Open-Domain Dialogue Quality Evaluation: Deriving Nugget-level Scores from Turn-level Scores

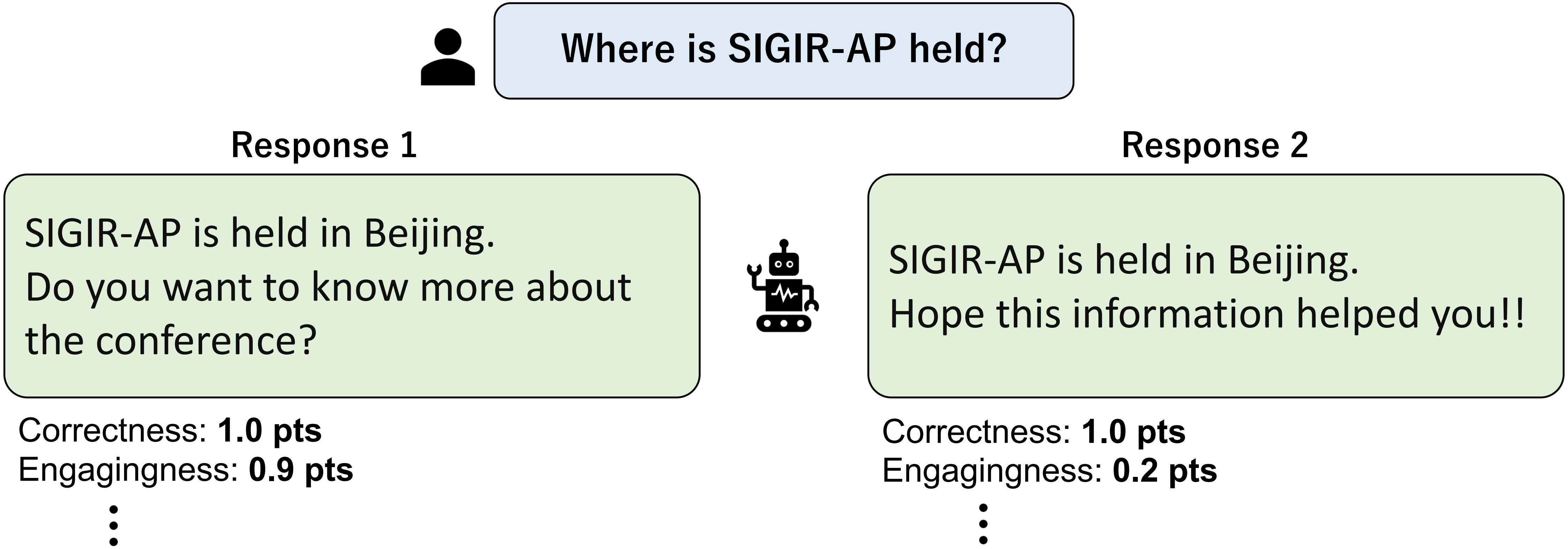
Rikiya Takehi, **Akihisa Watanabe**, Tetsuya Sakai

Waseda University

Evaluation of Dialogue Systems

Dialogue systems are evaluated on multiple metrics called **dialogue qualities**.

Examples of dialogue qualities: engagingness, correctness, fluency, interestingness, …



Where is SIGIR-AP held?

Response 1

SIGIR-AP is held in Beijing.
Do you want to know more about the conference?

