The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2022)

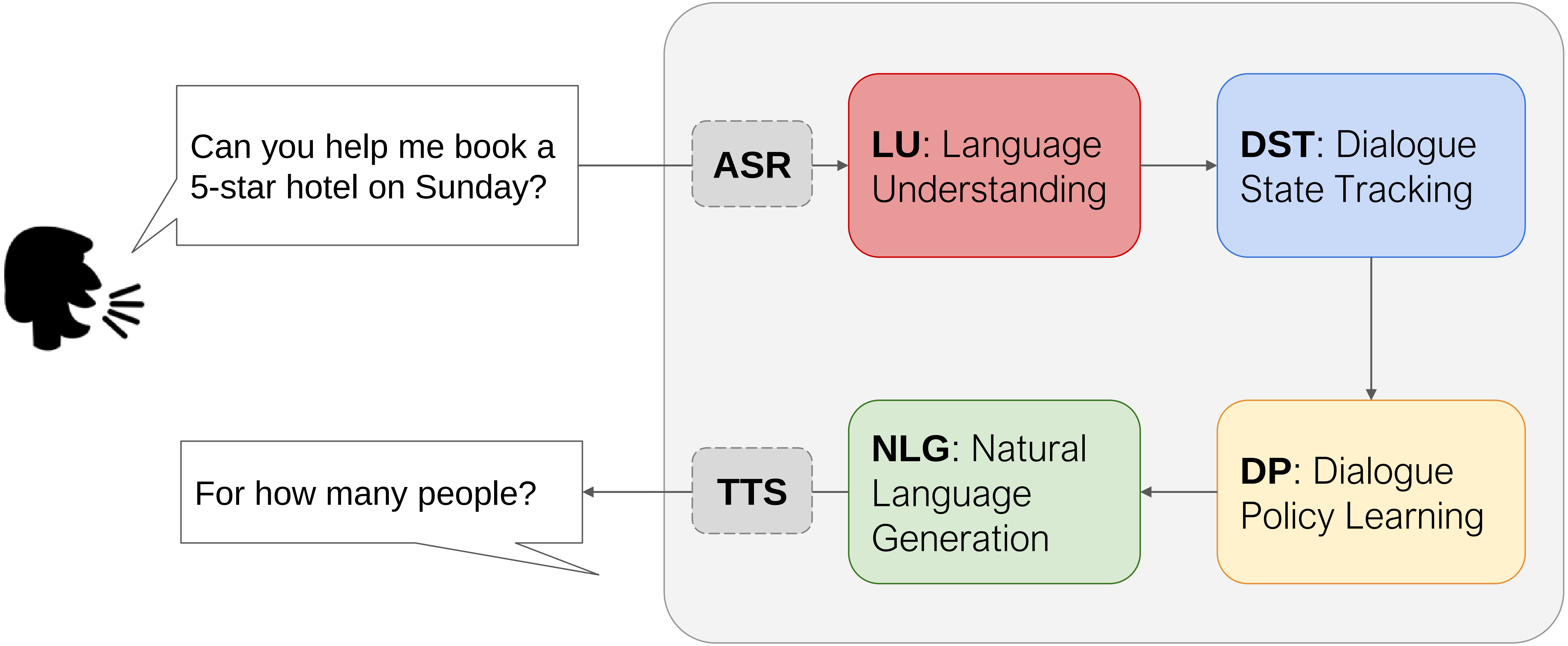
National Taiwan University
國立臺灣大學

陳縕儂 Yun-Nung (Vivian) Chen

September 7th, 2022

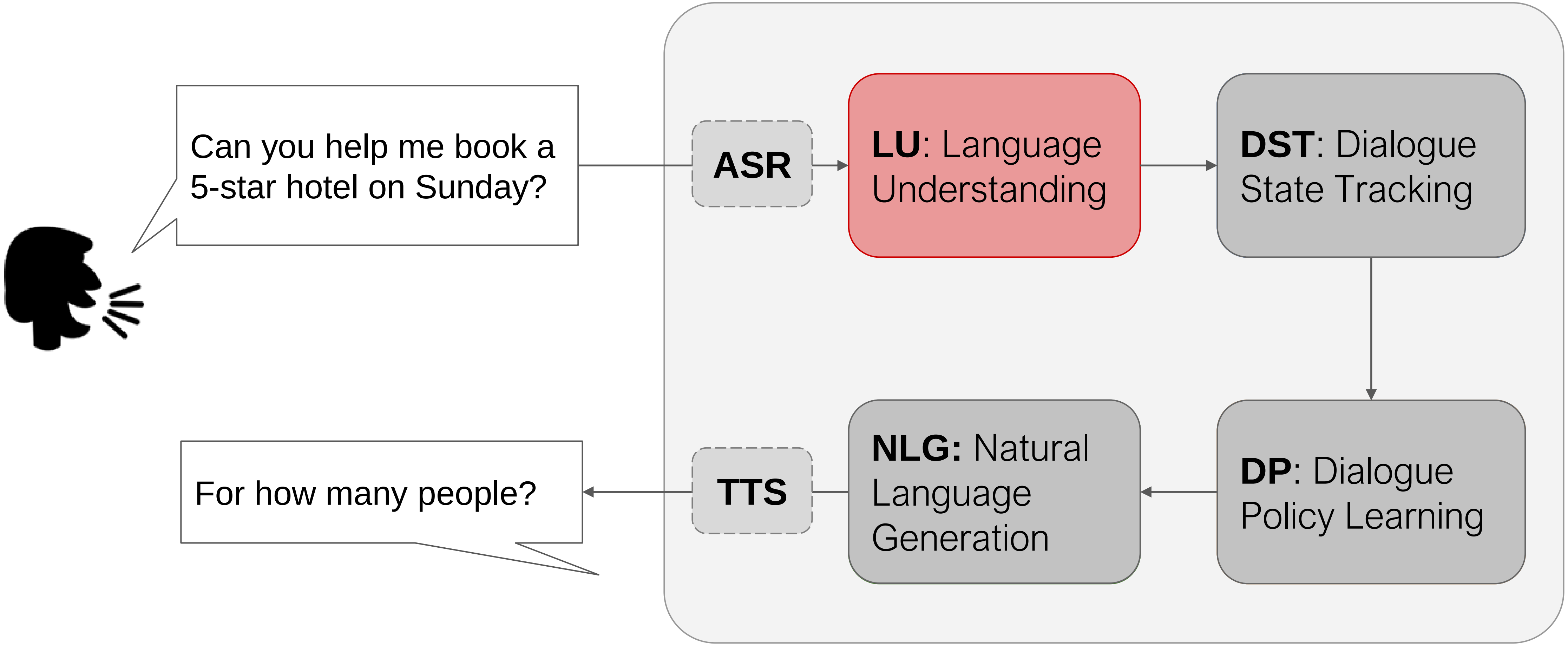
Robustness, Scalability & Practicality of Conversational AI

Task-Oriented Dialogue Systems (Young, 2000)

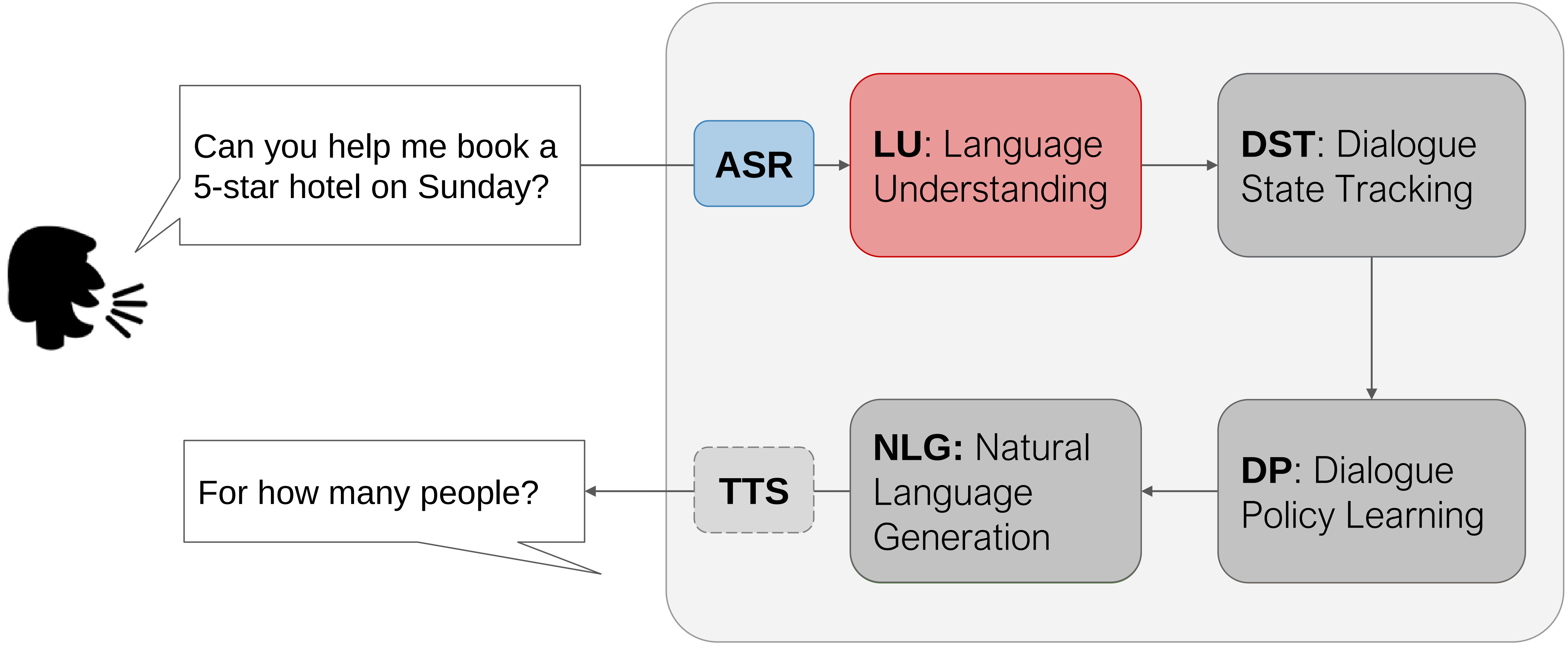


Robustness

Task-Oriented Dialogue Systems (Young, 2000)

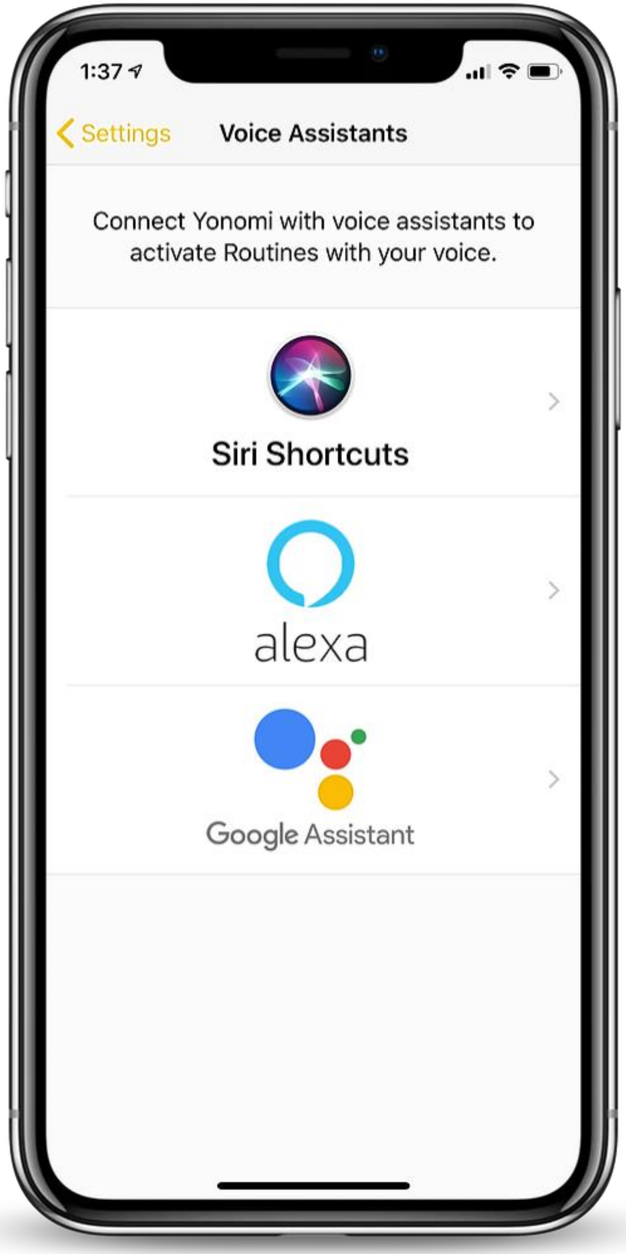


Task-Oriented Dialogue Systems (Young, 2000)



Recent Advances in NLP

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




Lift all lights to Morocco
List all flights tomorrow
Matt, what can I assist you with?

