Rethinking Binary Evaluation of Turn-Taking under Inherent Ambiguity

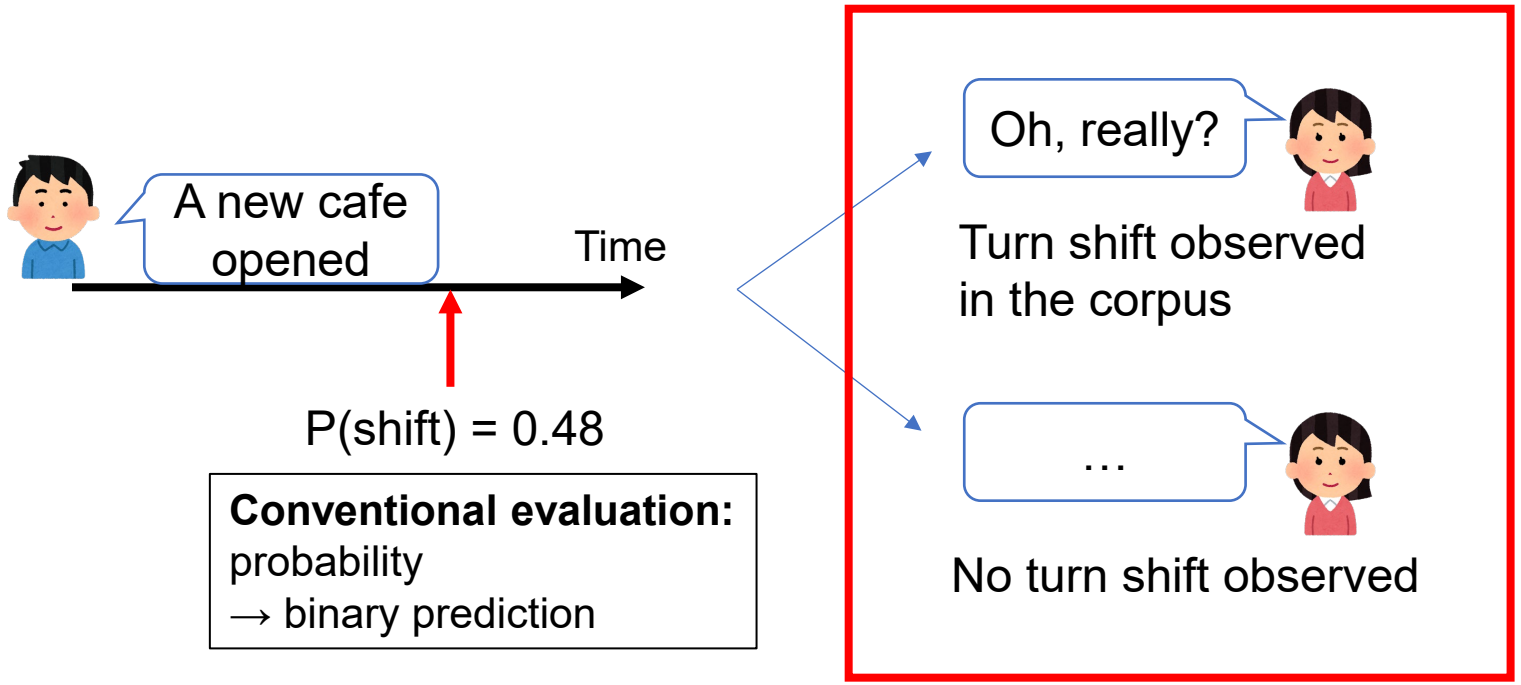
SIGDIAL2026

August 3, 2026

Yunosuke Kubo, Kenta Yamamoto, Ryu Takeda,
○**Kazunori Komatani**
(University of Osaka)

