山东大学 SHANDONG UNIVERSITY

When conversation meets information retrieval

Pengjie Ren (任鹏杰)

IRLab, Shandong University

renpengjie@sdu.edu.cn

https://pengjieren.github.io/

Joint work with

Jun Ma
Shandong University

Maarten de Rijke
University of Amsterdam

Evangelos Kanoulas
University of Amsterdam

Zhumin Chen
Shandong University

Pengjie Ren
Shandong University

Christof Monz
University of Amsterdam

Ming Zhou
Sinovation Ventures

Zhaochun Ren
Shandong University

Information retrieval

- **Technology to connect people to information**
  - Search engines
  - Recommender systems
  - Conversational Q&A
  - ……

Information goals

- **Navigational**, **informational**, and **resource** goals
  - Informational goals take up ~40–60%
  - More exploratory
    - When knowing little about the search target;
    - When wanting to know many aspects about the search target.

Table 1: The Search Goal Hierarchy. Queries are only assigned to leaf nodes. All examples are taken from actual AltaVista queries.

SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (I.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

Resource queries all represent a goal of obtaining something (other than information). If the resource is something I plan to use in the offline world, such as song lyrics, recipes, sewing patterns, etc., we call it an "**obtain**" goal. If the resource is something that needs to be installed on my computer or other electronic device to be useful, the goal is "**download**." If my goal is simply to experience (typically view or read) the resource for my enjoyment, the goal is "**entertain**." The most common example of queries with an entertainment goal were those that dealt with pornography. Finally, the "**interact**" goal occurs when the intended result of the search is a dynamic web service (such as a stock quote server or a map service) that requires further interaction to achieve the user's task.

The search goal framework described above proved to be both stable (requiring no major revisions as new queries were encountered) and comprehensive (encompassing the goals of all the queries we had seen). We were therefore able to move on to the second major task, associating goals with queries.

4. ASSOCIATING GOALS WITH QUERIES

There are two ways a search engine might associate goals with queries at runtime: either the user can identify the goal explicitly through the user interface, or the system can attempt to infer the goal automatically. Google's "I'm feeling lucky" feature [8], in which users implicitly identify their goal as "navigate to a specific web site," may be thought of as an early example of the first

15

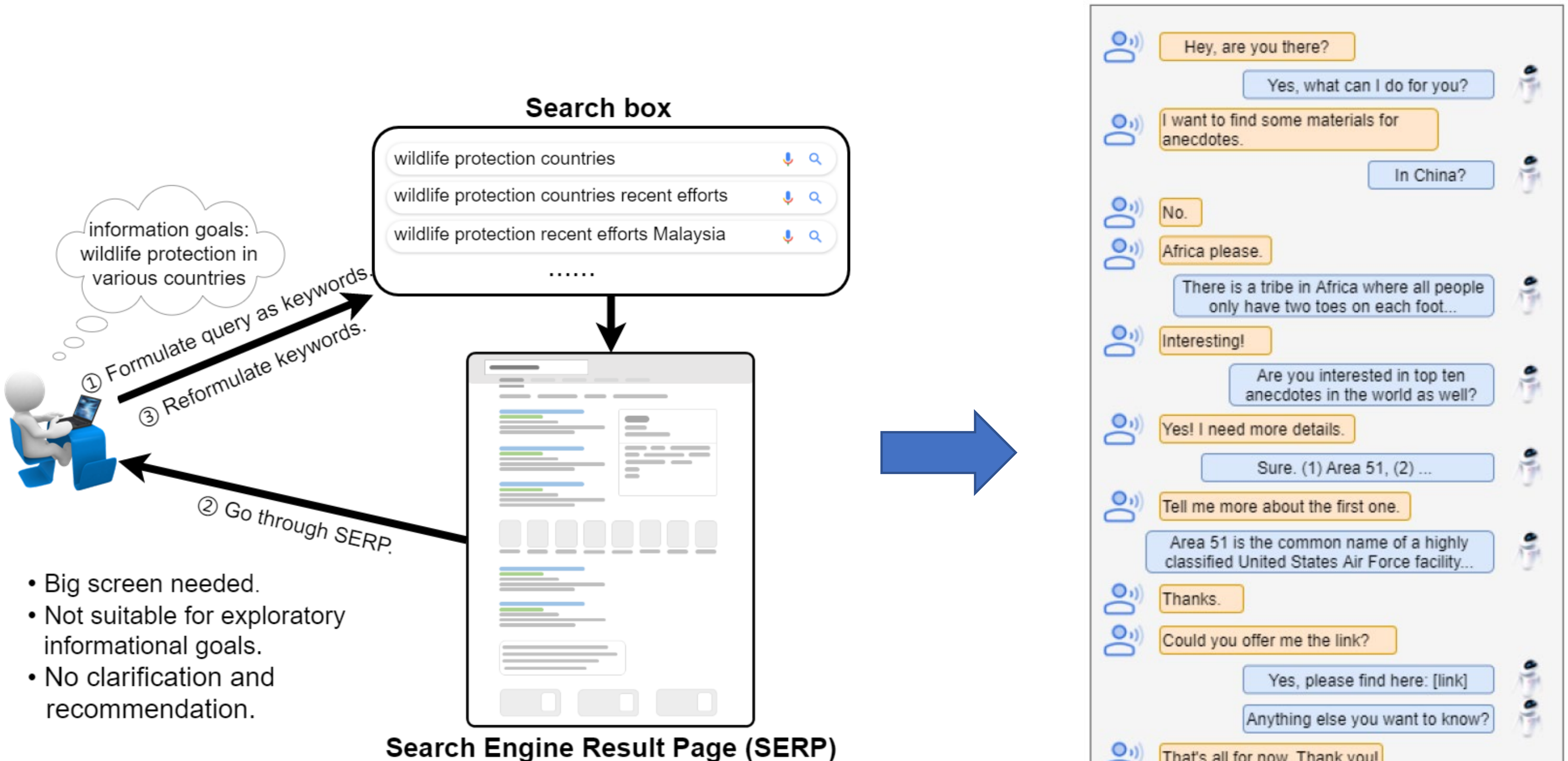
D.E. Rose and D. Levinson. Understanding User Goals in Web Search. In *WWW 2004*

Landscape is changing

- **More mobile queries**
  - At the start of 2019, over 60% of all queries submitted to Google were mobile
- **Spoken queries**
  - Exceeding 50% in some parts of the world
  - Spoken queries longer, sessions longer

https://www.statista.com/statistics/275814/mobile-share-of-organic-search-engine-visits/

Conversational search



- Big screen needed.
- Not suitable for exploratory informational goals.
- No clarification and recommendation.

