Conversational Response Re-ranking Based on Event Causality and Role Factored Tensor Event Embedding

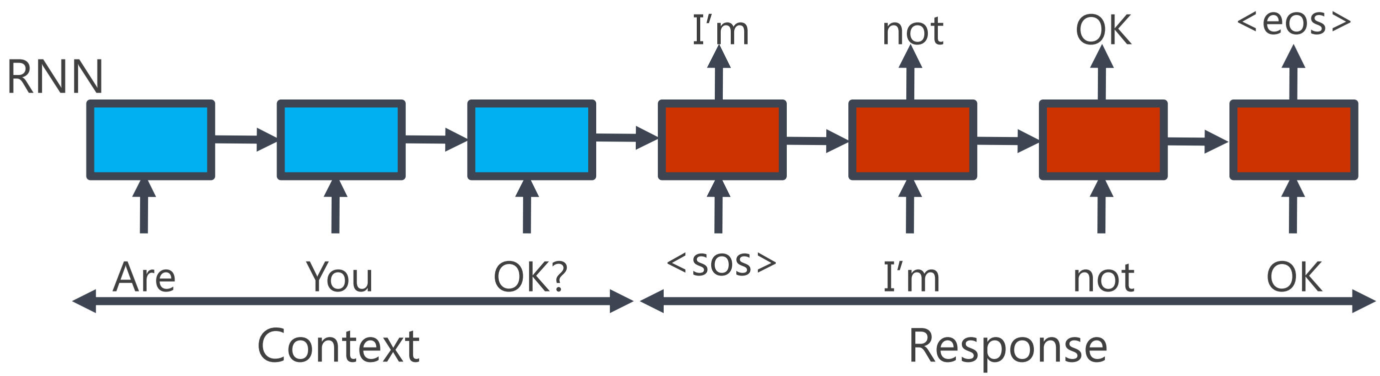
Shohei Tanaka[1], Koichiro Yoshino[1,2], Katsuhito Sudoh[1], Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology

[2]PRESTO, Japan Science and Technology Agency

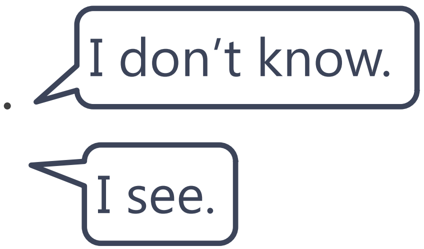
Introduction

Neural Conversational Model (NCM)



NCM [Vinyals et al., 2015] can generate responses flexibly.

Often generates simple and dull responses.

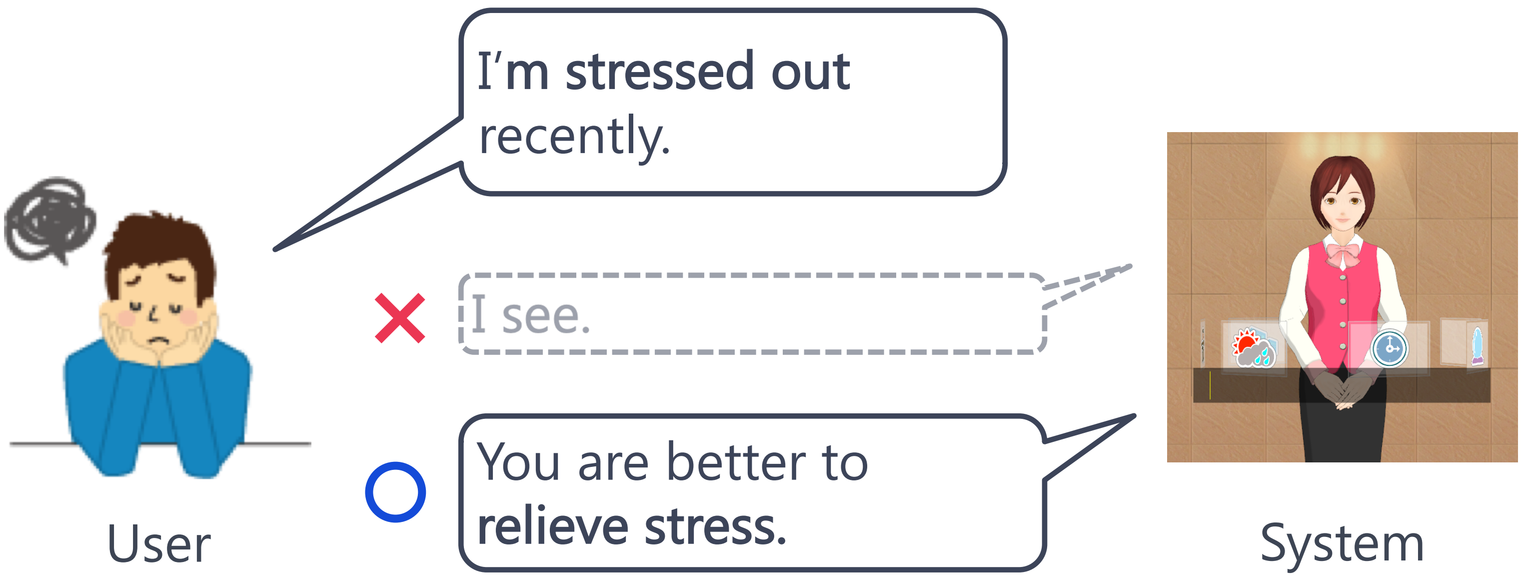


Users lose interest and finish dialogues.

