Conversations with Search Engines

Pengjie Ren
IRLab, Shandong University

renpengjie@sdu.edu.cn

Joint work with 👇 who have a PhD 😂

Maarten de Rijke
University of Amsterdam
Ahold Delhaize

Jun Ma
Shandong University

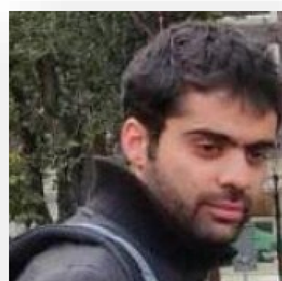
Evangelos Kanoulas
University of Amsterdam

Zhumin Chen
Shandong University

Ming Zhou
Sinovation Ventures

Christof Monz
University of Amsterdam

Pengjie Ren
Shandong University



Nikos Voskarides
University of Amsterdam

Zhaochun Ren
Shandong University

Information retrieval

- **Technology to connect people to information**
  - Search engines
  - Recommender systems
  - Conversational assistants

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/

Conversations with Search Engines

- Idea of **search as conversation** has been around since early 1980s.

- Making information retrieval interfaces feel more natural and convenient for their users.

- Ongoing research and development efforts heavily skewed towards:

  - ✓ Task-oriented dialogue systems
  - ✓ Question answering systems
  - ✓ Social bots
  - ✓ Question clarification
  - ✓ User studies
  - ✓ Theoretical/Conceptual frameworks

But there’s more …

Information goals

- **Navigational**, **informational**, and **resource** goals
  - Informational consistently ~40–60% of all goals
  - 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.

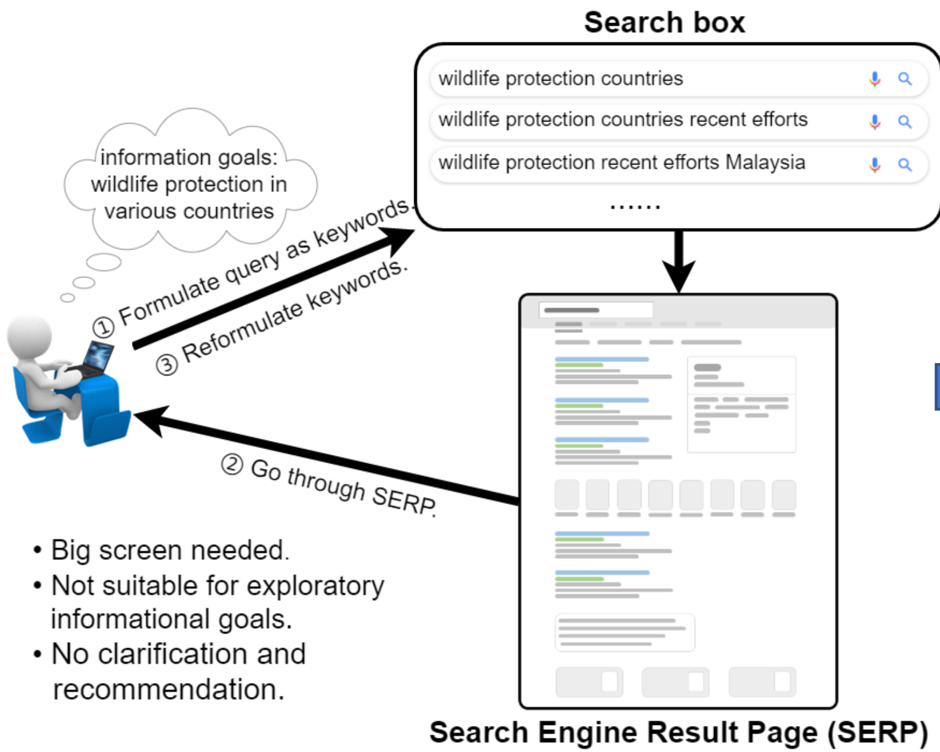
SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	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