Pengjie Ren et al. Conversations with Search Engines: SERP-based Conversational Response Generation. In TOIS 2021

Conversational recommender systems

New challenges

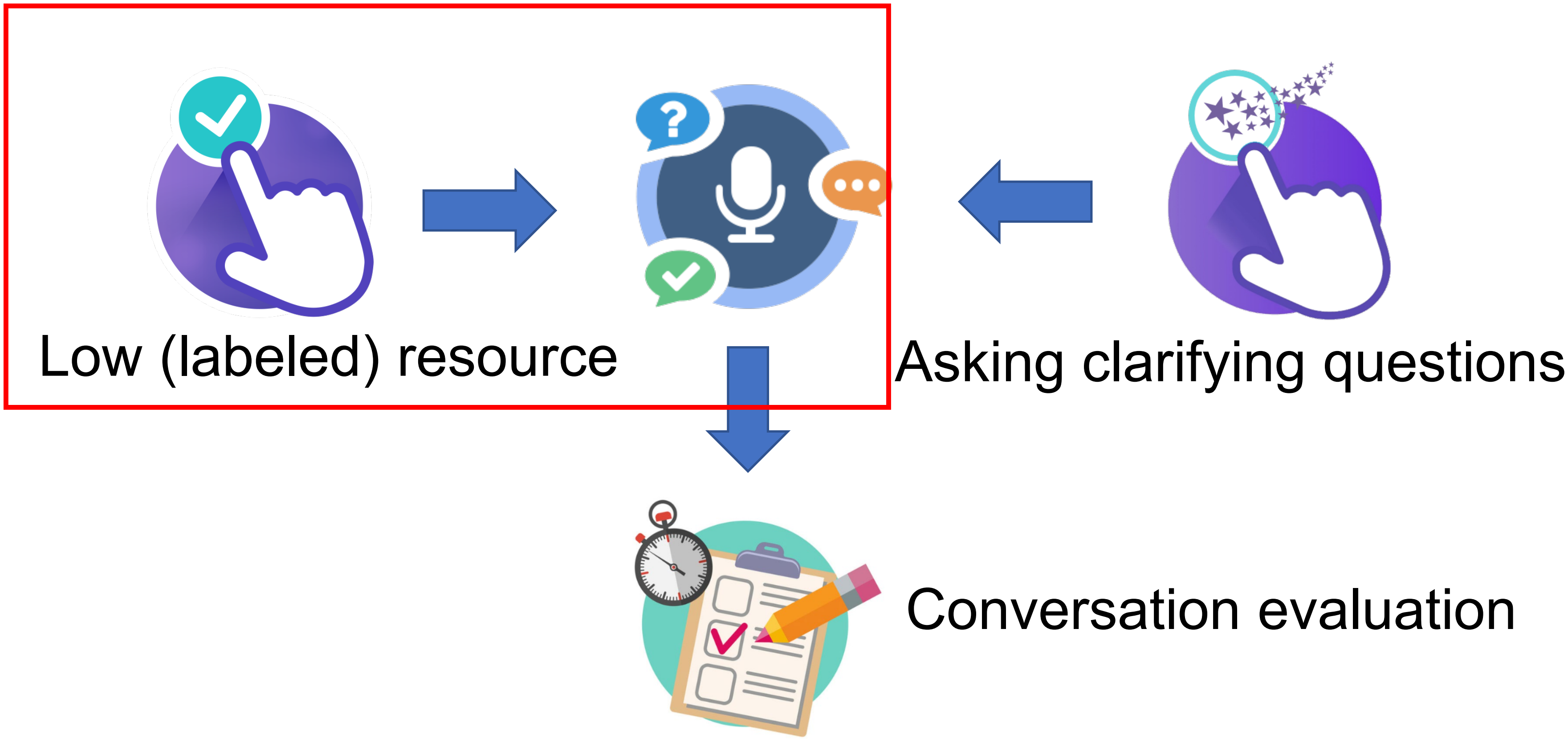
- Intents/actions increase
  - ✓Out-of-domain intents/actions
  - ✓Varying intent/action space
- Response presentation form
  - ✓Top n → Top 1
  - ✓Summary, steps, list, link, ...

……

- Cross-/Multi-Lingual conversations
  - ✓Leveraging available data better
- Multi-modal conversations
  - ✓Image, video, …
- Ethics control
  - ✓Safe AI

……

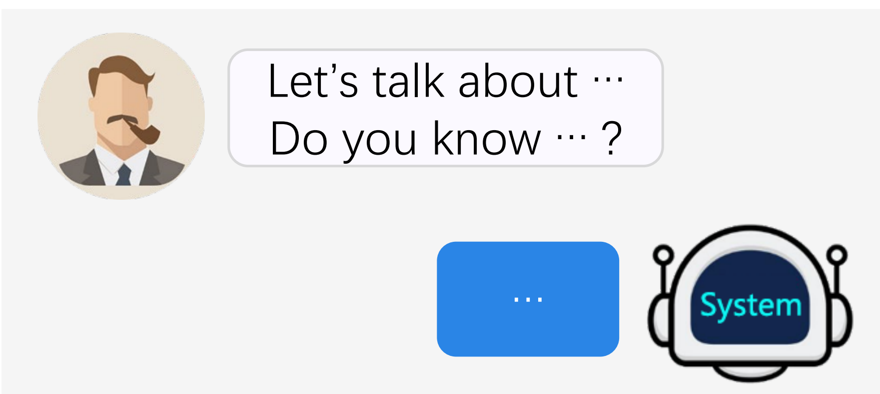
Three works we did in 2021



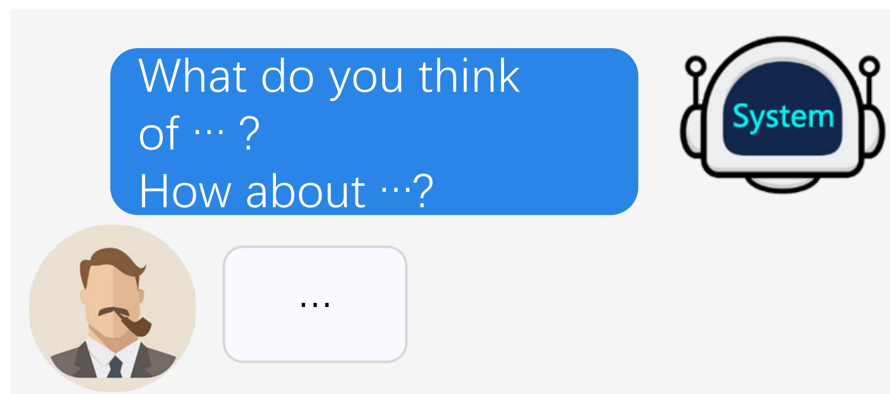
Mixed initiatives

Initiative is the ability to drive the direction of the conversation.
Mixed initiative is an intrinsic feature of human conversations.

