Suppressing Unnecessary Clarification Requests for Unknown Word Acquisition in Spoken Dialogue Using Syllable-Based ASR Confidence

@SIGDIAL 2026
2026/8/3

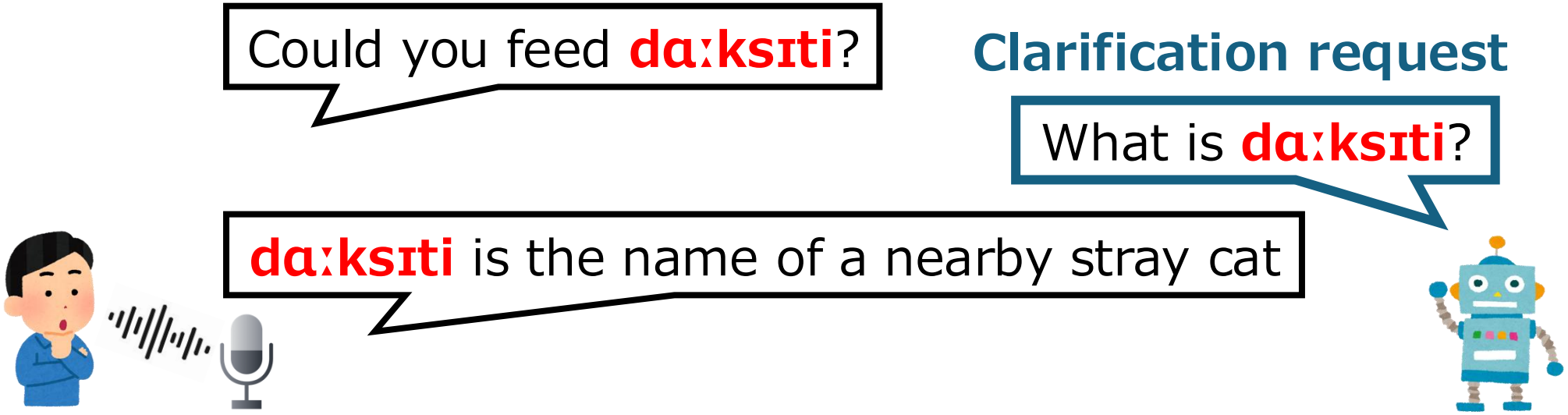
Takumi Furuta, Ryu Takeda, Kazunori Komatani
The University of Osaka, SANKEN

Background

Acquiring unknown words through clarification requests

- Direct acquisition of information from users
- Unknown words: User-coined names, local abbreviations