Correctness: **1.0 pts**
Engagingness: **0.9 pts**
⋮

Response 2

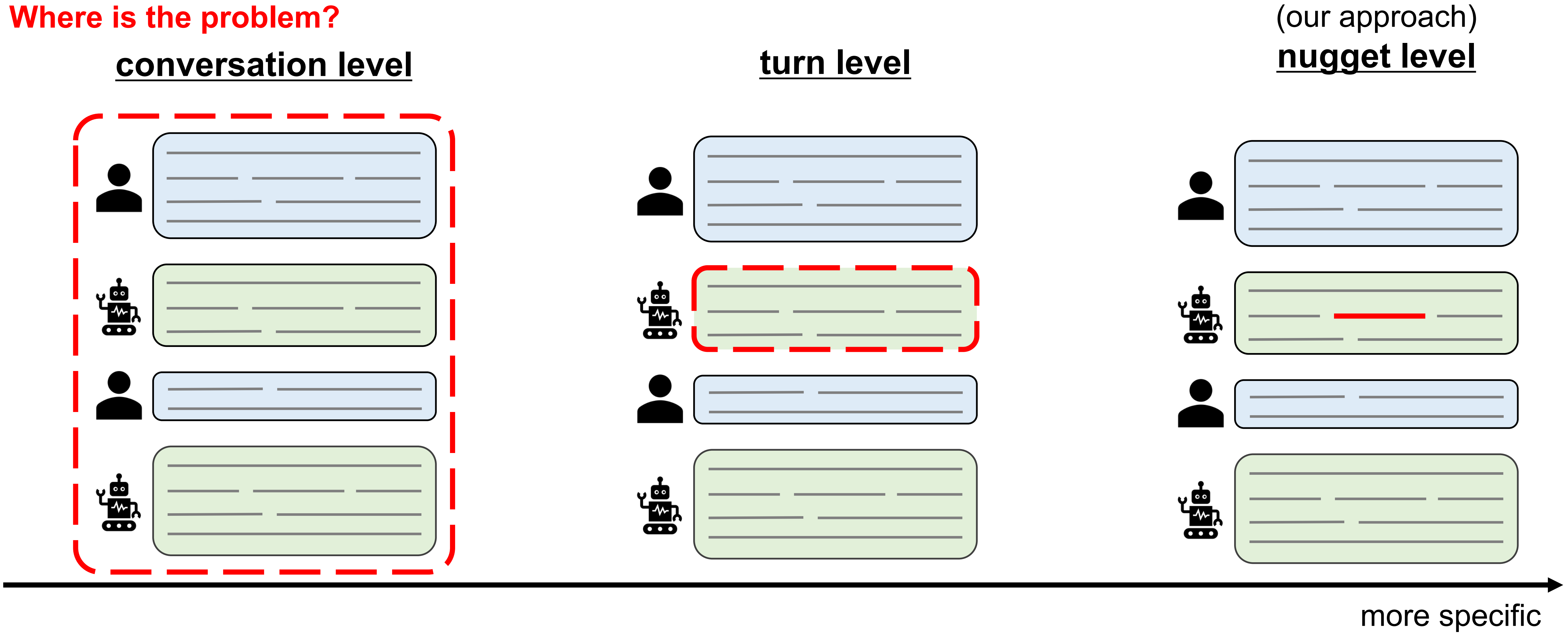
SIGIR-AP is held in Beijing.
Hope this information helped you!!

Correctness: **1.0 pts**
Engagingness: **0.2 pts**
⋮

Evaluation of Dialogue Systems

Dialogue qualities are measured at different levels.

Where is the problem?



What are nuggets?

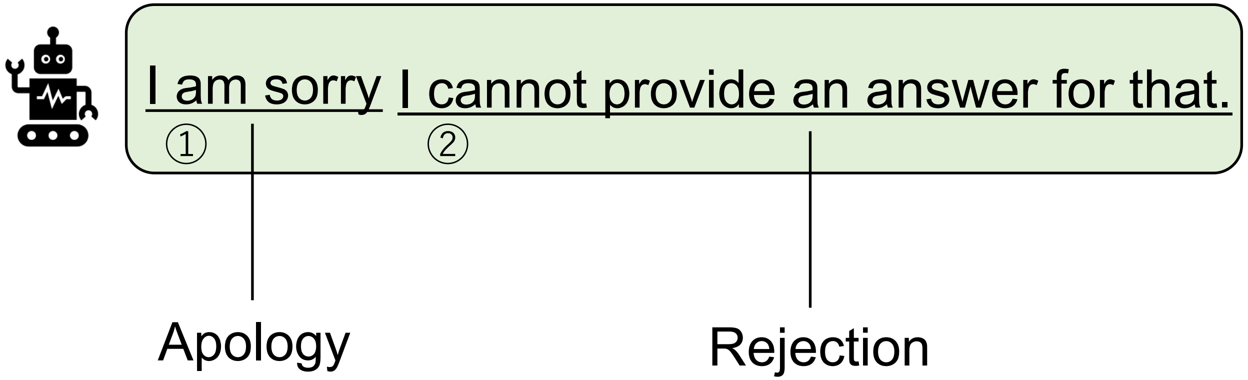
A 'nugget' is the smallest unit within a turn that represents a specific dialogue act.