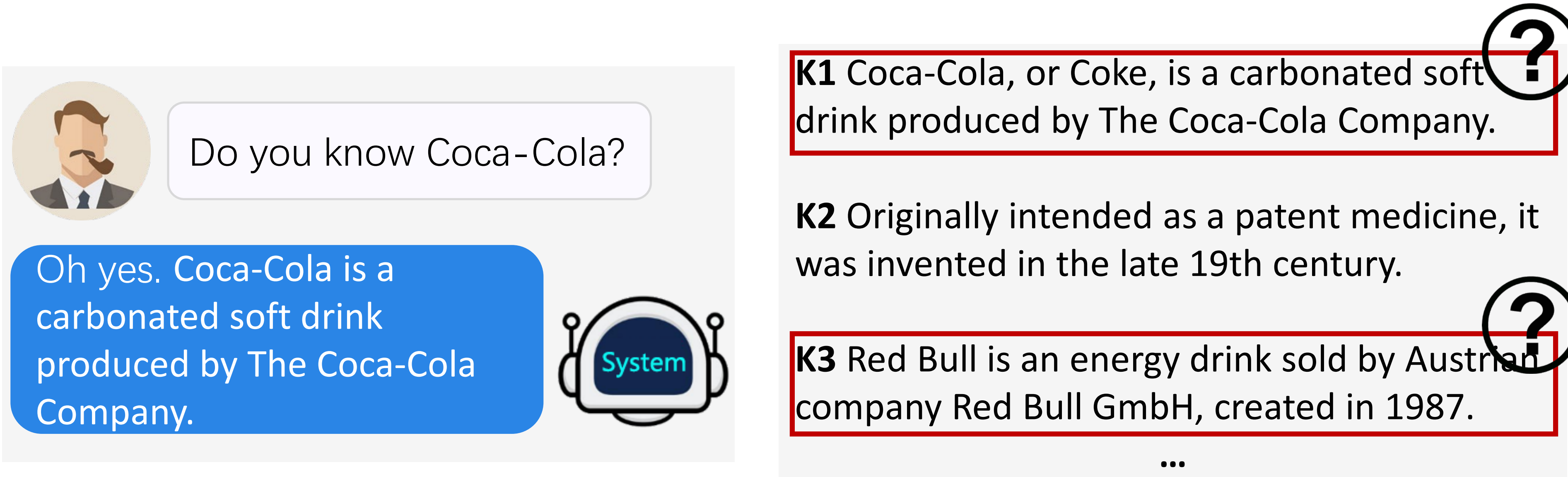
User Initiative



System Initiative



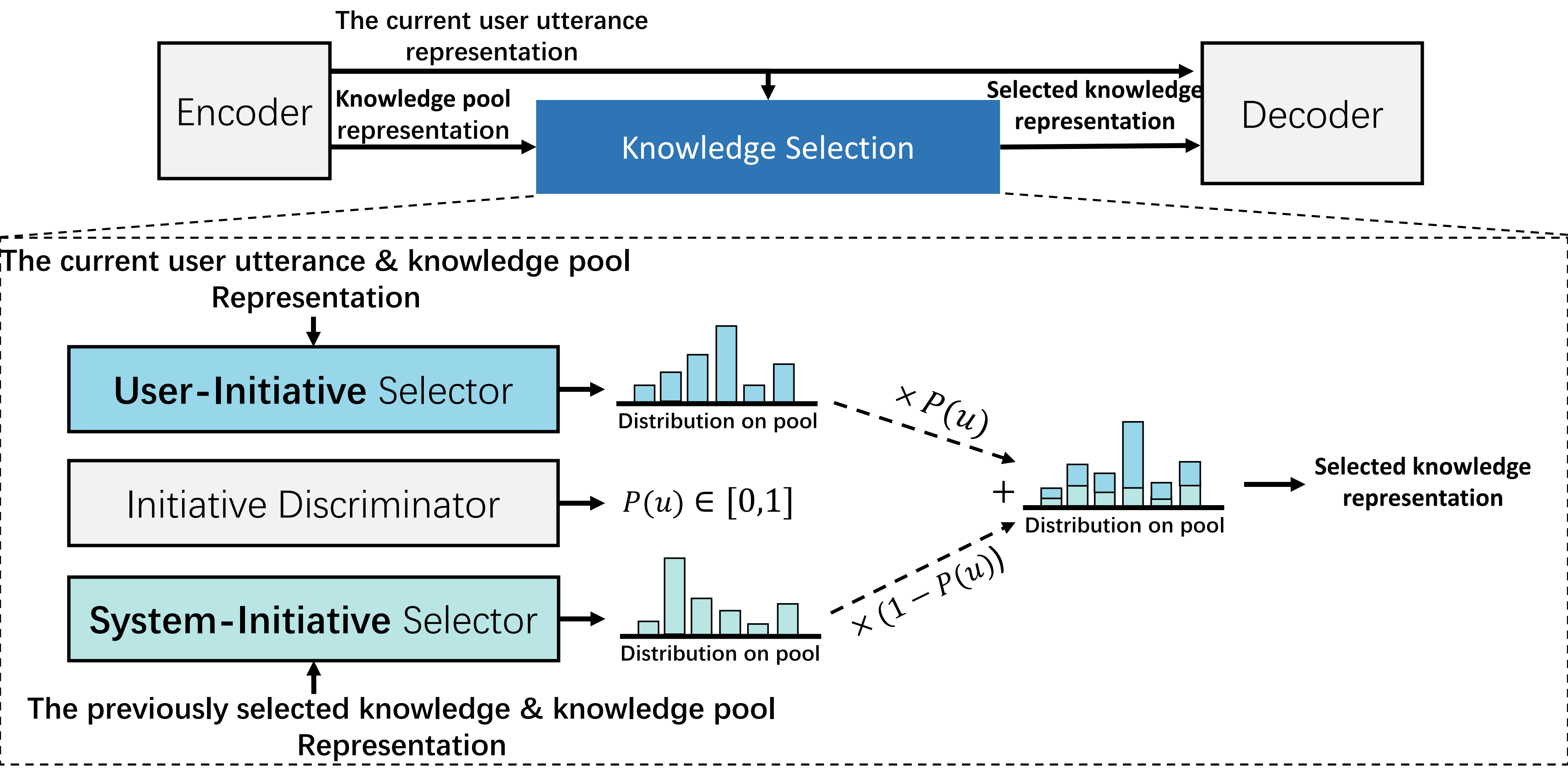
Low (labeled) resource



1. Online conversation data is unlabeled mostly.
2. Offline built data might not have the required labels.

How do we train mixed initiative systems without knowing the initiatives in training data?

Initiative-aware modeling



Chuan Meng et al. Initiative-Aware Self-Supervised learning for Knowledge-Grounded Conversations. In SIGIR 2021

Initiative-aware learning

Two hypotheses based on data:

1. If there is an **unsmooth knowledge shift** at the current turn, the current KS tends to be **user-initiative**, otherwise **system-initiative**.

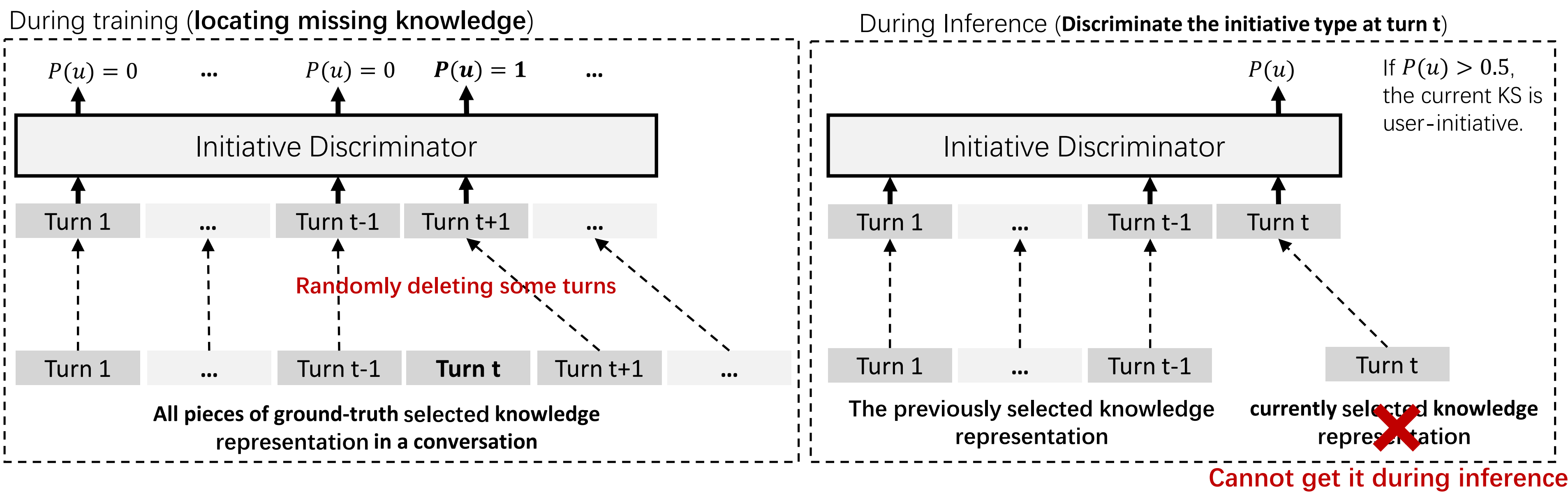
→ **(detecting the user-initiative ≈ detecting unsmooth knowledge shifts)**

