: leverage QA system's capability

Context

...

I never visit France, but I heard that it is a good place.

Question (FINDATTRactions)

Does the user want to travel there?

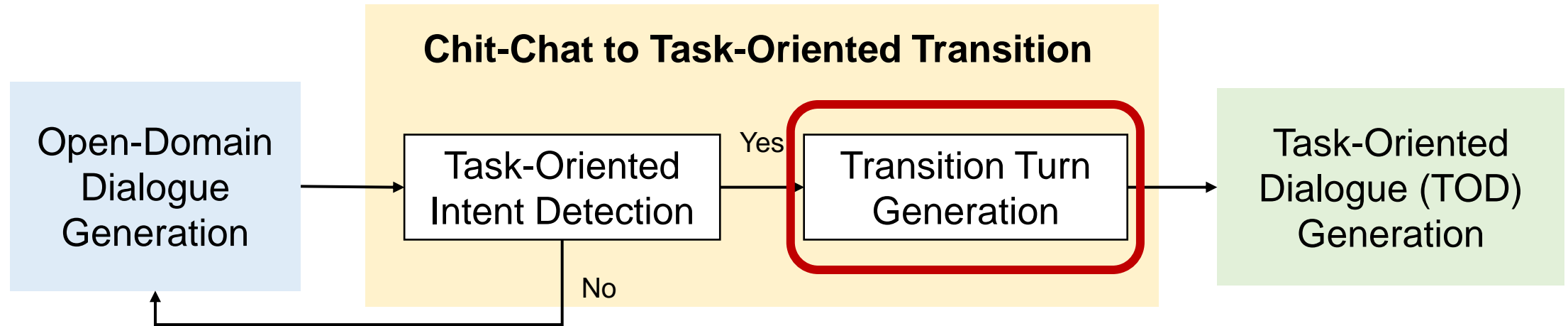
intent description: find attractions to visit

QA Model

Yes

Intent-associated questions are naively generated from their descriptions

SalesBot: Social to Task-Oriented



Challenges

- ① **When** to switch to the task-oriented dialogue system?
→ **Task-Oriented (Implicit) Intent Detection**
- ② **How** to smoothly switch from chit-chat to task-oriented dialogues?
→ **Transition Turn Generation**

Transition Turn Generation

Generative-based Generation:

- Training data: OTTers (Source Topic → Transition → Target Topic)

User A
User B

Entity Path: outside - garden – flower

User A **Source Topic:** I spend a lot of time **outside**. (Source Topic)

User B **Transition:** I like the outdoors as well, especially gardening. It destresses me.

Target Topic: I enjoy relaxing and getting **flowers**.

Entity Path: seafood - **Swedish fish** – candy

User A **Source Topic:** I like **seafood** a lot.

User B **Transition:** Since you like seafood, is Swedish fish a candy that you might enjoy?

Target Topic: I have no self control when it comes to **candy**.

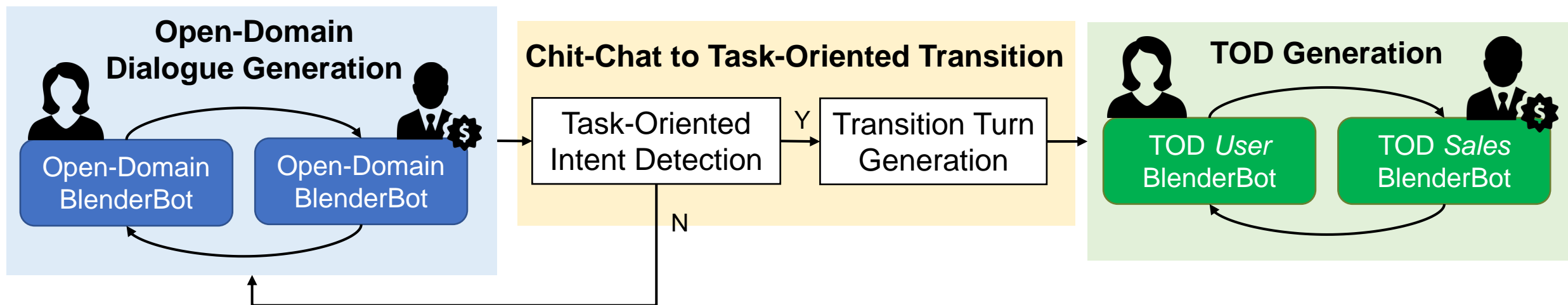
Entity Path: engagement - marriage - child

User A **Source Topic:** I think I am getting engaged soon.

User B **Transition:** I have two children from a previous marriage

Target Topic: My **children** are my life.

SalesBot Simulation Framework



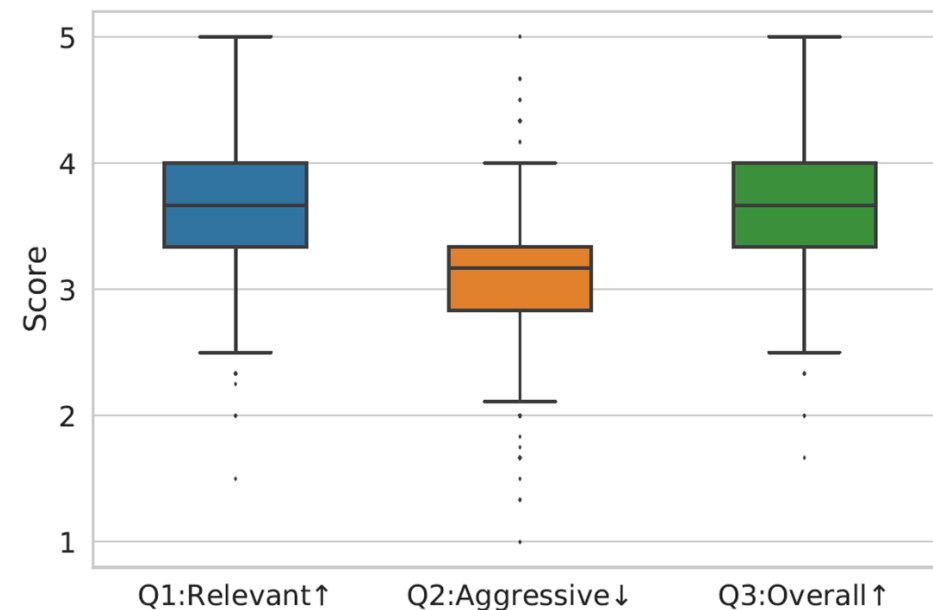
Contribution: simulate unlimited dialogues transitioning from chit-chat to task-oriented

Quality?

Human Evaluation

Overall dialogue quality

- **Q1 Relevance:** How relevant is the recommended service to the conversation context?
- **Q2 Aggressiveness:** How aggressive is the salesperson's communication strategy?
- **Q3 Overall:** Do you think the conversation is overall a good example of making a sales recommendation?



Average scores over 3 workers (4000 dialogues)

Can we further improve the dialogue quality?

Idea: LLMs for Scalability

Simulated User

My son is 2 and a half. He is handful, but I love him so much.

That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?

No, I don't. I work too much to have time for one. How about you?

I work a lot too. I'm a stay at home mom. What do you do for work?

I work at a library. It's a lot of fun, and I get to meet lots of interesting people.

That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?

I like to read a lot. I also like to go to the movies. What about yourself?

Are you interested in watching any movie?

I'm looking for a movie to watch. A regular showing would be fine.