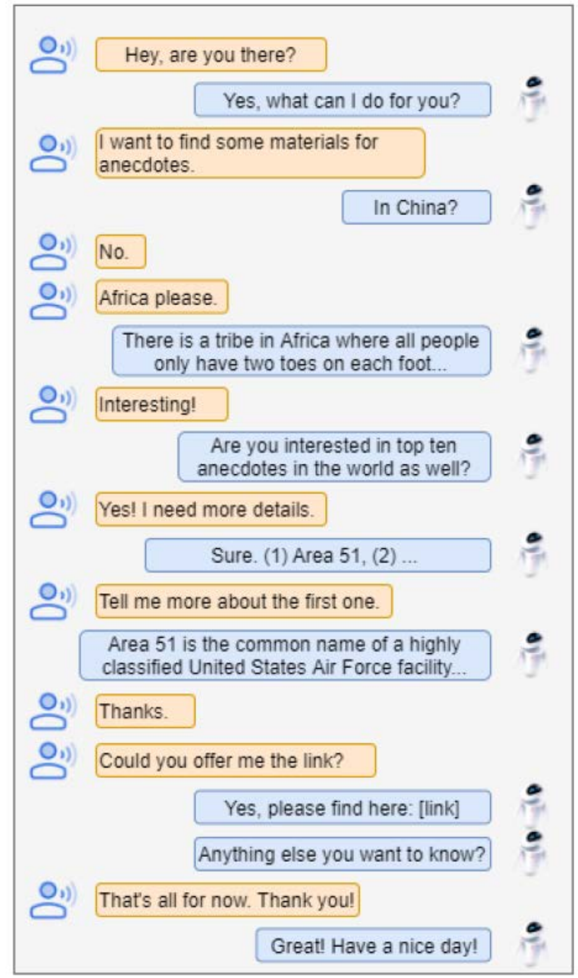
4. ASSOCIATING GOALS WITH QUERIES

15

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

Conversations with Search Engines





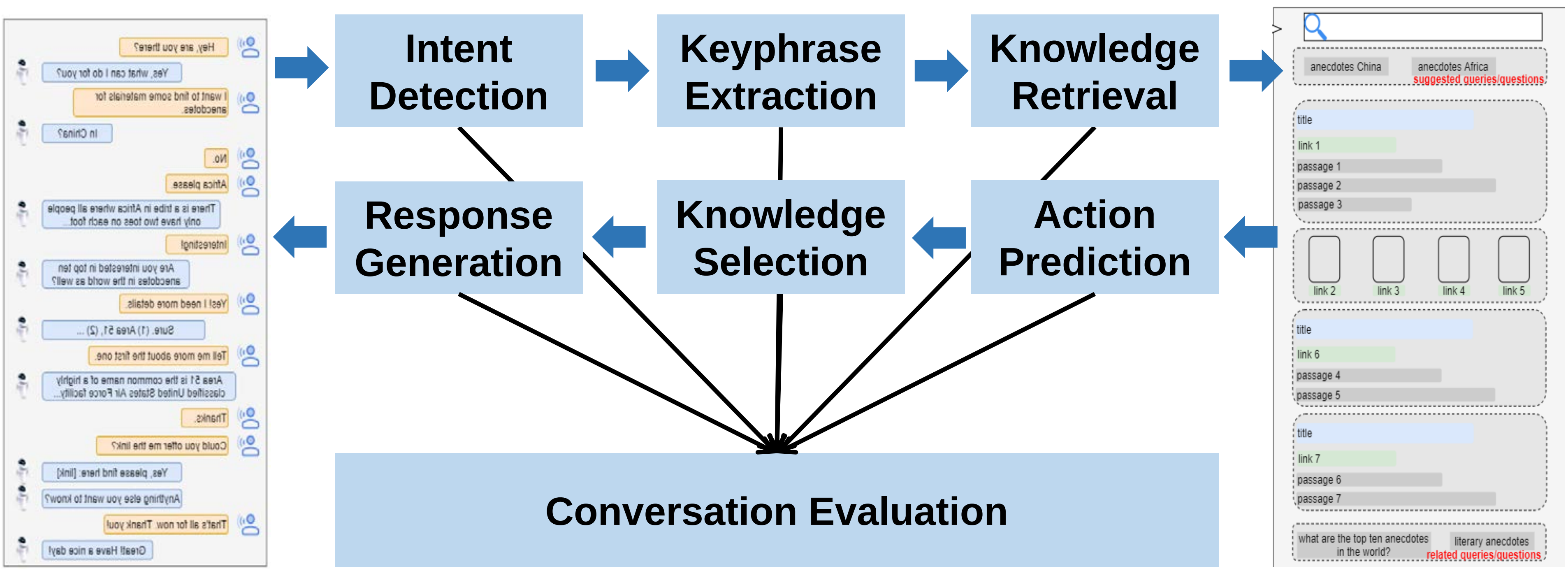
"Search with Search Engines" **"Conversations with Search Engines"**

