EMNLP 2022

PAPER QR CODE

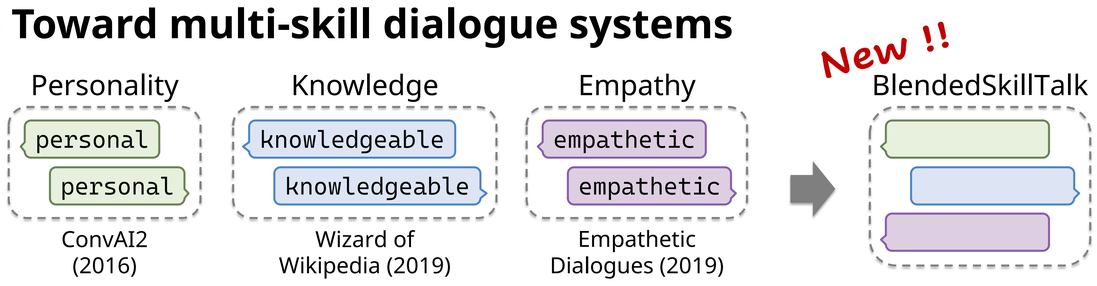
BotsTalk: Machine-Sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets

Minju Kim[1*], Chaehyeong Kim[1*], Yongho Song[1*], Seung-won Hwang[2], Jinyoung Yeo[1]

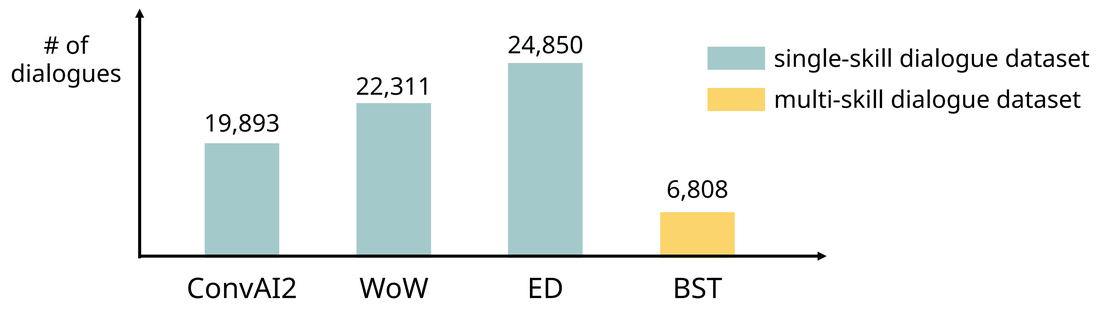
[1]Yonsei University [2]Seoul National University

Motivation

1) Toward multi-skill dialogue systems



2) Limitation of crowdsourcing: scale and cost



▶ Our idea: automatically collect a large-scale multi-skill dialogue dataset, which seamlessly blends various skills over the course of a multi-turn conversation, without additional costs or human efforts.