Where are you located?

I'm in San Francisco, please look for movies there.

There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

Open-Domain Task-Oriented

Idea: LLMs for Scalability

Simulated User

Simulated Salesperson

My son is 2 and a half. He is handful, but I love him so much.

That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?

No, I don't. I work too much to have time for one. How about you?

I work a lot too. I'm a stay at home mom. What do you do for work?

I work at a library. It's a lot of fun, and I get to meet lots of interesting people.

That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?

I like to read a lot. I also like to go to the movies. What about yourself?

Are you interested in watching any movie?

I'm looking

Expanding topics for **great coverage** via LLM rewriting

where are you located?

I'm in San Francisco, please look for movies there.

There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

Open-Domain

Idea: LLMs for Scalability

Simulated User

My son is 2 and a half. He is handful, but I love him so much.

That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?

No, I don't. I work too much to have time for one. How about you?

I work a lot too. I'm a stay at home mom. What do you do for work?

I work at a library. It's a lot of fun, and I get to meet lots of interesting people.

That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?

I like to read a lot. I also like to go to the movies. What about yourself?

Are you interested in watching any movie?

I'm looking for a movie to watch. A regular showing would be fine.

Increasing #turns for **smooth transition** via LLM rewriting

I'm in San Francisco, please look for movies there.

There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

Transition

Solution:

SalesBot 2.0: A Human-Like Intent Guided Chit-Chat Dataset (Chang & Chen, 2023)

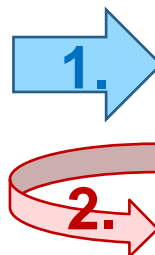
53

SalesBot 2.0: LLM Rewriting (Chang & Chen, 2023)

SalesBot 1.0

U: Yes, we grew up together. We are both going to U of Michigan this fall.
A: That's awesome! I'm hoping to start my own business soon. What do you do for a living?
U: I work at a grocery store. What kind of business are you looking to start? I hope it works out for you.
A: I'm thinking of starting a sports bar. I love sports and I think it would be a great place to start.

Chit-Chat Dialogue Revision



U: Yes, we grew up together. We are both going to U of Michigan this fall.
A: That's great! What are you planning to study at U of Michigan?
U: I haven't decided yet, but I'm leaning towards psychology. How about you?
A: I actually graduated a few years ago, but now I'm thinking of starting my own business. I want to open a sports bar.

Potential Intent Detection

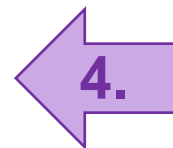
FINDRESTAURANT Triggered TOD Intent



Dialogue Continuation

U: That sounds exciting. Are you a big sports fan?
A: Definitely! Watching sports is one of my favorite pastimes. Speaking of which, have you visited any good sports bars around here yet
U: No, I haven't had a chance to check any out yet. Do you have any recommendations?
A: Definitely, there's a great place down the street called The Game Day Grill. They have amazing wings and a huge selection of beers on tap. You should definitely check it out sometime.

Transition Boundary Detector



SalesBot 2.0

U: Yes, we grew up together. ...
 ...
U: That sounds exciting. ...
A: Definitely! ...
U: No, I haven't had a chance to check any out yet. Do you have any recommendations?

A: Definitely, there's a great place ...

Additional Annotations



HasIntent?
Suggest?

SalesBot 2.0: LLM Rewriting (Chang & Chen, 2023)

SalesBot 1.0

U: Yes, we grew up together. We are both going to U of Michigan this fall.

A: That's awesome! I'm hoping to start my own business soon. What do you do for a living?

U: I work at a grocery store. What kind of business are you looking to start? I hope it works out for you.

A: I'm thinking of starting a sports bar. I love sports and I think it would be a great place to start.



1. Chit-Chat Dialogue Revision

U: Yes, we grew up together. We are both going to U of Michigan this fall.

A: That's great! What are you planning to study at U of Michigan?

U: I haven't decided yet, but I'm leaning towards psychology. How about you?

A: I actually graduated a few years ago, but now I'm thinking of starting my own business. I want to open a sports bar.

Dialogue revision improves naturalness and expand the topic coverage

FINDRESTAURANT

Intent detection improves the TOD triggering capability

SalesBot 2.0: LLM Rewriting (Chang & Chen, 2023)

U: Yes, we grew up together. We are both going to U of Michigan this fall.

A: That's great! What are you planning to study at U of Michigan?

U: I haven't decided yet, but I'm leaning towards psychology. How about you?

A: I actually graduated a few years ago, but now I'm thinking of starting my own business. I want to open a sports bar.



3. Dialogue Continuation

U: That sounds exciting. Are you a big sports fan?

A: Definitely! Watching sports is one of my favorite pastimes. Speaking of which, have you visited any good sports bars around here yet?

U: No, I haven't had a chance to check any out yet. Do you have any recommendations?

A: Definitely, there's a great place down the street called The Game Day Grill. They have amazing wings and a huge selection of beers on tap. You should definitely check it out sometime.

Dialogue continuation focuses on smoothly transitioning to the topic

SalesBot 2.0: LLM Rewriting (Chang & Chen, 2023)

U: Yes, we grew up together. We are both going to U of Michigan this fall.

A: That's great! What are you planning to study at U of Michigan?

U: I haven't decided yet, but I'm leaning towards psychology. How about you?

A: I actually graduated a few years ago, but now I'm thinking of starting my own business. I want to open a sports bar.

U: That sounds exciting. Are you a big sports fan?

A: Definitely! Watching sports is one of my favorite pastimes. Speaking of which, have you visited any good sports bars around here yet?

U: No, I haven't had a chance to check any out yet. Do you have any recommendations?

A: Definitely, there's a great place down the street called The Game Day Grill. They have amazing wings and a huge selection of beers on tap. You should definitely check it out sometime.

4. Transition Boundary Detector

Transition boundary detector avoids diverging topics and finds timing to TOD

Annotations

Suggest? Transitioning is reasonable but more aggressive

SalesBot 1.0 vs. SalesBot 2.0 (Chang & Chen, 2023)

Simulated User

Simulated Salesperson

SalesBot 1.0

Thank you. It was a long time ago, but I still miss her. What do you do for a living?

I am a stay at home mom. It is very rewarding. How about you?

I work at the zoo. It's a lot of fun, and I get to spend time with my family.

What about your family?

FINDATTRACTION

SalesBot 2.0

Thank you. It was a long time ago, but I still miss her deeply.

I'm sorry for your loss. It sounds like she was very important to you.
As for me, I'm also fortunate to have a family that I get to spend a lot of time with.

That's great to hear. I work at the zoo, and it's a fun job that allows me to support my family.

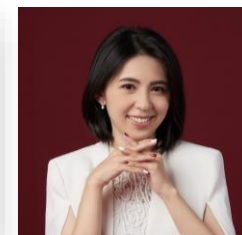
That sounds like a really cool job. What kind of animals do you work with at the zoo?

Mostly primates, but there are also elephants, lions, and bears.

Wow, that must be exciting to work with such a variety of animals.
Have you had the chance to explore the tourist attractions in the area?

Not really, I've been so focused on work and family lately that I haven't had much time to go out and explore.

Does SalesBot 2.0 have better quality? Human evaluation?



Solution:

LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models

(Lin & Chen, NLP4ConvAI 2023)

59

LLM-Eval (Lin & Chen, 2023)

LLM-Eval

{evaluation schema}

Score the following dialogue response generated on a continuous scale from 0.0 to 5.0.