2. If we remove some turns of a conversation, the adjacent conversation becomes **unsmooth**.

→ **(detecting unsmooth knowledge shifts ≈ locating missing knowledge)**

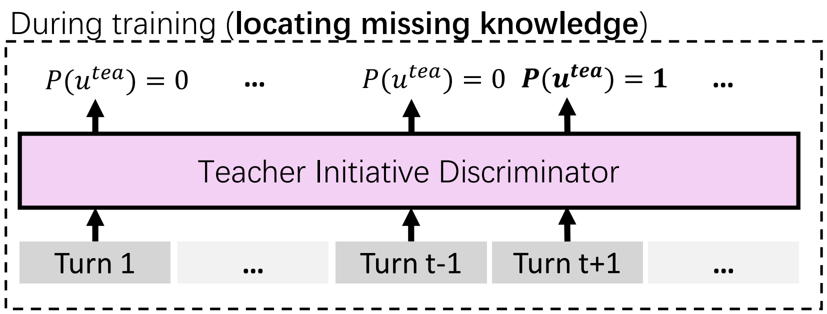
Coca-Cola, or Coke, is a carbonated soft drink produced by The Coca-Cola Company.	1st turn User-Initiative	Coca-Cola, or Coke, is a carbonated soft drink produced by The Coca-Cola Company.
↓ **Smooth knowledge shift**		↓
Originally intended as a patent medicine, it was invented in the late 19th century.	2nd turn	**Unsmooth knowledge shift**
↓ **Unsmooth knowledge shift**		↓
Red Bull is an energy drink sold by Austrian company Red Bull GmbH, created in 1987.	3rd turn User-Initiative	Red Bull is an energy drink sold by Austrian company Red Bull GmbH, created in 1987.

Initiative-aware learning

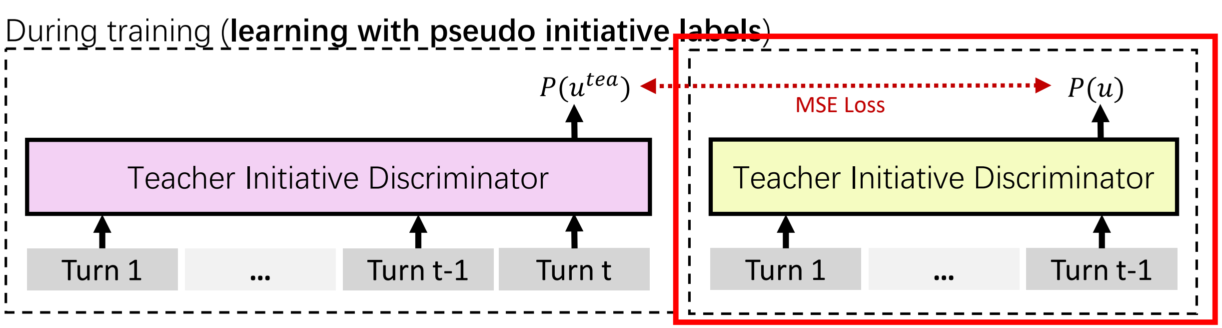


Incompatible issue: the knowledge currently selected cannot be fetched during inference.

Initiative-aware learning



- Further distinguish the initiative discriminator as **a teacher** and **a student initiative discriminator**
- Two tasks: **locating missing knowledge** and **learning with pseudo initiative labels**



Now it is compatible with inference.

Experiments

Methods	Test Seen (%)						Test Unseen (%)					
	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	R@1	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	R@1
PostKS + BERT	0.77	14.16	22.68	4.27	16.59	4.83	0.39	12.59	20.82	2.73	15.25	4.39
TMemNet + BERT	1.61	15.47	24.12	4.98	17.00	23.86	0.60	13.05	21.74	3.63	15.60	16.33
SKT	1.76	16.04	24.61	5.24	17.61	25.36	1.05	13.74	22.84	4.40	16.05	18.19
DiffKS + BERT	2.22	16.82	24.75	6.27	17.90	25.62	1.69	14.69	23.62	5.05	16.82	20.11
DukeNet	2.43	17.09	25.17	6.81	18.52	26.38	1.68	15.06	23.34	5.29	17.06	19.57
SKT+PIPM+KDBTS	2.47	17.14	25.19	7.01	18.47	27.40	1.71	14.83	23.56	5.46	17.14	20.20
MIKe (ours)	**2.78***	**17.76***	**25.40**	**7.11**	**18.78***	**28.41***	**2.00***	**15.64***	**23.78***	**5.61**	**17.41***	**21.47***

Wizard of Wikipedia dataset (R@1 denotes Recall@1)

Methods	Single golden reference (%)						Multiple golden references (%)					
	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	R@1	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	R@1
PostKS + BERT	6.54	19.30	28.94	9.89	22.15	3.95	8.49	23.97	32.85	13.10	26.17	6.40
TMemNet + BERT	8.99	24.48	31.65	13.24	25.90	28.44	12.36	28.61	35.29	16.14	29.51	37.30
SKT	17.81	29.41	35.28	21.74	30.06	28.99	24.69	35.78	41.68	28.30	36.24	39.05
DiffKS + BERT	19.08	30.87	36.37	22.88	31.30	29.39	26.20	37.32	42.77	29.57	37.53	38.99
DukeNet	19.15	30.93	36.53	23.02	31.46	30.03	26.83	37.73	43.18	30.13	38.03	40.33
SKT+PIPM+KDBTS	20.07	31.07	36.78	24.29	31.70	30.80	27.49	37.34	43.07	30.91	37.82	40.70
MIKe (ours)	**21.14***	**32.28***	**37.78**	**25.31***	**32.82***	**31.86***	**28.52***	**38.55***	**44.06**	**31.92***	**38.91***	**41.78***

Holl-E dataset (R@1 denotes Recall@1)

- According to Recall@1, MIKe significantly outperforms all baselines in terms of knowledge selection.
- According to BLEU-4, METEOR and ROUGE-1/2/L, MIKe significantly outperforms all baselines in terms of response generation.

Experiments

Methods	Test Seen (%)									Test Unseen (%)								
	Appropriateness			Informativeness			Engagingness			Appropriateness			Informativeness			Engagingness		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
MIKe vs DiffKS + BERT	32	59	9	18	76	6	26	62	12	27	67	6	19	77	4	24	64	12
MIKe vs DukeNet	27	64	9	18	75	7	22	65	13	30	66	4	18	74	8	24	61	15
MIKe vs SKT+PIPM+KDBTS	25	67	8	17	78	5	20	69	11	29	66	5	19	76	5	25	62	13

