

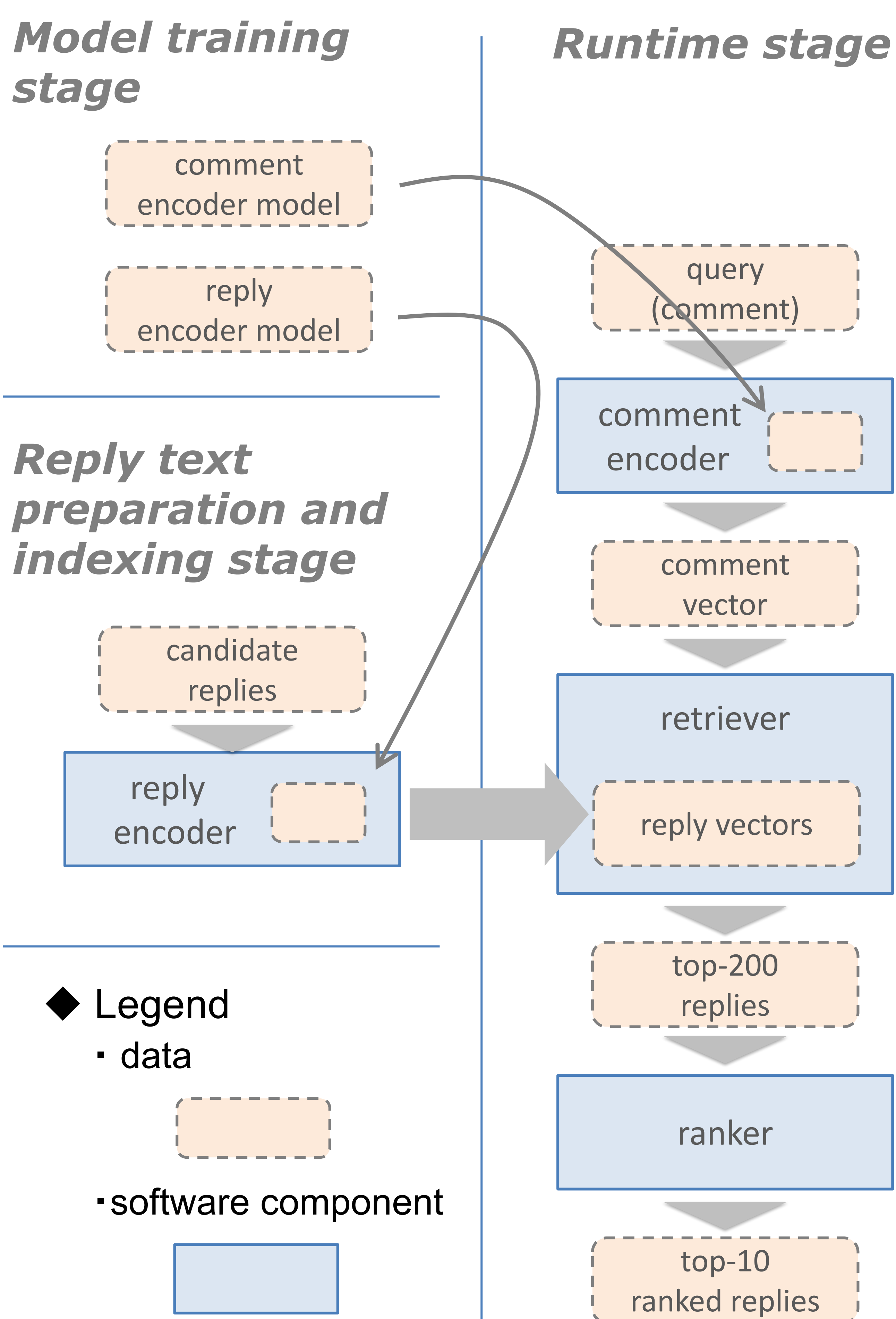
YJTI at the NTCIR-13 STC Japanese Subtask

Toru Shimizu, Yahoo Japan Corporation (toshimiz@yahoo-corp.jp)