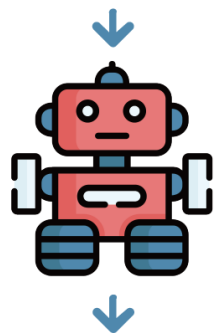
Context:

👤: My cat likes to eat cream.

👤: Be careful not to give too much, though.

Dialogue response :

👤: Don't worry, I only give a little bit as a treat.



Appropriateness: 3.0

Cotent: 2.5

Grammer: 4.0

Relevance: 2.0

- LLM has a reasonable capability of evaluating dialogue responses

r / ρ (%)	TopicalChat	PersonaChat	ConvAI2	DD	ED	DSTC6	Average
BLEU-4	21.6 / 29.6	13.5 / 9.0	0.3 / 12.8	7.5 / 18.4	-5.1 / 0.2	13.1 / 29.8	8.5 / 16.6
ROUGE-L	27.5 / 28.7	6.6 / 3.8	13.6 / 14.0	15.4 / 14.7	2.9 / -1.3	33.2 / 32.6	16.5 / 15.4
BERTScore	29.8 / 32.5	15.2 / 12.2	22.5 / 22.4	12.9 / 10.0	4.6 / 3.3	36.9 / 33.7	20.3 / 19.0
DEB	18.0 / 11.6	29.1 / 37.3	42.6 / 50.4	<u>33.7</u> / 36.3	35.6 / 39.5	21.1 / 21.4	30.0 / 32.8
GRADE	20.0 / 21.7	35.8 / 35.2	56.6 / 57.1	27.8 / 25.3	33.0 / 29.7	11.9 / 12.2	30.9 / 30.2
USR	41.2 / 42.3	44.0 / 41.8	50.1 / 50.0	5.7 / 5.7	26.4 / 25.5	18.4 / 16.6	31.0 / 30.3
USL-H	32.2 / 34.0	49.5 / 52.3	44.3 / 45.7	10.8 / 9.3	29.3 / 23.5	21.7 / 17.9	31.3 / 30.5
<i>without human reference</i>							
LLM-EVAL $0-5$	<u>55.7</u> / <u>58.3</u>	51.0 / 48.0	<u>59.3</u> / <u>59.6</u>	31.8 / 32.2	42.1 / 41.4	43.3 / 41.1	47.2 / 46.8
LLM-EVAL $0-100$	49.0 / 49.9	53.3 / 51.5	61.3 / 61.8	34.6 / <u>34.9</u>	<u>43.2</u> / <u>42.3</u>	44.0 / 41.8	47.6 / <u>47.0</u>
<i>with human reference</i>							
LLM-EVAL $0-5$	56.5 / 59.4	55.4 / 53.1	43.1 / 43.8	32.0 / 32.2	40.0 / 40.1	<u>47.0</u> / <u>45.5</u>	45.7 / 45.7
LLM-EVAL $0-100$	55.6 / 57.1	<u>53.8</u> / <u>52.7</u>	45.6 / 45.9	33.4 / 34.0	43.5 / 43.2	49.8 / 49.9	47.0 / 47.1

LLM-Eval better correlates with human-judged scores than all existing metrics

LLM-Eval (Lin & Chen, 2023)

- LLM-Eval works good on not only **single-turn** but **multi-turn** evaluation

$r / \rho (\%)$	DailyDialog-PE Turn-Level	FED Turn-Level	FED Dialog-Level	DSTC9 Dialog-Level	Average
DynaEval	16.7 / 16.0	31.9 / 32.3	50.3 / 54.7	9.3 / 10.1	27.1 / 28.3
USL-H	68.8 / 69.9	20.1 / 18.9	7.3 / 15.2	10.5 / 10.5	26.7 / 28.6
FlowScore	-	-6.5 / -5.5	-7.3 / -0.3	14.7 / 14.0	0.3 / 2.7
GPTScore	-	- / 38.3	- / 54.3	-	- / 46.3
LLM-EVAL ₀₋₅	<u>71.0</u> / 71.3	60.4 / 50.9	67.6 / 71.4	<u>15.9</u> / <u>16.5</u>	53.7 / 52.5
LLM-EVAL ₀₋₁₀₀	71.4 / <u>71.0</u>	<u>59.7</u> / <u>49.9</u>	<u>64.4</u> / <u>70.4</u>	16.1 / 18.6	<u>52.9</u> / <u>52.5</u>

Idea: LLM-Eval scores can be the proxy of human evaluation

Dialogue Quality Comparison

	#Chit-Chat Turn	#Transition Turn	#Total Turn	Naturalness ↑	Consistency ↑
SalesBot 1.0	4.49	1.00	5.49	3.574	2.656
SalesBot 2.0	5.22	4.55	9.29	4.258	4.026

SalesBot 2.0	Yes	No	Total
HasIntent?	4,149	1,197	5,391
Suggest?	5,167	224	5,391
Both		182	

less unsuitable dialogues (unnatural & aggressive)

SalesBots 2.0 has better quality than 1.0 for better usage

Future Conversational Ads



My son is 5 and my daughter is 2.
They are slightly naughty, but I love them so much. They are so cute!!



That's adorable. You must be very busy.



