



Predicting User Engagement Status for Online Evaluation of Intelligent Assistants

Rui Meng, Zhen Yue, Alyssa Glass



YAHOO!
RESEARCH

Introduction

- **Intelligent Assistants (IA)**

- New portal to Internet services
- Support a wide variety of functions

- **Challenges in Online Evaluation for IA**

- Large volume of daily user requests
- Real-time performance monitoring



Cortana



Siri



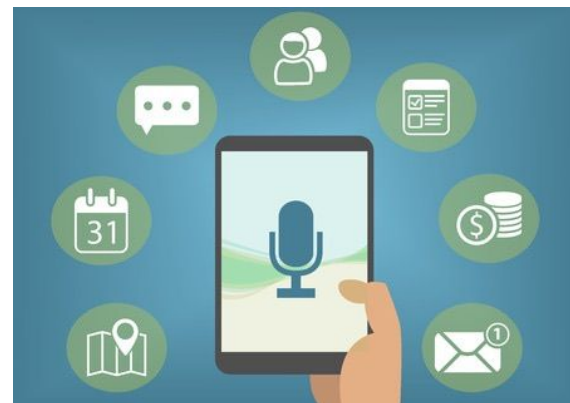
amazon echo



Google now

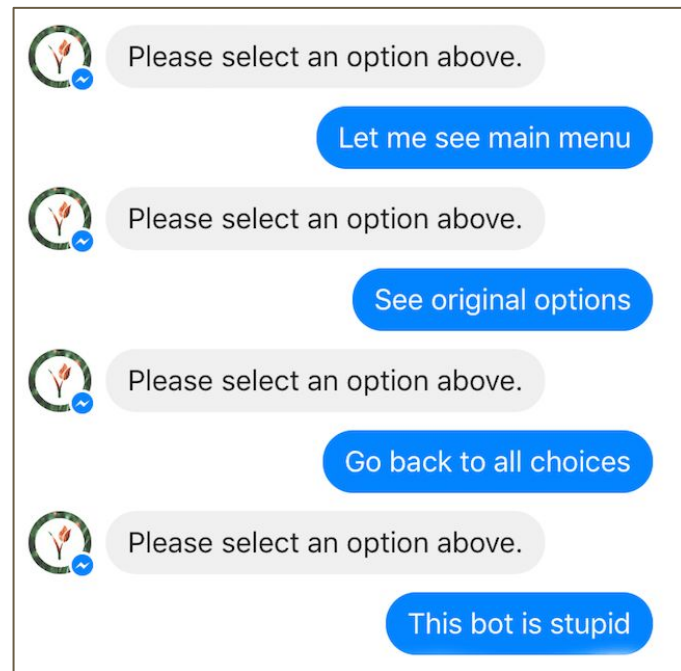


Facebook M



Introduction

- **Intelligent Assistants (IA)**
 - New portal to Internet services
 - Support a wide variety of functions
- **Challenges in Online Evaluation for IA**
 - Large volume of daily user requests
 - Real-time performance monitoring
 - Understanding user utterances



Online IA Evaluation

- **Desiderata**

- Reflect real-time system performance
- Can scale to billions of requests
- Indicative of end-user engagement/satisfaction
- Interpretable, easy for engineers to locate system errors

Previous Methods for IA Evaluation

1. Evaluation on individual components

- E.g. dialogue state tracking, dialogue act classification
- Does not reflect the overall system performance and user engagement

2. Evaluation by comparing system outputs with ground-truth responses

- Commonly used in dialogue generation systems, using metrics like BLEU, METEOR
- Correlated poorly with human rating

3. Human-based evaluation

- e.g. using Amazon Mechanical Turk
- Cannot scale up; Does not reflect real-time performance

Motivation

- **Lessons from Online Evaluation for IR Systems**
 - User behavioral signals can be effective and reliable indicators
 - E.g. long-dwell click, query reformulation, query abandonment
- **Lessons from PARADISE framework**
 - A classic framework for evaluating spoken dialogue agents
 - A successful dialogue system should maximize task success and minimize cost
 - Requires good ways to segment tasks and measure their success/cost
- **Can we find good user behavioral signals for evaluating intelligent assistants?**

Session from Yahoo Captain (Family Assistant)