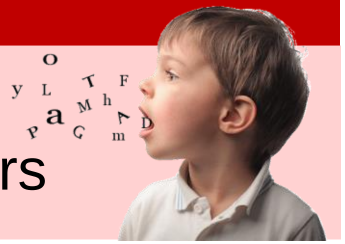
Mismatch between Written and Spoken Languages

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

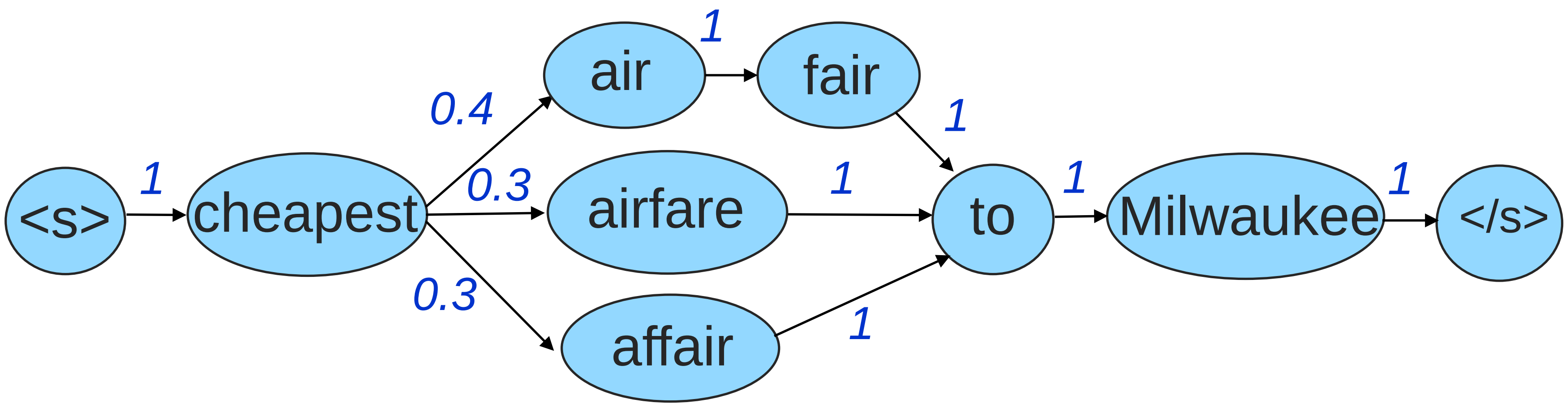
(Huang & Chen, ACL 2020)

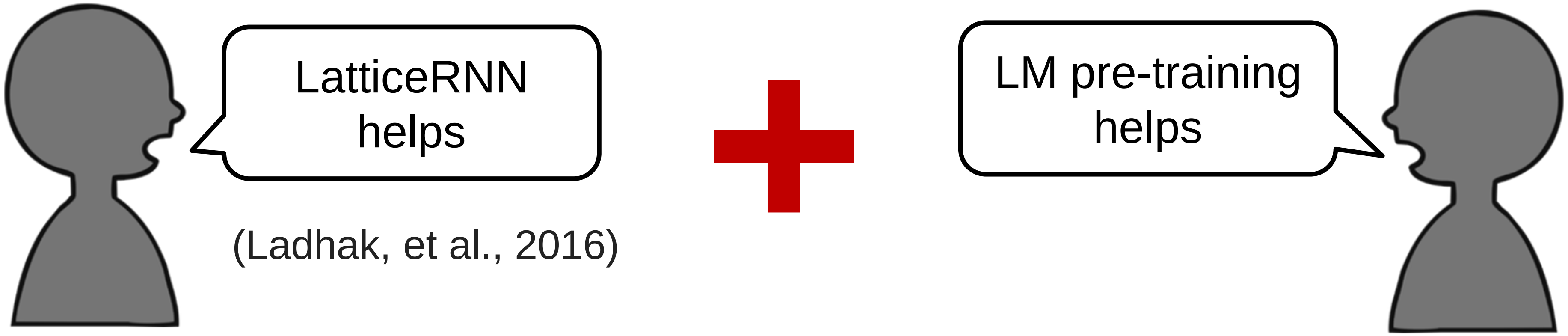
https://github.com/MiuLab/LatticeLM

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

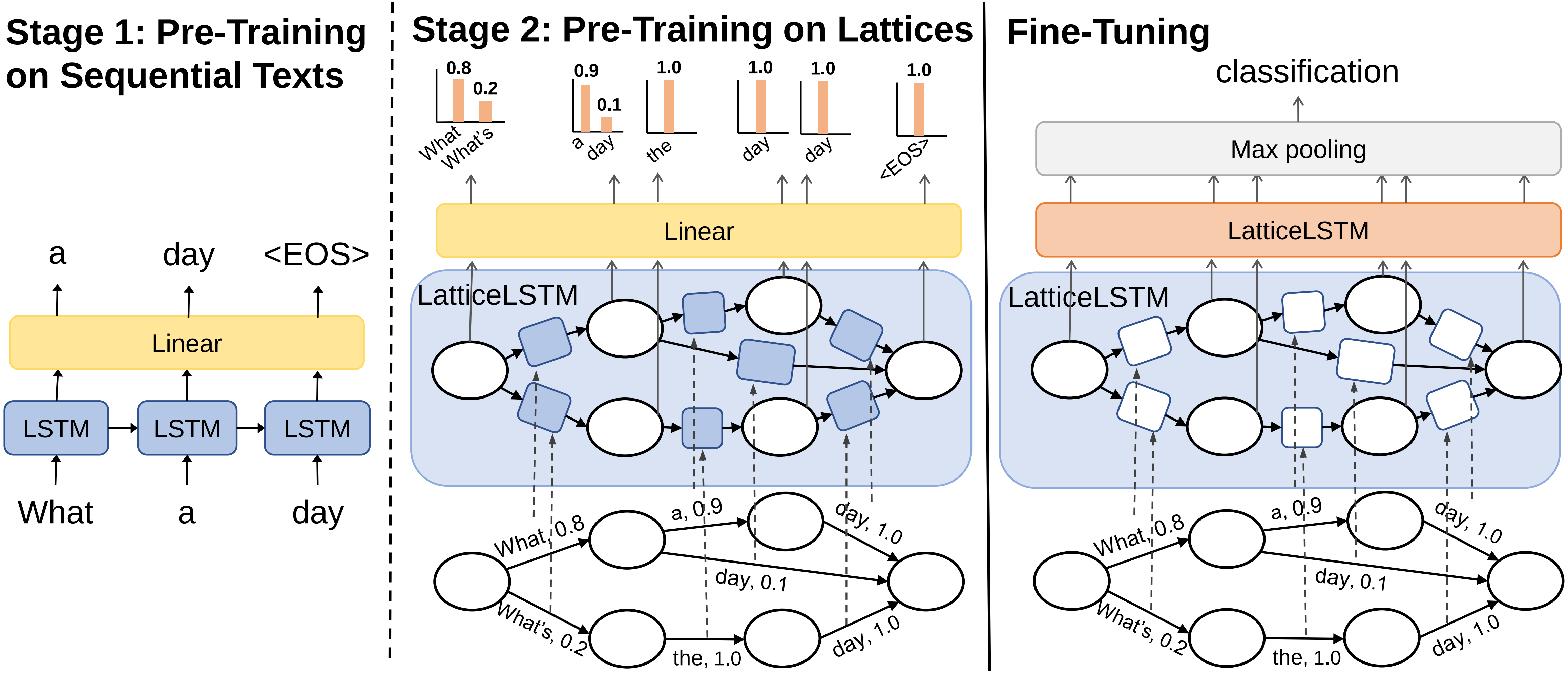
- Idea: lattices may include correct words





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.

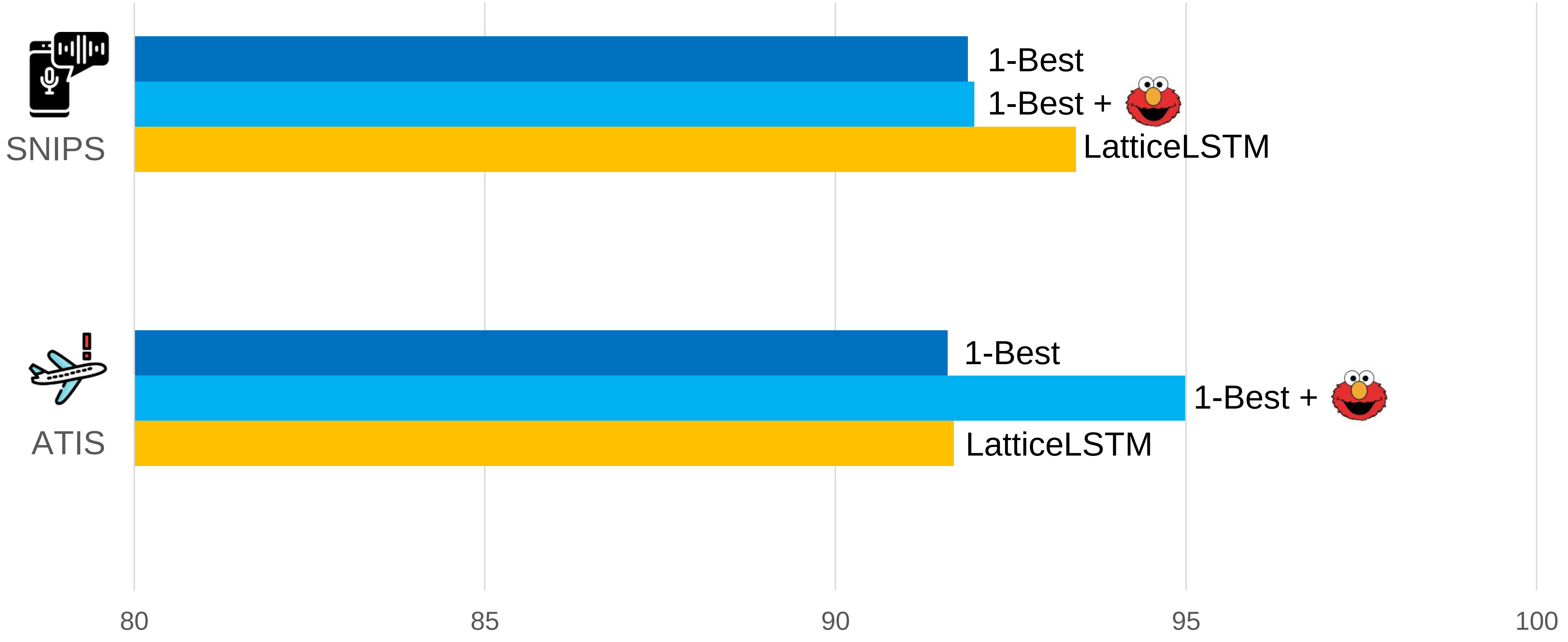
LatticeLM: Efficient Two-Stage Pre-Training



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.