1. Overview

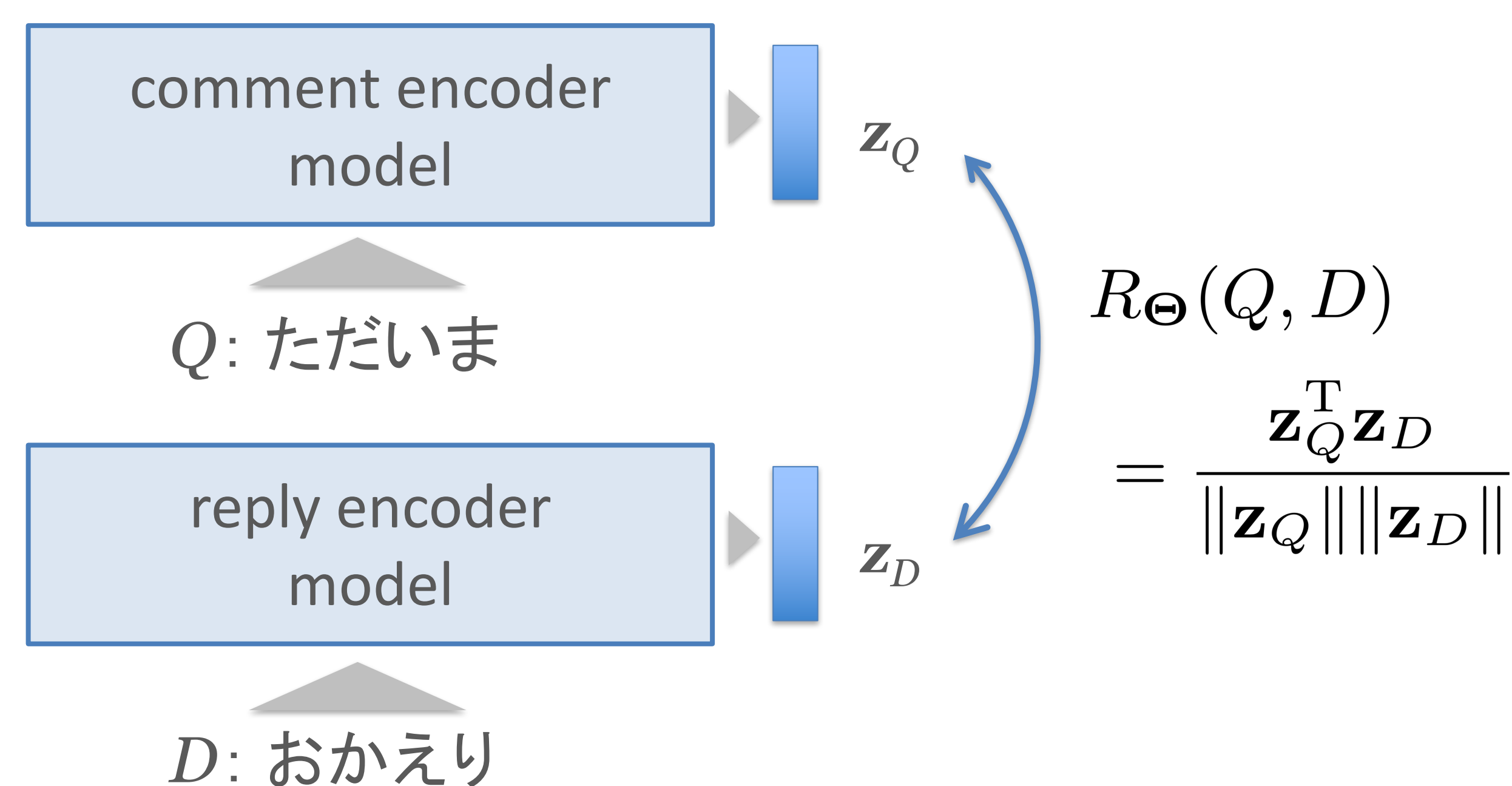
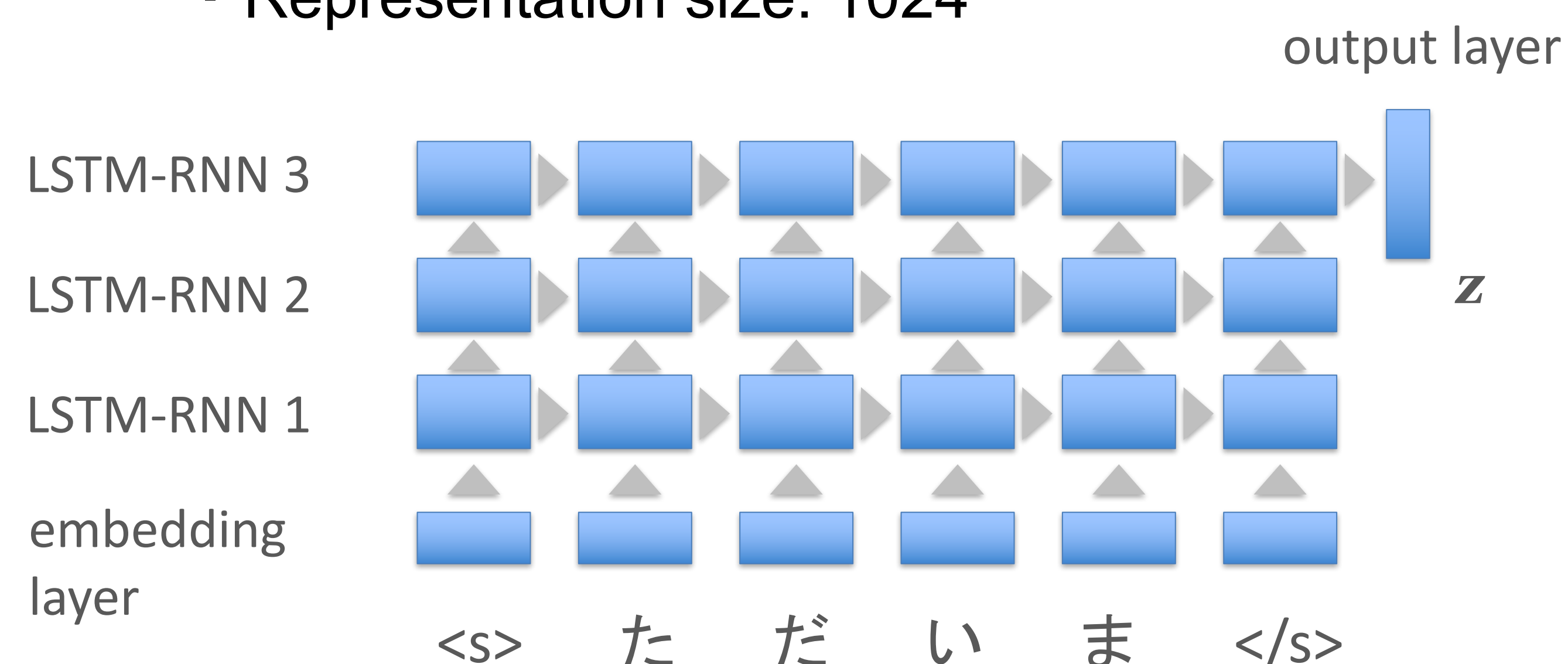
- ◆ The overall approach and the architecture
 - Retrieval-based approach, utilizing all the 1.2M comments in the training data
 - LSTM-DSSM
- ◆ Two runs
 - YJTI-J-R1
 - Trained by Twitter conversation data
 - YJTI-J-R2
 - Mainly trained by Yahoo! Chiebukuro QA data