Side	Content	Quality
User	"Chic peas, diced tomatoes"	
Bot	["I'm learning. Can you clarify?"]	Bad
User	"Add Chic peas, diced tomatoes"	
Bot	["I'm learning. Can you clarify?"]	Bad
User	"Add to list Chic peas, diced tomatoes"	
Bot	["Great! I've added your items, Sarah.\n\nHere's your list:\n1. potatoes\n2. sweet potatoes\n3. kale\n4. can crushed tomatoes\n5. almond butter\n6. broccoli\n7. beer\n8. onions\n9. peas\n10. hamburger\n11. cauliflower\n12. chic peas\n13. diced tomatoes"]	Good

- Task success \neq satisfactory user experience
 - Both utterance-level and task-level engagement matter

Session from DSTC-2

Side	Content	Quality
Bot	Thank you for calling the Cambridge Information system. Your call will be recorded for research purposes. You may ask for information about a place to eat, such as a restaurant, a pub, or a cafe. How may I help you?	
User	im looking for a turkish restaurant	
Bot	What kind of price range are you looking for? For example, cheap, or moderate.	G or B?
User	uh moderate	
Bot	Are you looking for somewhere near a particular venue? For example, the Museum Of Technology, Churchill College, or Clare Hall.	G or B?
User	no	
Bot	196 Meze Bar is an outstanding restaurant and it serves turkish food. It's in the moderate price range.	Good

- A successful task may need several necessary turns
 - They should not be treated as negative experience, nor positive
 - Measuring cost by number of turns only is not reasonable

Our Method

- **Utterance Classification by User Engagement Status**
 - Define a four-class utterance classification scheme
 - Each class represents the engagement status of a user after seeing system response
 - ***Fulfillment (F)***: current user request is understood and fulfilled by the system.
 - ***Continuation (C)***: current user request is understood by the system but more interactions are required to complete this request.
 - ***Reformulation (R)***: current user request is wrongly or incompletely understood by the system, and user repeats/rephrases this request in the next turn.
 - ***Abandonment (A)***: current user request is not understood or fulfilled by the system, and the user abandons this request by closing the conversation or starting a new request.

Annotated Session from DSTC-2

	# Turn	Utterance
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?
Request a restaurant	1	User: Moderately priced
		Bot: What part of town do you have in mind?
	2	User: North
		Bot: Would you like something in the cheap, moderate, or expensive price range?
	3	User: Moderate
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range
Ask address	4	User: Address
		Bot: Sure , golden wok is on 191 Histon Road Chesterton
Ask number	5	User: Phone number
		Bot: The phone number of golden wok is 01223 350688
Ask postcode	6	User: Postcode
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range
	7	User: Postcode
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range
	8	User: Thank you good bye

Annotated Session from DSTC-2

	# Turn	Utterance	Label
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
Request a restaurant	1	User: Moderately priced	R
		Bot: What part of town do you have in mind?	
	2	User: North	C
		Bot: Would you like something in the cheap, moderate, or expensive price range?	
	3	User: Moderate	F
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
Ask address	4	User: Address	F
		Bot: Sure , golden wok is on 191 Histon Road Chesterton	
Ask number	5	User: Phone number	F
		Bot: The phone number of golden wok is 01223 350688	
Ask postcode	6	User: Postcode	R
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	7	User: Postcode	A
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	8	User: Thank you good bye	

Annotated Session from DSTC-2

	# Turn	Utterance	Label
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
Request a restaurant	1	User: Moderately priced	R
		Bot: What part of town do you have in mind?	
	2	User: North	C
		Bot: Would you like something in the cheap, moderate, or expensive price range?	
	3	User: Moderate	F
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
Ask address	4	User: Address	F
		Bot: Sure , golden wok is on 191 Histon Road Chesterton	
Ask number	5	User: Phone number	F
		Bot: The phone number of golden wok is 01223 350688	
Ask postcode	6	User: Postcode	R
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	7	User: Postcode	A
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	8	User: Thank you good bye	

Annotated Session from DSTC-2

	# Turn	Utterance	Label	Task
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?		
Request a restaurant	1	User: Moderately priced	R	Success
		Bot: What part of town do you have in mind?		
	2	User: North	C	
		Bot: Would you like something in the cheap, moderate, or expensive price range?		
	3	User: Moderate	F	
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
Ask address	4	User: Address	F	Success
		Bot: Sure , golden wok is on 191 Histon Road Chesterton		
Ask number	5	User: Phone number	F	Success
		Bot: The phone number of golden wok is 01223 350688		
Ask postcode	6	User: Postcode	R	Failure
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
	7	User: Postcode	A	
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
	8	User: Thank you good bye		