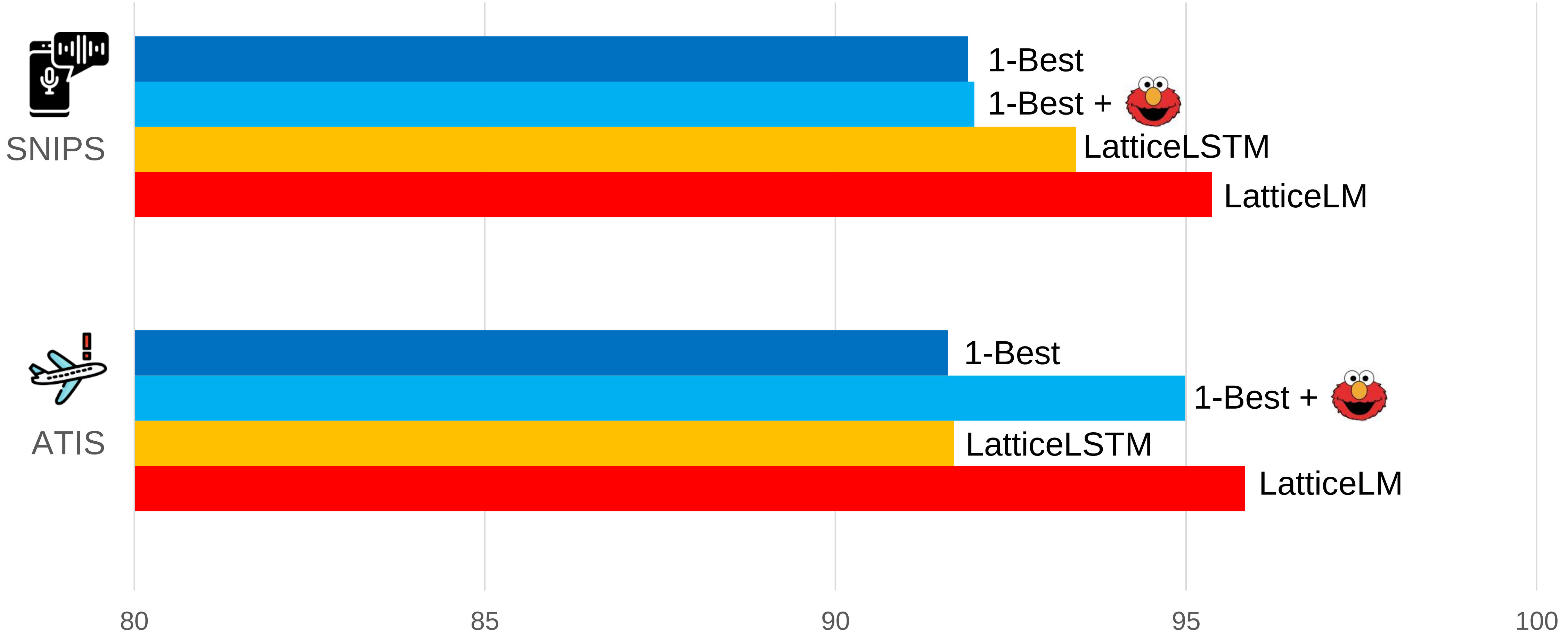
Spoken Language Understanding Results

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



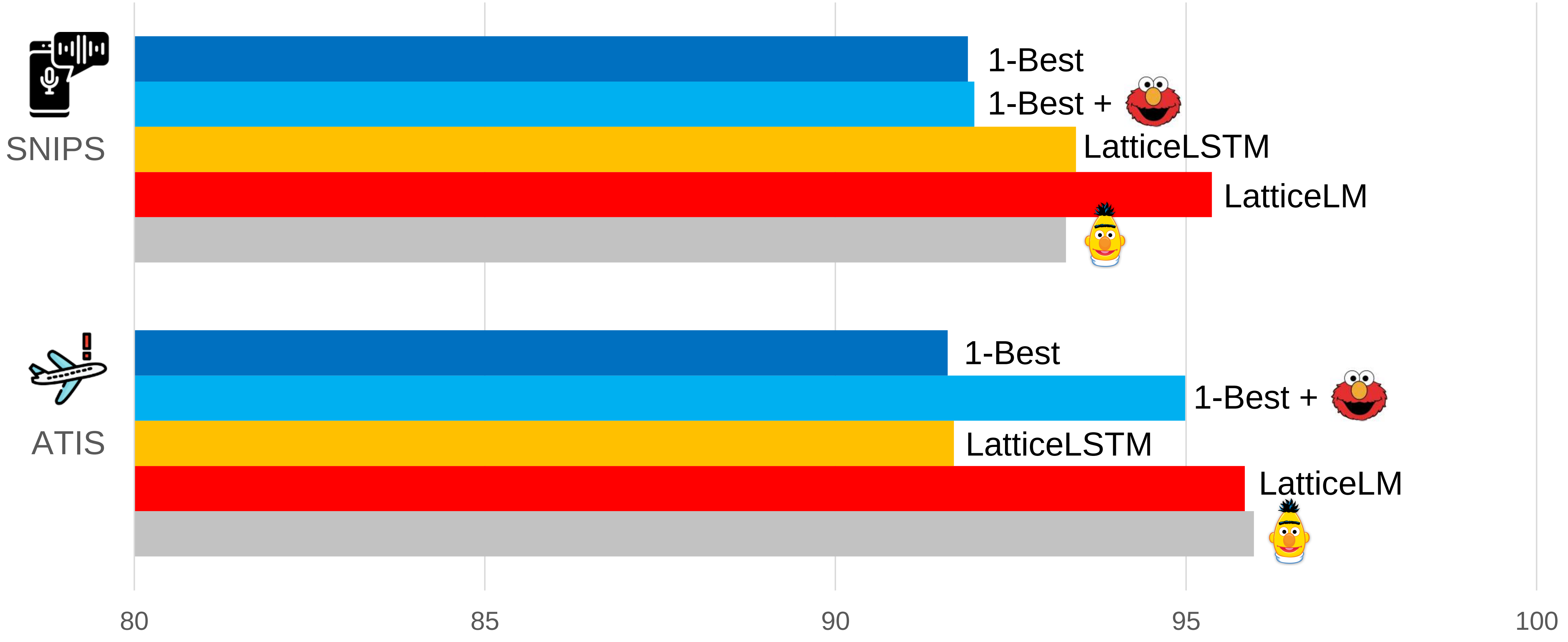
Spoken Language Understanding Results

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



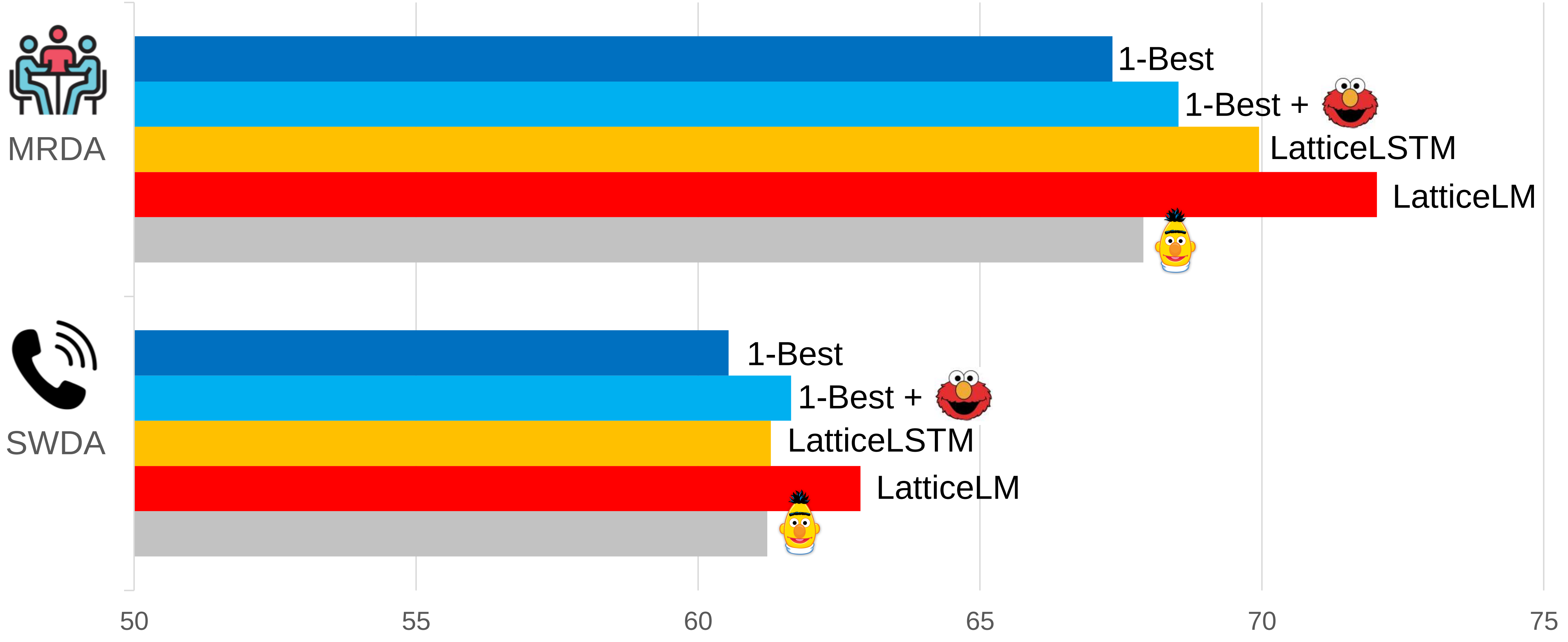
Spoken Language Understanding Results

- 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

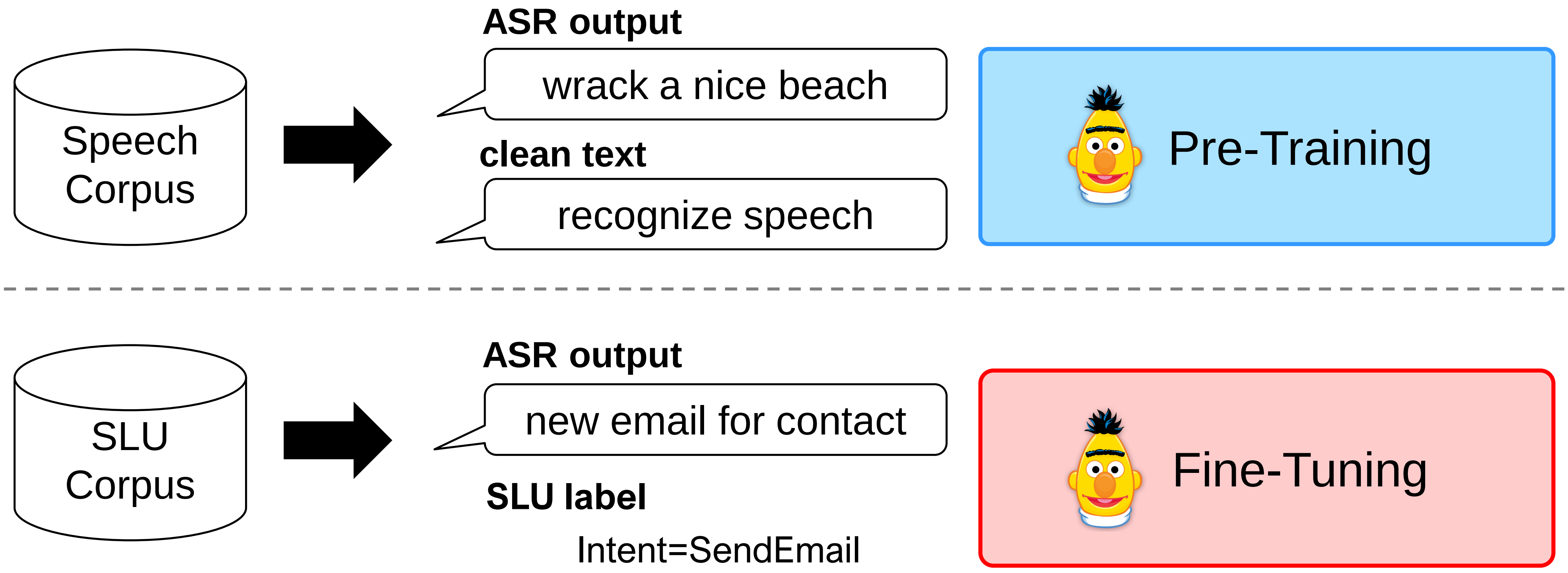
(Chang & Chen, INTERSPEECH 2022)

https://github.com/MiuLab/SpokenCSE

Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Improving ASR Robustness of Embeddings

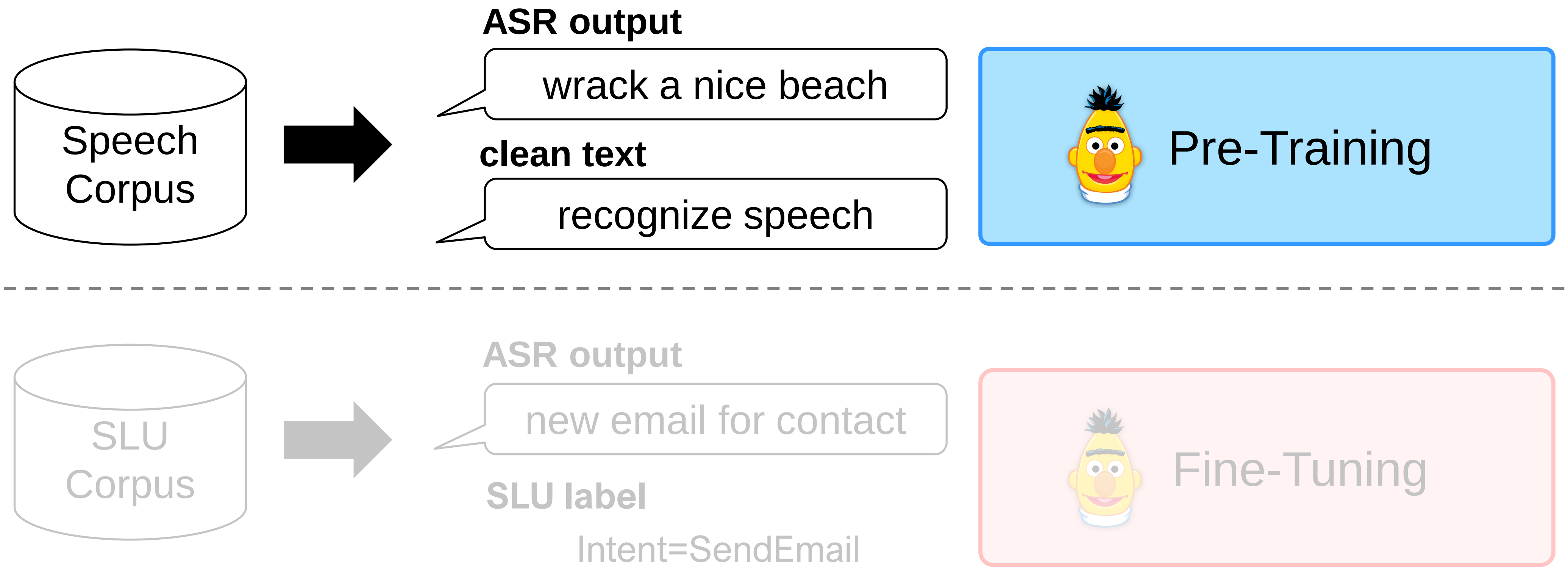
- Idea: adapt embeddings robust to errors <u>with only textual information</u>



Ya-Hsin Chang and Yun-Nung Chen, “Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding,” in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Improving ASR Robustness of Embeddings

- Idea: adapt embeddings robust to errors <u>with only textual information</u>

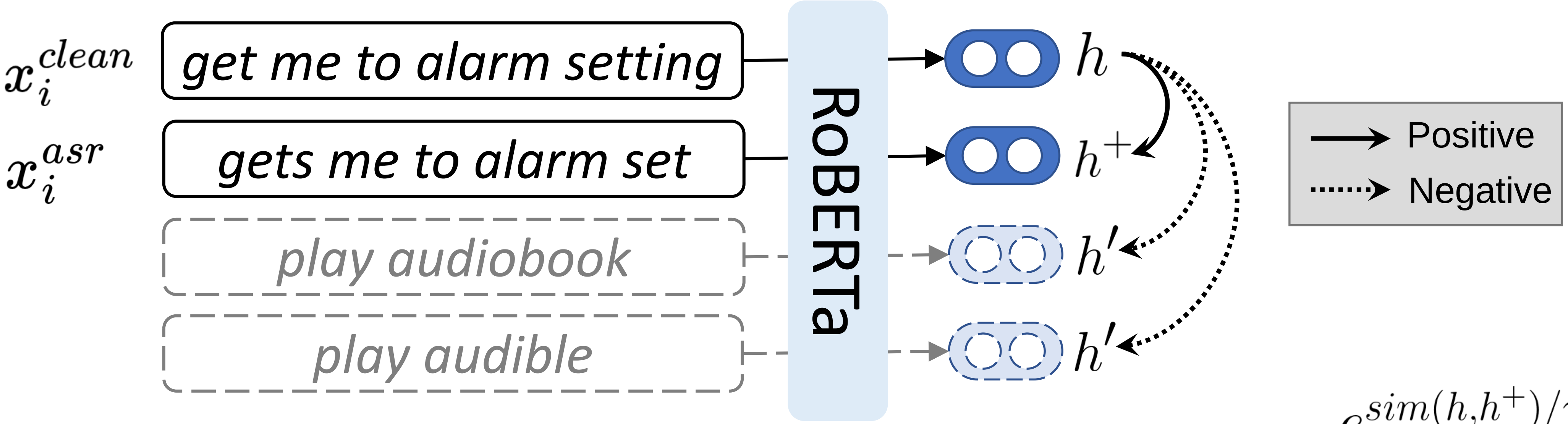


Idea: contrastive pre-training for spoken scenarios

Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Contrastive Pre-Training

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



$$\mathcal{L}_c = -\sum_h \log \frac{e^{sim(h,h^+)/\tau_c}}{\sum_{h' \neq h} e^{sim(h,h')/\tau_c}}$$