Doxitty (dɑːksɪti), a user-coined name for a stray cat

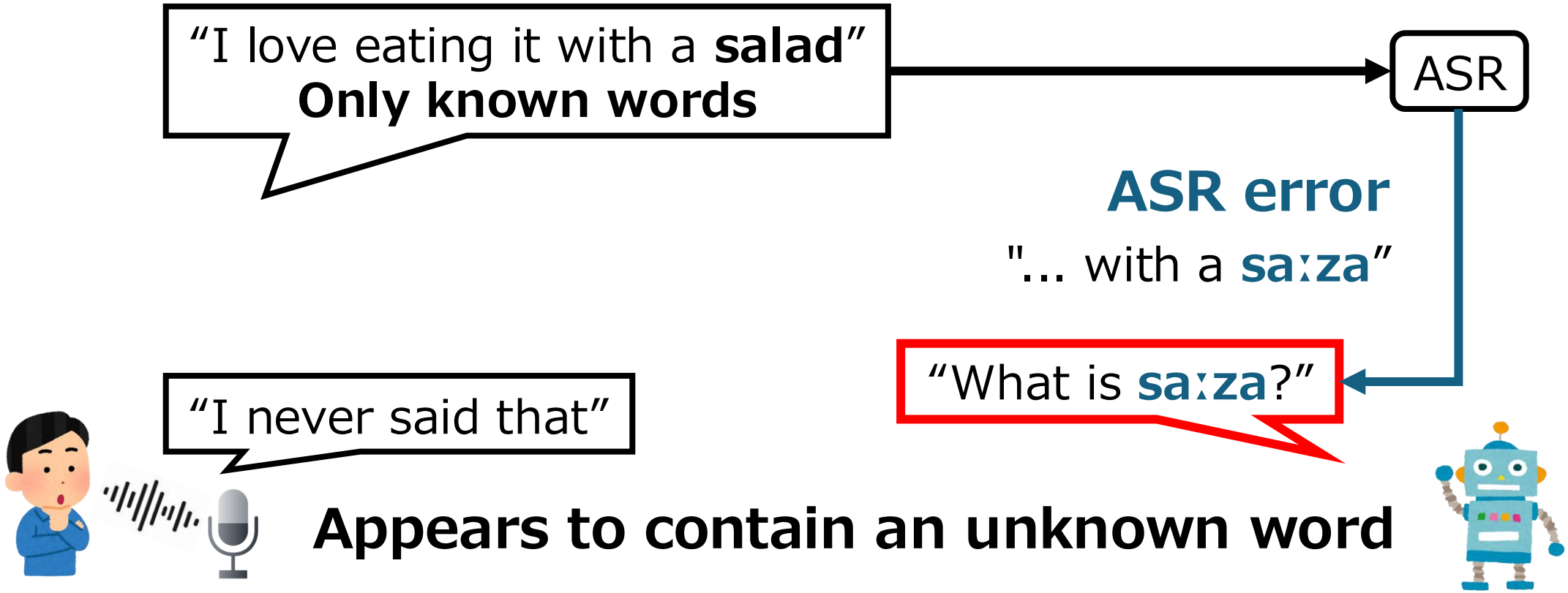


Requests only for utterances with unknown words [Waki2025]

Problem

Unnecessary clarification requests under speech input

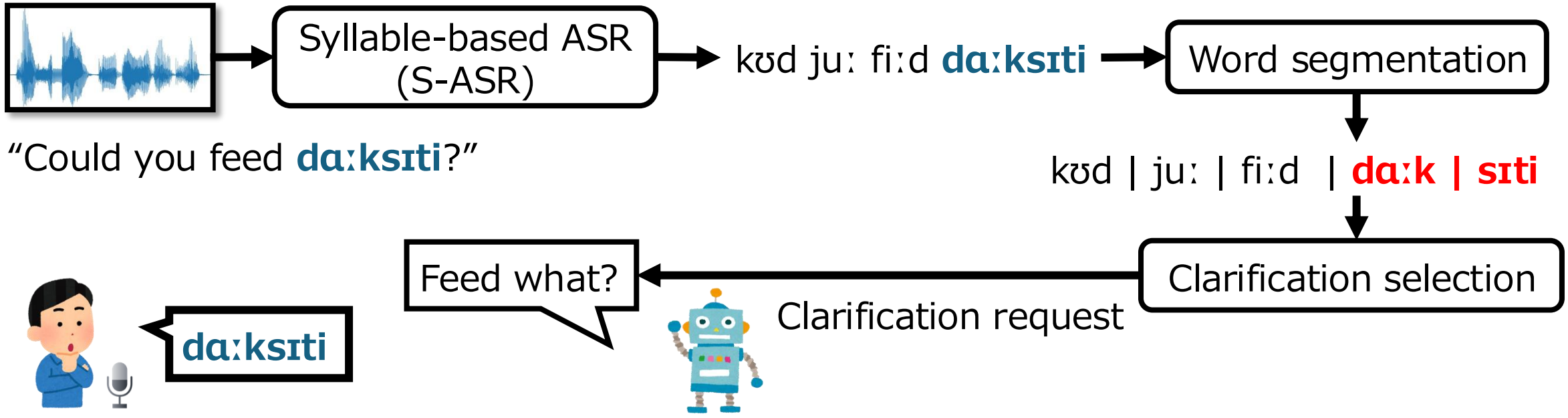
- Clarification requests burden users



Task setting

Clarification selection as stream-based active learning [Waki2025]

- Limited number of clarification requests
- Target information: correct syllable sequences + word boundaries

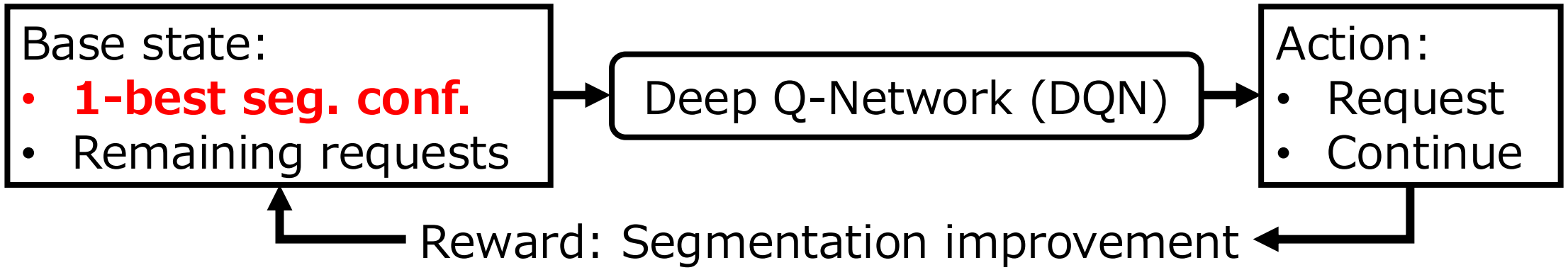


Update word segmentation with **target information**

Approach

Clarification selection using reinforcement learning (RL) [Waki2025]