Annotated Session from DSTC-2

	# Turn	Utterance	Label
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
Request a restaurant	1	User: Moderately priced	R
		Bot: What part of town do you have in mind?	
	2	User: North	C
		Bot: Would you like something in the cheap, moderate, or expensive price range?	
	3	User: Moderate	F
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
Ask address	4	User: Address	F
		Bot: Sure , golden wok is on 191 Histon Road Chesterton	
Ask number	5	User: Phone number	F
		Bot: The phone number of golden wok is 01223 350688	
Ask postcode	6	User: Postcode	R
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	7	User: Postcode	A
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	8	User: Thank you good bye	

Annotated Session from DSTC-2

	# Turn	Utterance	Label
		Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
Request a restaurant	1	User: Moderately priced	R
		Bot: What part of town do you have in mind?	C
	2	User: North	
		Bot: Would you like something in the cheap, moderate, or expensive price range?	
	3	User: Moderate	F
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
Ask address	4	User: Address	F
		Bot: Sure , golden wok is on 191 Histon Road Chesterton	
Ask number	5	User: Phone number	F
		Bot: The phone number of golden wok is 01223 350688	
Ask postcode	6	User: Postcode	R
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	7	User: Postcode	A
		Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
	8	User: Thank you good bye	

Our Method

- The classification implies two binary classification tasks
 1. Response Success
 - whether user's current utterance is reasonably responded or not

	Ongoing	Ending
Correctly responded	Continuation	Fulfillment
Wrongly responded	Reformulation	Abandonment

Our Method

- The classification implies two binary classification tasks
 1. Response Success
 - whether user's current utterance is reasonably responded or not
 2. Task completion (task boundary detection)
 - whether the current task has ended or not

	Ongoing	Ending
Correctly responded	Continuation	Fulfillment
Wrongly responded	Reformulation	Abandonment

Online Metrics w/ User Engagement Status

- We define an online metric **User Engagement Score** (UE_{SCORE}) taking into account both task success and dialogue efficiency.

$$UE_{SCORE} = \frac{SuccessRate + EfficiencyRate}{2}$$

- **Success Rate**

- The ratio of tasks that were successfully accomplished.

- **Efficiency Rate**

- Reflected by the number of reformulated utterances and average task length.

Annotated Session from DSTC-2

# Turn	Utterance	Label	Task
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?		
1	User: Moderately priced	R	Success
	Bot: What part of town do you have in mind?		
2	User: North	C	
	Bot: Would you like something in the cheap, moderate, or expensive price range?		
3	User: Moderate	F	Success
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
4	User: Address	F	
	Bot: Sure , golden wok is on 191 Histon Road Chesterton		
5	User: Phone number	F	Success
	Bot: The phone number of golden wok is 01223 350688		
6	User: Postcode	R	Failure
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
7	User: Postcode	A	
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
8	User: Thank you good bye		

$$\begin{aligned}
 SuccessRate &= \frac{\#(TASK_{success \in \mathbb{S}})}{\#(TASK_{\in \mathbb{S}})} \\
 &= \frac{3}{4} = 0.75
 \end{aligned}$$

Annotated Session from DSTC-2

# Turn	Utterance	Label	Task
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?		
1	User: Moderately priced	R	Success
	Bot: What part of town do you have in mind?		
2	User: North	C	
	Bot: Would you like something in the cheap, moderate, or expensive price range?		
3	User: Moderate	F	Success
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
4	User: Address	F	
	Bot: Sure , golden wok is on 191 Histon Road Chesterton		
5	User: Phone number	F	Success
	Bot: The phone number of golden wok is 01223 350688		
6	User: Postcode	R	
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		Failure
7	User: Postcode	A	
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
8	User: Thank you good bye		