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

prevent catastrophic forgetting

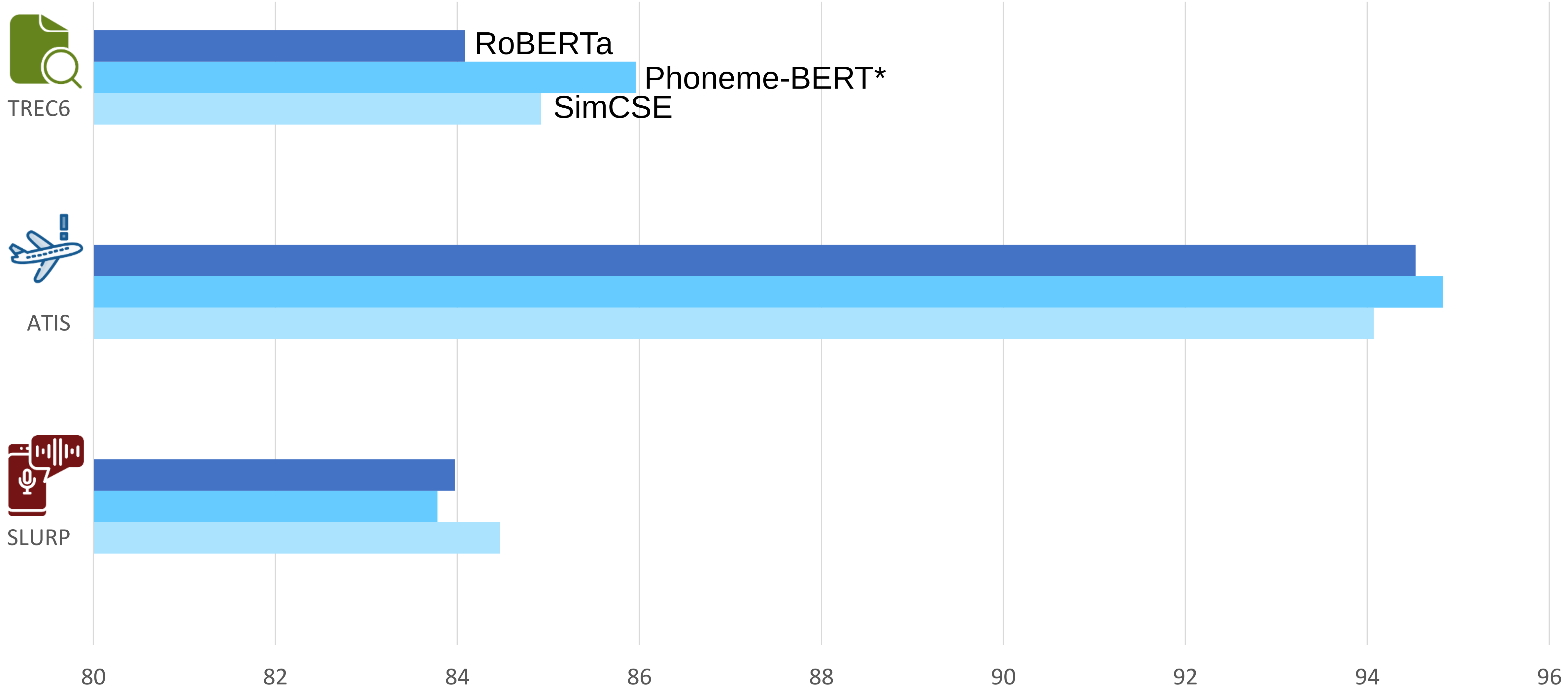
Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Spoken Language Understanding Results

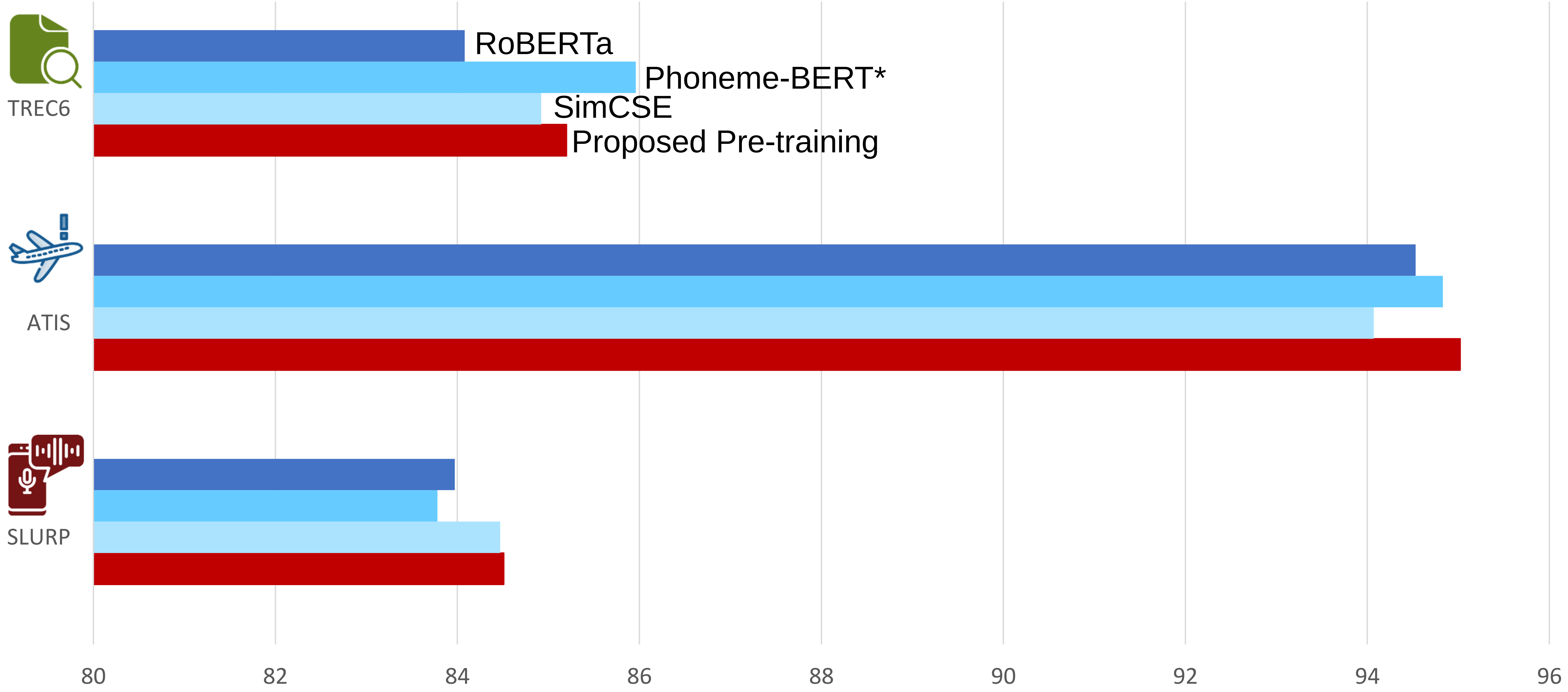
- SLU data
  - Synthesized TREC6 (WER=29%) & ATIS (WER=32%)
  - SLURP: Spoken Language Understanding Resources Package (WER=25%)

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



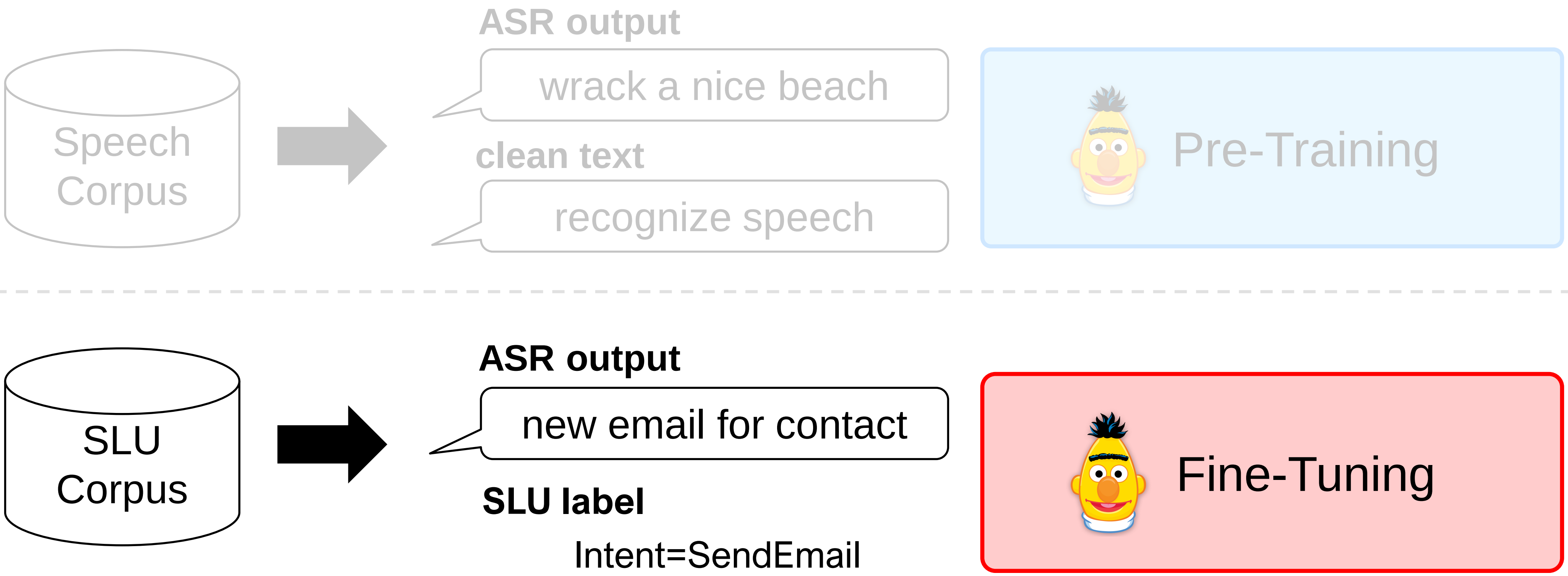
Spoken Language Understanding Results



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

Improving ASR Robustness of Embeddings

- Idea: adapting embeddings robust to misrecognitions

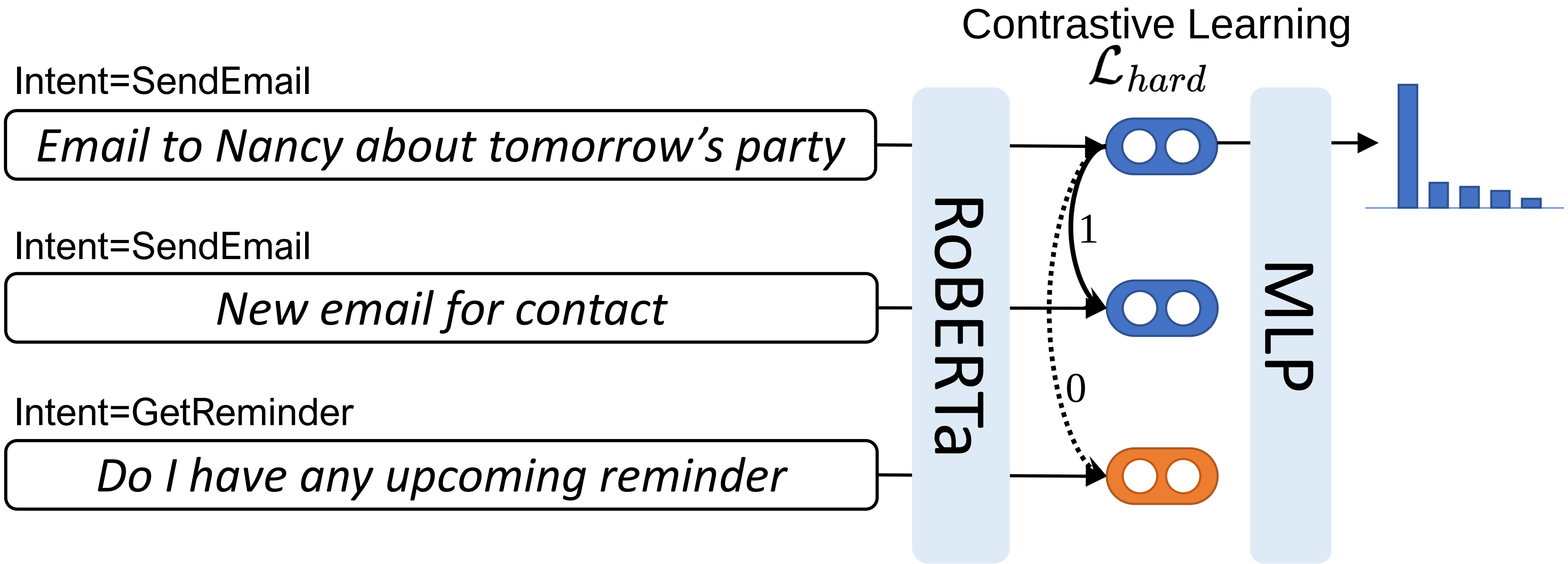


Idea: supervised contrastive learning with self-distillation

Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Supervised Contrastive Learning

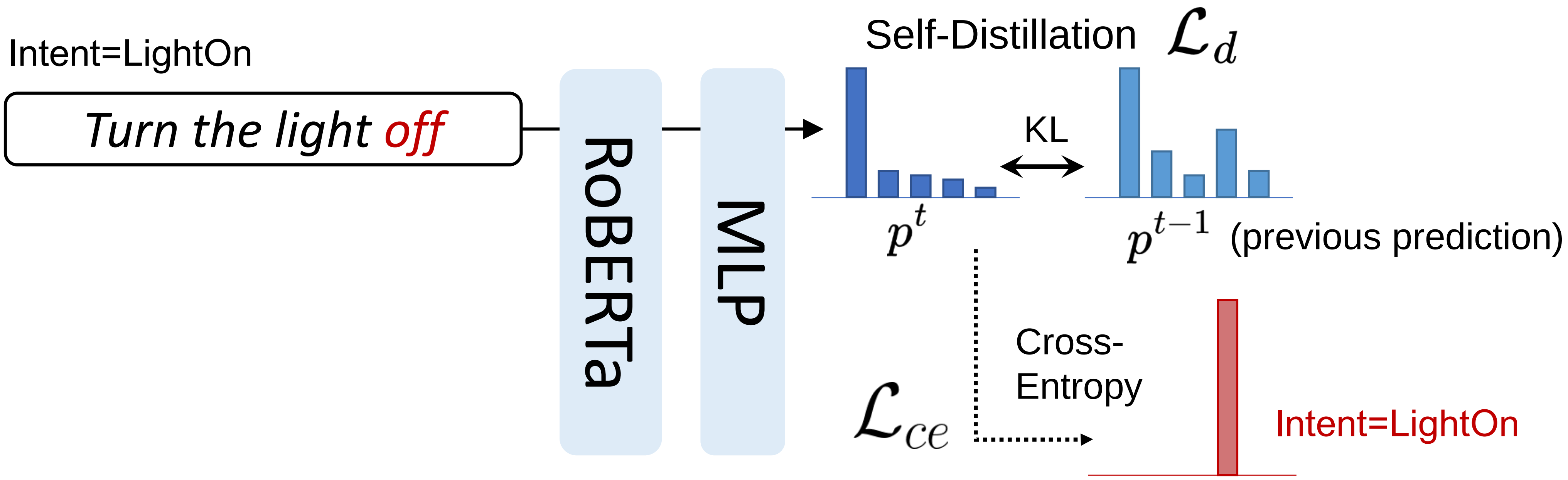
- Idea: data with the same label should be close to each other
- Objective: $\mathcal{L}_{hard} = -\sum_{i} \sum_{j \neq i} 1_{y_i = y_j} \log \frac{e^{sim(h_i, h_j)/\tau_{sc}}}{\sum_{k \neq i} e^{sim(h_i, h_k)/\tau_{sc}}}$



Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Self-Distillation

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



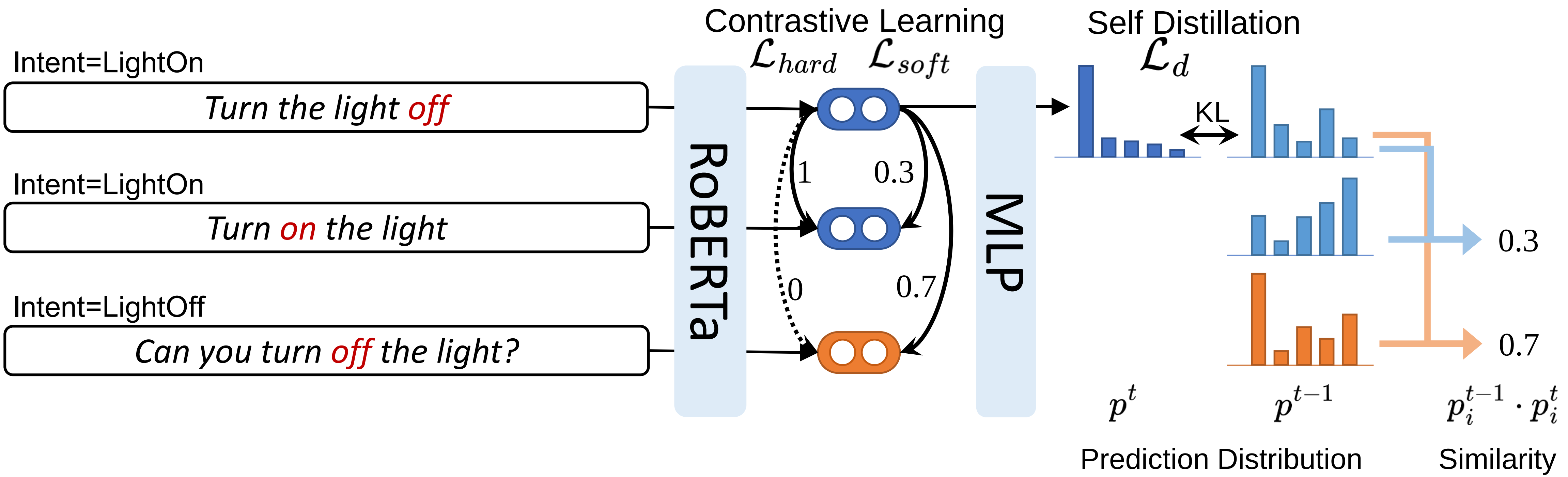
Ya-Hsin Chang and Yun-Nung Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *The 23rd Annual Meeting of the International Speech Communication Association (INTERSPEECH)*, 2022.

Supervised Contrastive with Self-Distillation

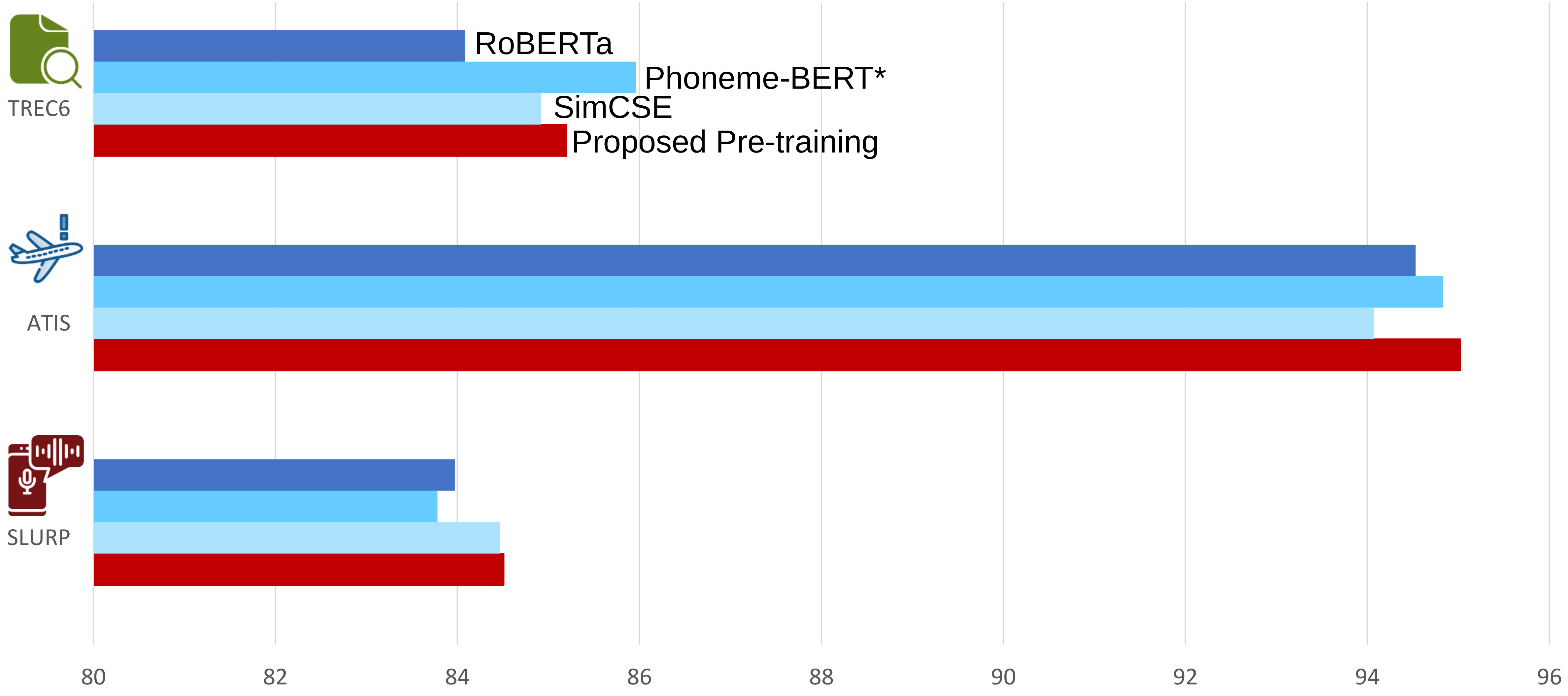
- Issue: noisy labels also affect $\mathcal{L}_{hard}$

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

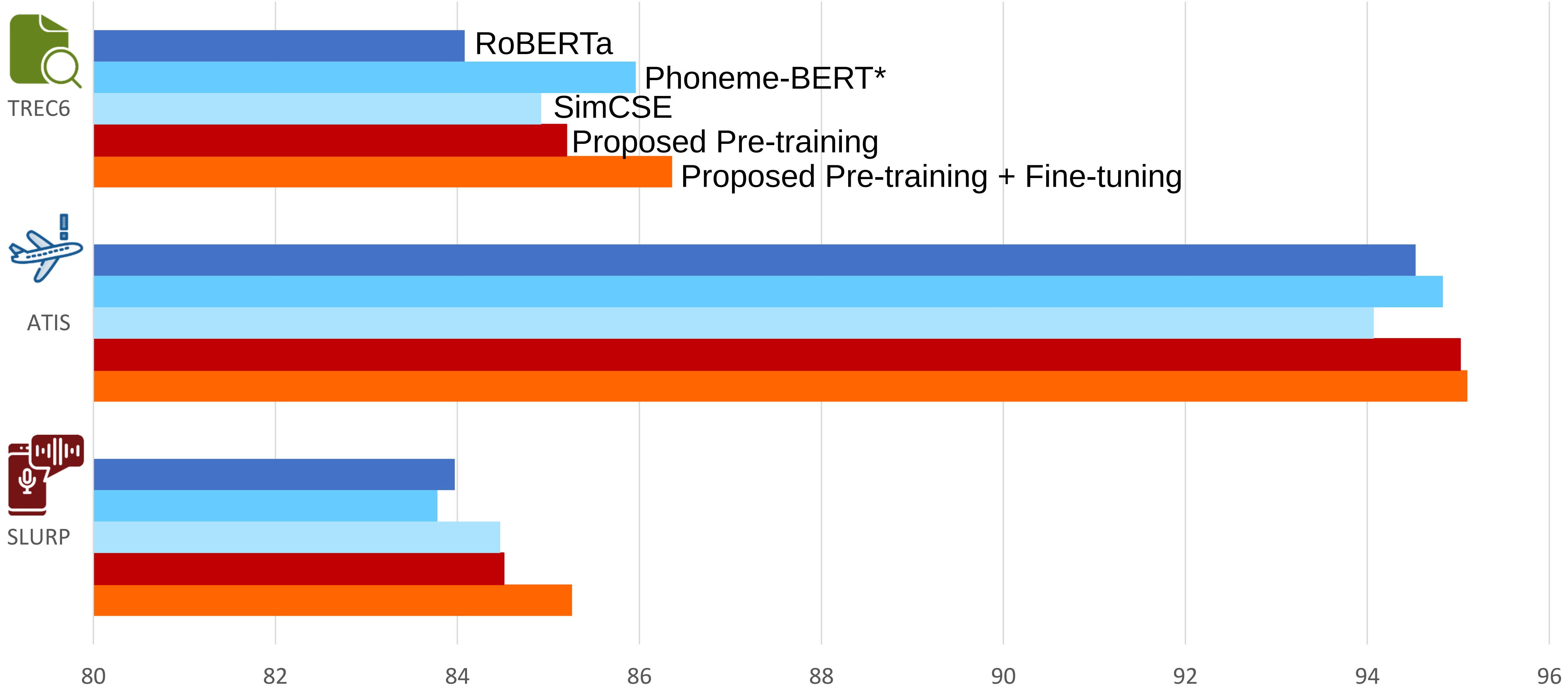
- Fine-tuning objective: $\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d\mathcal{L}_d + \lambda_{hard}\mathcal{L}_{hard} + \lambda_{soft}\mathcal{L}_{soft}$



Spoken Language Understanding Results



Spoken Language Understanding Results

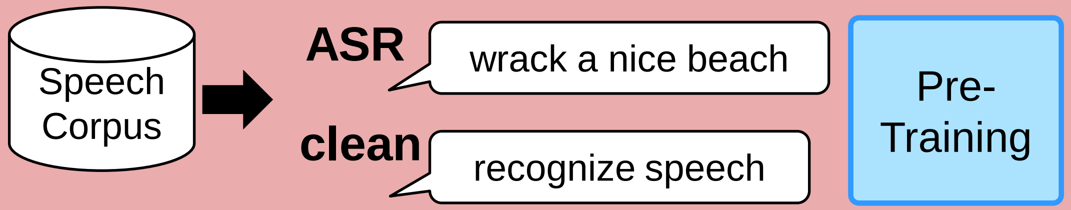


Proposed methods improve robustness to ASR errors and label noises

Robustness

✓ **LatticeLM** for preserving uncertainty

✓ **Contrastive learning** with only textual information



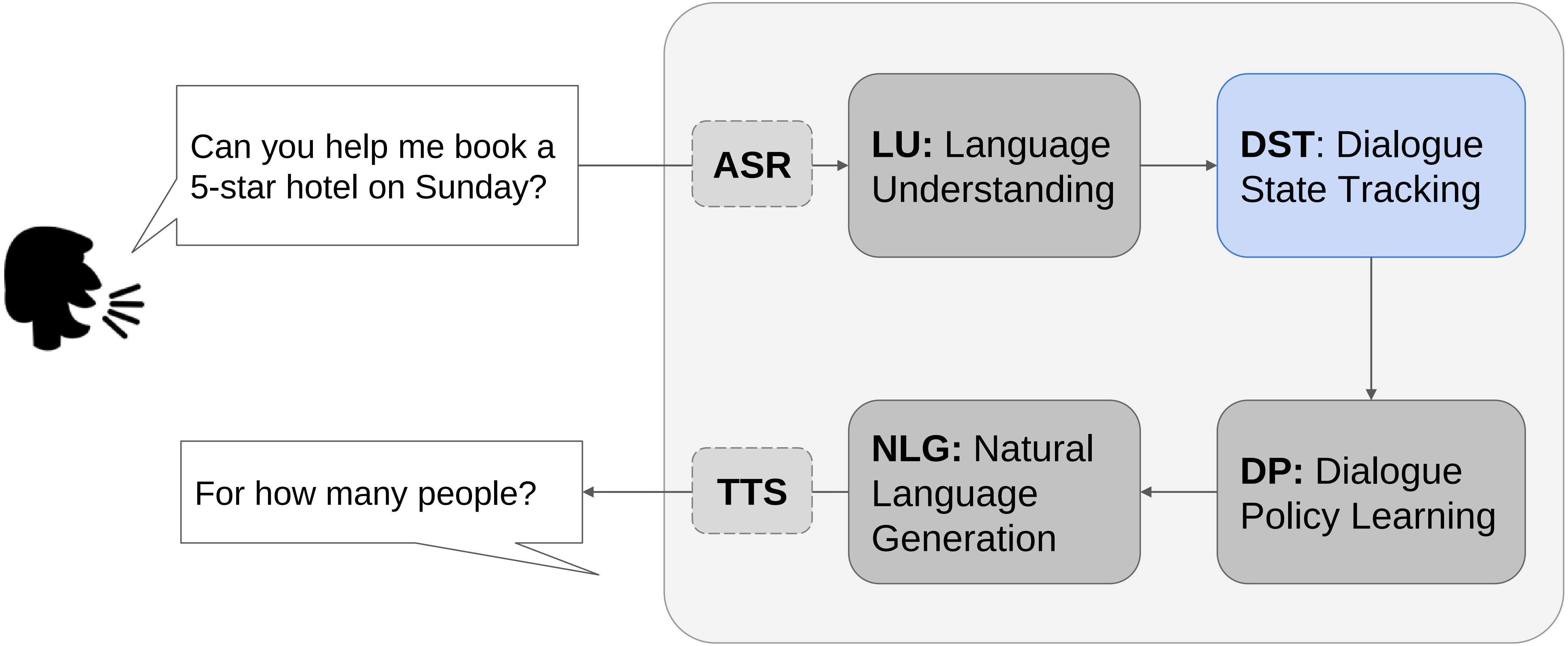
- **Contrastive Pre-training** learns error-invariant sentence embeddings
- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

Scalability

Practicality

Scalability

Task-Oriented Dialogue Systems (Young, 2000)



Issue: limited labelled data hinders scalability

Poster: Today 1pm (National Robotarium)

Solution: Data Augmentation

(Lai et al., SIGDIAL 2022)

https://github.com/MiuLab/CUDA

Chun-Mao Lai, Ming-Hao Hsu, Chao-Wei Huang, and Yun-Nung Chen, “Controllable User Dialogue Act Augmentation for Dialogue State Tracking,” in *Proceedings of The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2022.