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

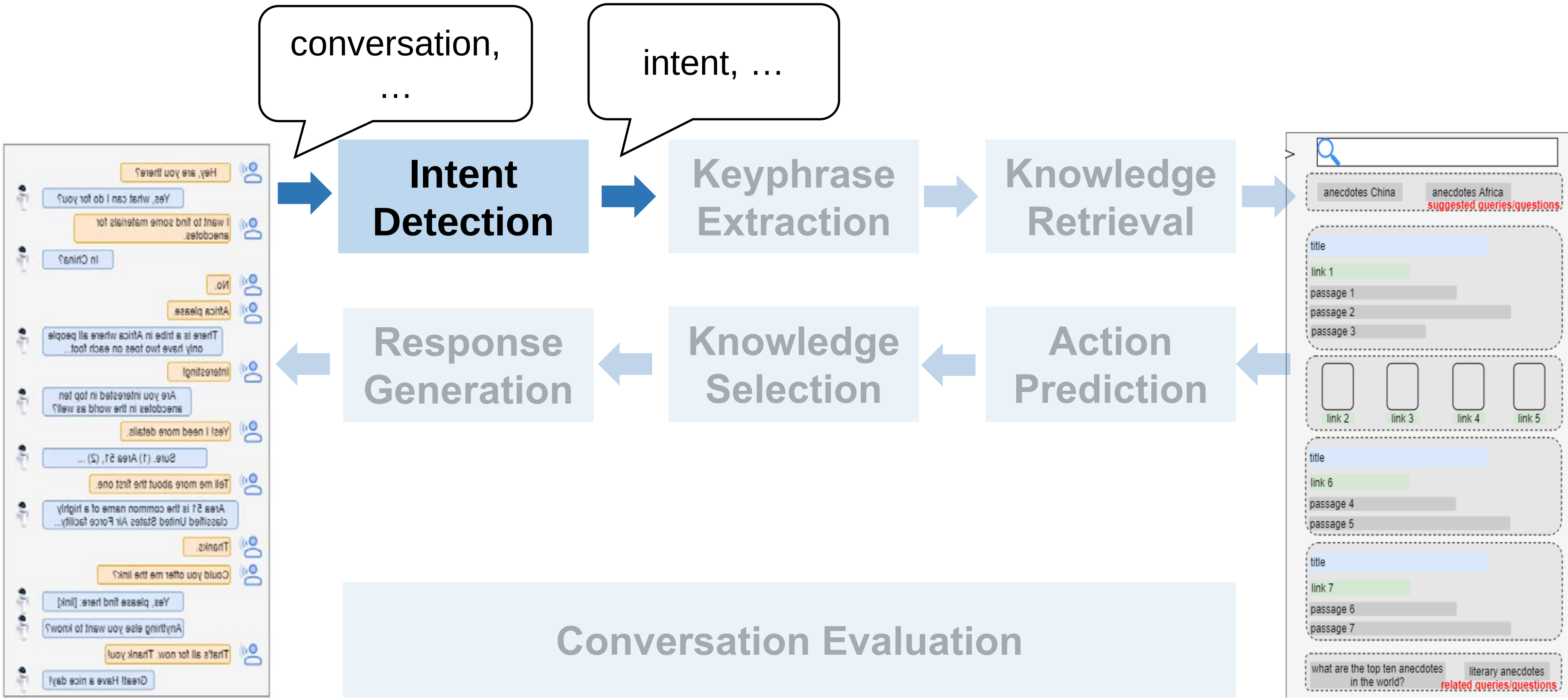
Conversations with Search Engines

- As our mode of interaction changes, how can we support information seeking through **conversations with search engines**?

Conversations with Search Engines



Conversations with Search Engines



Intent Detection

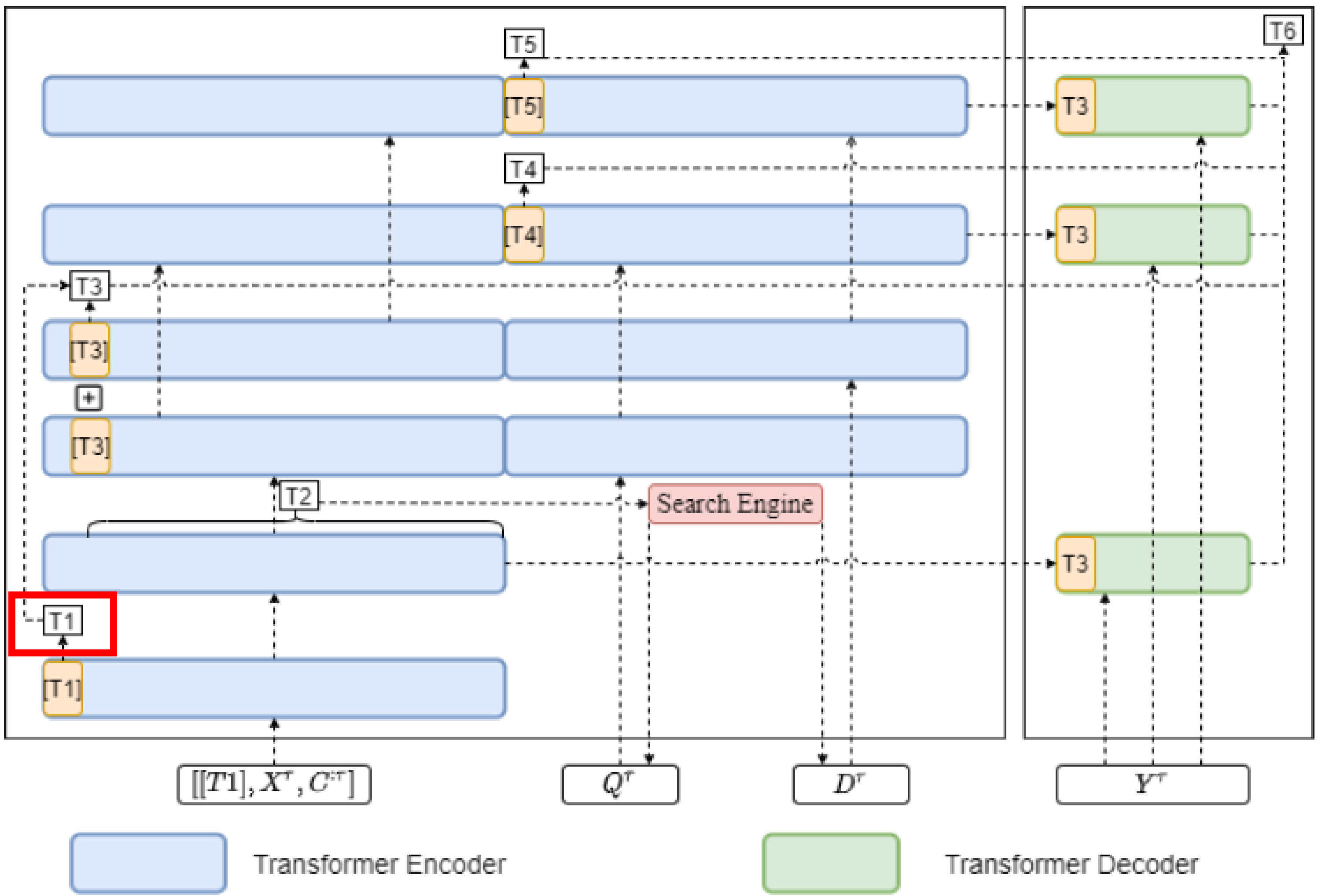
Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Intent Detection