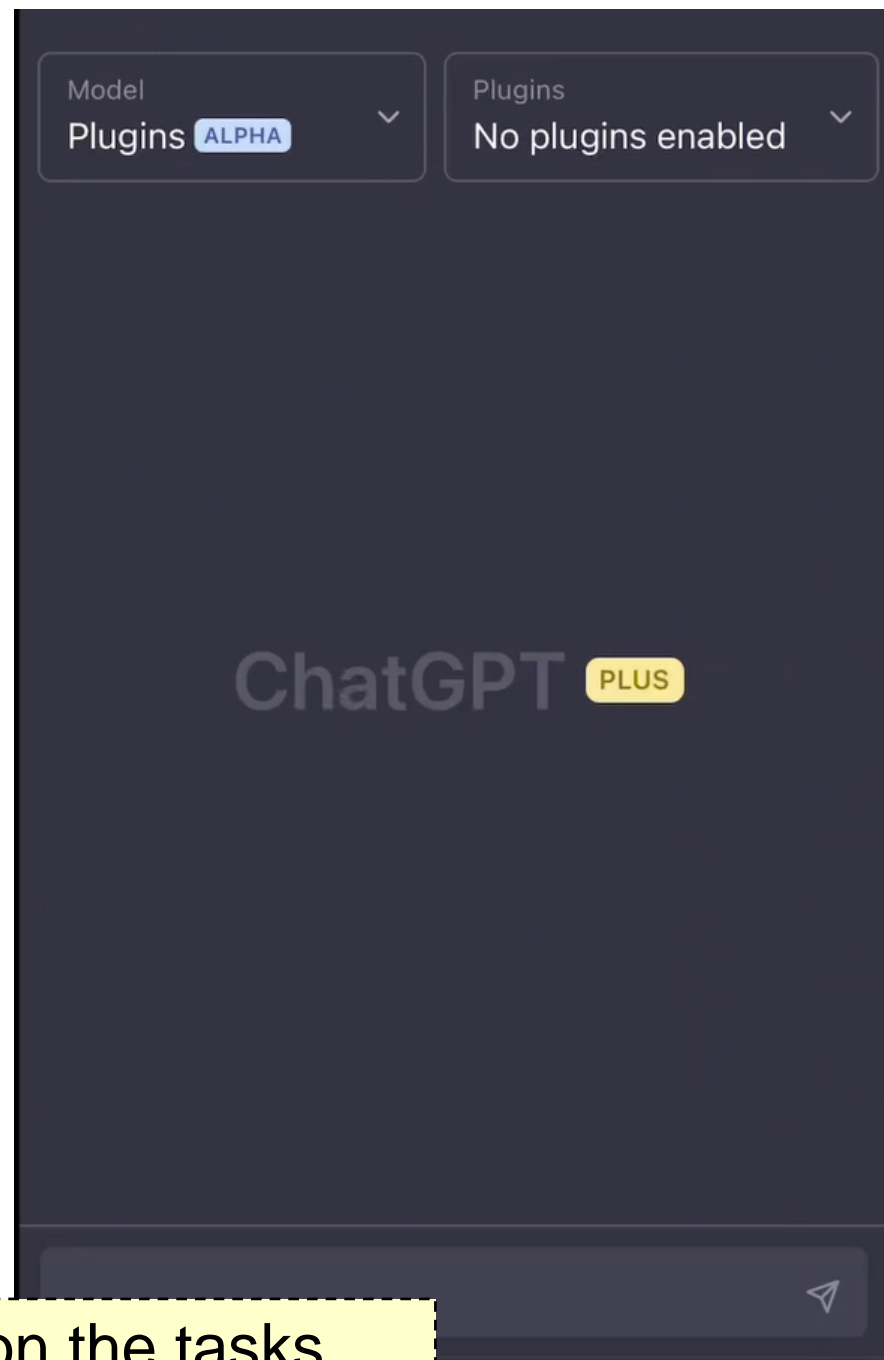
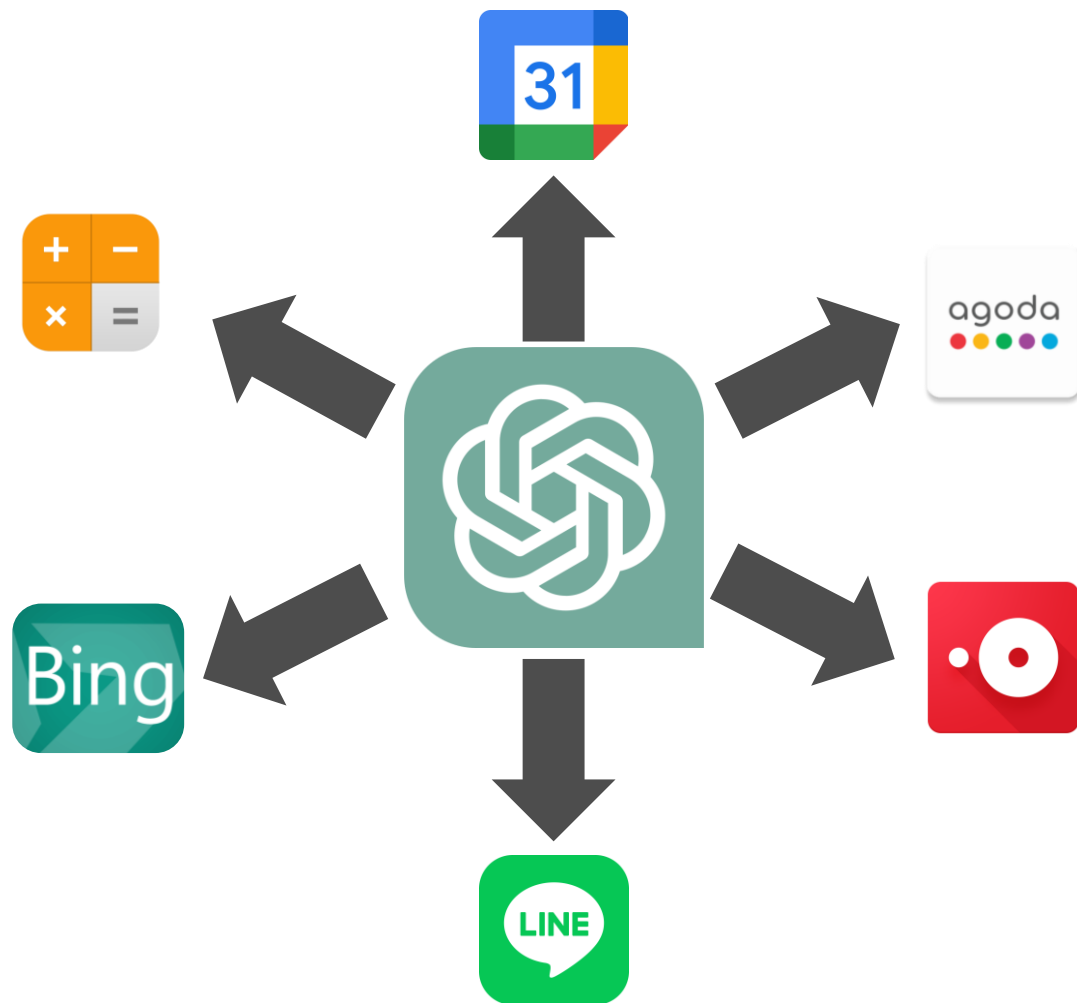
My husband and I enjoyed watching movies, but currently we don't have any time to go out for movie.



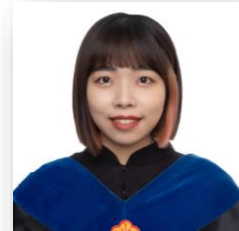
I have some alternatives for you if you really like watching movies.
You can still enjoy movies at home; for example, Disney+ and ...



ChatGPT Plugins



Human should explicitly mention the tasks



Solution:

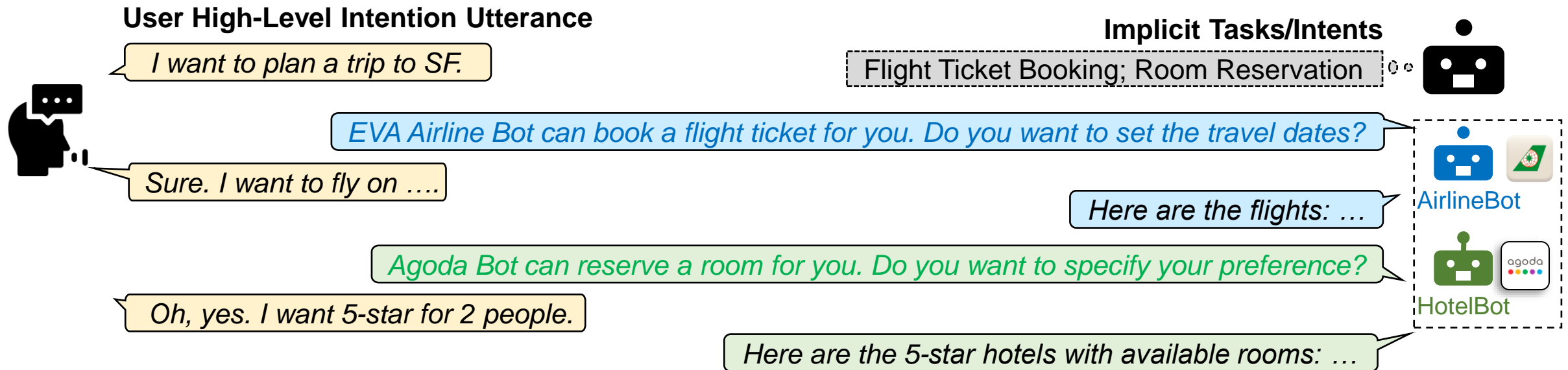
Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning

(Kuo & Chen, ACL Findings 2023)

65

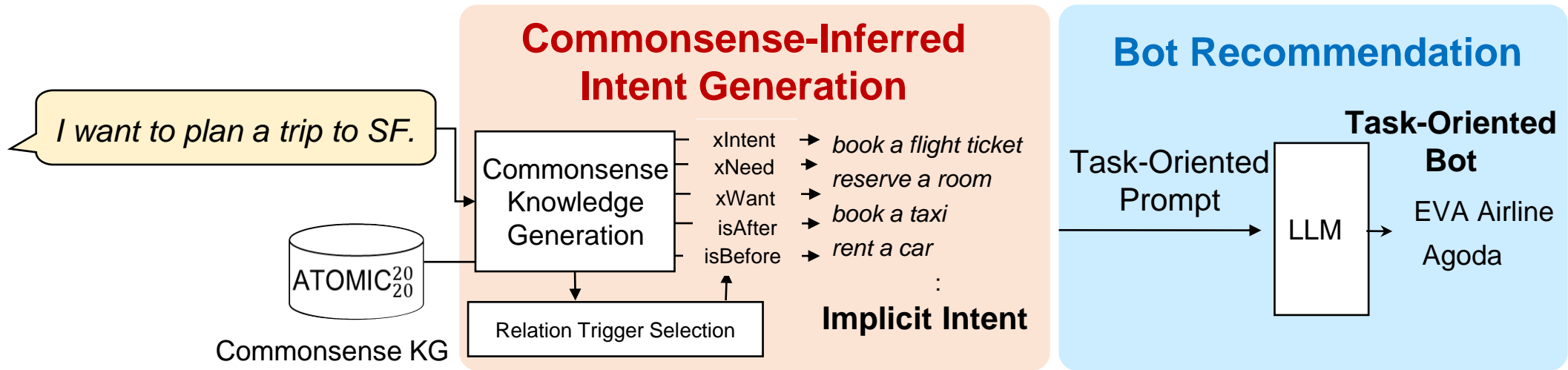
Human Understand High-Level Intention

- Goal: users interact in high-level descriptions and the agent learns how to plan dialogues



Idea: utilize implicit intents to bridge a high-level description and task-specific bots

Commonsense-Inferred Bot Recommendation



- 1st component: given a user's high-level intention, we generate implicit task-oriented intents
- 2nd component: we utilize the task-specific intents to recommend proper bots

Commonsense can provide the agent guidance for bot recommendation

Commonsense-Inferred Bot Recommendation

Q1: Is adding implicit intents (2-stage) for bridging useful?

A1: YES

Method	P	R	F1	Human Score	
1-Stage Baseline	30.3	20.6	23.7	1.73 ± 1.03	(1) Irrelevant
2-Stage GPT-3 Prompting	28.6	41.7	31.8	2.11 ± 0.46	(2) Acceptable
Our 2-Stage Commonsense	36.0	35.7	32.9	2.18 ± 0.34	(3) Useful

Q2: Is commonsense KG better than GPT-3 prompting?

A2: YES

Q3: Is adding reasons more convincing for recommendation?

A3: YES

Method	Win	Lose	Tie
Ours vs. 2-Stage GPT-3 Prompting	57.6	40.2	2.2
Ours vs. Ours w/o Reasons	55.1	38.8	6.1

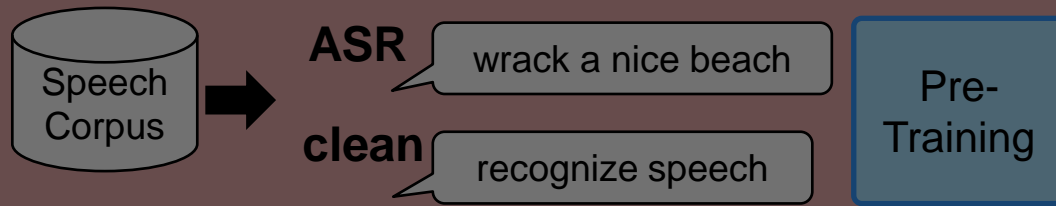
Recommendation Example

We are planning to celebrate friend's birthday at a restaurant in [City].		Score
1-Stage Prompting	Tinder (Lifestyle), Grindr (Lifestyle)	1.83
2-Stage Prompting	Zomato can help to book the restaurant in advance. WhatsApp can find out about their contact information.	2.00
Ours	WhatsApp can help have a good time and to celebrate a friend's birthday. OpenTable can help book a table at the restaurant and go to the restaurant.	2.67
Ours w/o Reasons	WhatsApp (Communication), OpenTable (Food & Drink)	2.17

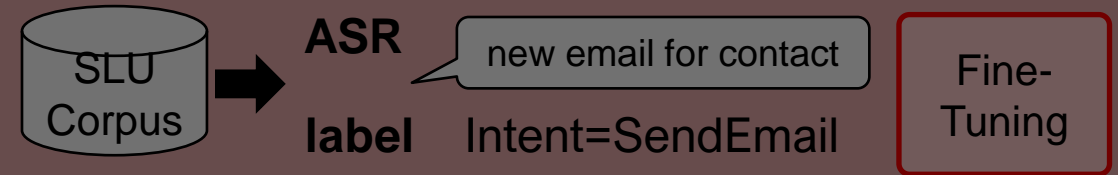
Using commonsense-inferred intents achieves a better precision and human score

Robustness

- ✓ **LatticeLM** for preserving uncertainty
- ✓ **Contrastive learning** with only textual information



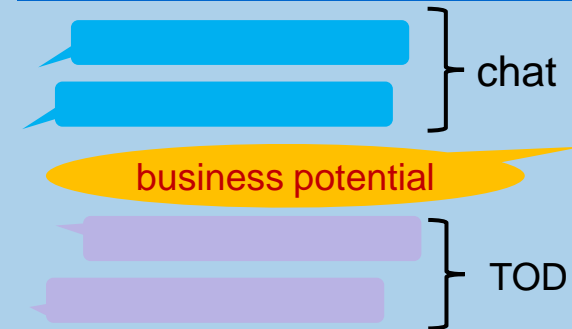
- **Contrastive Pre-training** learns error-invariant sentence embeddings



- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

- ✓ **SalesBot** simulates unlimited data bridging two types of agents
 - Reasonable quality
 - Data/simulators for learning (SL/RL) end-to-end agents with sales' behavior
- ✓ **Commonsense** bridges users' high-level intention and task bots
 - Better recommendation & more convincing

