Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills

Eric Michael Smith*, Mary Williamson*, Kurt Shuster, Jason Weston, Y-Lan Boureau

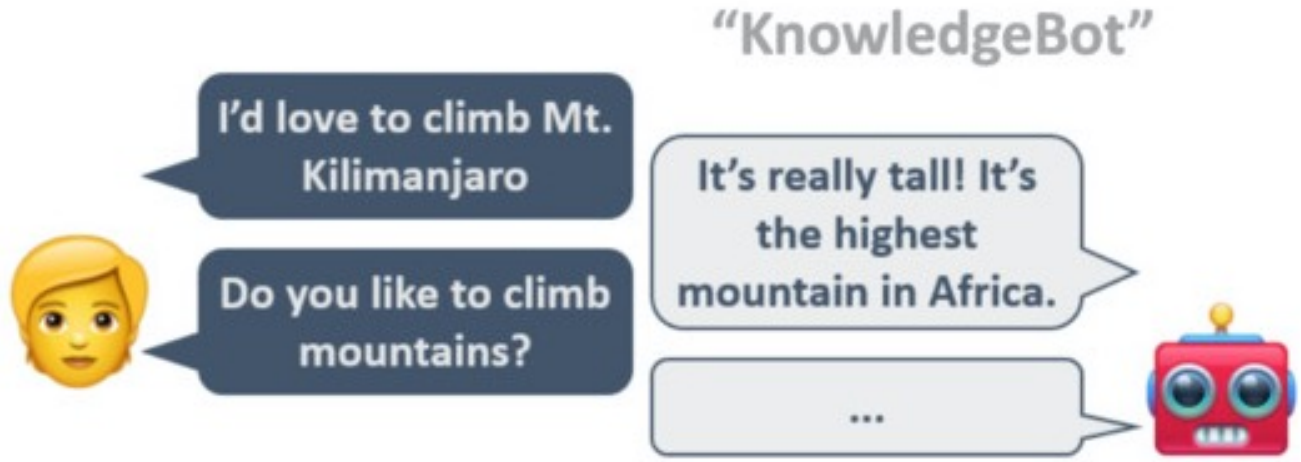
Presenter : Qian Jiang

ILLINOIS

Motivation & Background

Multi-skill Conversation

- Many conversational agents each good at only one thing?
  -> Not enough
- Good conversational agents should have different skills!