Intent	Explanation	Example	TSE operations
reveal	Reveal a new intent, or refine an old intent proactively.	User: I want to see a movie. (reveal) User: Can you tell me more about it? (reveal)	Issue a new query.
revise	Revise an intent proactively when there is wrong expression, e.g., grammatical issues, unclear expression.	User: Tell me some non-diary milks. User: I mean dairy not diary. (revise)	Revise the query.
interpret	Interpret or refine an intent by answering a clarification question from the system.	User: Do you know The Avengers? System: Do you mean the movie, novel or game? User: The movie (interpret)	Select suggested queries.
request-rephrase	Request the system to rephrase the response if it is not understandable.	Sorry, I didn't get it. (request-rephrase)	–
chitchat	Greetings or other utterances that are not related to the information need.	I see. (chitchat) Are you there? (chitchat)	–

Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Intent Detection



Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Intent Detection

	ID (%)		
	P	R	F1
-ID	–	–	–
-KE	**66.2**	28.2	30.6
-AP	52.5	32.2	**35.3**
-QS	51.9	**32.6**	32.6
-PS	51.3	30.8	32.8
WISE	45.2	32.5	34.1

Results of joint learning.

	ID (%)		
	P	R	F1
test (unseen)	38.4	28.5	29.3
test (seen)	**48.3**	**36.1**	**37.4**
test	45.2	32.5	34.1

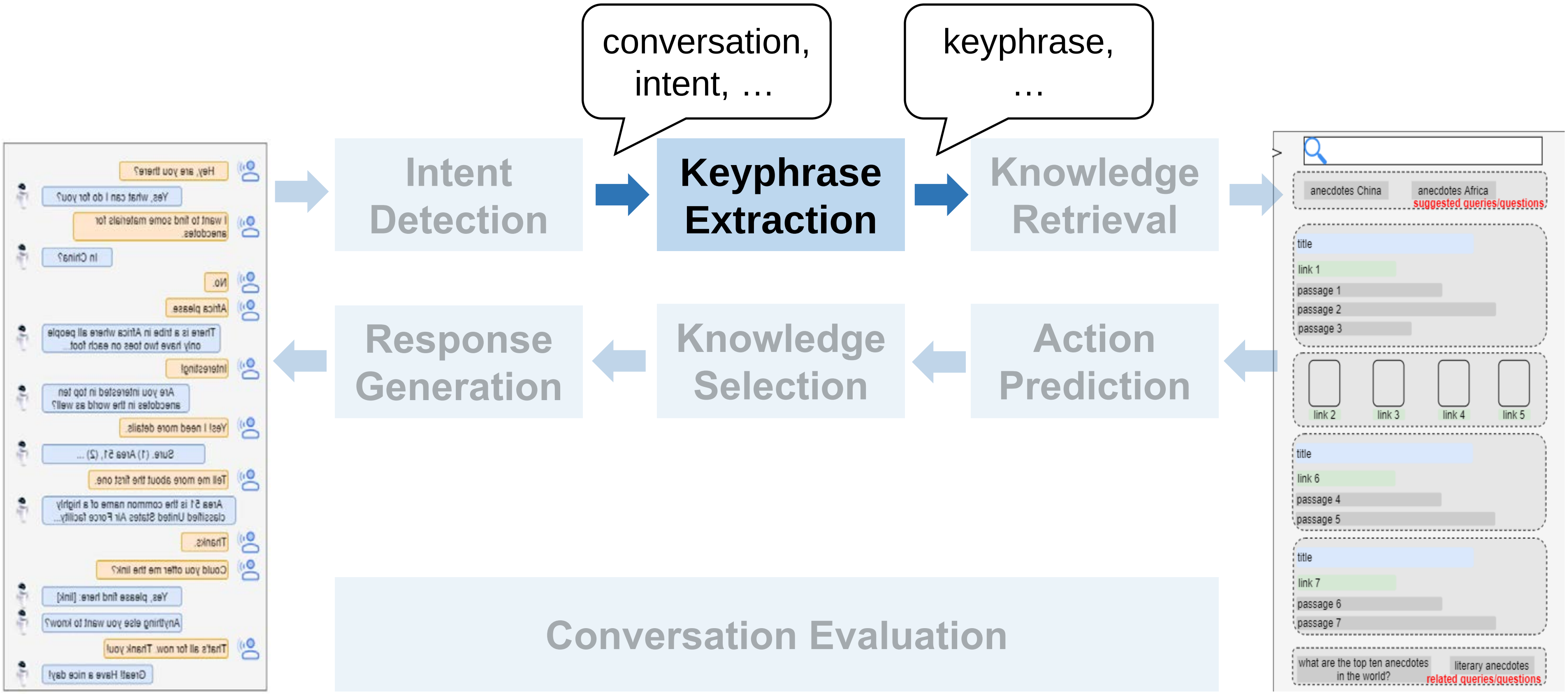
Results on seen/unseen data.