A dialogue act represents the intention behind a speaker's utterance.

Example

nugget ① (apology) ：I'm sorry.

nugget ② (rejection)：I cannot provide an answer for that.



I am sorry I cannot provide an answer for that.

① ②

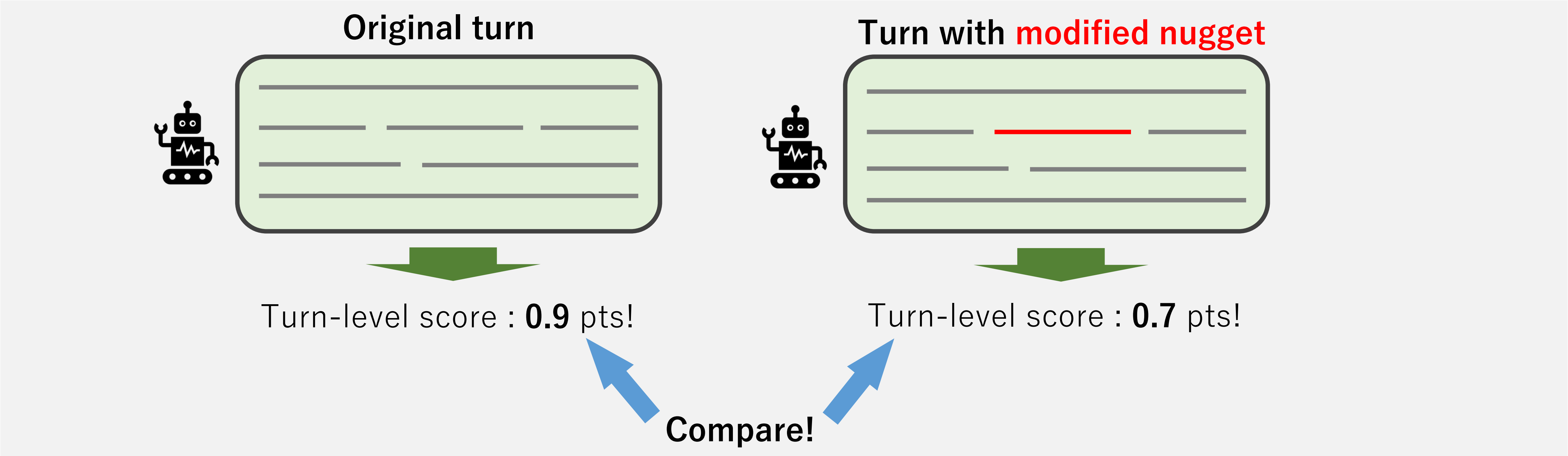
Apology Rejection

Dialogue Act	Example
Agreement	I agree
Disagreement	I disagree
Yes Answer	Yes, you are correct
No Answer	No, that is wrong
Opening	Hello
Closing	It was nice talking with you.
Apology	I am sorry
Thanking	Thank you
Rejection	I cannot provide an answer.
Applause	Well done.
Declarative Question	What do you mean by ...?
Confusion	I don't understand
Reasoning	This is because ...
Downplayer	That's all right.
Assumption	I assume you meant ...
Acknowledgment	Ok.
Clarification	The pdf you provided me is ....
Non-Declarative Question	Isn't it exciting?
User instruction	Please click on ....
Recommendation	I would recommend....
Citation	According to ...
Example	For example, ...
Commissive	I am happy to help ...
Opinion	I think ...