Background:
Turn-taking is inherently ambiguous

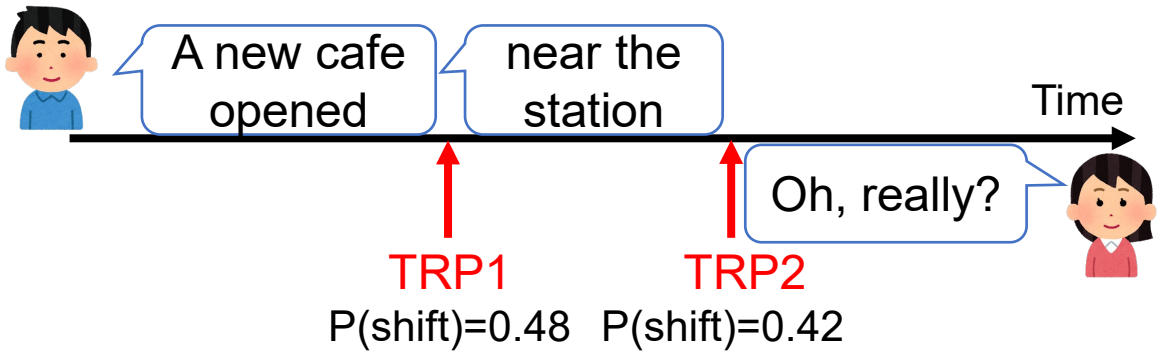
- Turn-taking prediction models output turn-shift probabilities



Was the corpus-observed outcome the only reasonable outcome?

A corpus observation is one realized outcome

- The corpus records what happened in that interaction
  - Other outcomes may also be reasonable in the same context

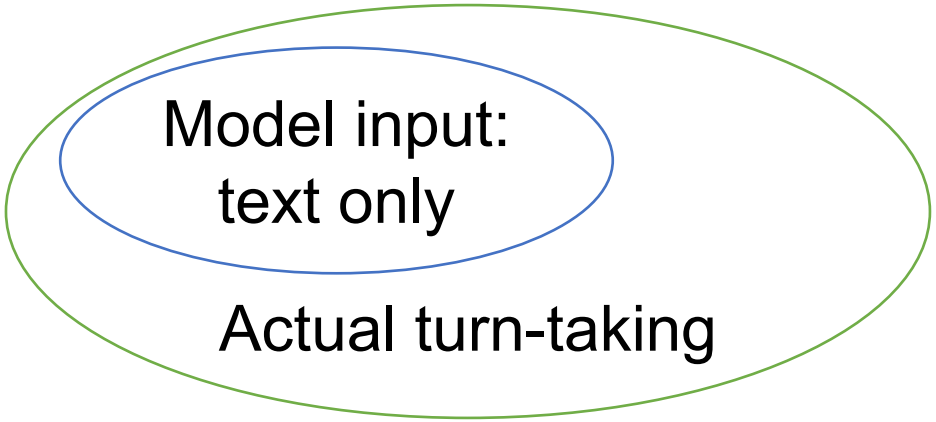


TRP: Transition Relevance Place

Yet conventional evaluation treats the realized outcome as a binary reference

The model observes only part of the evidence

- Actual turn-taking also depends on prosody, timing, overlap, and interruption
- Text alone may not uniquely determine whether the turn should shift or continue



A text-only model should be confident only when **the text clearly supports its prediction**

Evaluation:
- *Neutral* for uncertain cases

Training data:
- Diagnose observations that are difficult to explain from text alone