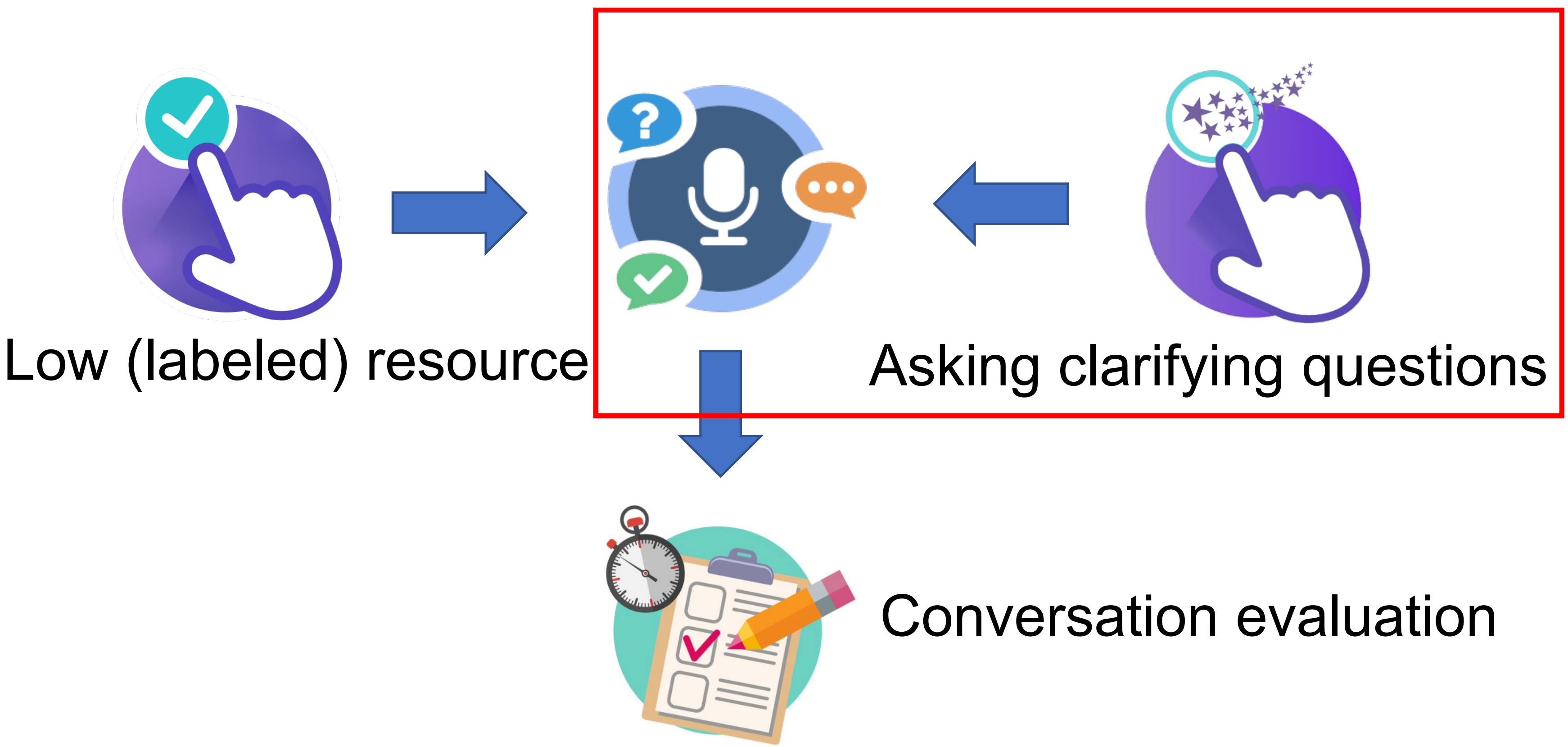
- MIke achieves the best performance in terms of all metrics compared to the three most competitive baselines.
- The results are consistent with the automatic evaluation results.

Experiments

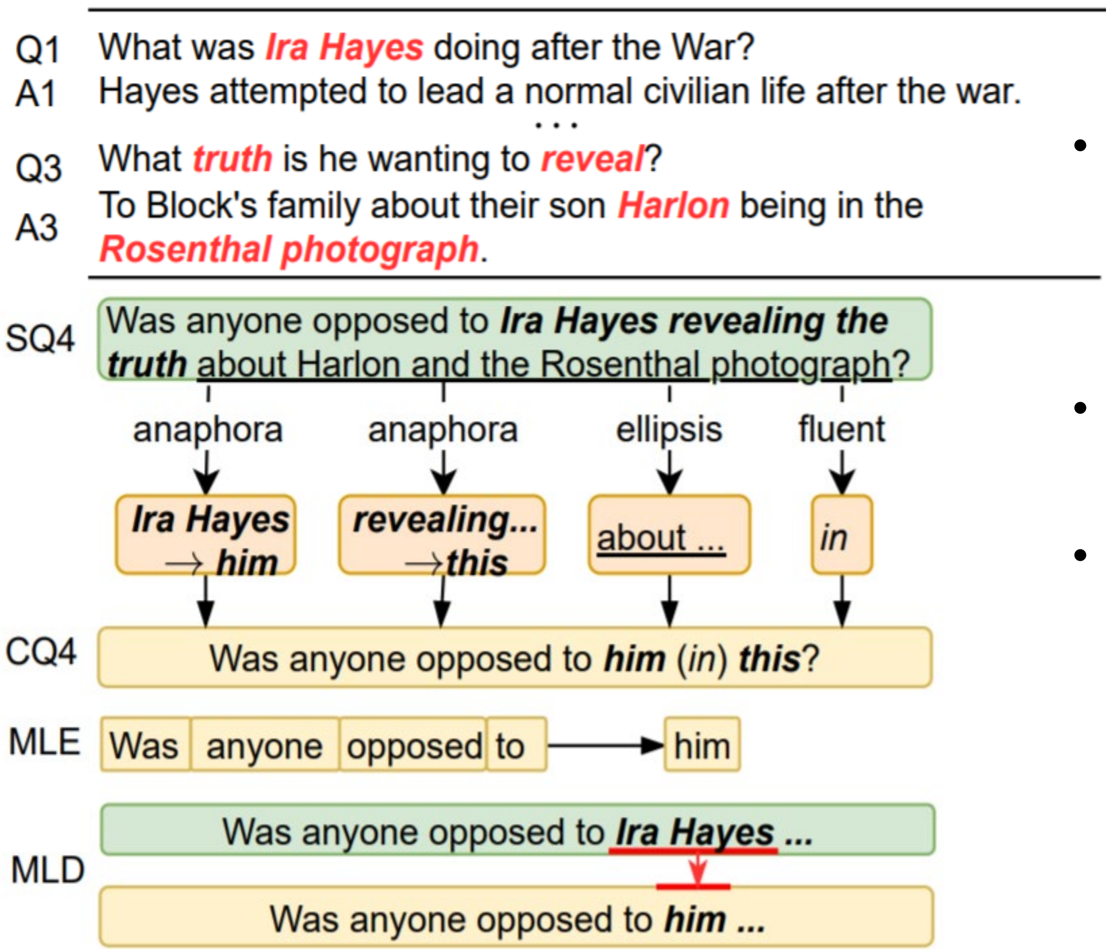
	Example 1 (Test seen)
Knowledge pool	K_1: no knowledge used . K_2: while basketball is most often played as a team sport with five players on each side , two-on-two , and one-on-one competitions are also common . K_3: basketball is a limited contact sport played on a rectangular court . K_4: jordan played 15 seasons in the nba for the chicago bulls and washington wizards . …
Context	User: are you a basketball fan ? System: (K_2) yes , i am a fan of the five player sport . are you ? User: not as much as i used to be . i watched the bulls in the 90s when they were the dream team . (the current user utterance)
Initiative type	User-initiative KS
Response	DiffKS + BERT: (K_3 ✗) i do know that basketball is a limited contact sport played on a rectangular court . DukeNet: (K_2 ✗) i agree . i like to play basketball . i like the sport with five players on each side . SKT+PIPM+KDBTS: (K_2 ✗) i ' m not sure but i know that while basketball is most played as a team sport with five players . MIKe: (K_4 ✓) i know that jordan played 15 seasons in the nba for the chicago bulls and washington wizards .

- MIKe identifies the current turn as user-initiative and then selects the knowledge about "Jordan."