- Applied to speech using the 1-best S-ASR hypothesis
- State includes 1-best segmentation confidence (1-best seg. conf.) [Lewis1995]



Low 1-best seg. conf. may indicate:

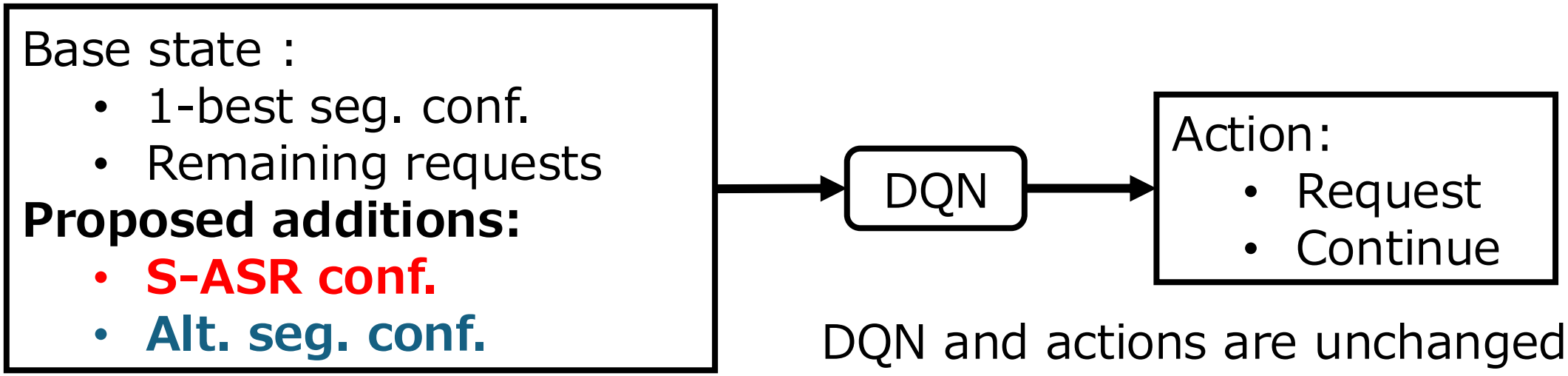
- **Unknown word** → Request
- **S-ASR error in known words** → Continue

Same low confidence, different actions

Proposed Method

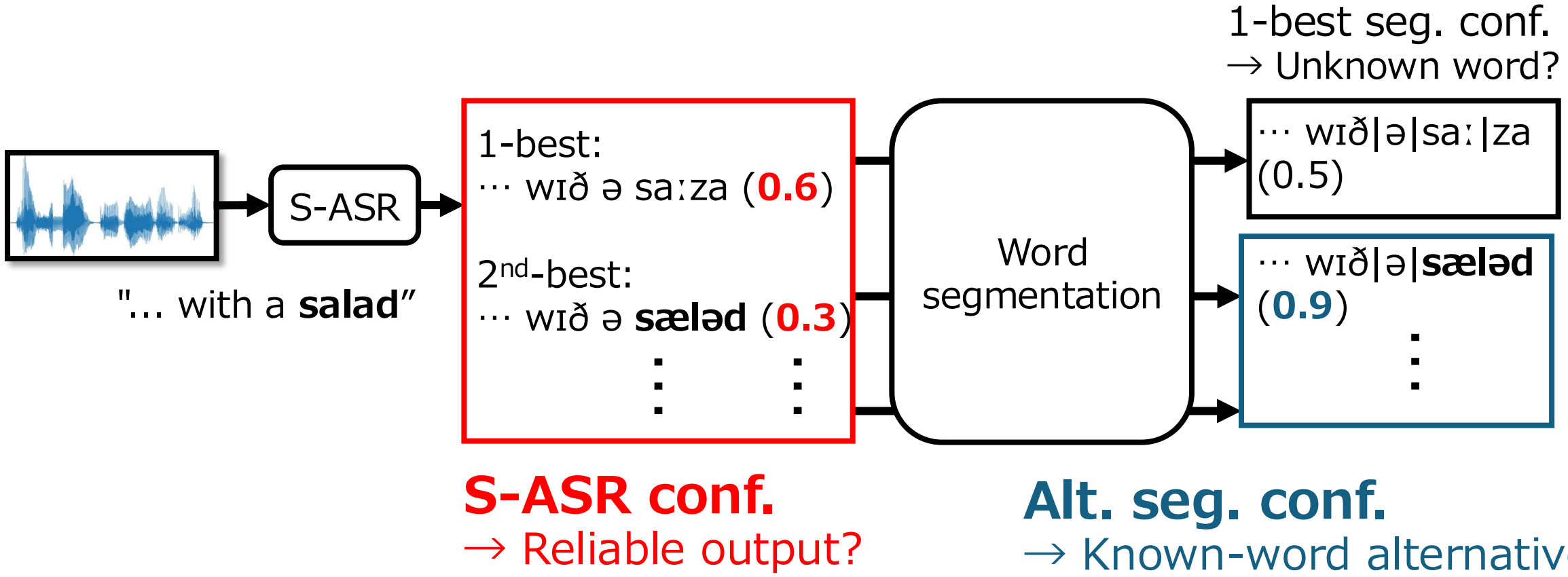
Extend RL state with two S-ASR-related confidence signals