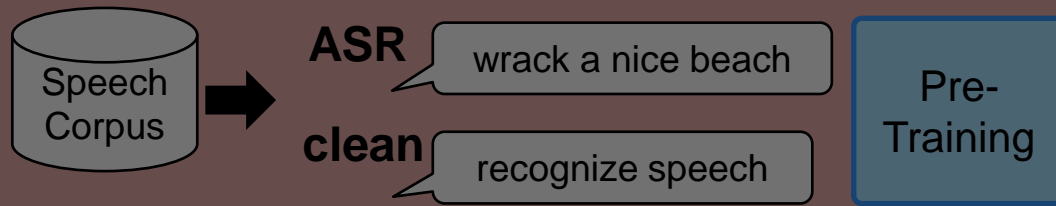
Practicality



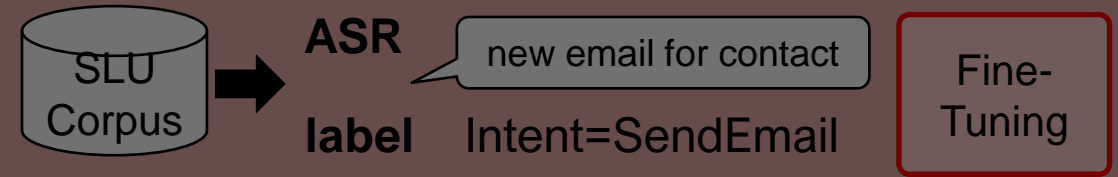
Scalability

Robustness

- ✓ **LatticeLM** for preserving uncertainty
- ✓ **Contrastive learning** with only textual information



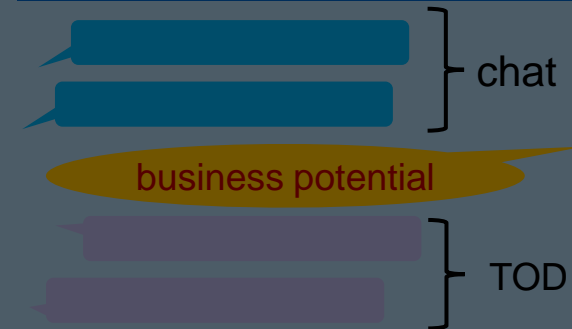
- **Contrastive Pre-training** learns error-invariant sentence embeddings



- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

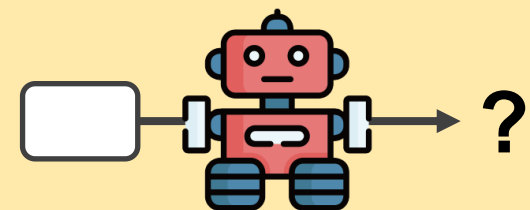
- ✓ **SalesBot** simulates unlimited data bridging two types of agents
 - Reasonable quality
 - Data/simulators for learning (SL/RL) end-to-end agents with sales' behavior
- ✓ **Commonsense** bridges users' high-level intention and task bots
 - Better recommendation & more convincing

Practicality



Scalability

- ✓ LLMs rewrites data towards better **naturalness** and **consistency**
 - Diverse domain coverage
 - Smooth transition
- ✓ LLM-Eval better correlates with human-judged scores
 - Addressing an important issue in NLG tasks