Our approach

SANKEN
THE UNIVERSITY OF OSAKA

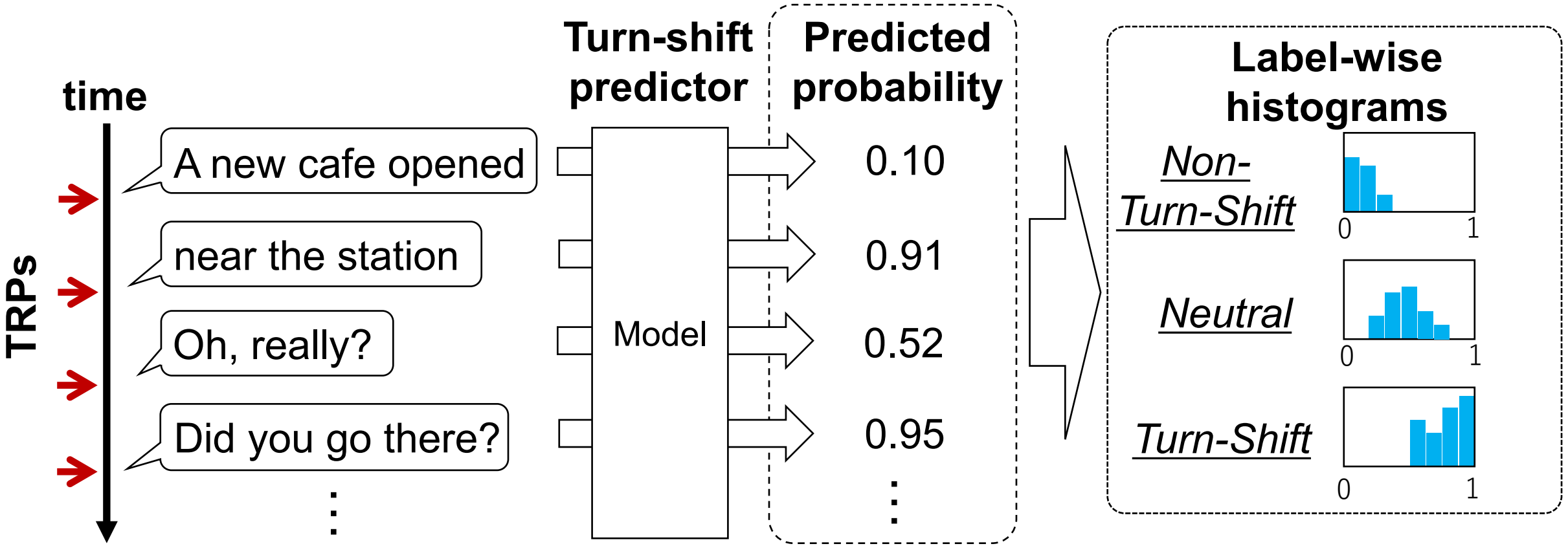
We address this issue from two perspectives:

① **Evaluation**
Binary reference → ambiguity-aware reference with *Neutral*
- Evaluate predicted probabilities at the distribution level

② **Training data**
Prediction-corpus discrepancy → refinement candidate
- Use discrepancies between model predictions and corpus observations as diagnostic signals for training-data refinement