Needs to maintain response coherency and diversity to continue dialogues.

Selecting Response Based on Event Causality

Re-ranks response candidates generated from NCM based on event causality.



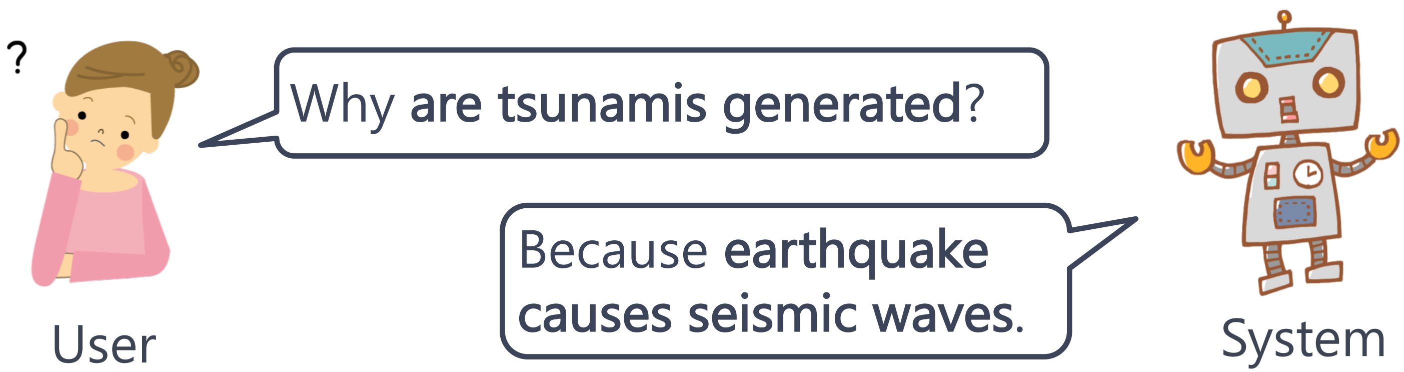
Selects a response with an event causality ("be stressed out" -> "relieve stress") related to the dialogue history.

What is Event Causality?

Cause-effect relation between two events

e.g. be stressed out (cause) -> relieve stress (effect)

Used in why-QA system [Oh et al., 2013].



Generates an answer related to the question based on a causality (“earthquake causes seismic waves” -> “tsunamis are generated”).

Why is Event Causality Useful?

Selects a conversational response based on causality.

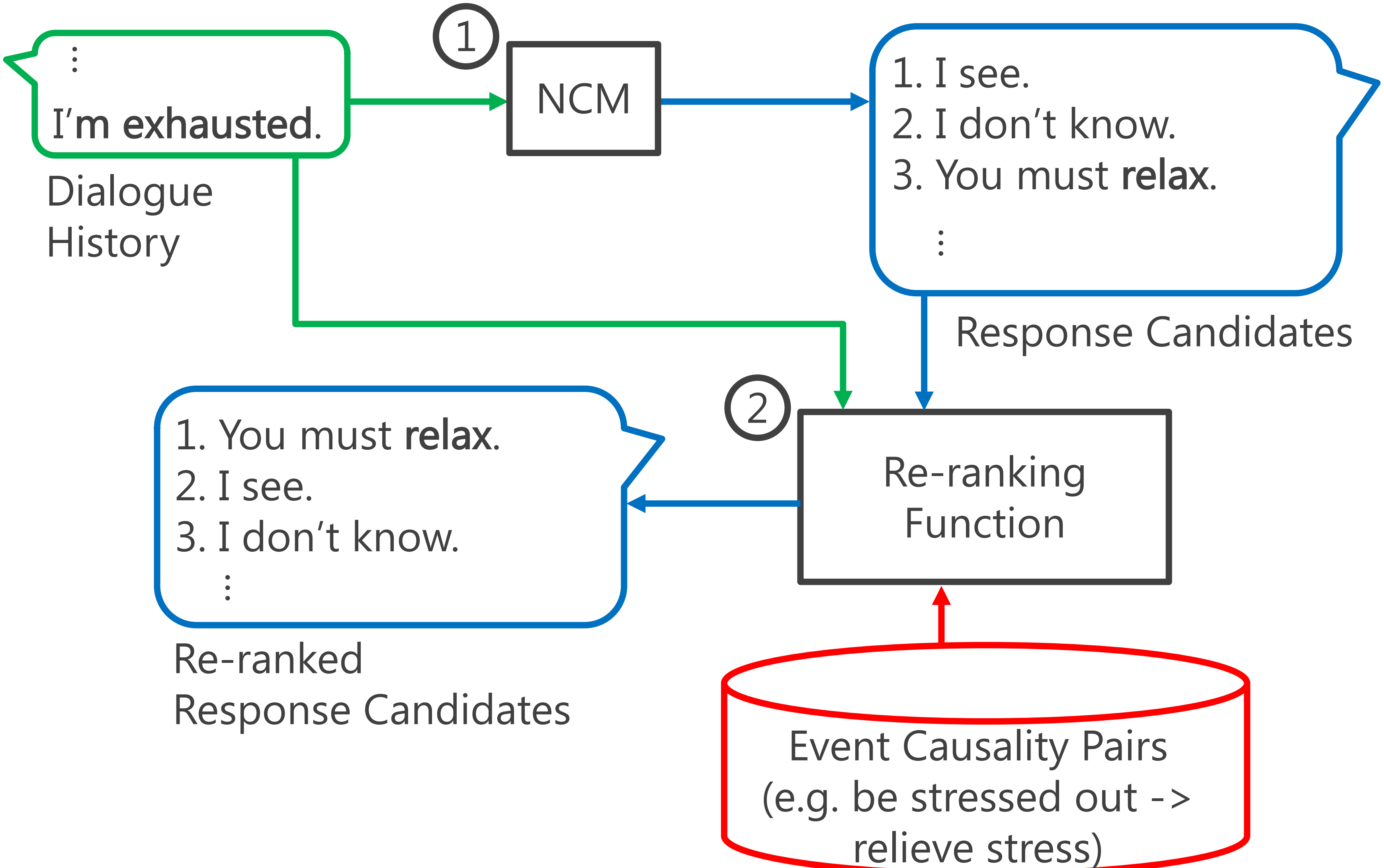
Event in the response is related to its dialogue history.
-> Coherency will be improved.