	ID (%)		
	P	R	F1
-DuReader	**47.5**	25.7	27.7
-KdConv	41.1	27.7	28.1
-DuConv	43.9	**35.5**	**35.8**
-WebQA	39.0	30.6	32.0
WISE	45.2	32.5	34.1

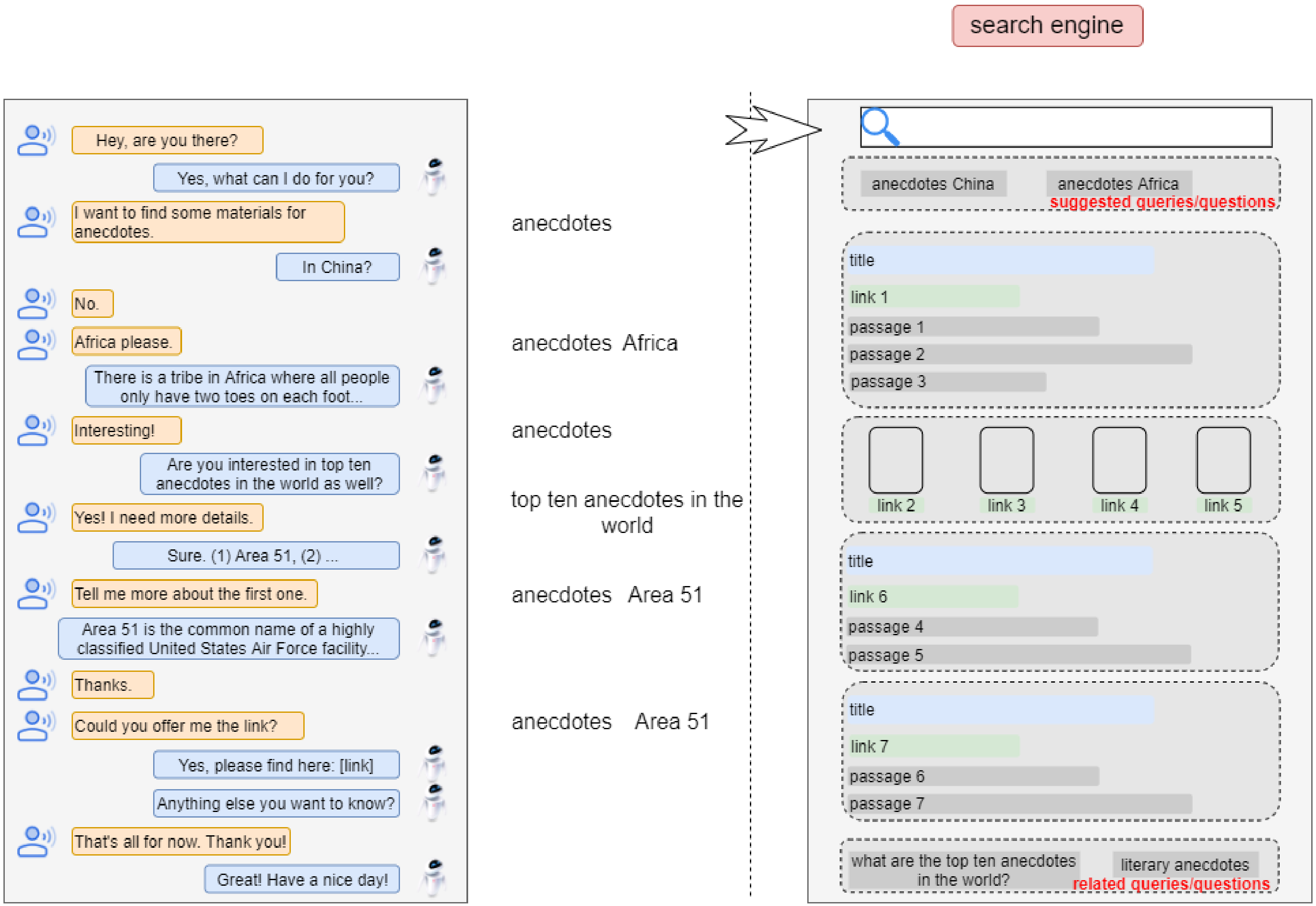
Results with different pretraining data.

- ✓ The joint learning tasks seem incompatible with the current architecture.
- ✓ Better performance on seen data.
- ✓ Not all pretraining data is helpful for ID performance.