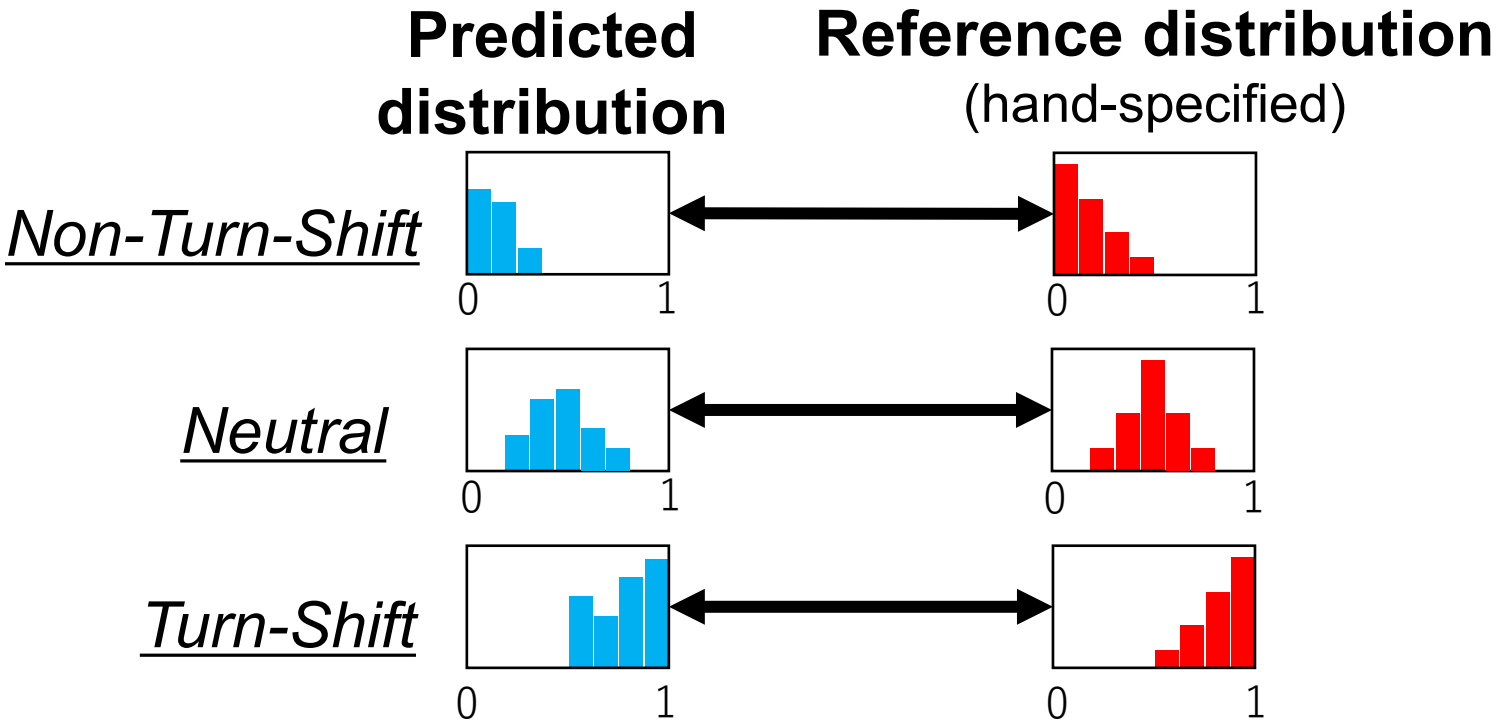
① Evaluation approach

- Ambiguity-aware labels: *NTS* / *Neutral* / *TS*
  - Manually annotated on a test set: 500 turns



Evaluation:
Reference distributions and Wasserstein distance

- Compare predicted distributions with reference distributions
  - using Wasserstein distance (optimal transport)
- Smaller distance indicates closer alignment



② Data refinement as diagnostic exploration

- Identify refinement candidates from prediction–corpus discrepancies

Remove candidate instances
or relabel FN candidates: *TS* → *NTS*

Candidate definitions

	Corpus observation	**Initial prediction (text only)**
FN (False Negative)	*Turn-Shift*	Low shift prob.
FP (False Positive)	*Non-Turn-Shift*	High shift prob.