Response has a high mutual information.
-> Diversity will be improved.

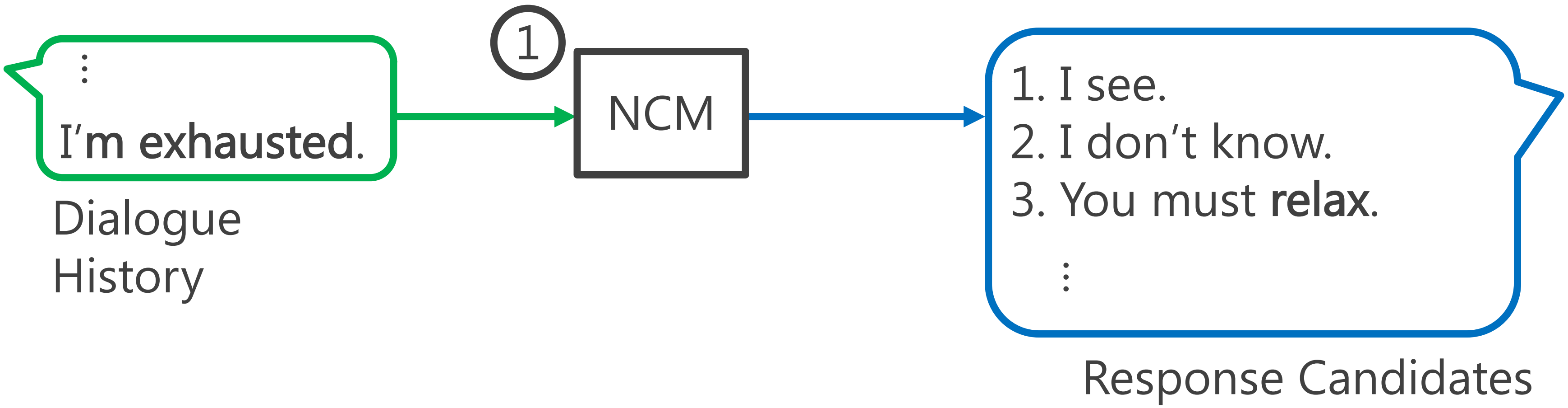
Dialogue continuity will be improved.

Response Re-ranking Using Event Causality Relations

Overview of Re-ranking



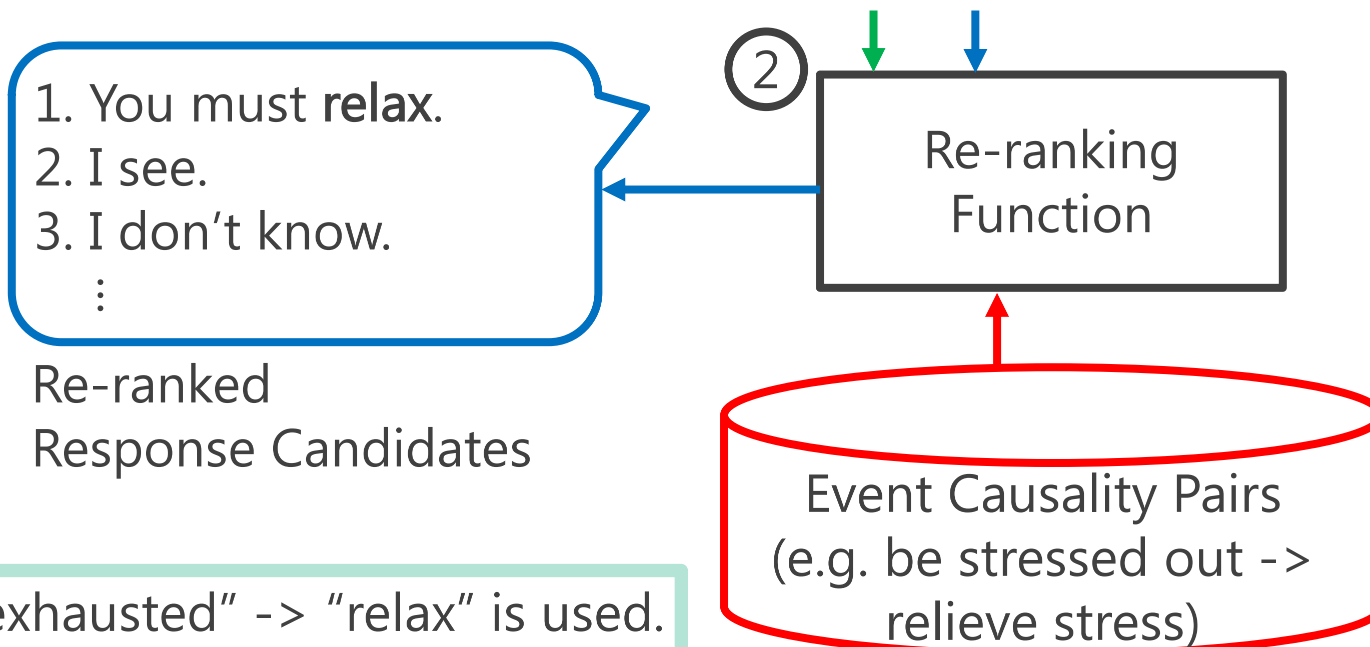
Response Candidates Generation



Generates response candidates from a dialogue history.