Diverse User Dialogue Acts in States

Hello, how can I help you?

I need to find a restaurant in the center.

I recommend Pho Bistro, a popular restaurant in the center.

Recommendation { rest-name=pho bistro, rest-area=center }

No, it needs to serve British food and I'd like a reservation for 18:00.

Confirm=False, Inform{ rest-area=center, rest-food=British, rest-time=18:00 }

Value Substitution

No, it needs to serve Chinese food and I'd like a reservation for 17:00.

Confirm=False, Inform{ rest-area=center, rest-food=Chinese, rest-time=17:00 }

CoCo
(Li et al., ICLR '21)

No, it should serve Chinese food and I need to book a table for 2 people.

Confirm=False, Inform{ rest-area=center, rest-food=Chinese, rest-people=2 }

Slot Combination

Proposed

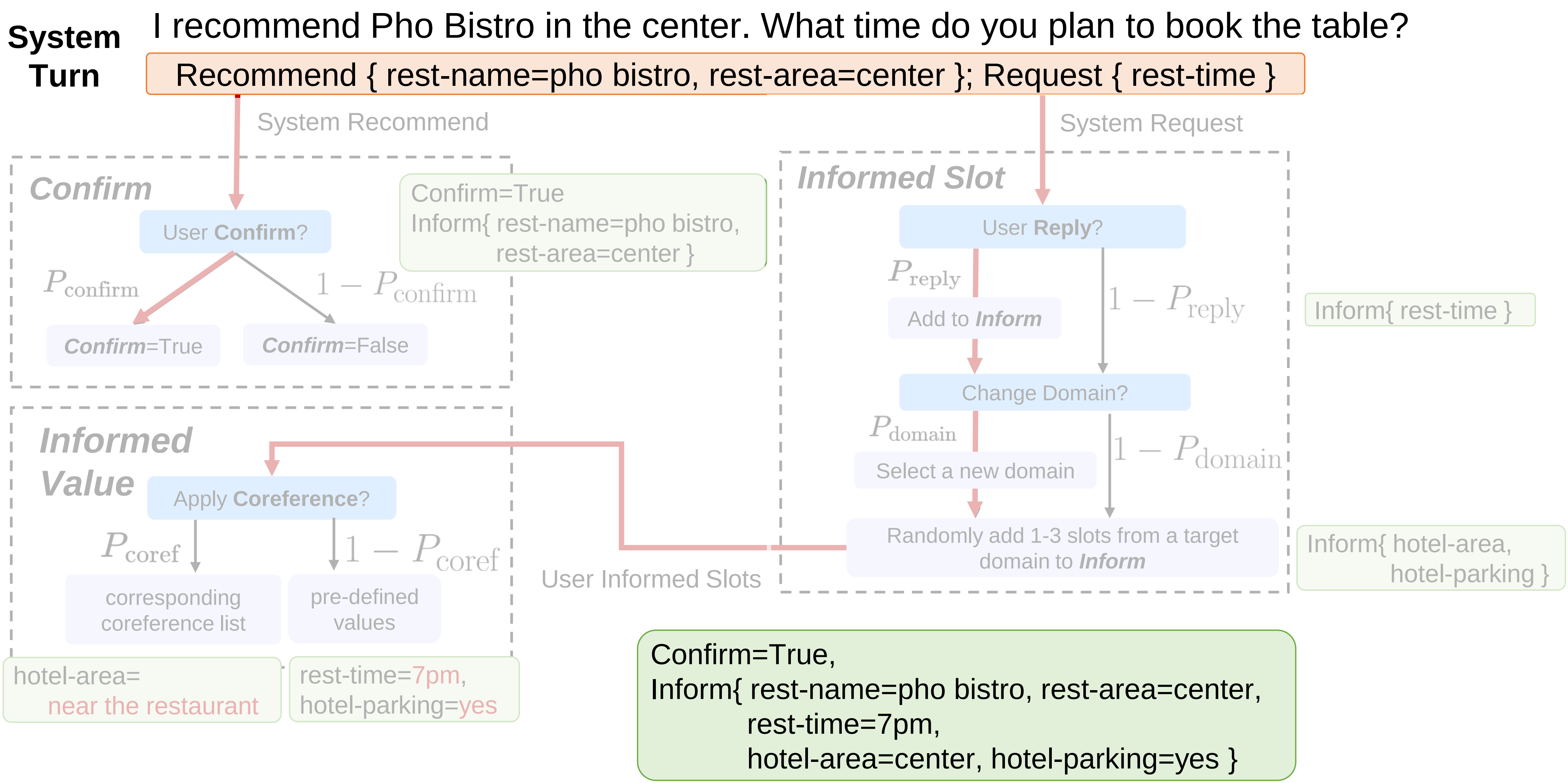
Thank you, can you also find me a hotel with parking near the restaurant?

Confirm=True, Inform{ rest-area=center, rest-name=pho bistro, hotel-area=center, hotel-parking=yes }

Confirm **Domain Change** **Coreference** **Boolean**

Chun-Mao Lai, Ming-Hao Hsu, Chao-Wei Huang, and Yun-Nung Chen, "Controllable User Dialogue Act Augmentation for Dialogue State Tracking," in *Proceedings of The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2022.

User Dialogue Act Generation



User Utterance Generation

System Utterance I recommend Pho Bistro in the center. What time do you plan to book the table?

User Dialogue State

Confirm=True,
Inform{ rest-name=pho bistro, rest-area=center,
rest-time=7pm, hotel-area=center, hotel-parking=yes }

beam search

User Utterance

1.Thank you, 7pm works for me. Can you also find me a hotel with parking near the restaurant?

Slot Appearance: ✓ Value Consistency: (span) ✓ (boolean) ✓

~~2.Thank you, 7pm works for me. Can you also find me a hotel without parking near the restaurant?~~

Slot Appearance: ✓ Value Consistency: (span) ✓ (boolean) ✗

~~3.Thank you, 7pm works for me. Can you also find me a hotel with parking in the center of the town?~~

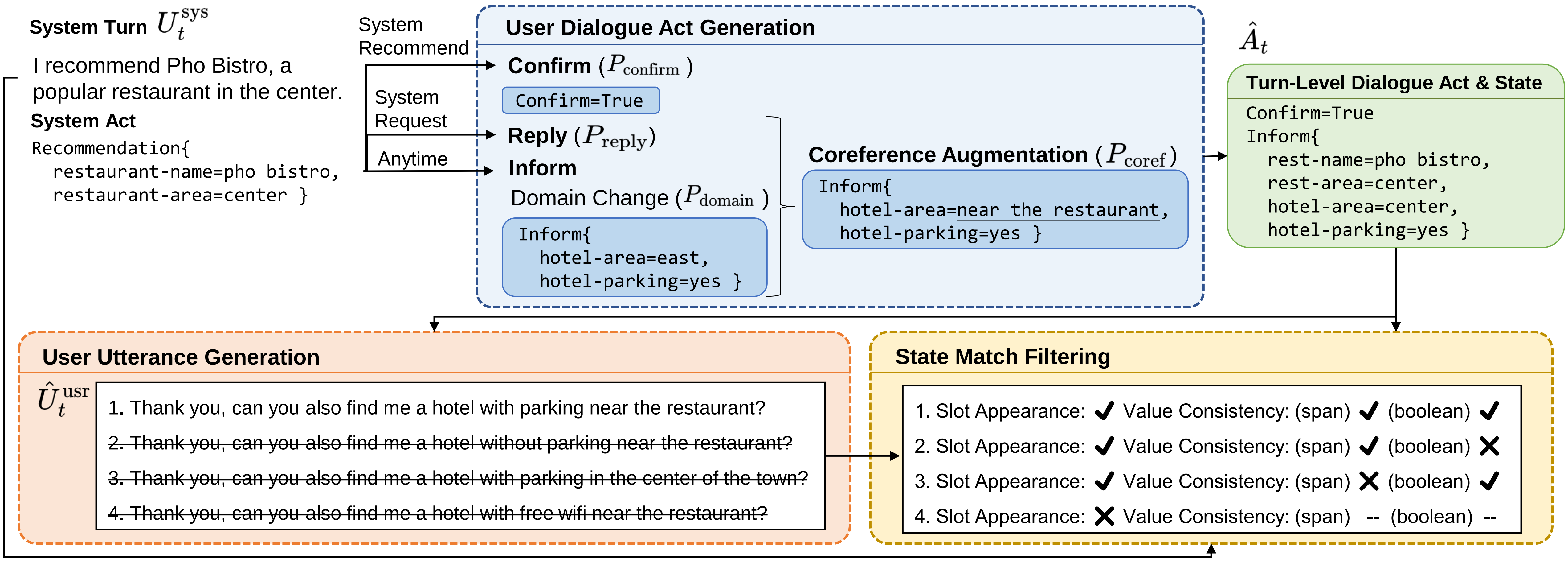
Slot Appearance: ✓ Value Consistency: (span) ✗ (boolean) ✓

~~4.Thank you. Can you also find me a hotel with free wifi near the restaurant?~~

Slot Appearance: ✗ Value Consistency: (span) -- (boolean) --

→ **78% success rate for generation** (57% for CoCo)

CUDA: Controllable User Dialogue Act



Chun-Mao Lai, Ming-Hao Hsu, Chao-Wei Huang, and Yun-Nung Chen, “Controllable User Dialogue Act Augmentation for Dialogue State Tracking,” in *Proceedings of The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2022.

DST Performance

MultiWOZ 2.1	TripPy	TRADE
Original	57.72	44.08
Value Substitution	59.48	43.76
CoCo (Li et al., 2021)	60.46	43.53
CUDA	**62.93**♦	**44.86**♦

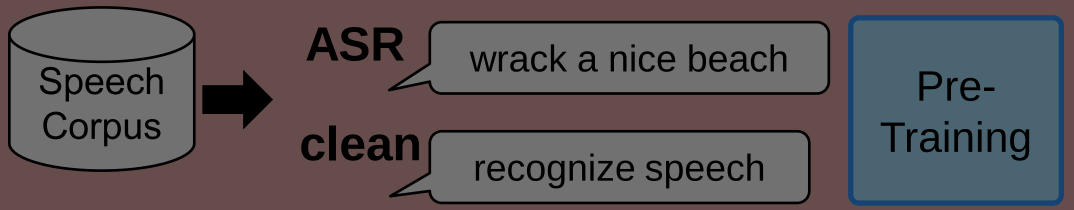
CUDA improves many trackers and achieves SOTA in MultiWOZ 2.1

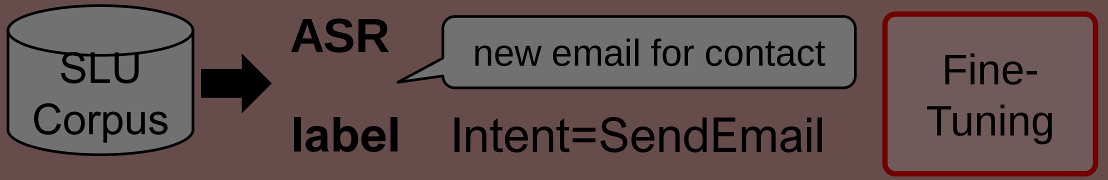
CoCo+ (rare)	TripPy	TRADE
Original	28.38	16.65
Value Substitution	39.42	16.42
CUDA	**48.83**♦	**17.79**♦

CUDA shows better robustness for rare state combinations

Robustness

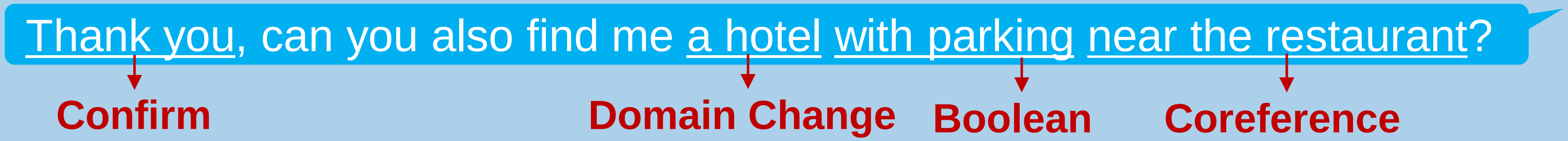
- ✓ **LatticeLM** for preserving uncertainty
- ✓ **Contrastive learning** with only textual information





- **Contrastive Pre-training** learns error-invariant sentence embeddings
- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

Scalability



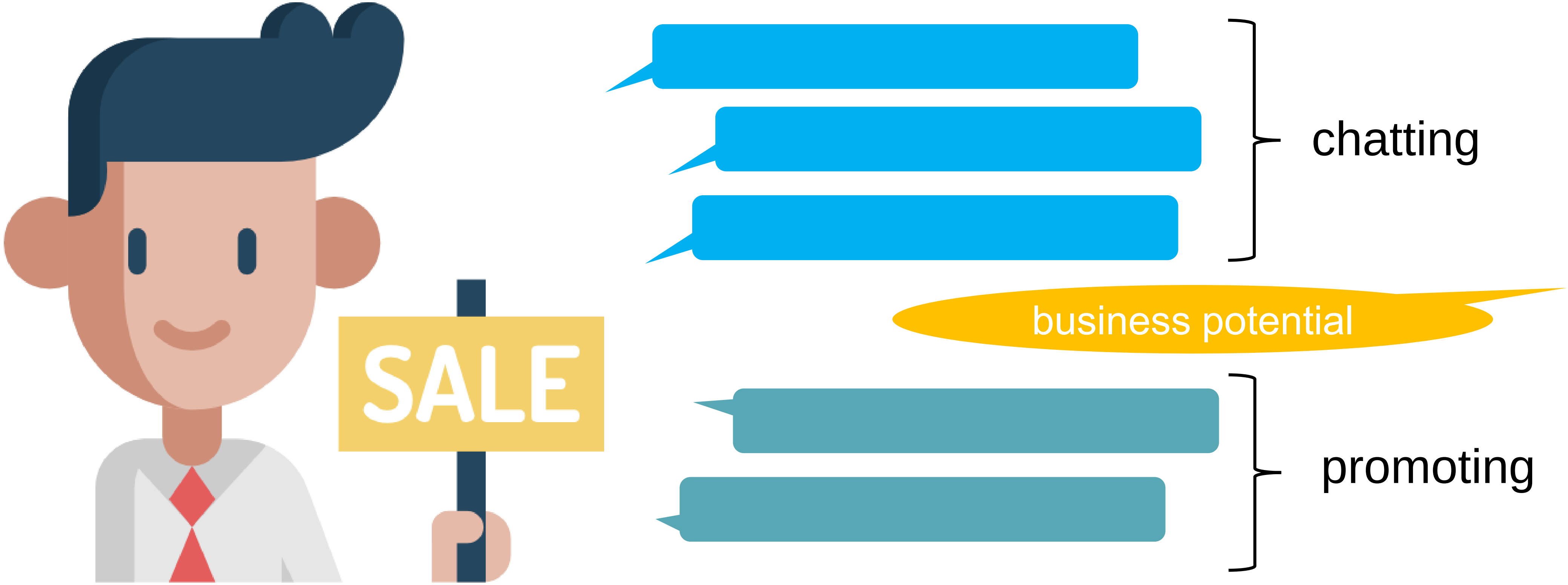
- ✓ **CUDA** augments DST data with diverse user behaviors
  - Efficiency of augmentation
  - Generalizability to different datasets/domains
  - SOTA in MultiWOZ 2.1
  - Potential for user simulators