- Conditions:
  - FN Removal / FP Removal / FN+FP Removal / FN Relabeling

Experimental setup

SANKEN
THE UNIVERSITY OF OSAKA

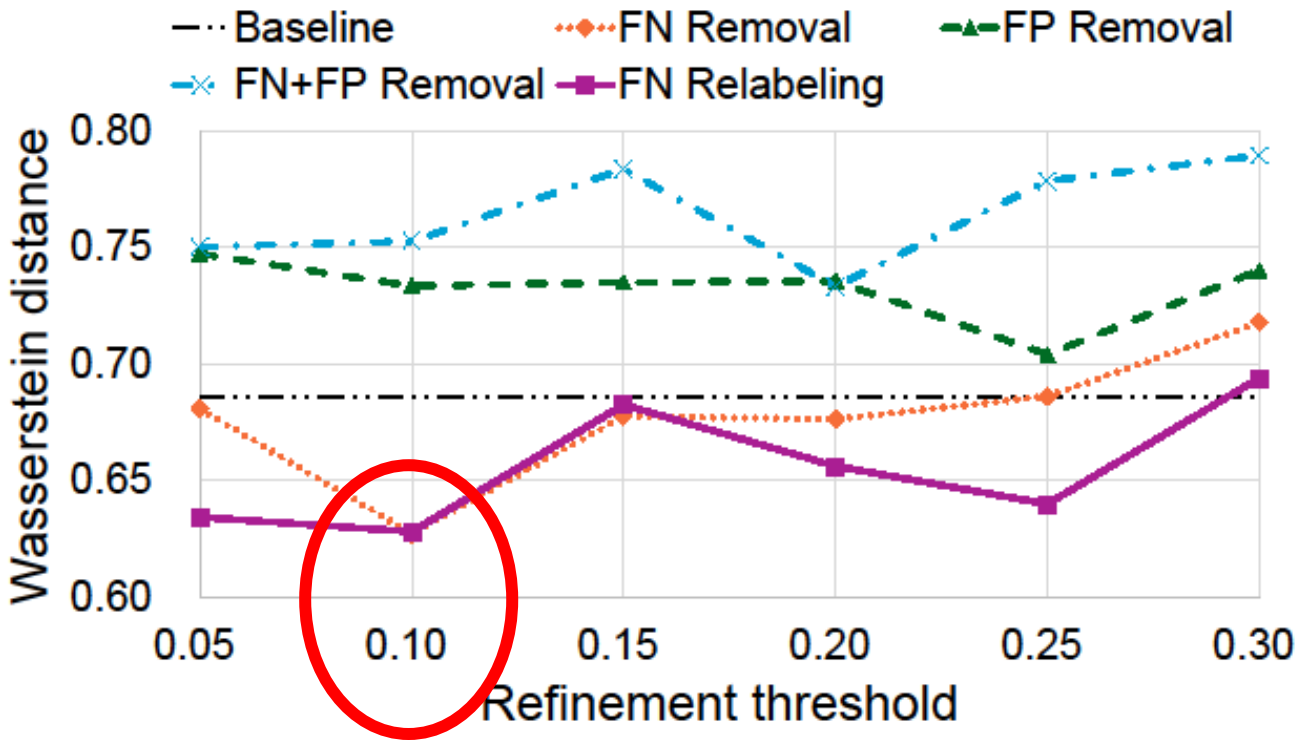
Case study:

Use the proposed evaluation to analyze how data refinement changes predicted distributions

- Dataset:
  - Corpus of Everyday Japanese Conversation (CEJC) [Koiso+ 2022]
    - 577 dialogues (~200 hours total)
  - Train: 520 dialogues / Valid: 29 dialogues
  - Evaluation set: 500 manually annotated turns (*TS* 249 / *Neutral* 163 / *NTS* 88)
- Input: text only
- Model:
  - TurnGPT-style predictor [Ekstedt and Skantze, 2020]
  - Fine-tuned from rinna/japanese-gpt-1b