- The more reformulated utterances, the more cost

$$ReformRate = \sum_{T \in \mathcal{S}} \frac{\#(UTT_{reform \in T})}{\#(UTT_{\in T})}$$

$$= \frac{2}{7} = 0.286$$

- The longer task length, the more user fatigue
 - α is a hyperparameter (=1), acceptable number of turns

$$FatigueValue = \frac{\sum_{T \in \mathcal{S}} \max(1, \#(UTT_{\in T}) - \alpha)}{\#(TASK_{\in \mathcal{S}})}$$

$$= \frac{2 + 1 + 1 + 1}{4} = 1.25$$

$$EfficiencyRate = \frac{1 - ReformRate}{FatigueValue} = \frac{1 - 0.286}{1.25} = 0.5712$$

Annotated Session from DSTC-2

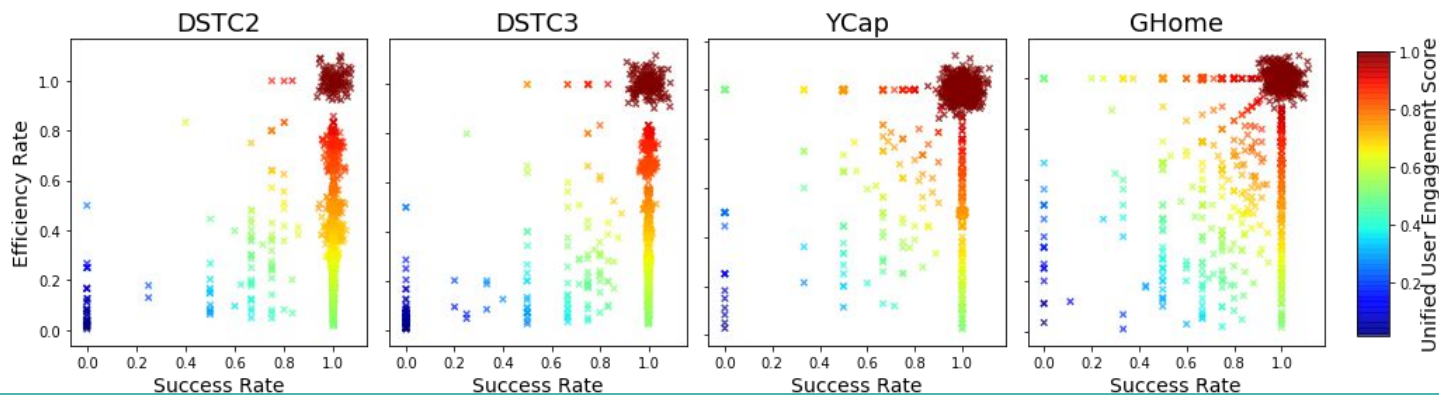
# Turn	Utterance	Label	Task
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?		
1	User: Moderately priced	R	Success
	Bot: What part of town do you have in mind?		
2	User: North	C	
	Bot: Would you like something in the cheap, moderate, or expensive price range?		
3	User: Moderate	F	Success
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
4	User: Address	F	
	Bot: Sure , golden wok is on 191 Histon Road Chesterton		
5	User: Phone number	F	Success
	Bot: The phone number of golden wok is 01223 350688		
6	User: Postcode	R	Failure
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
7	User: Postcode	A	
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range		
8	User: Thank you good bye		

- Average of task success and dialogue efficiency

$$\begin{aligned}
 UE_{SCORE} &= \frac{SuccessRate + EfficiencyRate}{2} \\
 &= \frac{0.75 + 0.5712}{2} = 0.661
 \end{aligned}$$

Case Study on Four Annotated Datasets

- Datasets
 - 4 IA systems: **DSTC2**, **DSTC3**, Yahoo Captain (**YCap**), Google Home (**GHome**)
 - 4,000 dialogue sessions, labelled by professional annotators, kappa=0.790.
 - Identified 35/63/9/13 low-quality sessions whose UE score is less than 0.2
 - **DSTC2/DSTC3**: poor ASR and language understanding ability
 - **YCap**: only takes user commands matching particular templates
 - **GHome**: Problems are more diverse, e.g. failed to answer open-domain questions



Automatic Engagement Status Prediction

- **Goal**

- Preliminary exploration of automatic prediction using machine learning methods
- Identify potential technical challenges in this task

- **Experiment setting**

- An utterance-level four-way classification
- Models
 - Traditional model (Logistic Regression, SVM, Random Forest) + feature engineering
 - Convolutional Neural Networks
- Extract 7 groups of features from utterances
 1. Basic Feature (utterance length)
 2. User Action Feature
 3. Time Interval Feature
 4. Lexical Feature
 5. Phrasal Feature
 6. Syntactic Feature
 7. Semantic/Topic Feature

Results

- **Models**

- CNNs perform well without any manual feature.