Andreas Stolcke, Klaus Ries and et al. Dialogue act modeling for automatic tagging and recognition of conversational speech.2000

Proposal Method

Our method evaluates the dialogue quality of a nugget by **implementing a change on the nugget** and **see how the turn-level score changes from the original score**.



Original turn

Turn with modified nugget

Turn-level score : **0.9** pts!

Turn-level score : **0.7** pts!

Compare!

- ➢ What **modifications** should be done??

(1)Deletion

(2)Substitution with nuggets of *different* dialogue act

(3)Substitution with nuggets of *same* dialogue act nuggets

Proposal Method: Nugget Operation

- **Deletion:**

  Q. Is this nugget necessary?

- **Substitution with Different Dialogue Act Nuggets (Diff):**

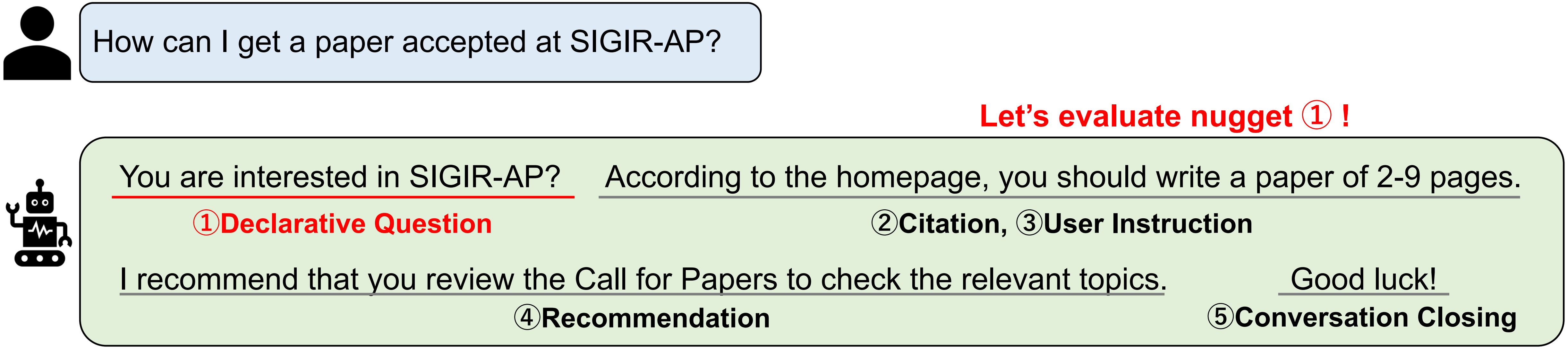
  Q. Is this the appropriate dialogue act?

- **Substitution with Same Dialogue Act Nuggets (Same):**

  Q. Is this the most suitable expression within the dialogue act?

Proposal Method

We will show an example of how our method works using the following example conversation.



How can I get a paper accepted at SIGIR-AP?

Let’s evaluate nugget ① !

You are interested in SIGIR-AP? According to the homepage, you should write a paper of 2-9 pages.