Practicality

Practicality

Salesperson: Discover Business Potential

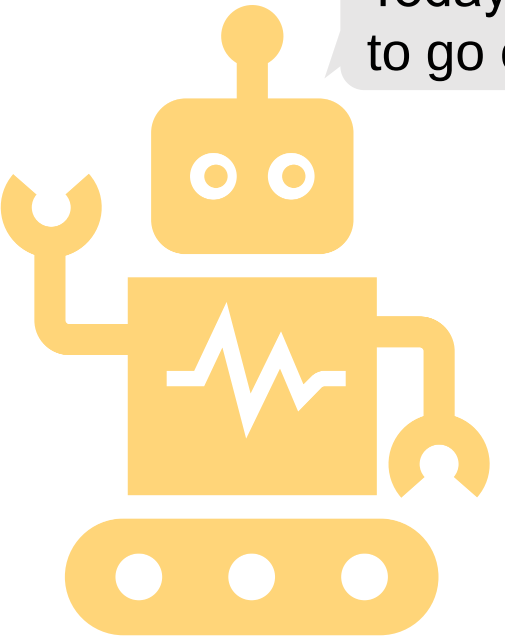
- Human salesperson is capable of
  - *mining business-related intentions* from customers *via chatting*
  - *smoothly promoting* the related product *through conversations*



Two Types of Dialogue Systems

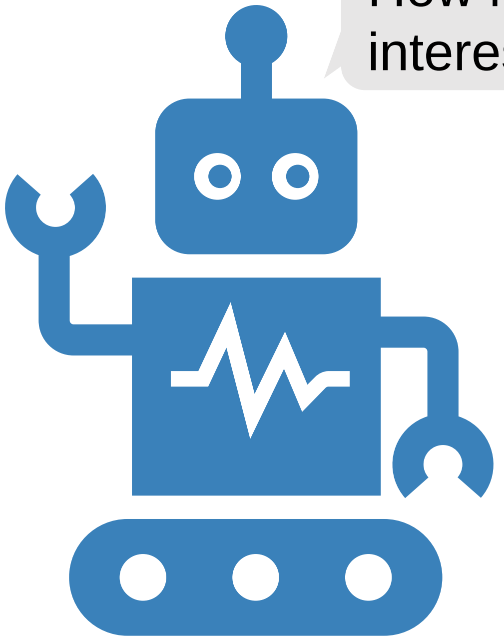
- Open-Domain Chatting

Today is a sunny day. It’s great to go on a picnic.



- Task-Oriented

How may I help you? Are you interested in listening some songs?



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

(Chiu et al., ACL 2022)

https://github.com/MiuLab/SalesBot

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 | **Simulated Salesperson**

Open-Domain

- User: My son is 2 and a half. He is handful, but I love him so much.
- Salesperson: That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?
- User: No, I don't. I work too much to have time for one. How about you?
- Salesperson: I work a lot too. I'm a stay at home mom. What do you do for work?
- User: I work at a library. It's a lot of fun, and I get to meet lots of interesting people.
- Salesperson: That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?
- User: I like to read a lot. I also like to <u>go to the movies</u>. What about yourself?

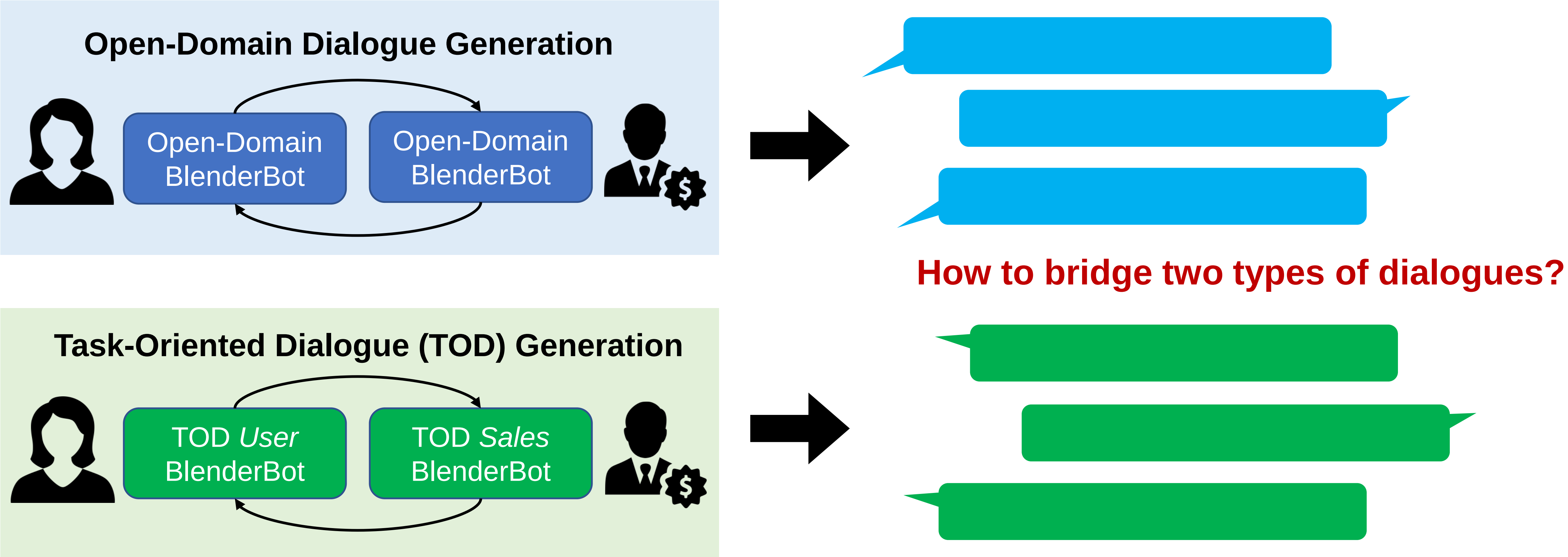
Task-Oriented

- Salesperson: Are you interested in watching any movie?
- User: I'm looking for a movie to watch. A regular showing would be fine.
- Salesperson: Where are you located?
- User: I'm in San Francisco, please look for movies there.
- Salesperson: There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

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

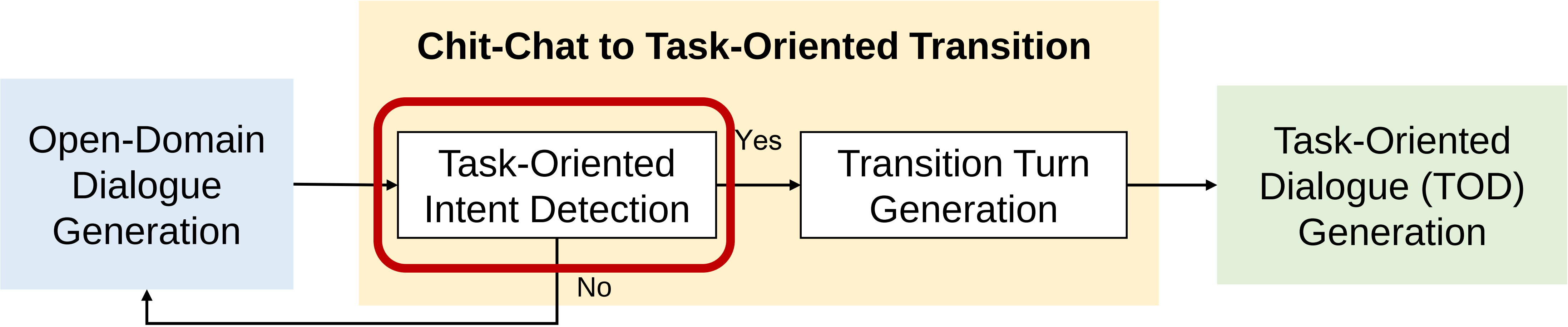
SalesBot: Social to Task-Oriented

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



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



- 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**

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.

Task-Oriented (Implicit) Intent Detector

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

I want to check some landmarks in Paris.

FindAttractions

Explicit

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

FindAttractions

Implicit

Issue: no data with annotated implicit intents

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.

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

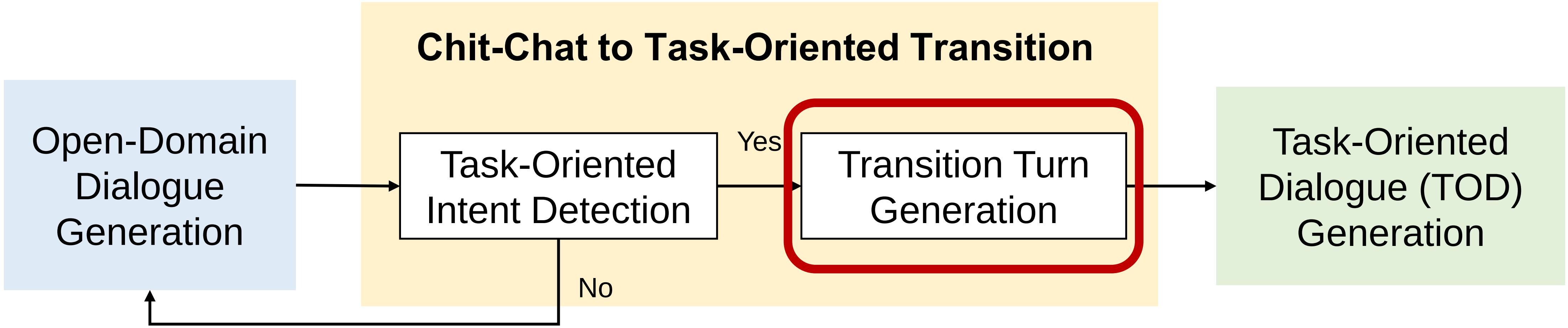
QA Model

Yes

Intent-associated questions are naively generated from their descriptions

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



- 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**

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.

Transition Turn Generation

Generative-based Generation:

- Training data: OTTers $\underbrace{\text{(Source Topic}}_{\textbf{User A}} \rightarrow \underbrace{\text{Transition} \rightarrow \text{Target Topic)}}_{\textbf{User B}}$

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

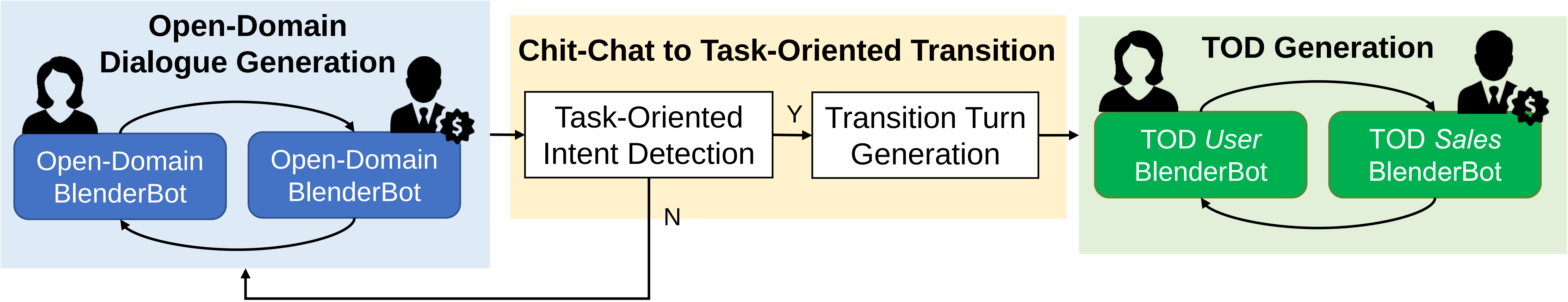
User A **Source Topic:** I like **seafood** a lot.
User B **Transition:** Since you like seafood, is **Swedish fish** a candy that you might enjoy?
Target Topic: I have no self control when it comes to **candy**.

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.

OTTers: One-turn Topic Transitions for Open-Domain Dialogue, ACL2021

SalesBot Simulation Framework



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

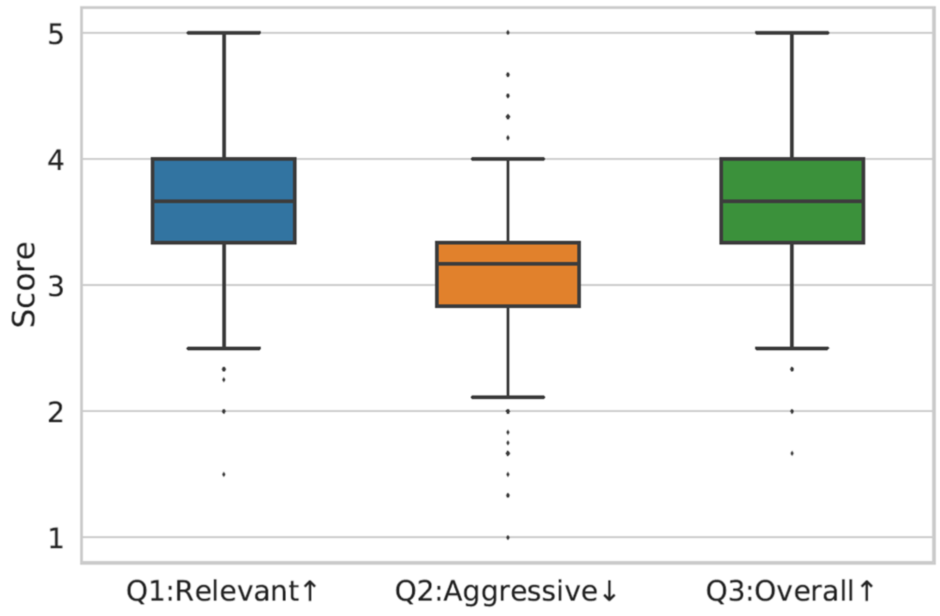
Quality?