Re-ranking Based on Event Causality

Gives higher scores to response candidates that have event causality relations to the dialogue history.



"be exhausted" -> "relax" is used.

Event Causality Pairs

Each event consists of a predicate and arguments.

Predicate: **required**, Argument: **optional**

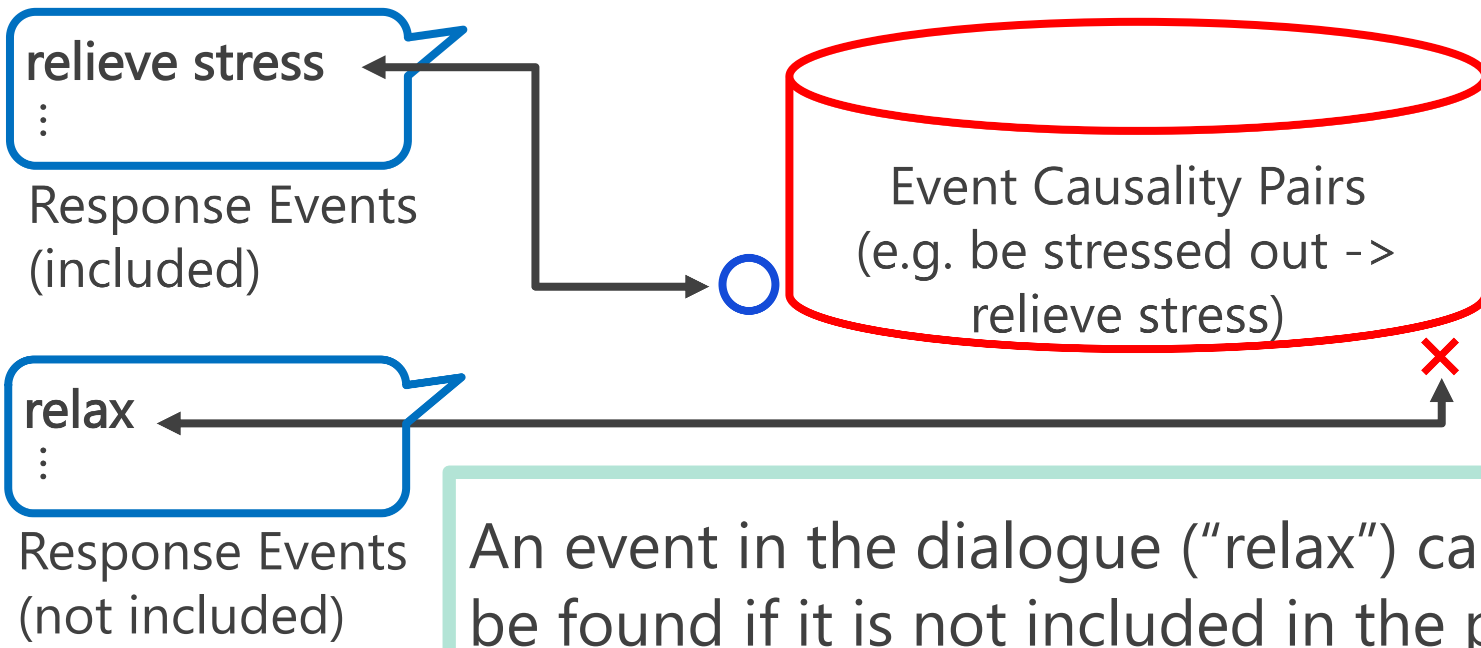
Event Causality			
Cause Event		Effect Event	
Predicate	Arguments	Predicate	Arguments
be stressed out	-	relieve	stress

Uses event causality pairs to find causalities between a dialogue history and response candidates.