- **S-ASR confidence (S-ASR conf.)**
  → Is the S-ASR output itself reliable?
- **Alternative segmentation confidence (Alt. seg. conf.)**
  → Can another S-ASR hypothesis be segmented into known words?



Proposed Method

Confidence cues for possible S-ASR errors



The DQN learns how to use these cues from rewards

Experiments: Objective and dataset

Objective:

Can proposed method reduce unnecessary clarification requests?

Dataset:

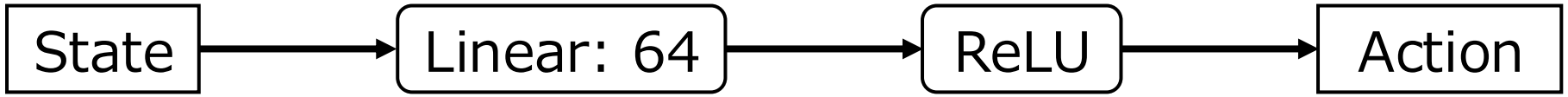
- Japanese utterances from multi-turn clarification dialogues
- **14 food names treated as unknown words**
  → Target words include Eisbein, Dal Bhat, and Semifreddo

Dialogue example:

Experiments: Setup and metrics (1/2)

Setup

- Evaluation target: DQN
- Stream data: 336 training / 261 test utterances
- Word segmentation: Nested Pitman-Yor language model [Mochihashi2009]
- S-ASR: Hybrid CTC/Attention model [Watanabe2017]