Conversations with Search Engines



Keyphrase Extraction



Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Keyphrase Extraction

Turn	Query
1	who formed **saosin**?
2	when was the **band** founded?
3	what was their **first** album?
4	when was the album released? *resolved:* when was saosin 's first album released?

Relevant passage to turn #4: The original lineup for **Saosin**, consisting of Burchell, Shekoski, Kennedy and Green, was formed in the summer of 2003. On June 17, the **band** released their **first** commercial production, the EP Translating the Name.

- ✓ Keyphrase extraction bridges the gap between traditional search engines with conversational search.
- ✓ Labelling keyphrase is label intensive.

Nikos Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020

Keyphrase Extraction

Turn	Query
1	who formed **saosin**?
2	when was the **band** founded?
3	what was their **first** album?
4	when was the album released?
	resolved: when was saosin 's first album released?

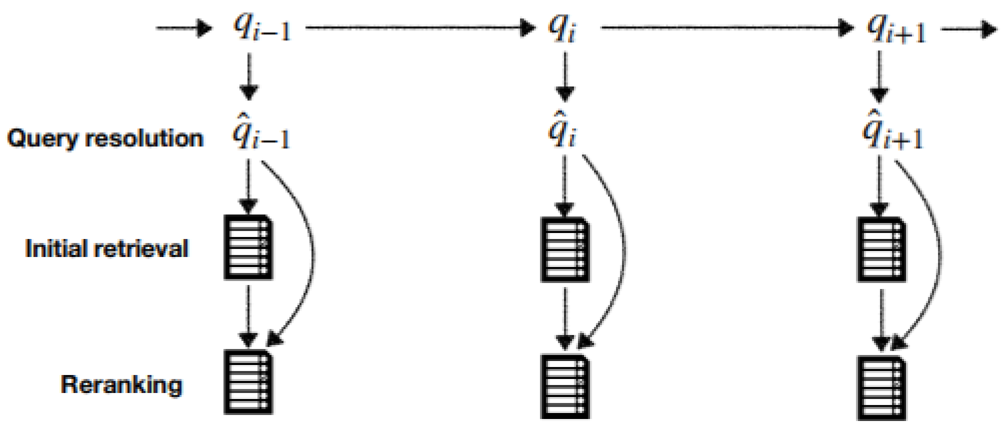
Relevant passage to turn #4: The original lineup for **Saosin**, consisting of Burchell, Shekoski, Kennedy and Green, was formed in the summer of 2003. On June 17, the **band** released their **first** commercial production, the EP Translating the Name.

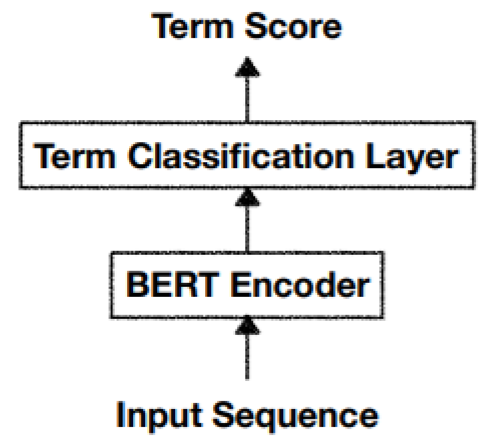
- ✓ Keyphrase extraction bridges the gap between traditional search engines with conversational search.
- ✓ Labelling keyphrase is label intensive.

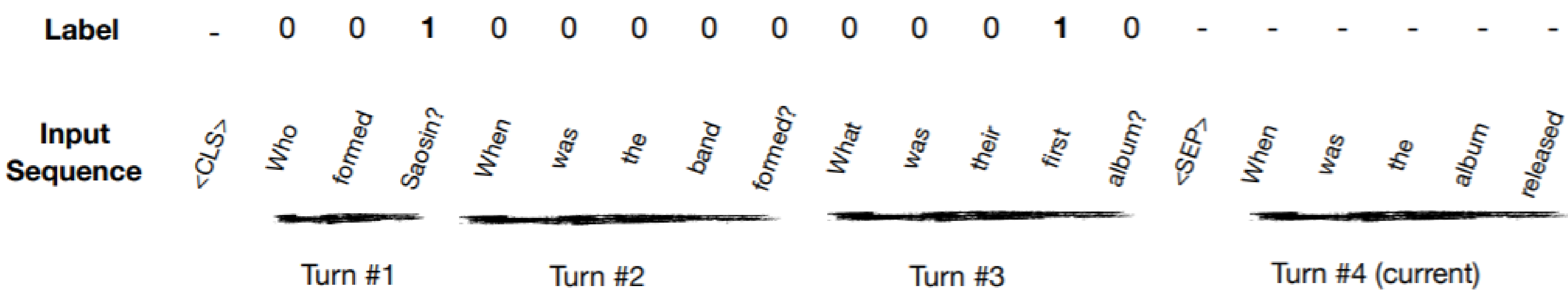
So how we address keyphrase extraction in an unsupervised/weak-supervised/self-supervised manner?