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.

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?

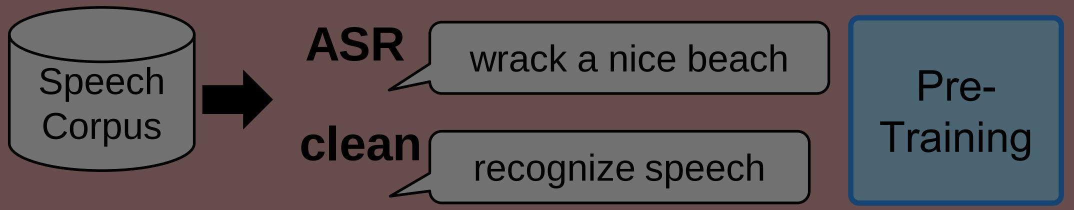


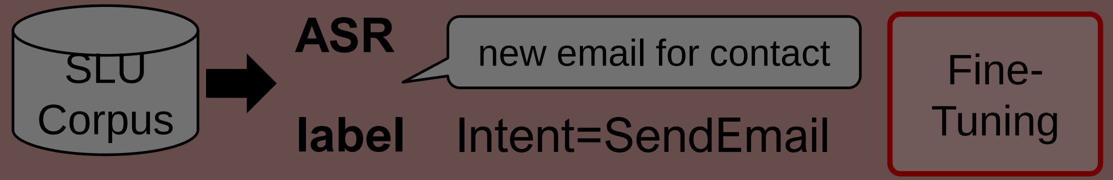
Average scores over 3 workers (4000 dialogues)

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.

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

Scalability

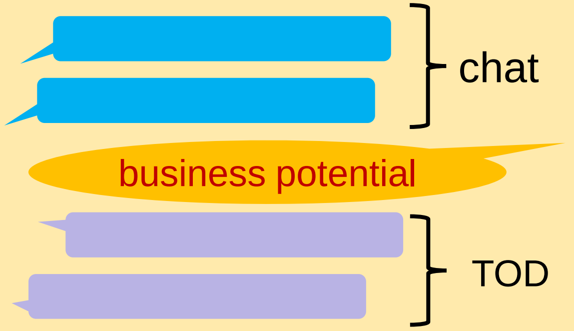
Thank you, can you also find me a hotel with parking near the restaurant?

Confirm — Domain Change — Boolean — Coreference

- ✓ **CUDA** augments DST data with diverse user behaviors
  - ○ Efficiency of augmentation
  - ○ Generalizability to different datasets/domains
  - ○ SOTA in MultiWOZ 2.1
  - ○ Potential for user simulators

Practicality

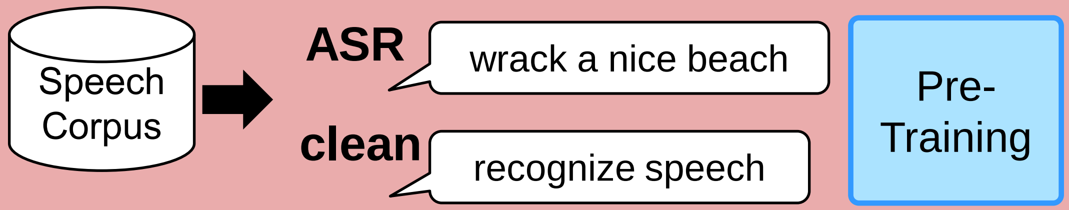
- ✓ **SalesBot** simulates unlimited data bridging two types of agents
  - ○ Reasonable quality
  - ○ Data/simulators for learning (SL/RL) end-to-end agents with sales' behavior



Robustness

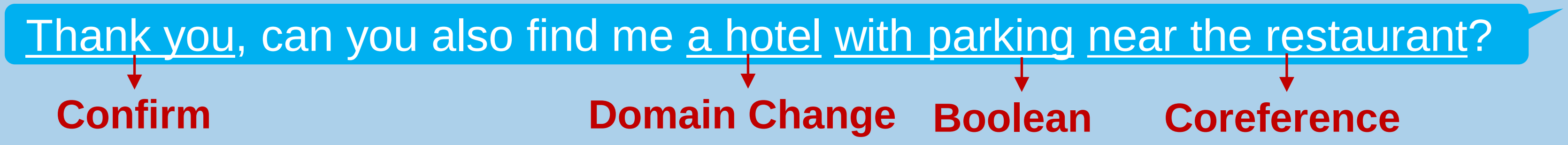
✓ **LatticeLM** for preserving uncertainty

✓ **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

Scalability



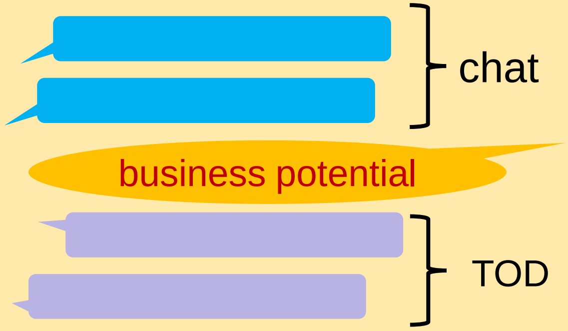
✓ **CUDA** augments DST data with diverse user behaviors

- Efficiency of augmentation
- Generalizability to different datasets/domains
- SOTA in MultiWOZ 2.1
- Potential for user simulators

Practicality

✓ **SalesBot** simulates unlimited data bridging two types of agents

- Reasonable quality
- Data/simulators for learning (SL/RL) end-to-end agents with sales' behavior



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

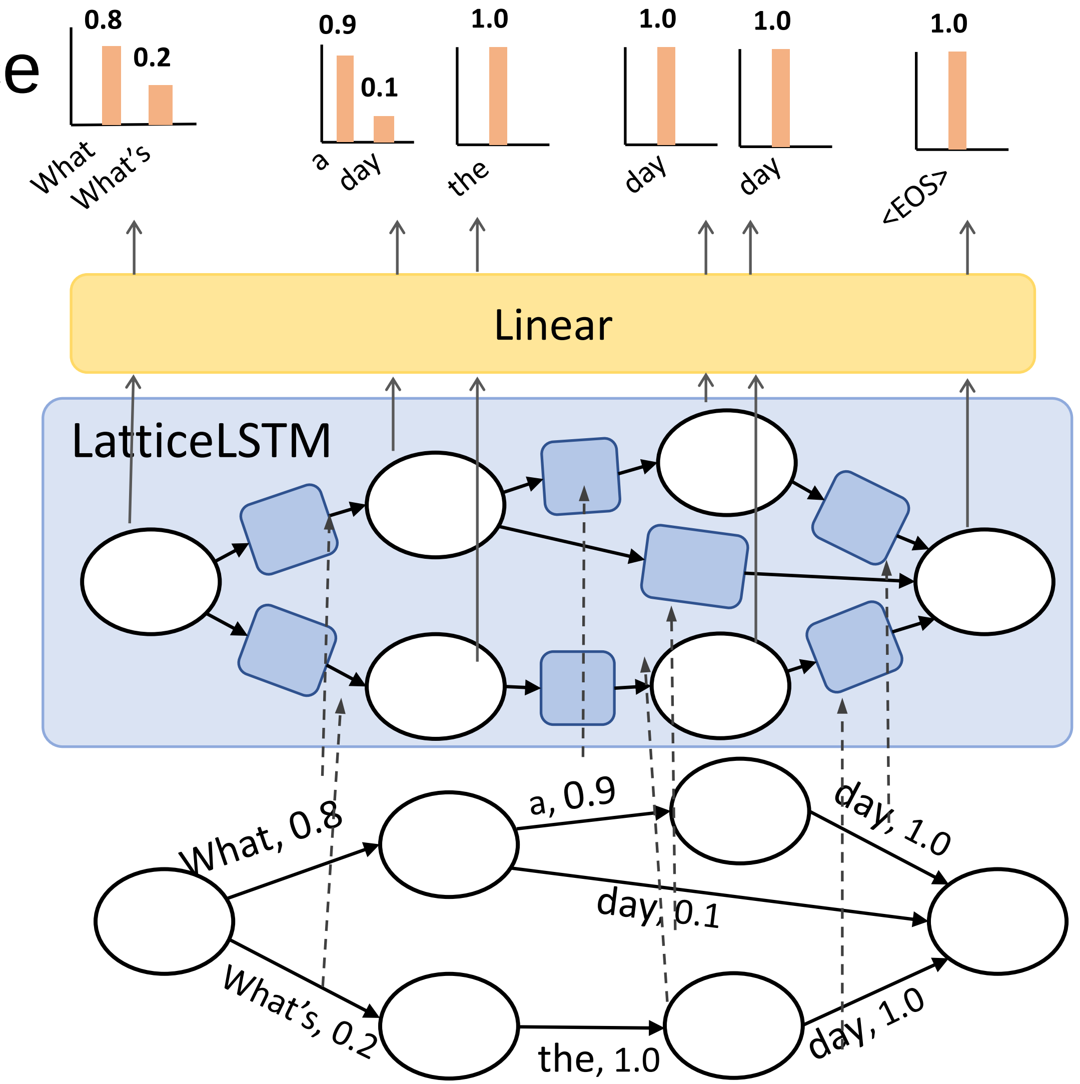
Appendix

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



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.

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

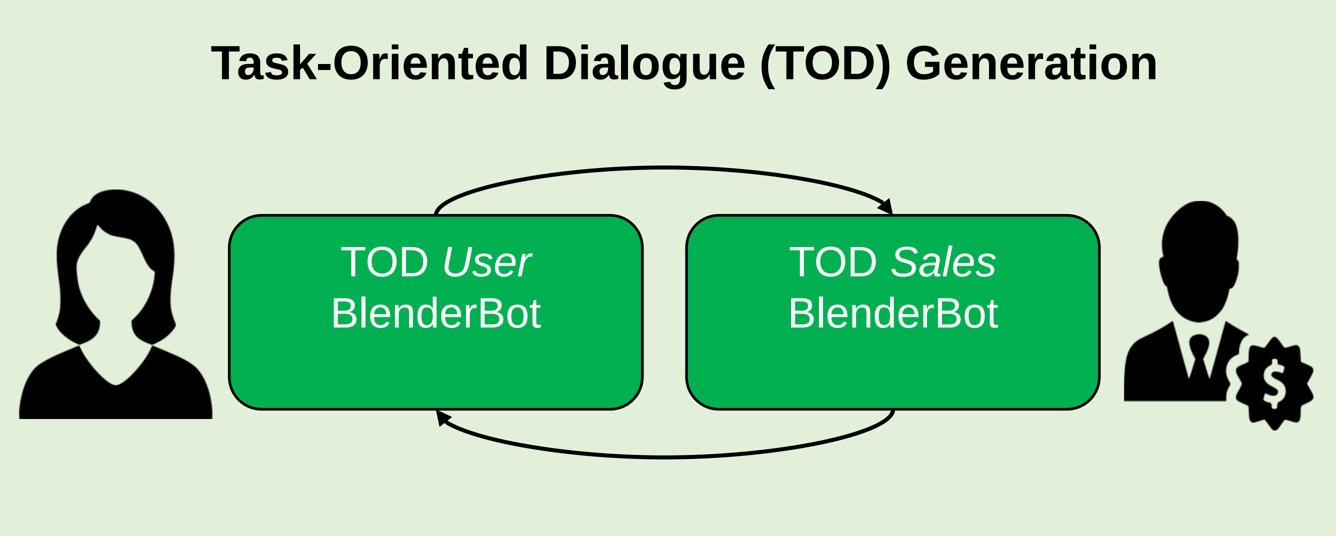
Pre-Training	Fine-Tuning	SLURP WER Interval Clean =0	Low (0, 0.16]	Medium (0.16, 0.40]	High >0.4	All
RoBERTa	Direct	**95.69**	92.41	85.89	56.71	83.97
Phoneme-BERT	Direct	94.97	92.34	85.87	57.20	83.78
SimCSE	Direct	95.55	93.47	**86.82**	57.59	84.47
Proposed	Direct	95.54	**93.86**	86.68	**57.72**	**84.51**
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:	*Do you want to find movies by genre and optionally director?*
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

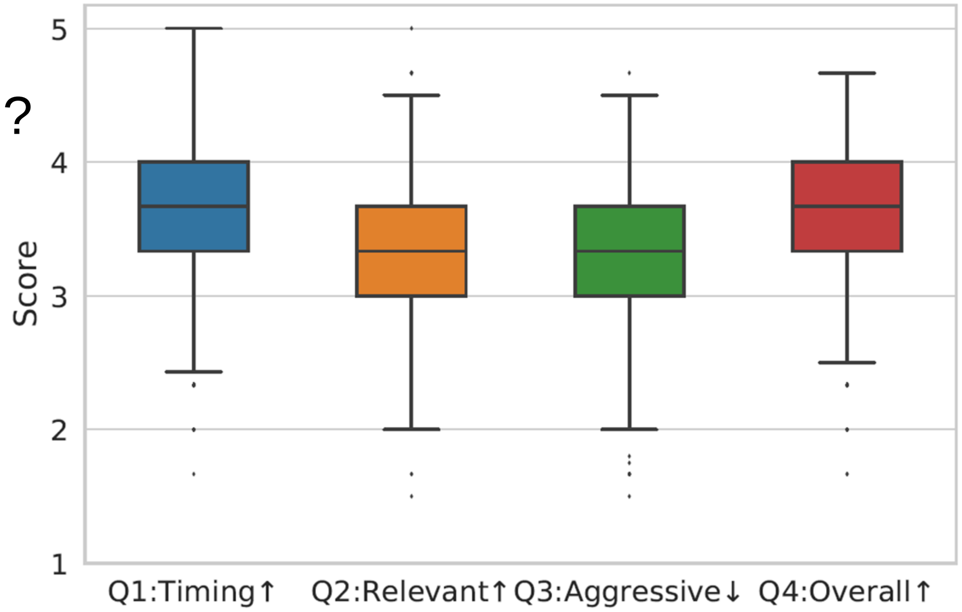
User:	I like to read a lot. I also like to go to the movies. What about yourself?
Sales:	*Are you interested in watching any movie?*
User:	I'm looking for a movie to watch. A regular showing would be fine.

Generated Transition

Human Evaluation

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