Taiwan-LLaMa: LMs for Taiwanese Culture

<https://github.com/MiuLab/taiwan-llama>

Language Models for Taiwanese Culture

 Online Demo •  HF Repo •  Twitter •  [Paper Coming Soon] •  Yen-Ting Lin



Code License

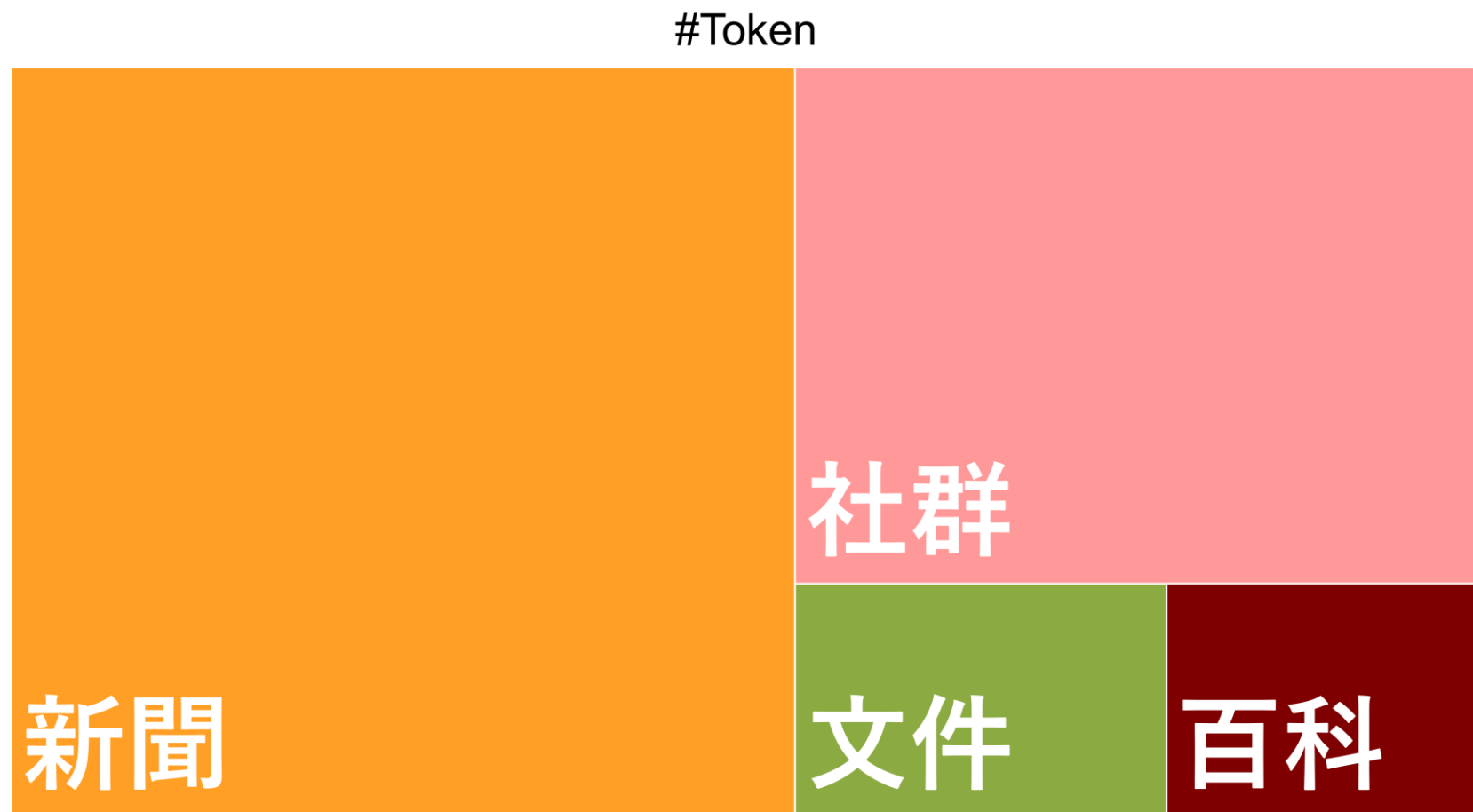
Apache 2.0

Data License

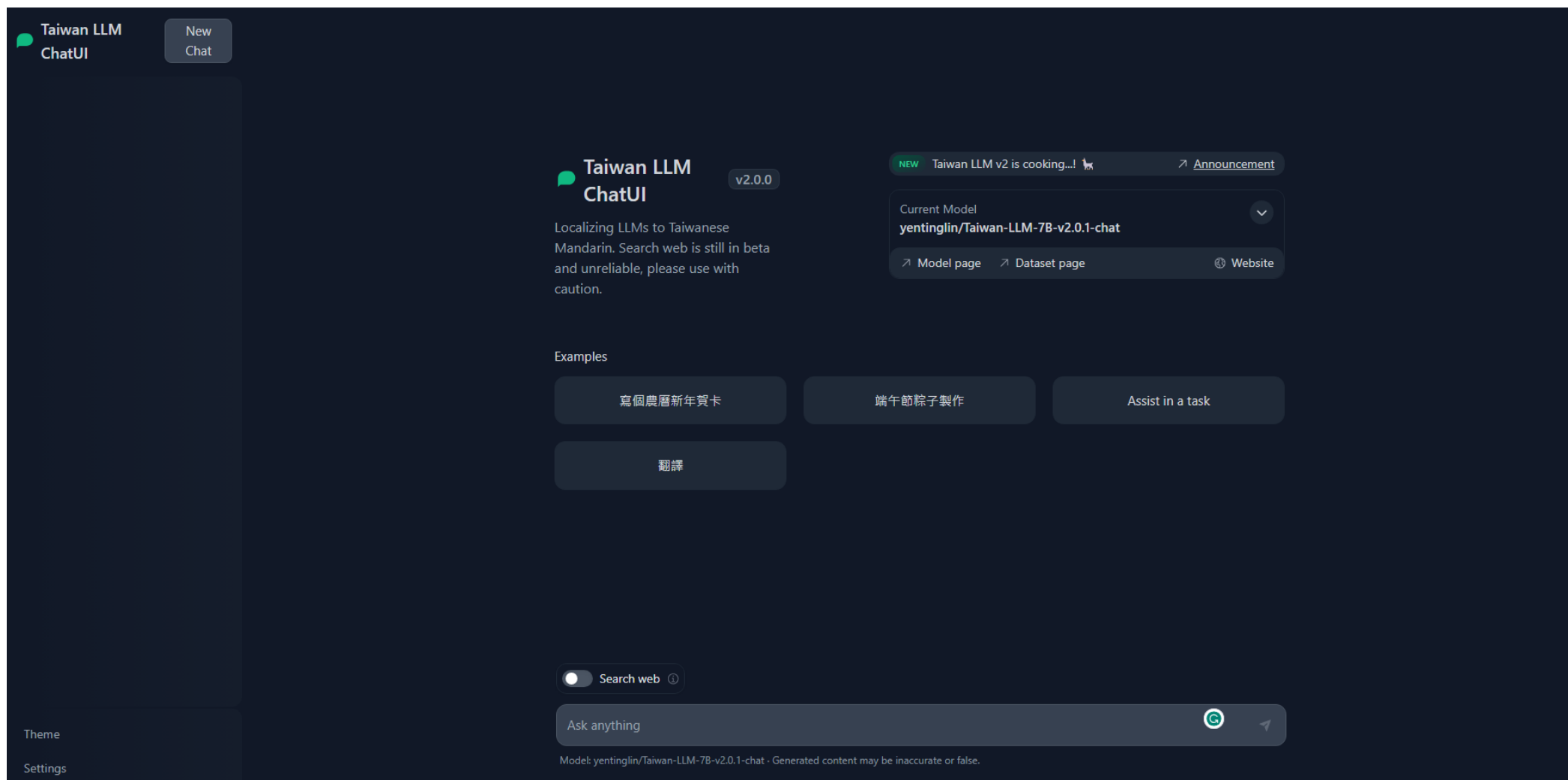
CC By NC 4.0

Pretraining Data (v2.0)

- News
- Social Media
- Legal Documents
- Wikipedia-zh
- ~~○ Web Data~~



Try Taiwan-LLaMa v2.0 <http://twllm.com/>





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



Appendix



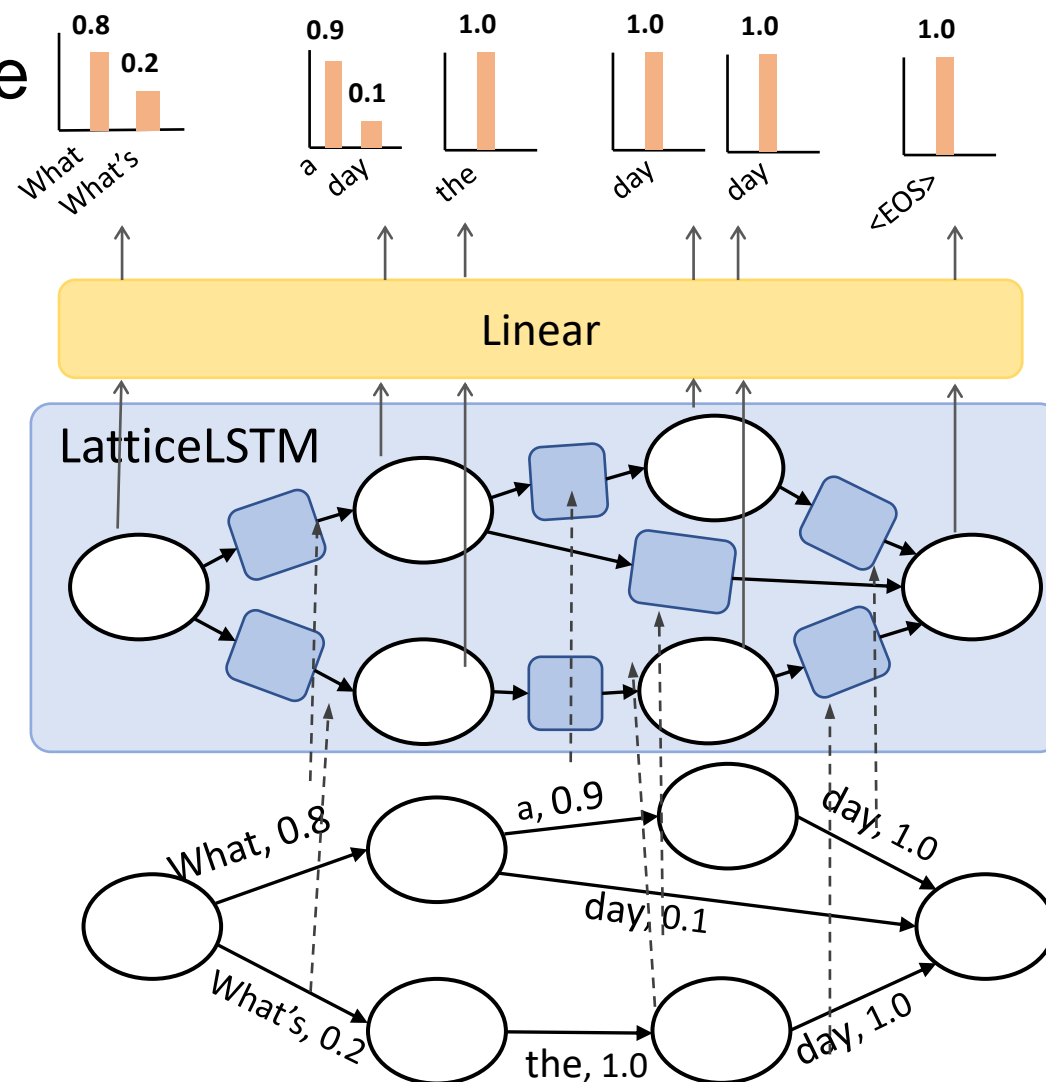
“

A thin vertical line extends downwards from the bottom of the yellow circle.