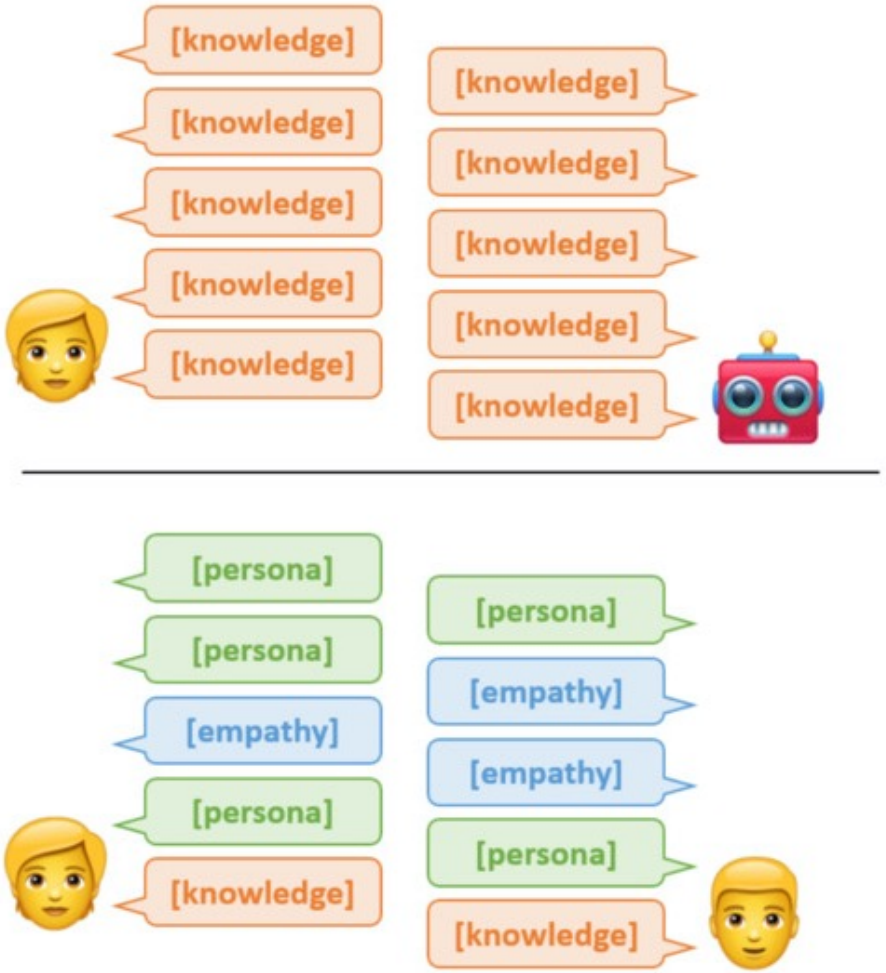
Existing benchmarks

- Existing datasets tailored for specific individual skills:
  - **PersonaChat/ConvAI2 :** showing **personality**
  - **Wizard of Wikipedia(WOW):** being **knowledgeable**
  - **EmpatheticDialogues(ED):** showing **empathy**

BlendedSkillTalk Dataset (BST)

Blended-skill dataset

- Crowdsourced dataset of 5k conversations
- Workers are instructed to be **knowledgeable**, **empathetic**, or talking about **personal** details



Persona for Unguided Speaker:	**Persona for Guided Speaker**:
My son plays on the local football team.	My eyes are green.
I design video games for a living.	I wear glasses that are cateye.

Wizard of Wikipedia topic: Video game design
Previous utterances (shown to speakers):
U: What video games do you like to play?
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget

Actual utterances:
U: Exactly! I think many people fail to notice how beautiful the art of video games can be. **(PB)**
(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")
G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. **(K)**
U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids **(K)**
G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. **(S)**
U: I use games to relax after a stressful day, the small escape is relaxing. **(PB)**

- **Knowledge (K)**
- **Empathy (E)**
- **Personal situations (S)**
- **Personal background (PB)**

Dataset Analysis - Guided workers choice of suggestions

Chosen suggestion	Initial Context	Count	Total
none	ConvAI2	7280	21468
	ED	7257	
	WoW	6931	
ConvAI2	ConvAI2	567	1599
	ED	496	
	WoW	536	
ED	ConvAI2	766	2221
	ED	773	
	WoW	682	
WoW	ConvAI2	634	1730
	ED	494	
	WoW	602	

- Overall balanced
- More likely to choose the same suggestion as initial context

Dataset Analysis – Unguided workers response related to seed context

% classified as:	Source of Seed Context ConvAI2	WoW	ED
ConvAI2	**29.6**	25.3	25.5
WoW	49.6	**57.5**	30.3
ED	20.8	17.1	**44.2**

- A three-class classifier on top of BERT that assigns an utterance to the dataset it came from
- More likely to be classified as same as seed context

Dataset Analysis – number of modes

Mode Count	Conversations	Pct (%)
1	51	6.9%
2	167	22.6%
3	290	39.2%
4	232	31.4%

- **Knowledge (K)**
- **Empathy (E)**
- **Personal situations (S)**
- **Personal background (PB)**

Methods

Single-task approaches

- Train on a single-skill dataset and evaluate on all skills datasets
- + With/Without finetuning on BlendedSkillTalk (BST) dataset

Multi-task approaches

- Train in a multi-task way (MT Single-Skills)
- Random selection from three single models (Random-Skill)
- Train a top-level classifier to select from three single models (MT Two-Stage) -- a three-class classifier on top of BERT that assigns an utterance to the dataset it came from
- + With/Without finetuning on BlendedSkillTalk (BST)

Bias for MT Single-Skills

- Sample training data from each task during updates
- However, each dataset contains different pre-context
  - **PersonaChat/ConvAI2 :** persona context
  - **Wizard of Wikipedia(WOW):** topic context
  - **EmpatheticDialogues(ED):** None
- This introduce bias!