Coverage Problem of Event Causality Pairs

Event causality pairs do not include all causalities in dialogue

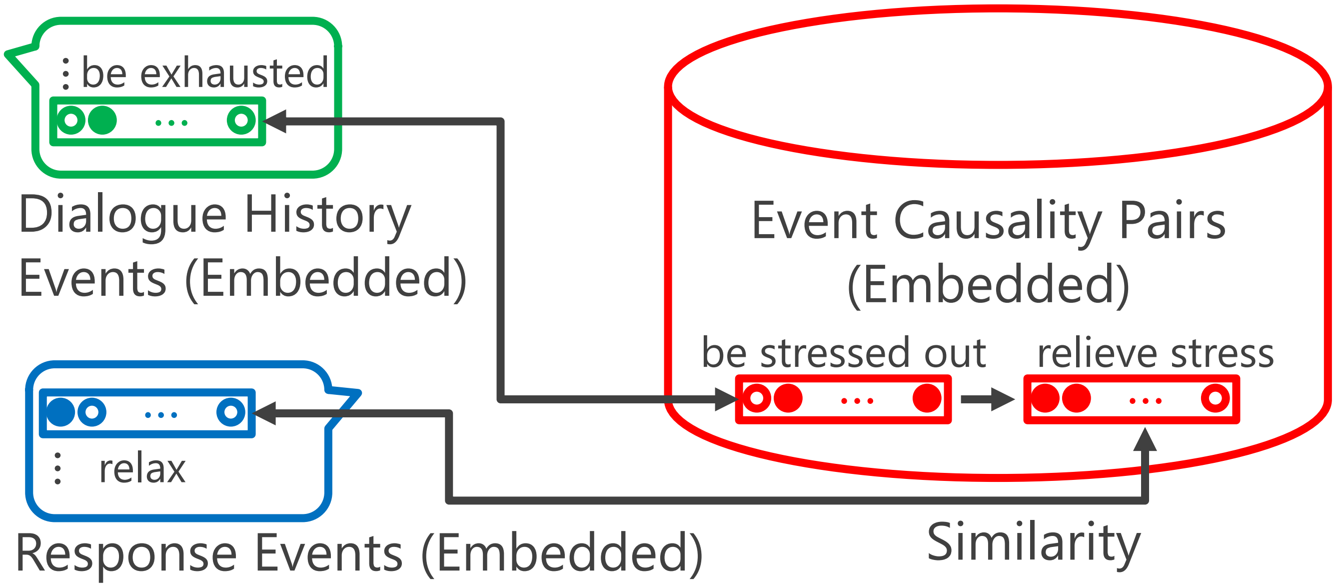
because they are obtained from limited Web corpus.



An event in the dialogue ("relax") can not be found if it is not included in the pairs.

Matching Based on Event Embedding

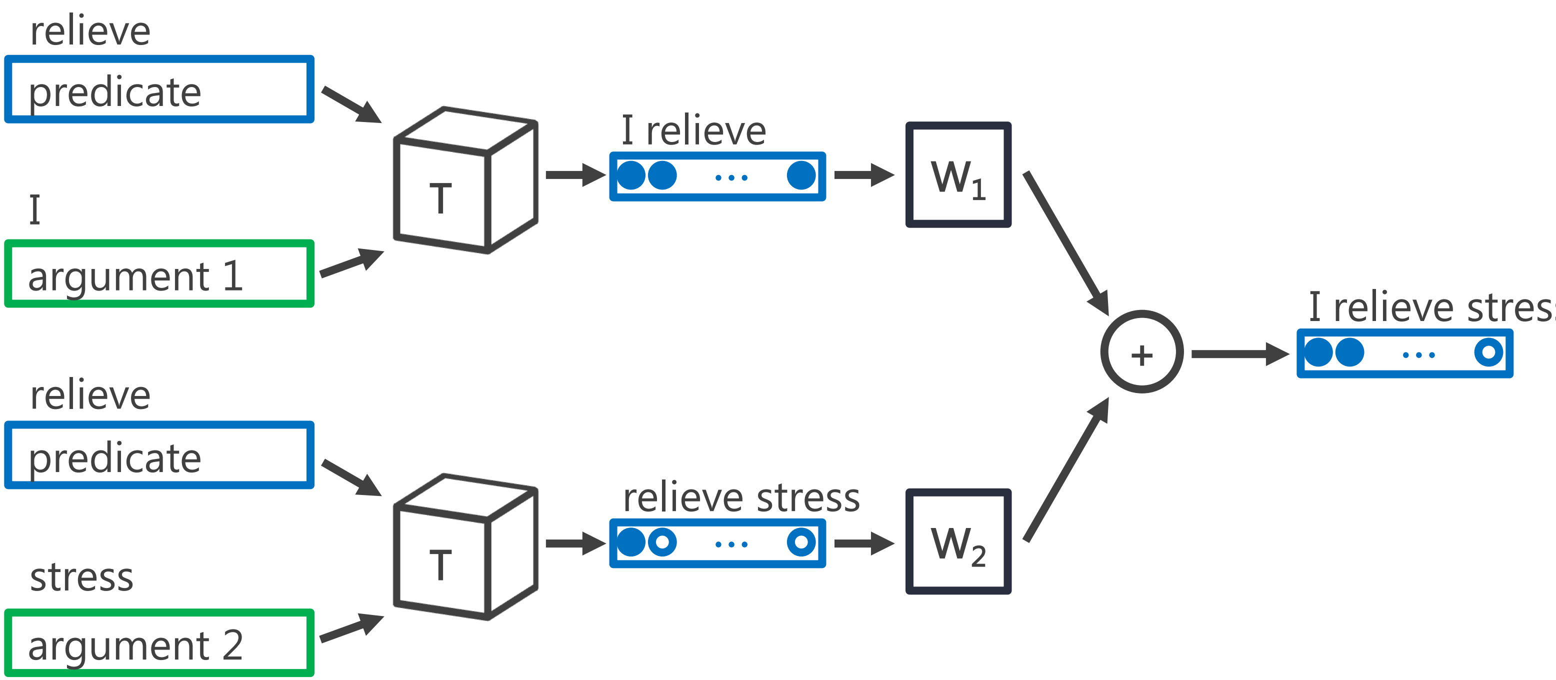
Finds a similar event causality pair on vector space.



A causality in the dialogue (“be exhausted” -> “relax”) is found if a similar causality (“be stressed out” -> “relieve stress”) is included in the pairs.