Main result

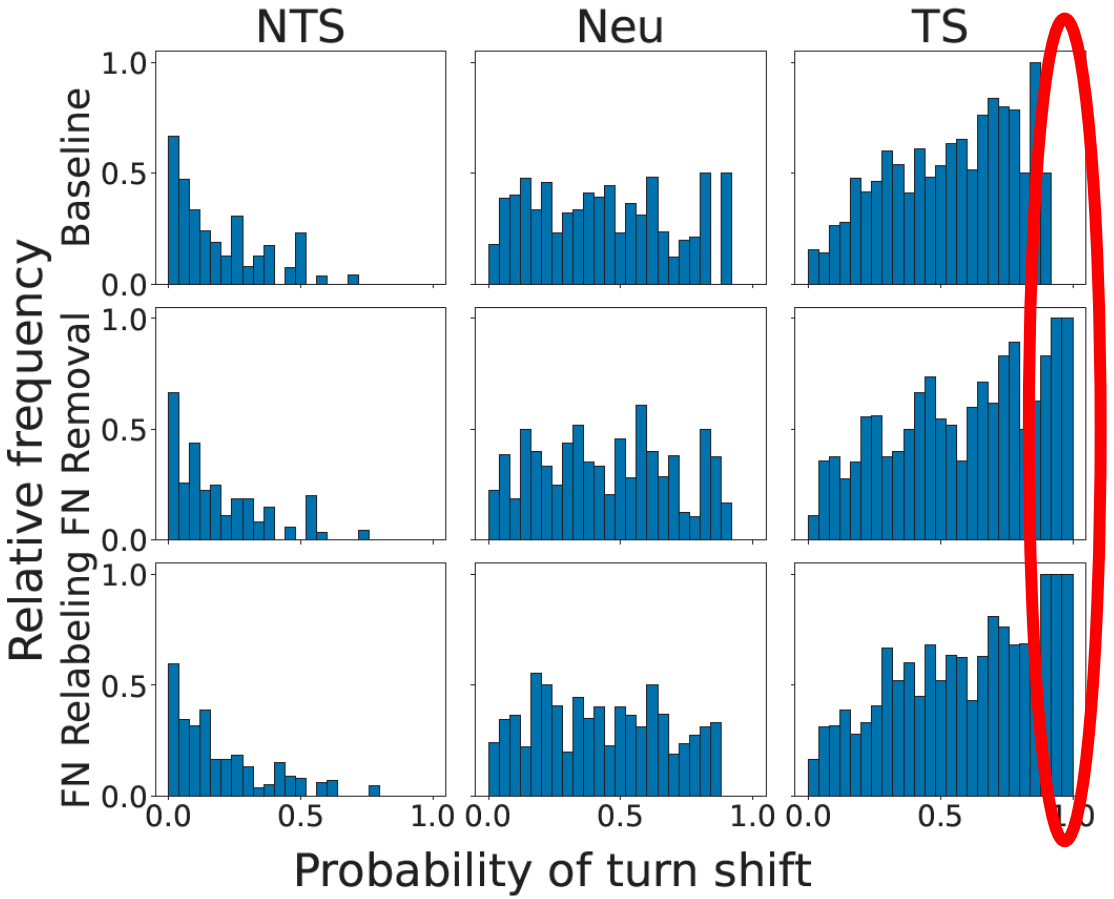
- FN-based refinement improves distribution-based evaluation
  - reducing Wasserstein distance



Where does it improve?

Distribution-based evaluation reveals where the change occurs

- The largest improvement is for *Turn-Shift*



Why might FN-based refinement help?

- FN-based refinement produced the largest improvement

	Corpus observation	**Initial prediction (text only)**
FN (False Negative)	*Turn-Shift*	Low shift prob.

- Some FN cases may be only weakly supported by the text alone
  - e.g., prosody, timing, overlap, or interruption

- Reducing their influence may make the textual patterns of the remaining *Turn-Shift* cases more consistent

Conventional binary metric shows the same trend

SANKEN
THE UNIVERSITY OF OSAKA

- FN-based refinement also achieved slightly higher balanced accuracy than the baseline

Model	Balanced Acc.
Baseline	0.777
FN Removal	0.788
FN Relabeling	0.799

  - This is a supplementary consistency check, not the main evaluation

Take-home:
observed outcome ≠ uniquely correct outcome

SANKEN
THE UNIVERSITY OF OSAKA

1. Corpus observations are realized outcomes, not absolute binary targets
2. A text-only model should be confident only when the text clearly support its prediction
3. Distribution-based evaluation captures ambiguity and reveals where probabilistic behavior changes

Limitations

- Single-author annotation / hand-specified reference distributions
- Text-only input, Japanese corpus, and a single model
- No downstream dialogue-system evaluation