Three works we did in 2021



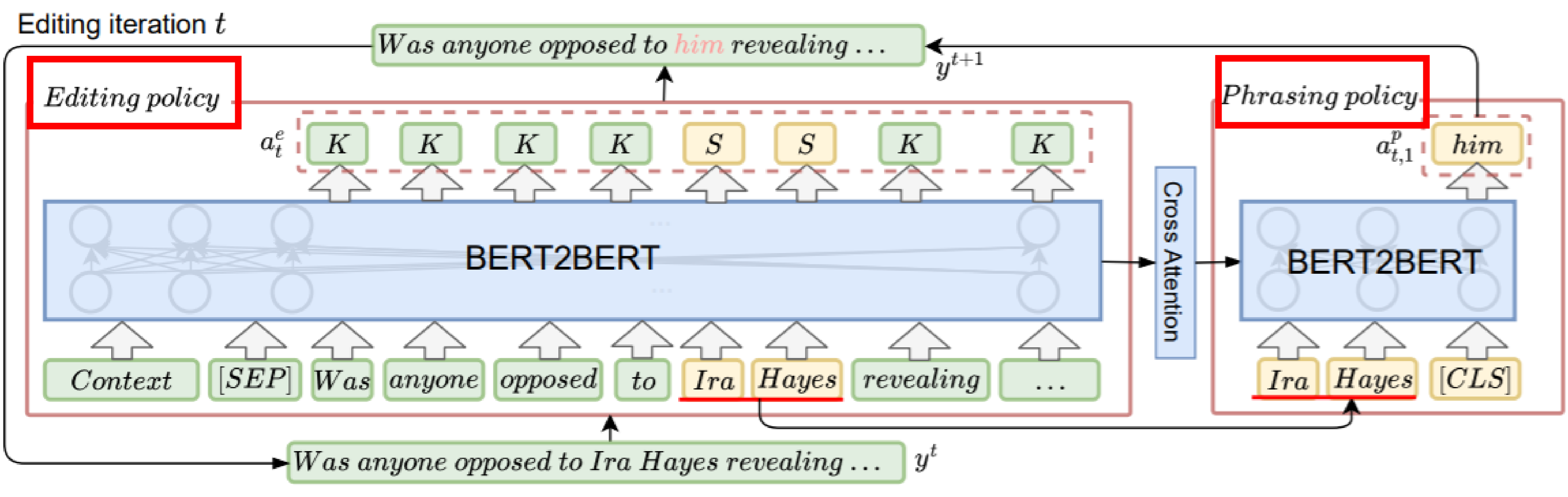
Asking clarifying questions



- Asking clarifying questions is one of the most important characteristics of mixed initiatives.
- Pure generation vs. Retrieval + Reranking + Rewriting
- MLE gives equal attention to generate each question token, stuck in easily learned tokens, i.e., tokens appearing in input, ignoring conversational tokens, e.g., him, which is a small but important portion of output.

Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

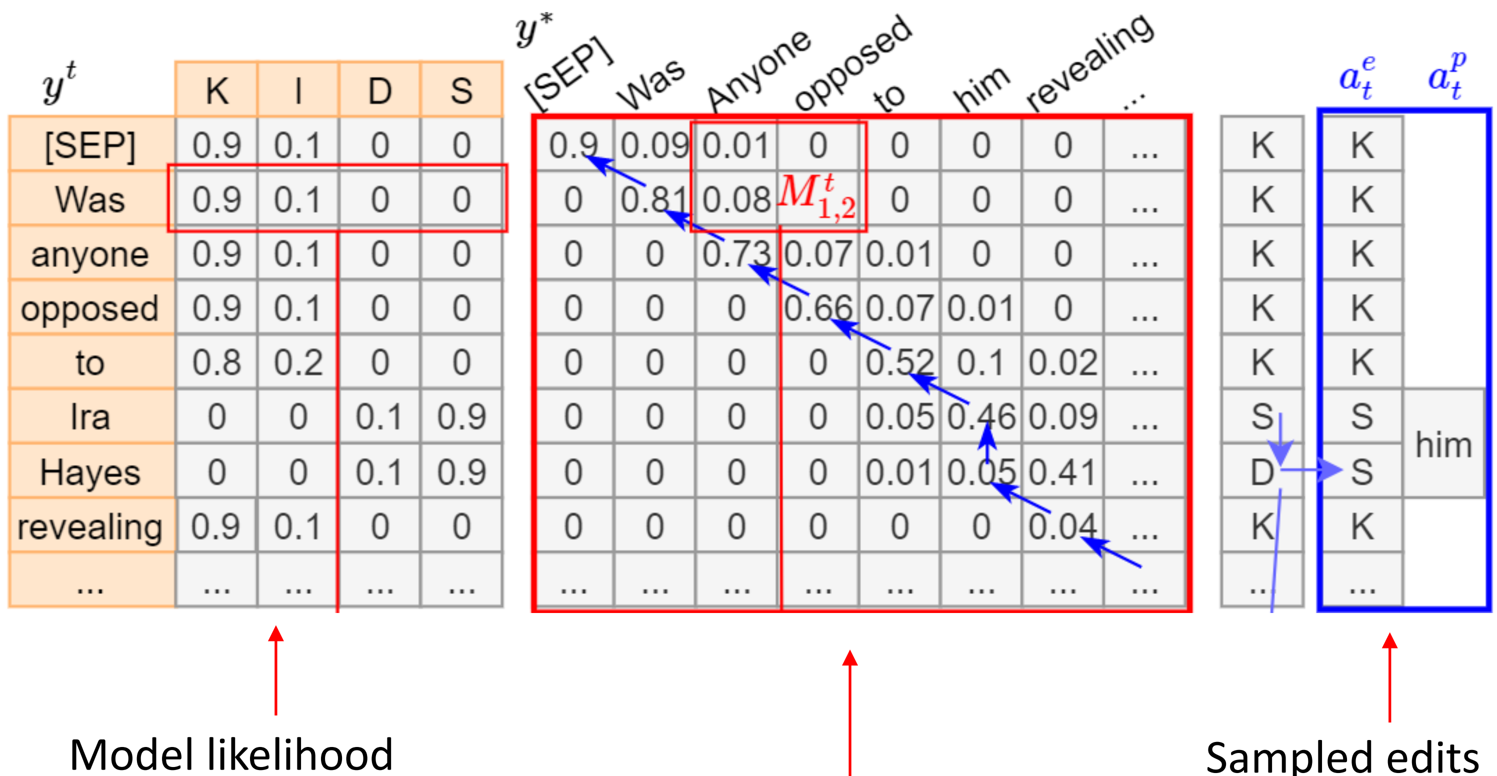
Iterative sequence editing



Four edits: ‘K’: keep, ‘D’: delete, ‘I’: insert, ‘S’: substitute.

Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

Dynamic programming based sampling



M^t: $M^t_{i,j}$ tracks the expectation of converting $y^t_{:i}$ to $y^*_{:j}$

Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

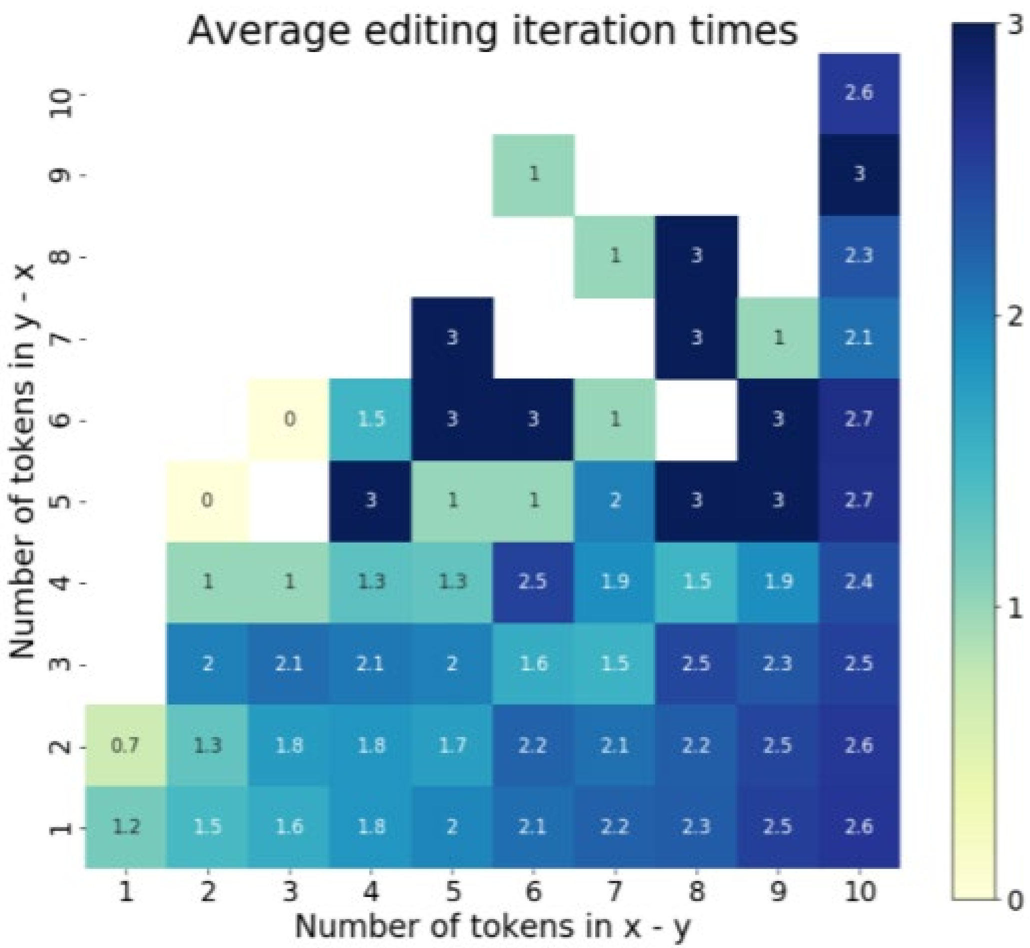
Experiments

	CANARD (%)						CAsT (%) (unseen)					
Method	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**
Origin	54.7	47.0	40.6	35.3	70.9	3.460	75.9	69.2	62.9	57.6	85.0	5.946
Rule	55.0	47.0	40.2	34.8	70.5	3.420	78.0	71.4	65.3	60.0	86.1	6.220
Trans++	84.3	77.5	72.1	67.5	84.6	6.348	76.0	64.3	54.8	47.2	76.5	4.258
QGDiv	85.2	78.6	73.3	68.9	85.2	6.469	75.9	65.3	56.7	59.6	78.0	4.694
QuerySim	83.1	78.5	74.5	71.0	82.7	6.585	80.6	75.3	70.2	65.5	83.3	6.345
RISE	**86.3***	**80.5***	**75.6**	**71.6***	**86.2***	**6.759**	**85.1***	**78.4**	**72.2**	**66.8**	**87.8***	**6.543**

Results on CANARD and CAsT.

- ✓ RISE has a better ability to emphasize conversational tokens, rather than treating all tokens equally.
- ✓ RISE is more robust, which generalizes better to unseen data of CAsT.

Experiments



- ✓ As the number of different tokens between x and y increases, the number of editing iterations increases too.

Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

Experiments

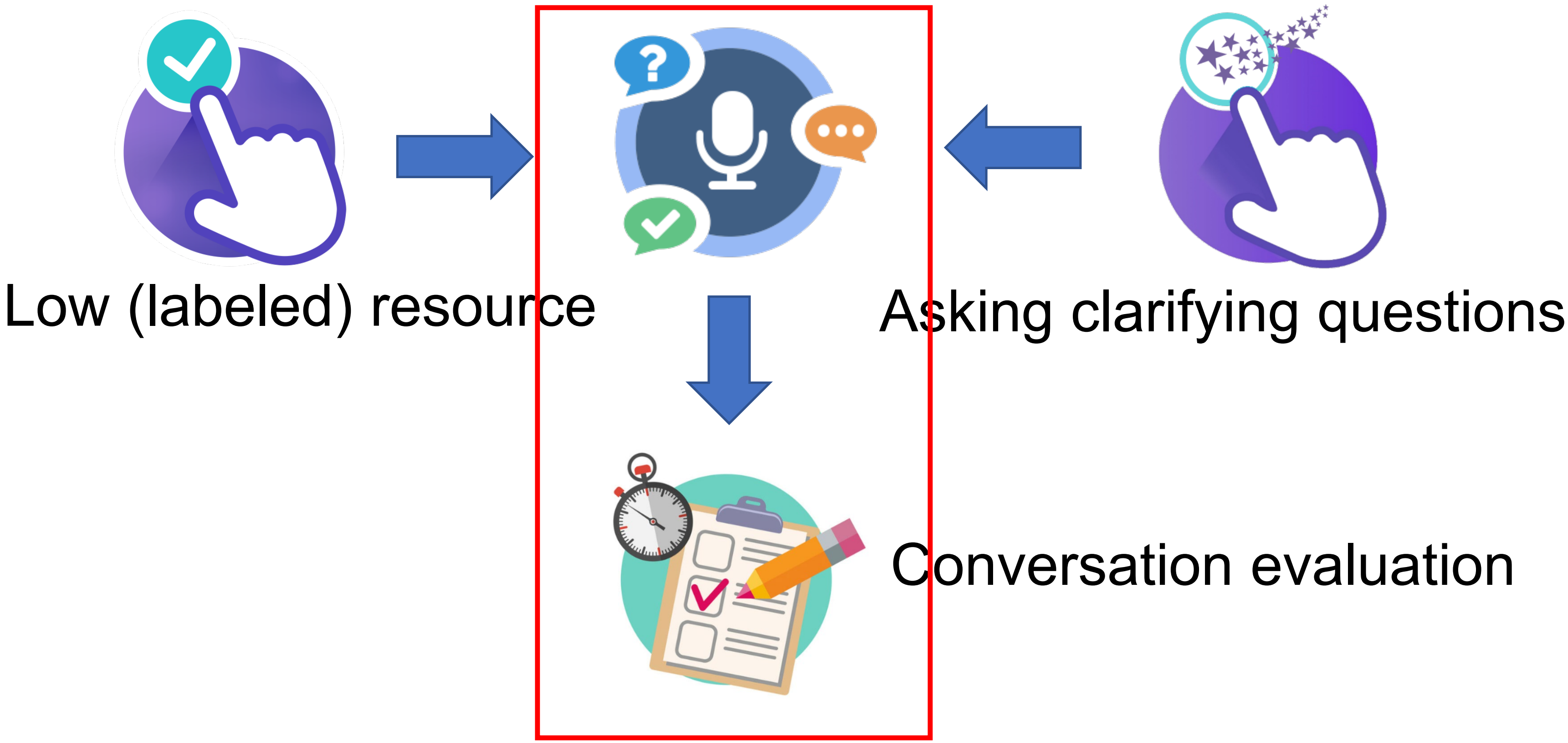
Example 1 Context	1. At Tabuk the standard of the army was entrusted to Abu Bakr. 2. Where was Tabuk located? 3. Tabuk on the Syrian border.
Question	What did Abu Bakr do during the expedition of Tabuk?
Rewrite#1	What did he bakr do during expedition?
Rewrite#2	What did he do during expedition?
Target	What did abu bakr do during the expedition?