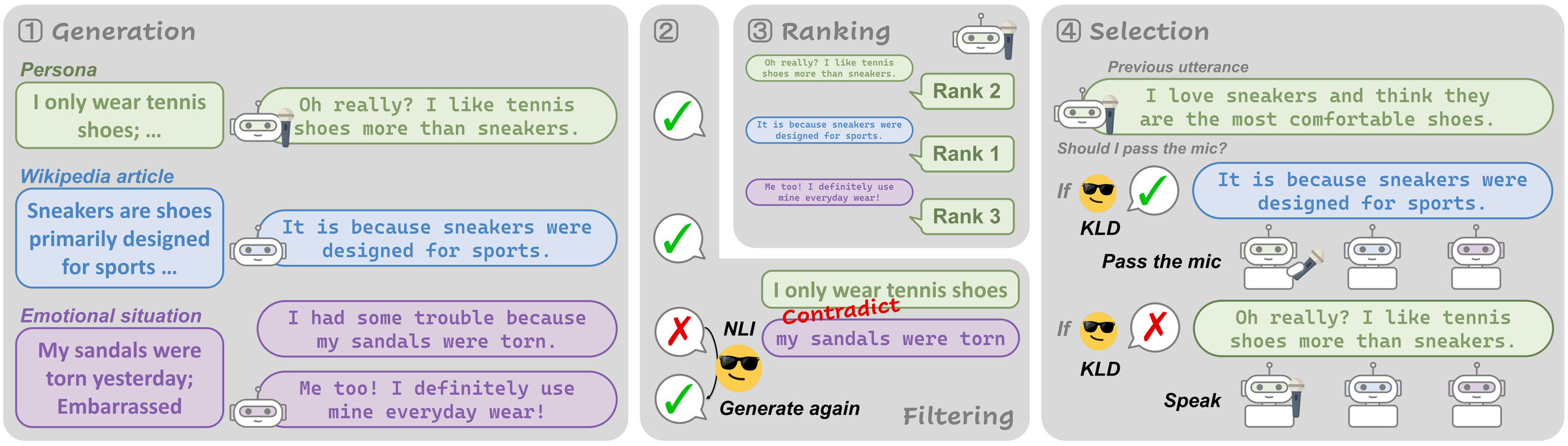
Problem Formulation

Multi-skill dialogue annotation

- Inputs of task: single-skill datasets separately collected on M skills (e.g. personality, knowledge, empathy).
- Output of task: a new multi-skill dialogue dataset, which covers all targeted M skills.
- **Desirable characteristics**
  - ✓ Skill blending: dialogue models should learn to exhibit different dialogue skills in a conversation.
  - ✓ Skill grounding: dialogue models should learn to maintain each dialogue skill when appropriate.

BotsTalk Framework

- Skill agents 🤖 annotate skill-grounded utterances.
- Active agent 🤖🎤 refers to the only one skill agent with a priority (mic 🎤) for the current conversational flow. The agent is willing to pass the mic to other agents if necessary.
- Moderator agent 😎 is an omniscient oracle which controls the overall conversational flow and mediates the skill agents.



Experiments

Blended Skill BotsTalk (BSBT)

- Using BotsTalk, we construct a multi-skill dialogue dataset, BSBT, comprising 300K dialogues with 3M utterances.

	Engaging	Interesting	Natural
BST	43	47	44
BSBT	57	53	56

▲ In human evaluation, machine-written BSBT dataset achieves higher win percentages over human-written BST dataset.

► Our BotsTalk framework can be an effective/efficient alternative to crowdsourcing when collecting multi-skill conversations.

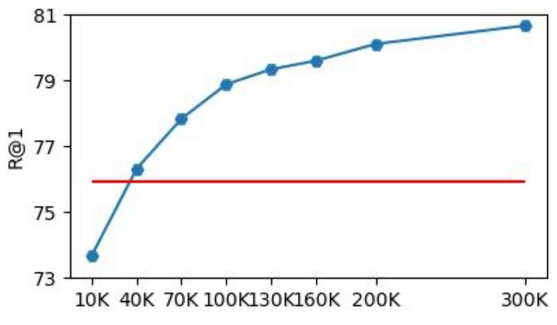
Automatic Evaluation on BST benchmark

	Retrieval (poly-encoder)			Generative (bart)		
	R@1	R@5	MRR	BLEU-1	BLEU-2	BLEU-4
BST	75.92	94.76	84.14	12.19	3.65	0.37
BSBT	80.68	95.79	87.39	11.92	3.74	0.57

▲ BSBT model outperforms all baselines on all automatic metrics.

► Our BSBT dataset works properly as the training resource to learn the ability of blending skills as well as grounding to various skills.

Analysis on scalability



▲ The effect on performance by varying the number of dialogues in training set.

► Large-scale training is important.

► This indicates the potential of BSBT, as our dataset is collected by automatic approach without human intervention.

► Our BotsTalk framework is scalable with respect to data size and increasing skill types.

Analysis on multi-task learning

	R@1	MRR
BSBT	80.68	87.39
MTL	78.95	86.23
+ BSBT100K	80.94	86.92
+ BSBT200K	82.01	87.83
+ BSBT300K	82.10	88.04

An overlap between parameterized (BSBT) and materialized (MTL) knowledge for multi-skill dialogues.

Why?

The performance gain becomes marginal.