①Declarative Question

②Citation, ③User Instruction

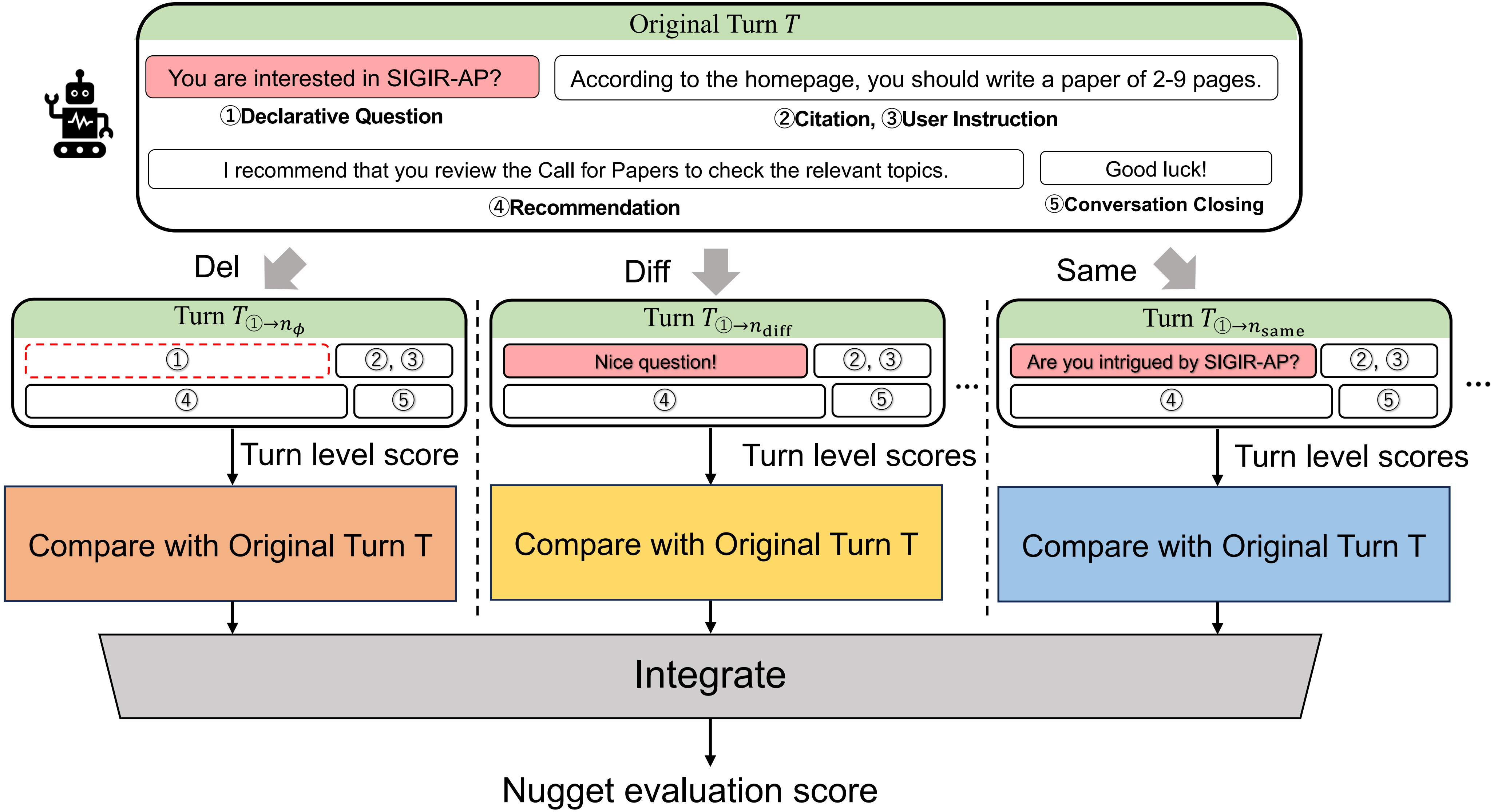
I recommend that you review the Call for Papers to check the relevant topics. Good luck!

④Recommendation

⑤Conversation Closing

Proposal Method:

Identify the score changes between the original turn-level score and those after each nugget modification.



Original Turn T

You are interested in SIGIR-AP?

①Declarative Question

According to the homepage, you should write a paper of 2-9 pages.

②Citation, ③User Instruction

I recommend that you review the Call for Papers to check the relevant topics.

④Recommendation

Good luck!