Role Factored Tensor Model (RFTM) [Weber et al., 2018]

Converts events to distributed representations based on the relationship between a predicate and arguments.



Captures the specific meaning of the predicate.

Experiments

Experiment Settings

	Setting
NCM	EncDec, HRED
Re-ranking	1-best (w/o re-ranking), w/o embedding, w/ embedding
Data	2.6 million Twitter dataset (60 thousand test data)

Re-ranked ratio of response candidates

Indicates how much re-ranking is applicable.

Re-ranking	NCM	Re-ranked
w/o embedding	EncDec	6,469 (12.72)
	HRED	6,231 (12.25)
w/ embedding	EncDec	35,284 (69.39)
	HRED	36,373 (71.53)

12 %
↓
70 %

Ratios were improved drastically by introducing the event embedding method.

Dist and Pointwise Mutual Information (PMI)

Dist and PMI indicate diversity and coherency

NCM	Re-ranking	dist-1	dist-2	PMI
EncDec	1-best	0.06	0.18	1.77
	w/o embedding	0.06	0.19	1.78
	w/ embedding	0.07	0.21	1.77
HRED	1-best	0.07	0.20	1.84
	w/o embedding	0.06	0.20	1.84
	w/ embedding	0.06	0.20	1.86

Models with the embedding have the highest scores.

Diversity (dist) and coherency (PMI) were improved.

NCM Used in Human Evaluation

Baseline model: HRED

v.s.

Our models:
HRED-based models that re-rank w/o or w/ embedding

Human Evaluation

Ten crowd-workers compared hundred responses selected by two of three models in the two criteria.

- **Word coherency**

  Which words in a response are more related to a dialogue history.

- **Dialogue continuity**

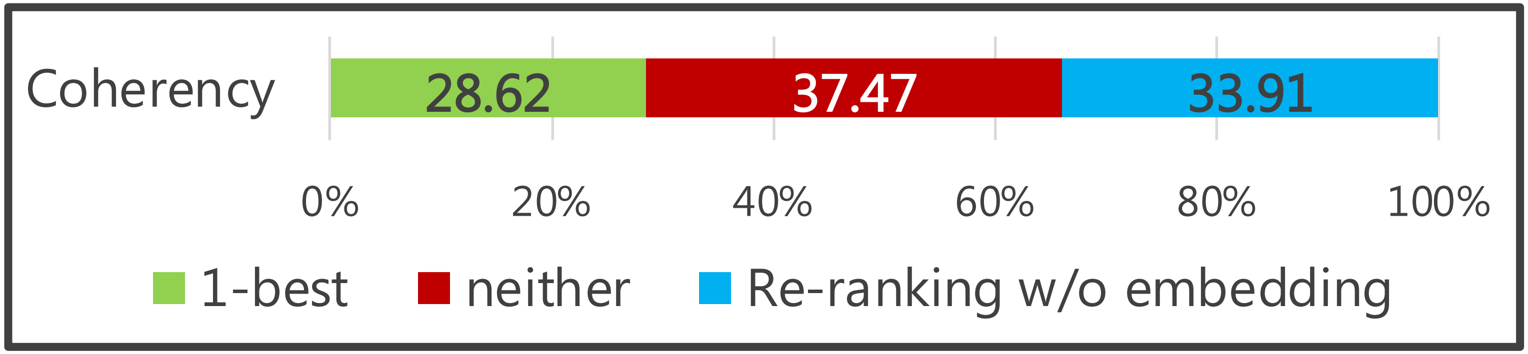
  Which response is easier to respond to.

To reduce the workload, we removed the following data.

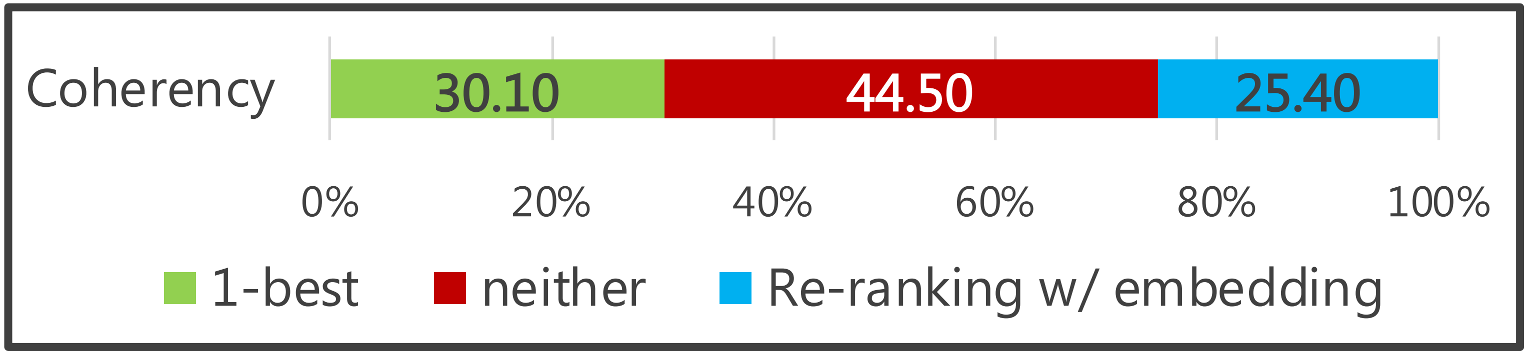
- Number of user utterances is more than two.
- Needs external knowledge to evaluate.