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



Ablation Study

$$\mathcal{L}_{pt} = \mathcal{L}_c + \lambda_{mlm} \cdot \mathcal{L}_{mlm}$$

$$\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d \mathcal{L}_d + \lambda_{hard} \mathcal{L}_{hard} + \lambda_{soft} \mathcal{L}_{soft}$$

Pre-Training	Fine-Tuning	SLURP	ATIS	TREC6
Full	Full	85.26	95.10	86.36
No \mathcal{L}_{mlm}	Full	84.83	93.75	85.32
No \mathcal{L}_c	Full	85.15	95.00	85.53
Full	No $\mathcal{L}_{hard} + \mathcal{L}_{soft}$	85.14	94.83	86.08
Full	No $\mathcal{L}_d + \mathcal{L}_{soft}$	84.77	94.75	85.60
Full	No \mathcal{L}_{soft}	84.81	94.65	86.20

All parts in the proposed approach are necessary to achieve better SLU performance.

Improvement of Different WER

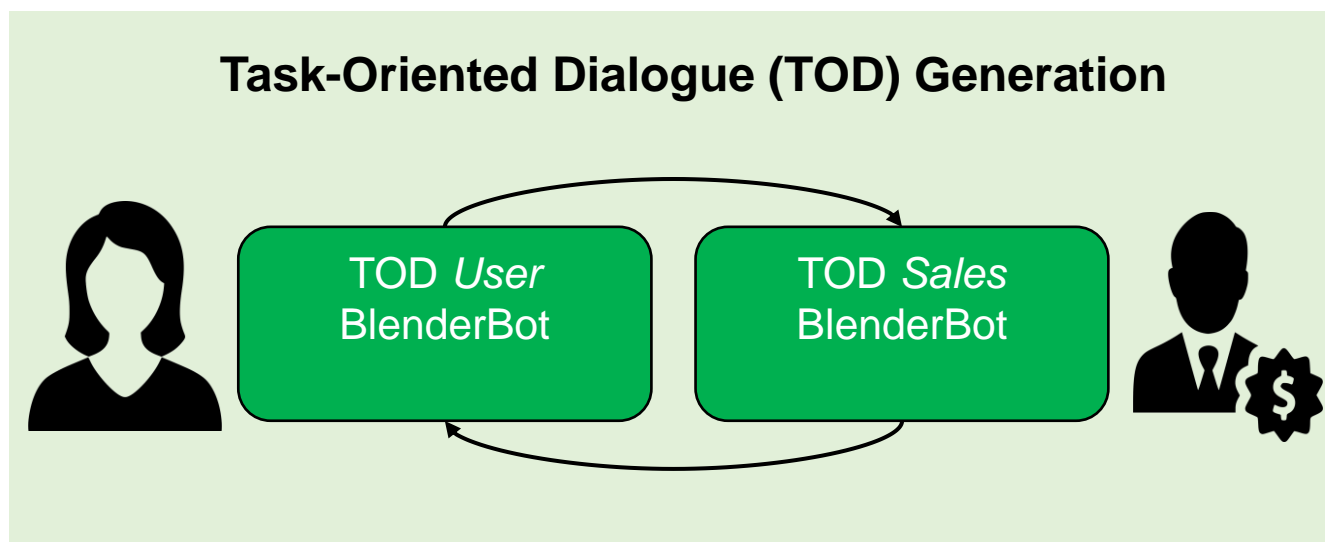
Pre-Training	Fine-Tuning	SLURP WER Interval				All
		Clean =0	Low (0, 0.16]	Medium (0.16, 0.40]	High >0.4	
RoBERTa	Direct	95.69	92.41	85.89	56.71	83.97
Phoneme-BERT	Direct	94.97	92.34	85.87	57.20	83.78
SimCSE	Direct	95.55	93.47	86.82	57.59	84.47
Proposed	Direct	95.54	93.86	86.68	57.72	84.51
RoBERTa	Proposed	96.59	94.27	86.70	57.24	84.87
Phoneme-BERT	Proposed	95.61	93.42	86.87	57.50	84.48
SimCSE	Proposed	96.57	94.54	87.39	58.01	85.25
Proposed	Proposed	96.08	94.41	87.63	58.72	85.26

Proposed approach is more effective when WER is higher

Proposed fine-tuning can generalize to diverse pre-training strategies for better SLU results

Task-Oriented Dialogue Generation

- Task-Oriented Simulation
 - Two BlenderBot simulators are additionally trained on
 - user turns to simulate **users**
 - agent turns to simulate **salespersons**
 - These turns are taken from task-oriented dialogues.



Transition Turn Generation

Template-based Generation:

- Use a template sentence to trigger the corresponding task-oriented user reaction

Template-based generation	
User:	I like to read a lot. I also like to go to the movies. What about yourself? - FindMovies
Sales:	<i>Do you want to find movies by genre and optionally director?</i>
User:	I'm looking for a movie to watch. A regular showing would be fine.

Detected Intent

Template Transition

Generative-based Generation:

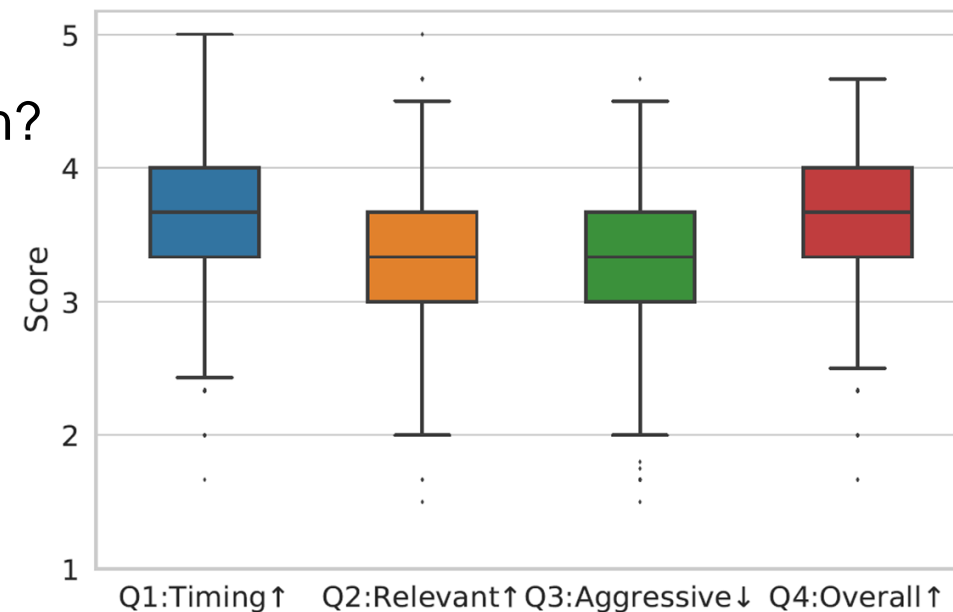
- Re-generate the transition turn for better *fluency* and *diversity*

Generative-based Re-generation	
User:	I like to read a lot. I also like to go to the movies. What about yourself?
Sales:	<i>Are you interested in watching any movie?</i>
User:	I'm looking for a movie to watch. A regular showing would be fine.

Generated Transition

Transition turn quality

- **Q1 Timing:** Is it a good timing to make the transition?
- **Q2 Relevance:** Is the transition relevant to the conversation context?
- **Q3 Aggressiveness:** Is the transition aggressive?
- **Q4 Overall:** Do you think it is overall a good transition?



Average scores over 3 workers (4000 dialogues)

All scores above 3 (neutral) demonstrates reasonable quality of the generated data