Why there is bias and what to do?

- Deep models like shortcuts ☹
- Recall three datasets contain different pre-contexts
- Model will try to make decisions based on the pre-context instead of the dialogue itself!

- Make all the data have topic and persona context ☺

Debias Results

	MT Single-Skills		MT S.-S. + BST	
Utt. Selected	orig.	debiased	orig.	debiased
ConvAI2	64.4%	38.9%	61.1%	48.1%
WoW	11.3%	29.4%	10.0%	21.3%
ED	24.2%	31.6%	28.8%	30.5%

- Debias results in the multi-task retrieval models selecting utterances more evenly

Results

Results on single-skill datasets

- Base model: 256-million parameter transformer-based model pretrained on reddit dataset
- Metric: Hits@1 (accuracy at retrieving right response from set)

	Single-skill benchmarks			
Model	ConvAI2	WoW	ED	Avg.
SOTA Reported	87.3	87.4	**66.0**	*80.2*
ConvAI2	**89.4**	78.4	42.6	70.1
WoW	57.3	**91.8**	47.7	65.6
ED	63.3	81.0	**65.1**	69.8

- Single-task can match SOTA on corresponding task but suffer on others

Model	Single-skill benchmarks			
	ConvAI2	WoW	ED	Avg.
SOTA Reported	87.3	87.4	**66.0**	*80.2*
ConvAI2	**89.4**	78.4	42.6	70.1
WoW	57.3	**91.8**	47.7	65.6
ED	63.3	81.0	**65.1**	69.8
BST model	78.5	84.1	52.0	71.5
Random-Skill	71.0	83.9	52.0	69.0
MT Two-Stage	84.7	90.1	63.4	79.4
MT Single-Skills	**88.8**	**92.8**	**63.2**	**81.6**
	Added-context benchmarks			
MT Single-Skills	88.9	92.8	63.2	81.6
	Mixed-candidates evaluation			
Single-task	82.1	88.2	60.2	*76.8*
MT Two-Stage	77.2	86.6	59.0	74.3
MT Single-Skills	**85.2**	**92.1**	**61.1**	**79.5**

- MT Single-Skills achieves the best performance in MT models, yet worse than single task on corresponding task
- Debias barely change numbers
- Best respective single-task models suffers, while the MT Single-Skills model proves more resilient

Results on BlendedSkillTalk (BST) dataset

- Tested directly on BST without any additional training in a zero-shot setting
- Fine-tuned on the BST training set

Model	BST, zero-shot	+BST, FT
ConvAI2	76.8	81.7
WoW	67.5	79.4
ED	69.0	80.4
BST	-	79.2
Random-Skill	71.2	-
MT Two-Stage	71.9	-
MT Single-Skills	**80.1**	**83.8**

- MT Single-Skills achieve good performance even without pre-training

Human Evaluation

- Workers chat with various models and then rate the conversation along several axes

Model	Knowledge	Empathy	Personal	Overall quality
ConvAI2	3.2	3.1	**3.4**	3.0
WoW	3.3	2.9	2.7	2.6
ED	3.4	3.3	3.0	3.0
BST	**3.5**	**3.6**	3.1	3.3
Random-Skill	3.2	2.9	**3.2**	2.7
MT Two-Stage	**3.7**	**3.6**	**3.3**	**3.5**
MT Single-Skills	**3.7**	**3.6**	3.0	**3.4**
MT Single-Skills +BST fine-tuning	**3.7**	**3.8**	**3.2**	**3.6**

Take aways

- Collect a new dataset blending conversational skills
- Train a model multi-task on multiple single-purpose conversational datasets
- Show good performance on new dataset even without fine-tuning

- Future work: Expand to other conversational skills

Thank You