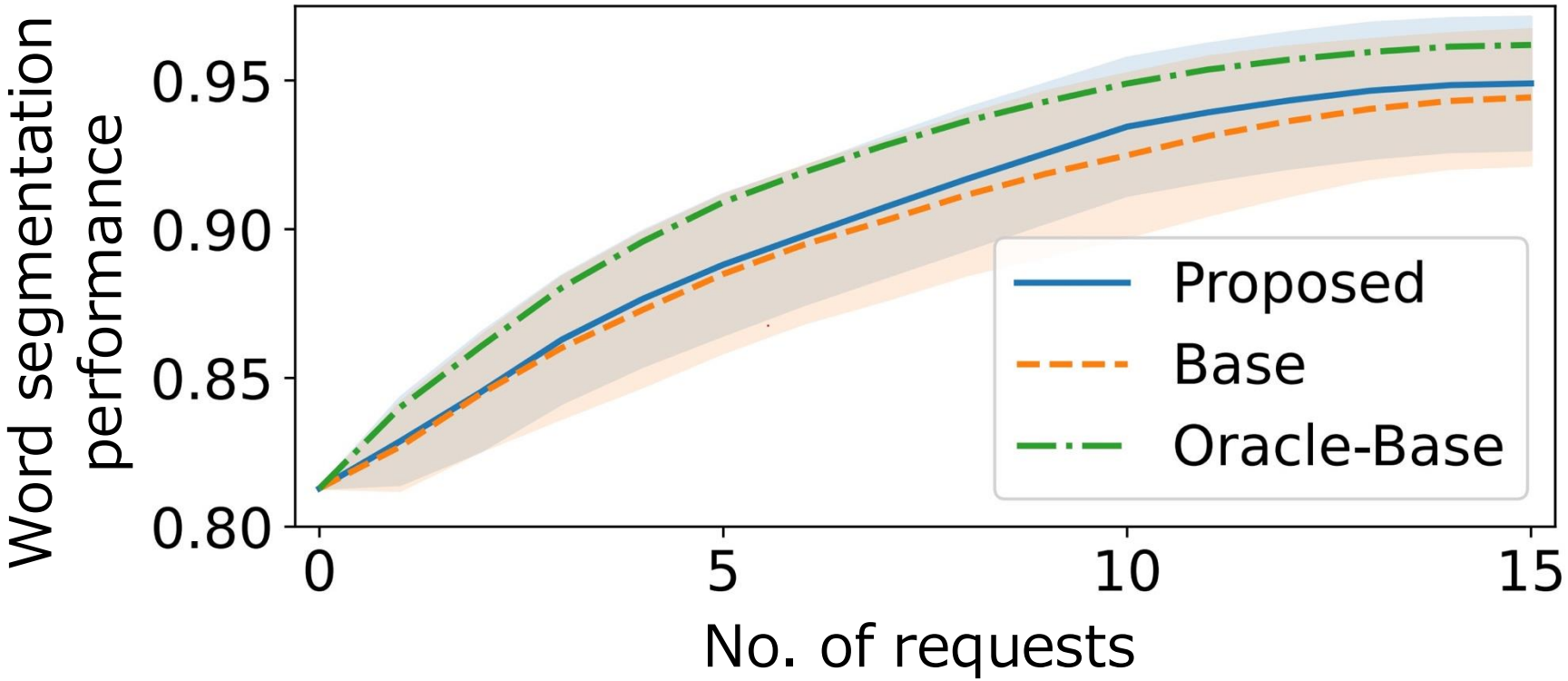
Primary Metric

- **No. of unnecessary requests (↓)**:
  The number of requests on utterances with S-ASR errors

Experiments: Metrics (2/2)

Unknown-word acquisition performance metric

- AUC (↑):
  Area under the learning curve of word segmentation



Experiments: Main results

400 runs: 20 policy trainings × 20 stream orders

*Oracle-Base: Base using reference syllable sequences without S-ASR errors

Method	No. of unnecessary requests (↓)	No. of requests (Max: 15)	AUC (↑)
Base	8.31 ± 1.34	13.38 ± 1.66	13.47 ± 0.18
Proposed	6.97 ± 1.14	12.75 ± 1.78	13.54 ± 0.11
***Oracle-Base**	–	12.37 ± 2.11	13.78 ± 0.12
	Fewer (−1.34)	Fewer (−0.63)	**Maintained (+0.07)**

Fewer unnecessary requests while maintaining performance

A gap to the oracle condition remained

Experiments: Ablation results

400 runs: 20 policy trainings × 20 stream orders

Method	No. of unnecessary requests (↓)	No. of requests (Max: 15)	AUC (↑)
Base	8.31 ± 1.34	13.38 ± 1.66	13.47 ± 0.18
+ S-ASR conf.	7.84 ± 1.54	13.34 ± 1.98	13.51 ± 0.22
+ Alt. seg. conf.	8.58 ± 1.02	13.82 ± 1.14	13.50 ± 0.08
Proposed	6.97 ± 1.14	12.75 ± 1.78	13.54 ± 0.11

Fewer (−1.34)

Lower variability

The two confidence signals played different roles

- S-ASR conf. → fewer unnecessary requests
- Alt. seg. conf. → lower variability

Summary

Purpose

Clarification requests only for utterances with unknown words

Problem

Unnecessary clarification requests caused by S-ASR errors

Approach

RL-based clarification selection for speech input

Proposed

RL state with two additional S-ASR-related confidence signals

Results

Fewer requests for utterances with S-ASR errors
Maintained unknown-word acquisition performance

References

[Waki2025] Issei Waki et al.
Learning to Ask Efficiently in Dialogue: Reinforcement Learning Extensions for Stream-based Active Learning. In Proc. SIGDIAL, pp. 431–440, 2025.

[Lewis1995] David D. Lewis.
A Sequential Algorithm for Training Text Classifiers: Corrigendum and Additional Data. ACM SIGIR Forum, vol. 29, no. 2, pp. 13–19, 1995.

[Watanabe2017] Shinji Watanabe et al.
Hybrid CTC/Attention Architecture for End-to-End Speech Recognition.
IEEE JSTSP, vol. 11, no. 8, pp. 1240–1253, 2017.

[Mochihashi2009] Daichi Mochihashi et al.
Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling.In Proc. ACL-IJCNLP, pp. 100–108, 2009.