Word Coherency

w/o embedding



w/ embedding

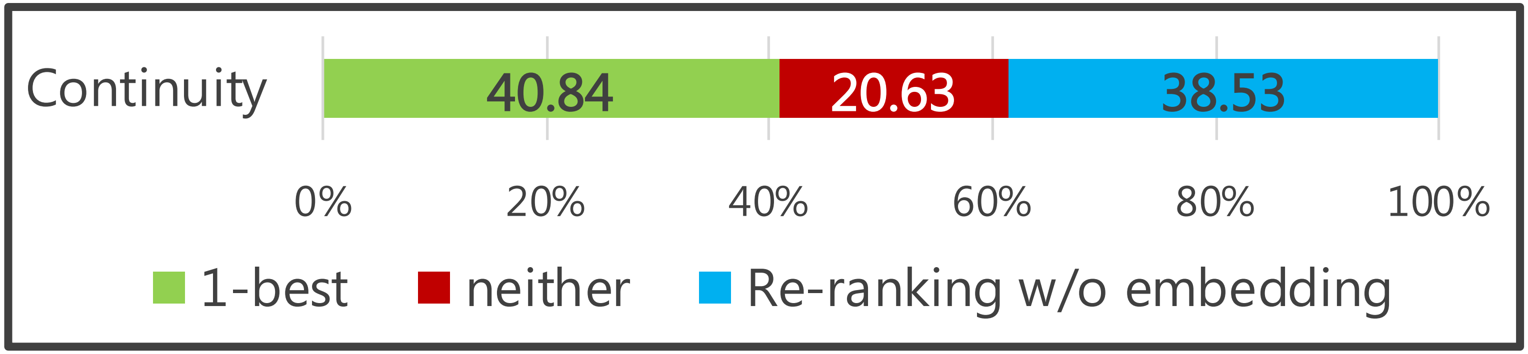


Word coherency was improved by our model without the embedding, but worsened with the embedding.

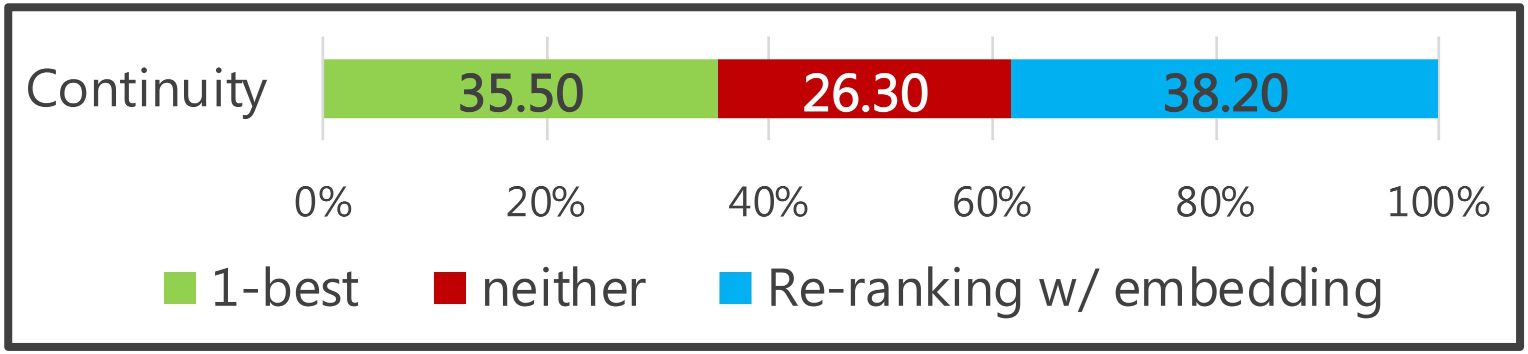
Event embedding over-generalized events.

Dialogue Continuity

w/o embedding



w/ embedding



Dialogue continuity was improved by our model with the embedding because dull responses were reduced.

Case Study

Appropriate Re-ranking

Conversation 1:

User: Because of my fears, I have **been stressed out**.

System (1-best): Are you OK?

System (Re-ranked): Are you OK? Don’t **work too hard**.

Appropriate causality, **“work too hard”** -> **“be stressed out,”** was used to select a coherent response.

Over-generalization

Conversation 2:

User: Our president gives me beer, but he **can**'t **drink**...

System (1-best): I see.

System (Re-ranked): I **drink**.

"drink" -> "can drink" was obtained by generalizing
"enter restaurant" -> "order beer", which is included in the pairs.

Event embedding over-generalized events.

Need to **prevent over-generalization** by improving the embedding architecture.

Lack of Response Naturalness

Conversation 3:

User: I can't get up… I'm gonna **take a day off**… I have a fever…

System (1-best): Take care of yourself.

System (Re-ranked): You must **be tired**.

Proposed method focuses on only word coherency ("be tired" -> "take a day off").

It is not natural.

Need to **maintain the response naturalness.**

Conclusion

Conclusion

We proposed a method to select coherent and diverse responses based on event causality.