Model	Accuracy	F1-score
Majority	0.6020	0.1858
Random	0.2503	0.2029
SVM	0.8410	0.6440
LR	0.8398	0.6413
RF	0.8415	0.6192
CNN.Rand	0.8287	0.6549
CNN.MultiCh	0.8367	0.6674

Results

• Models

- CNNs perform well without any manual feature design.
- But classic models can improve significantly with careful feature engineering, e.g. combining multiple groups and feature selection.

Model	Accuracy	F1-score
Majority	0.6020	0.1858
Random	0.2503	0.2029
SVM	0.8410	0.6440
LR	0.8398	0.6413
RF	<u>0.8415</u>	0.6192
CNN.Rand	0.8287	0.6549
CNN.MultiCh	0.8367	0.6674

Model	w/o FeatSelect w\Sim	w\FeatSelect	
		w/o Sim	w\Sim
(a) Basic	0.4105	-	0.4105
(b) Phrasal	0.6316	-	0.6318
(c) Syntactic	0.6280	-	0.6402 ^{†‡}
(d) N-grams	0.6573	-	0.6770 ^{†‡}
(e) Topic model	0.6346	-	0.6358
(f) Word2Vec	0.6521	-	0.6523
(g) Doc2Vec	0.5968	-	0.5969
(h) Skip-thought	0.6216	-	0.6216
(i) (a) + (b) + (c) + (d)	0.6694	0.6511	0.7085 ^{†‡}
(j) + Topic Model	0.6720	0.6617	0.7152 ^{†‡}
(k) + Word2Vec	0.6790	0.6617	0.7135 ^{†‡}
(l) + Doc2Vec	0.6713	0.6631	0.7100 ^{†‡}
(m) + Skip-thought	0.6747	<u>0.6666</u>	0.7124 ^{†‡}
(n) All	<u>0.6825</u>	0.6589	0.7140 ^{†‡}
(o) CNN.Rand	<u>0.6647</u>	-	-
(p) CNN.MultiCh	0.6685	-	-

Results

- Context Range
 - Five settings of context window
 - **CUR_UTT** = {user_utt_i}
 - **CUR** = {user_utt_i, bot_utt_i}

# Turn	Utterance	Label
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
1	User: Moderately priced Bot: What part of town do you have in mind?	
2	User: North Bot: Would you like something in the cheap, moderate, or expensive price range?	
3	User: Moderate Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	F
4	User: Address	

CUR_UTT (green dashed box) highlights the User utterance in Turn 3.

CUR (orange dashed box) highlights both the User and Bot utterances in Turn 3.

The **Target Label** 'F' is circled in blue in the Label column for Turn 3.

Results

- Context Range
 - Five settings of context window
 - **PREV** = {user_utt_{i-1}, bot_utt_{i-1}, user_utt_i}
 - **NEXT** = {user_utt_i, bot_utt_i, user_utt_{i+1}}

# Turn	Utterance	Label
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
1	User: Moderately priced	
	Bot: What part of town do you have in mind?	
2	User: North	
	Bot: Would you like something in the cheap, moderate, or expensive price range?	
3	User: Moderate	F
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	
4	User: Address	

PREV (yellow dashed box) covers turns 1, 2, and 3.

NEXT (red dashed box) covers turns 3, 4, and 5.

The label **F** is circled in blue and labeled **Target Label**.

Results

- Context Range

- Five settings of context window

- **ALL** = {user_utt_{i-1}, bot_utt_{i-1}, user_utt_i, bot_utt_i, user_utt_{i+1}}

# Turn	Utterance	Label
	Bot: Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?	
1	User: Moderately priced	
	Bot: What part of town do you have in mind?	
2	User: North	
	Bot: Would you like something in the cheap, moderate, or expensive price range?	
3	User: Moderate	
	Bot: Golden wok is a nice restaurant in the north of town in the moderate price range	F
4	User: Address	

Target Label

Results

- Context Range

- Including the user's next utterance is critical to predict user engagement (**CUR** -> **NEXT**)