2. Runtime System



3. Model

- ◆ 3-layer LSTM RNN with a fully-connected layer
 - Two models: a comment encoder model and a reply encoder model
 - LSTM's hidden layer size: 1024
 - Embedding layer size: 256
 - Representation size: 1024



run	model type	data name	records consumed
YJTI-J-R1	DSSM	Twitter conversation	135.0M
YJTI-J-R2	LM	Y! Chiebukuro LM	171.5M
	DSSM	Twitter conversation	85.8M
	DSSM	Y! Chiebukuro QA	42.9M

4. Data

name	type	no. of records
Twitter LM	posts	100.0M
Twitter conversation	pairs	65.1M
Y! Chiebukuro LM	posts	202.0M
Y! Chiebukuro QA	pairs	66.3M

5. Analysis and Conclusions

- ◆ Comparison of matching task performances

matching task	YJTI-J-R1	YJTI-J-R2
Twitter conversation	0.835	0.759
Chiebukuro QA	0.864	0.967

- ◆ Official STC results of our runs (Rule-2)

metrics	YJTI-J-R1	YJTI-J-R2
Mean $nG@1$	0.4322	0.4893

- Effectiveness of DSSM-like approaches combined with large-scale linguistic resources
- Social QA data can be useful for modeling topic-oriented conversations