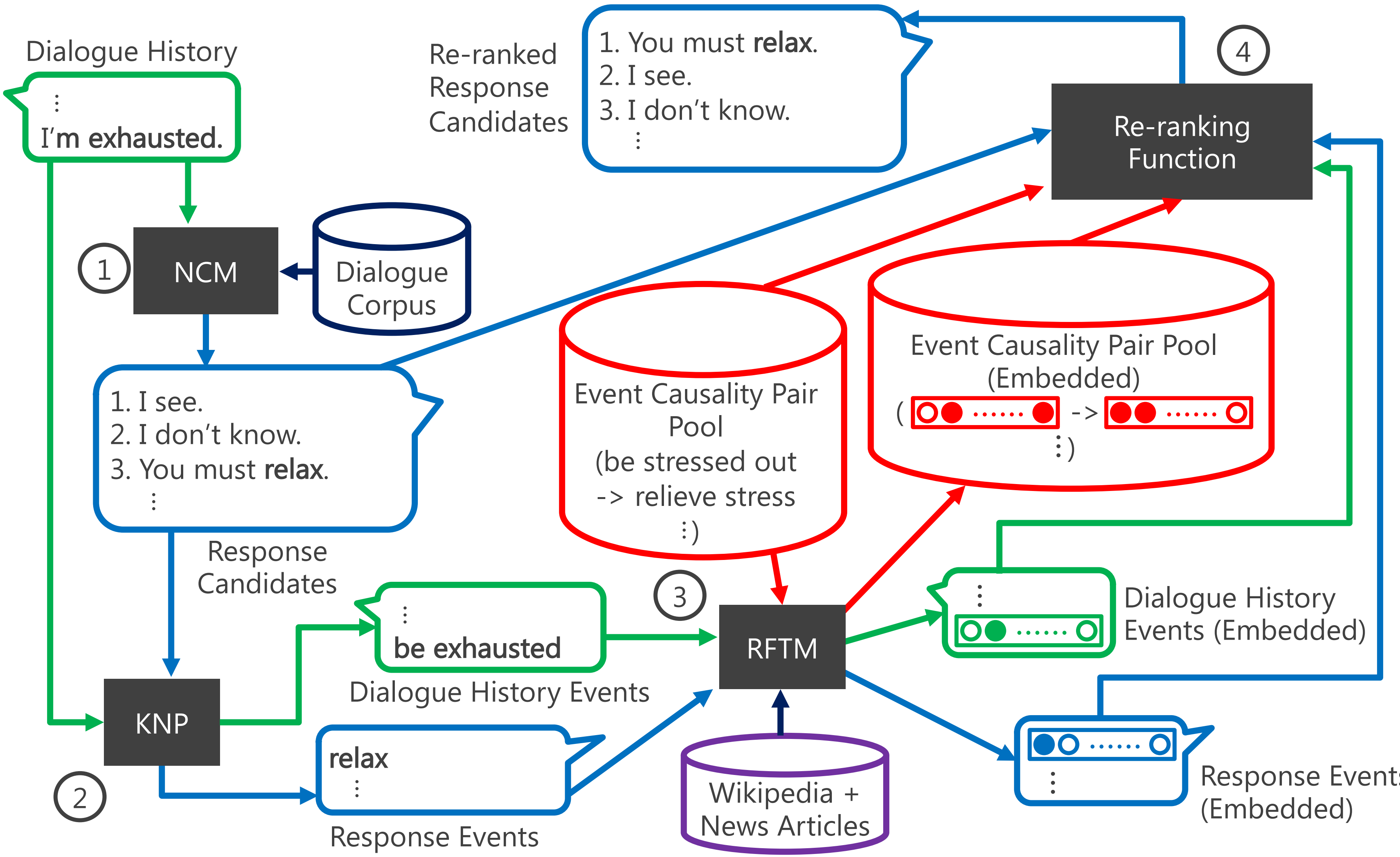
Proposed method improved coherency, diversity, and dialogue continuity by re-ranking.

Future Work

- **Updating the event embedding**
- **Maintaining response naturalness**

Appendix

Details of Re-ranking



Similarity Scores to References

- BLEU

  N-gram coincidence rate of references and generated responses.

  Actual responses are coherent to dialogue histories.

  BLEU correlates with response coherency to some extent.

- NIST

  Based on BLEU, but heavily weights less frequent N-grams to focus on content words.

Similarity Scores to References (Cont.)

- Vector Extrema

Cosine similarity between sentence vectors of a reference and a generated response.

Each sentence vector e_s is computed by taking extrema of Skip-gram word vectors e_w in each dimension d as,

$$e_{sd} = \begin{cases} \max_{w \in s} e_{wd} & \text{if } e_{wd} > |\min_{w' \in s} e_{w'd}| \\ \min_{w \in s} e_{wd} & \text{otherwise} \end{cases}$$

BLEU, NIST, and extrema

NCM	Re-ranking	BLEU	NIST	extrema
EncDec	1-best	1.12	1.19	0.42
	w/o embedding	1.09	1.17	0.42
	w/ embedding	1.00	1.04	0.39
HRED	1-best	1.34	2.74	0.42
	w/o embedding	1.33	2.73	0.42
	w/ embedding	1.28	2.74	0.41

Re-ranking worsened similarity scores to the references.

NCMs generate similar responses to the references.

1-best responses should have the highest scores.

Diversity/coherency evaluation

- Dist-1, 2

Ratio of distinct N-grams in all responses.

Indicates response diversity.

- Pointwise Mutual Information (PMI)

Word in a dialogue history

$$\boxed{\mathrm{PMI} = \frac{1}{|response|} \sum_{wr}^{|response|} \max_{wh} \mathrm{PMI}(wr, wh)}$$

Word in a response

Indicates response coherency.

Dists and PMI are **unrelated to references.**

Summary of Experimental Results

In the human evaluation...

- Word coherency was improved.
- Dialogue continuity was improved.

Diversity (dists) and Coherency (PMI) were also improved in the automatic evaluation.