Nikos Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020

Keyphrase Extraction

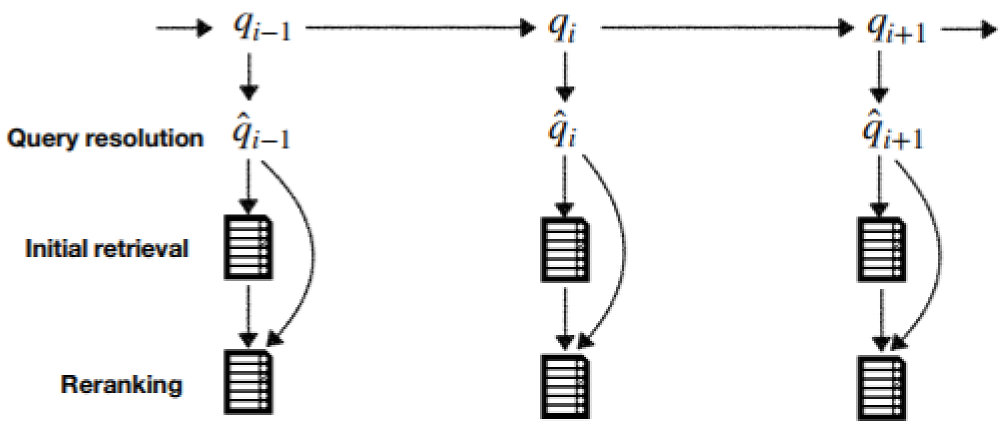


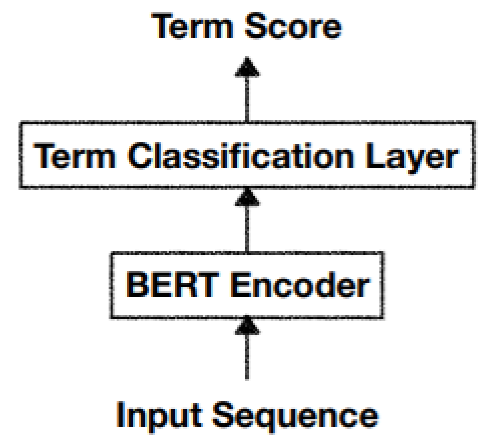




Nikos Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020

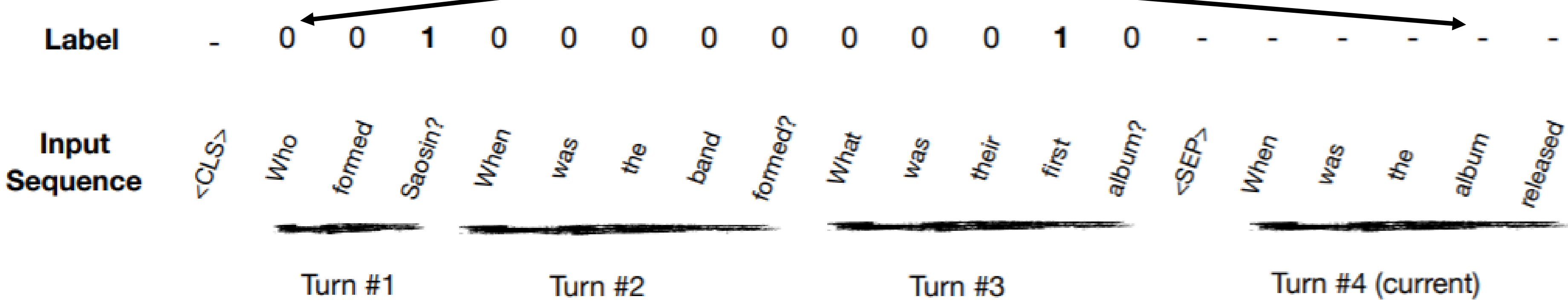
Keyphrase Extraction





Sim (Context, Passage, Answer)

Distant supervision



Nikos Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020

Keyphrase Extraction

Method	P	R	F1
Original (cur+prev)	22.3	46.4	30.1
Original (cur+first)	41.1	49.5	44.9
Original (all)	12.3	**100.0**	21.9
NeuralCoref	65.5	30.0	41.2
BiLSTM-copy	67.0	53.2	59.3
QuReTeC	**71.5**	66.1	**68.7**

Intrinsic evaluation - results on QuAC.

Method	P	R	F1
Original (cur+prev)	32.5	43.9	37.4
Original (cur+first)	43.0	74.0	54.4
Original (all)	18.6	**100.0**	31.4
RM3 (cur)	35.8	8.3	13.5
RM3 (cur+prev)	34.6	32.5	33.5
RM3 (cur+first)	40.9	32.9	36.5
RM3 (all)	41.5	38.8	40.1
NeuralCoref	**83.0**	28.7	42.7
BiLSTM-copy	51.5	36.0	42.4
QuReTeC	77.2	79.9	**78.5**

Intrinsic evaluation - results on CAsT.

- ✓ QuReTeC outperforms all the variations of Original and the baselines.
- ✓ Original (all) has perfect recall but at the cost of very poor precision.
- ✓ QuReTeC generalizes well to CAsT (even though it was only trained on QuAC).