Example 2 Context	1. When did Clift start his film career? 2. His first movie role was opposite John Wayne in Red River, which was shot in 1946 and released in 1948.
Question	Did Montgomery Clift win any awards for any of his films?
Rewrite#1	Did he win any awards for and?
Rewrite#2	Did he win any awards?
Target	Did he win any awards for any of his films?

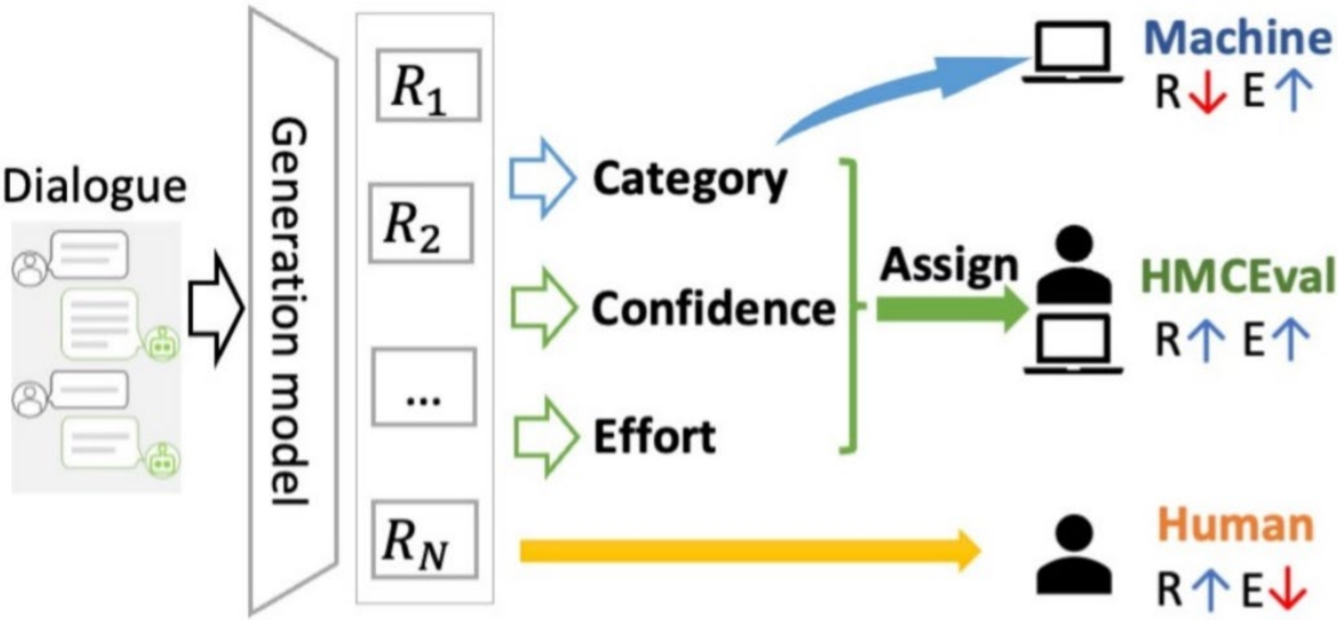
- It is helpful to edit iteratively.
- RISE can generate more conversational questions than human sometimes.

Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

Three works we did in 2021



Conversation evaluation



- ✓ Automatic Evaluation: Efficient but not reliable usually.
- ✓ Human Evaluation: Mostly reliable but not efficient.

Yangjun Zhang et al. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

Sample assignment execution

Sample Assignment Execution (SAE)

$$\max \sum_{i=1}^{M} \hat{a}_i z_i + \sum_{i=1}^{M} b_i (1 - z_i),$$

$$\min \sum_{i=1}^{M} k_i z_i + \sum_{i=1}^{M} \hat{l}_i (1 - z_i),$$

$$z_i = \begin{cases} 0, & \text{sample } i \text{ is assigned to a human;} \\ 1, & \text{sample } i \text{ is assigned to machine.} \end{cases}$$

M The number of all samples.

$\hat{a}_i$ The model confidence for evaluating sample i.

b_i The human confidence for evaluating sample i.

k_i The machine effort for evaluating sample i.

$\hat{l}_i$ The human effort for evaluating sample i.

Sample assignment execution

Sample Assignment Execution (SAE)

$$\max \left[\sum_{i=1}^{M} \hat{a}_i z_i + \sum_{i=1}^{M} b_i (1 - z_i) - \lambda \left(\sum_{i=1}^{M} k_i z_i + \sum_{i=1}^{M} \hat{l}_i (1 - z_i) \right) \right],$$

subject to

$$\sum_{i=1}^{M} z_i \geq M - N$$

$$b_i = 1 \text{ for } i = 1, \ldots, M$$

$$k_i = 0 \text{ for } i = 1, \ldots, M$$

$$\lambda \geq 0.$$

N The number of samples assigned to human.

(a) The number of samples assigned to a human is less than or equal to N.

(b) Human confidence is assumed to be 1.

(c) Machine effort is assumed to be 0.

(d) λ is to balance confidence and effort.

Yangjun Zhang et al. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

Model confidence estimation

Model Confidence Estimation (MCE)

- Maximum Class Probability (MCP)
  - Use the classification probabilities to measure the confidence.
- Trust Score (TS)
  - Estimate whether the predicted category of a test sample by a classifier can be trusted, i.e., the ratio between the Hausdorff distance from the sample to the non-predicted and the predicted categories.
- True Class Probability (TCP)
  - Similar to TS, except that the estimation is obtained by a learning-based method, BERT + ConfidNet.

Yangjun Zhang et al. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021
Heinrich Jiang et al. To Trust or Not to Trust a Classifier. In NIPS 2018
Charles Corbiere et al. Addressing Failure Prediction by Learning Model Confidence. In NIPS 2019

Human effort estimation

Human Effort Estimation (HEE)

- Use time cost, i.e., the time spent for each annotation, to represent human effort.
- Use random forest regression to estimate the time cost.
- Dialogue related features
  - total turns, malevolent turns, non-malevolent turns, first submission or not, paraphrased turns, total length, FK score (readability), DC score (readability), contains malevolent turn or not, perplexity score…
- Worker related features
  - worker test score, approval rate ranking…

Experiments

Metric	Machine	Human	HMCEval
Reliability			
Precision	0.818	1	0.983
Recall	0.803	1	0.976
F1-score	0.810	1	0.980
Accuracy	0.862	1	0.985
Efficiency			
Human ratio	0	1	0.500
Time cost	0	1	0.500

N/M=0.5

HMCEval achieves around 99% evaluation accuracy with half of the human effort spared.

Yangjun Zhang et al. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

References

- **Zhongkun Liu, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Maarten de Rijke and Ming Zhou. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. The 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.**
- **Yangjun Zhang, Pengjie Ren and Maarten de Rijke. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. The 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.**
- **Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi and Maarten de Rijke. Initiative-Aware Self-Supervised learning for Knowledge-Grounded Conversations. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.**
- Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren and Maarten de Rijke. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.
- Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen and Maarten de Rijke. Conversations Powered by Cross-Lingual Knowledge. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.
- Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, Maarten de Rijke. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Transactions on Information Systems (TOIS), 2021.
- Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Maarten de Rijke. Conversations Powered by Cross-Lingual Knowledge. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.
- Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, Maarten de Rijke.Few-Shot Variational Reasoning for Medical Dialogue Generation. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2021.

Thank you for your attention!

Pengjie Ren
renpengjie@sdu.edu.cn
https://pengjieren.github.io/