⑤Conversation Closing

Del

Turn $T_{\text{①}\rightarrow n_{\phi}}$

① | ②, ③ | ④ | ⑤

Turn level score

Compare with Original Turn T

Diff

Turn $T_{\text{①}\rightarrow n_{\text{diff}}}$

Nice question! | ②, ③ | ④ | ⑤

Turn level scores

Compare with Original Turn T

…

Same

Turn $T_{\text{①}\rightarrow n_{\text{same}}}$

Are you intrigued by SIGIR-AP? | ②, ③ | ④ | ⑤

Turn level scores

Compare with Original Turn T

…

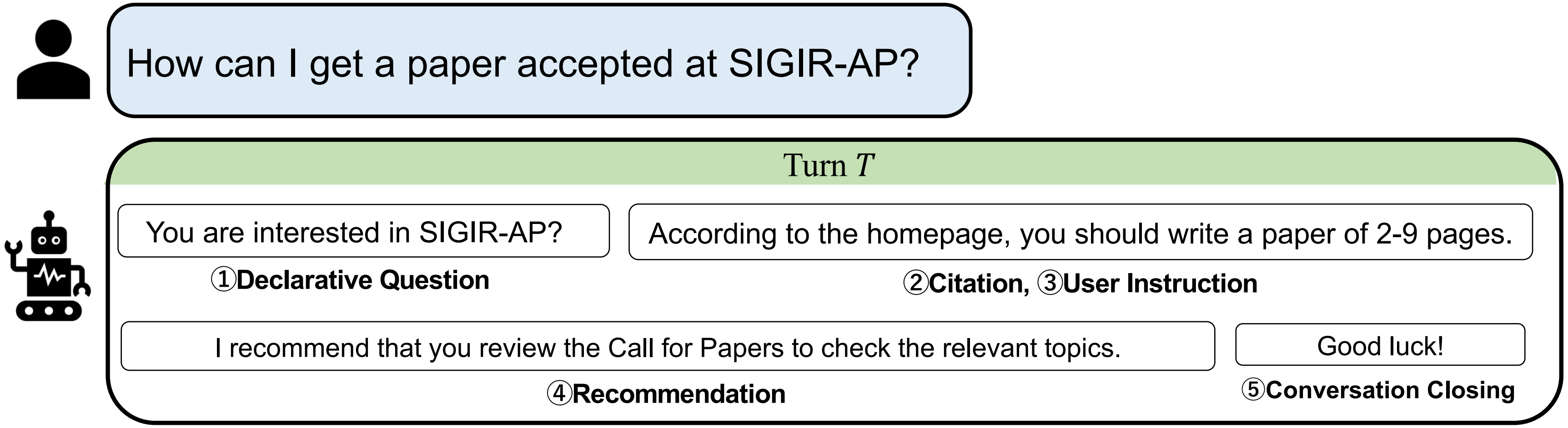
Integrate

Nugget evaluation score

Case Study: Measuring Engagingness

We evaluated dialogue engagingness using our proposed method.

Turn-Level Scoring Framework: We used EnDex framework to calculate turn-level score.



How can I get a paper accepted at SIGIR-AP?

Turn T

You are interested in SIGIR-AP?

①**Declarative Question**

According to the homepage, you should write a paper of 2-9 pages.

②**Citation,** ③**User Instruction**

I recommend that you review the Call for Papers to check the relevant topics.

④**Recommendation**

Good luck!

⑤**Conversation Closing**

Nugget level score $NS(T, n)$ score of the five nuggets in the turn.

Nugget	$NS(T, n)$
①	0.6231
②	0.6294
③	0.3229
④	0.7599
⑤	0.3892

Summary and Contributions

➢ Methodology Overview:

- We proposed a method for nugget-level evaluation of open-domain dialogue quality.

- Analyzes impact on turn-level scores after modifying nuggets.

➢ Key Contributions:

- **First to Evaluate at Nugget Level:**

  Our research is the first to evaluate open-domain dialogue systems at the nugget level.

- **Nugget-Level Score Derivation:**

  Developed a method to derive nugget-level dialogue quality scores from turn-level evaluations.

- **Dialogue Act Taxonomy:**

  Introduced a new taxonomy for dialogue acts, enhancing analysis structure.