Model	CUR_UTT	CUR	NEXT	PREV	ALL
Basic	0.3425	0.3503	0.3836	0.3501 ^{†‡}	0.3963^{†‡}
Phrasal	0.3679	0.5521	0.5913	0.3709	0.5661 ^{†‡}
Syntactic	0.3485	0.5530	0.6078	0.3671 ^{†‡}	0.5867 ^{†‡}
N-grams	0.3839	0.5694	0.6113	0.3788	0.5984 ^{†‡}
Topic Model	0.2982	0.5255	0.5803	0.3464 ^{†‡}	0.5829
Word2Vec	0.3704	0.5723	0.6162	0.3827 ^{†‡}	0.6032 ^{†‡}
Doc2Vec	0.3427	0.5379	0.5858	0.3722 ^{†‡}	0.5740 ^{†‡}
Skip-thought	0.3648	0.5545	0.6063	0.3692	0.6008 [†]
CNN.Rand	<u>0.4252</u>	<u>0.5862</u>	0.6647	0.4153	0.6549 [†]
CNN.MultiCh	0.4207	<u>0.5829</u>	0.6685	<u>0.4288</u>	<u>0.6674</u>

Results

- Context Range

- Including the user's next utterance is critical to predict user engagement (CUR -> NEXT)
- Previous history is not helpful, can be even detrimental (NEXT -> ALL)

Model	CUR_UTT	CUR	NEXT	PREV	ALL
Basic	0.3425	0.3503	0.3836	0.3501†‡	0.3963 †‡
Phrasal	0.3679	0.5521	0.5913	0.3709	0.5661†‡
Syntactic	0.3485	0.5530	0.6078	0.3671†‡	0.5867†‡
N-grams	0.3839	0.5694	0.6113	0.3788	0.5984†‡
Topic Model	0.2982	0.5255	0.5803	0.3464†‡	0.5829
Word2Vec	0.3704	0.5723	0.6162	0.3827†‡	0.6032†‡
Doc2Vec	0.3427	0.5379	0.5858	0.3722†‡	0.5740†‡
Skip-thought	0.3648	0.5545	0.6063	0.3692	0.6008†
CNN.Rand	<u>0.4252</u>	<u>0.5862</u>	0.6647	0.4153	0.6549†
CNN.MultiCh	0.4207	0.5829	0.6685	<u>0.4288</u>	<u>0.6674</u>

Results

- **Similarity Features**

- Similarities (cosine, jaccard distance etc.) between two adjacent user utterances (user_utt_i , user_utt_{i+1})
- Huge boost for detecting reformulations.

Model	w/o Similarity	w\Similarity
Basic	0.3836	0.4105 (+2.69%)
Phrasal	0.5913	0.6316 (+4.03%)†‡
Syntactic	0.6078	0.6280 (+2.02%)†‡
N-grams	0.6113	0.6573 (+4.60%)†‡
Topic Model	0.5803	0.6346 (+5.43%)†‡
Word2Vec	<u>0.6162</u>	0.6521 (+3.59%)†‡
Doc2Vec	0.5858	0.5968 (+1.10%)
Skip-thought	0.6063	0.6216 (+1.53%)

Failure Cases

- Manually examine 50 random wrongly-predicted examples from GHome dataset
 - **R**eformulation / **A**bandonment / **F**ulfillment / Continuation=44% / 42% / 8% / 6%
 - Common failure reasons:
 - i. (32%) model fails to understand the relevance of a system response to a user's request
 - ii. (30%) needs long-term contextual information
 - iii. (18%) model fails to understand reformulated utterances
 - "I want the stair lights" vs. "turn on the stair lights"

Future Work

- **Verify the effectiveness of proposed metrics**
 - User study (offline)
 - A/B testing (online, large-scale)
- **Improve automatic prediction**
 - Increase NLU with pretrained language models e.g. BERT
 - Pretraining with large amounts of un-annotated dialogues

Thank you!

rui.meng@pitt.edu
memray.me

Deep Keyphrase Generation

- Use neural networks to summarize keywords/keyphrases of a long text.
 - Deep Keyphrase Generation, ACL 2017
 - One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases, ACL 2020
 - An Empirical Study on Neural Keyphrase Generation, NAACL 2021
- Studies have shown its effectiveness for improving IR performance
 - By Florian Boudin & Ygor Gallina:
 - Keyphrase Generation for Scientific Document Retrieval
 - Redefining Absent Keyphrases and their Effect on Retrieval Effectiveness, arXiv 2103.12440