Keyphrase Extraction

Method	Recall	MAP	MRR	NDCG@3
Original (cur)	0.438	0.129	0.310	0.155
Original (cur+prev)	0.572	0.181	0.475	0.235
Original (cur+first)	0.655	0.214	0.561	0.282
Original (all)	0.694	0.190	0.552	0.256
RM3 (cur)	0.440	0.140	0.320	0.158
RM3 (cur+prev)	0.575	0.200	0.482	0.254
RM3 (cur+first)	0.656	0.225	0.551	0.300
RM3 (all)	0.666	0.195	0.544	0.266
Nugget	0.426	0.101	0.334	0.145
QCM	0.392	0.091	0.317	0.127
NeuralCoref	0.565	0.176	0.423	0.212
BiLSTM-copy	0.552	0.171	0.403	0.205
QuReTeC	**0.754**▲	**0.272**▲	**0.637**▲	**0.341**▲
Oracle	0.785	0.309	0.660	0.361

Extrinsic evaluation – retrieval results on CAsT.

- ✓ QuReTeC outperforms all the baselines achieving performance close to Oracle.
- ✓ Nugget and QCM perform poorly, which indicates that session search is different in nature than conversational search.
- ✓ BiLSTM-copy performs poorly, which means that it does not generalize well to CAsT.

Nikos Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020

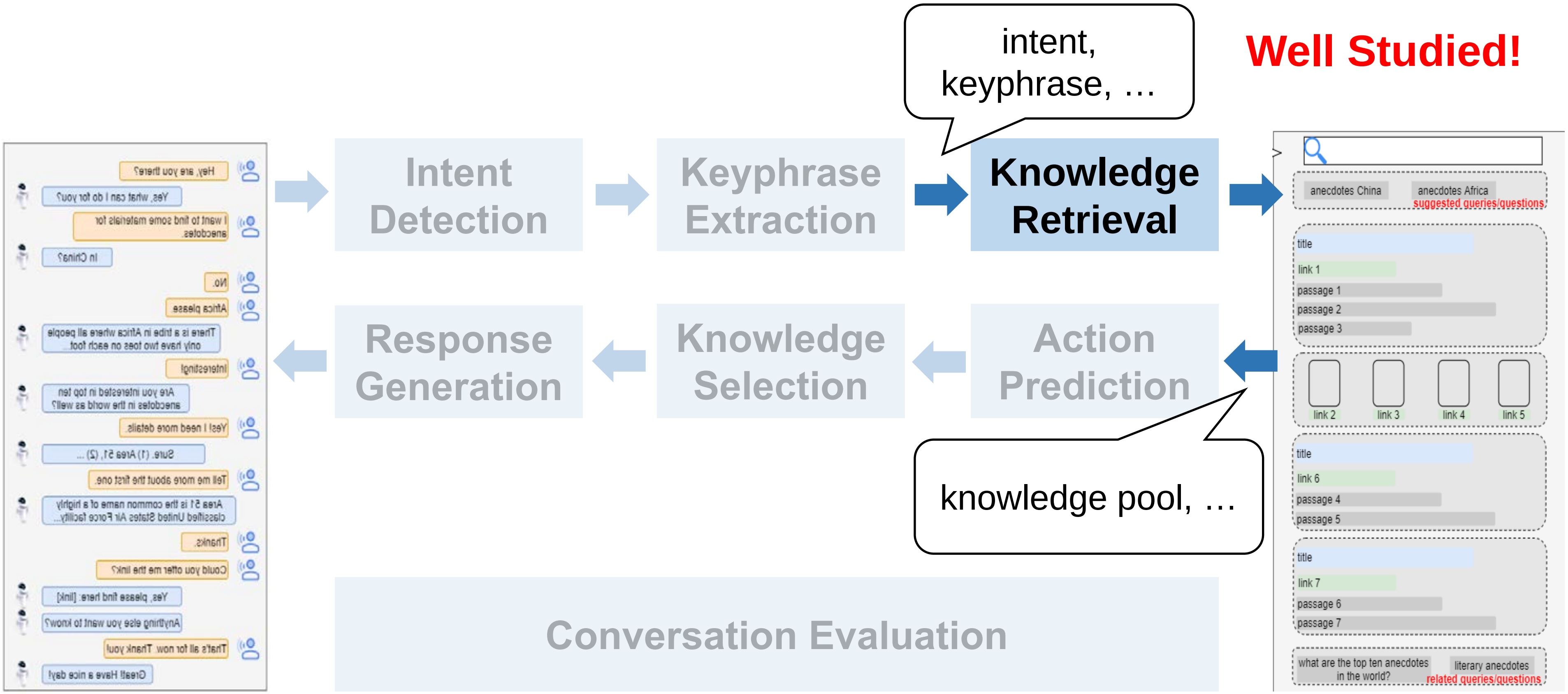
Keyphrase Extraction

	Method	MAP	MRR	NDCG@3
QeReTeC	Initial	0.272	0.637	0.341
	BERT-base	0.272	0.693	0.408
	RRF (Initial + BERT-base)	**0.355**▲	**0.787**▲	**0.476**▲
	Oracle	0.754	0.956	0.926
	TREC-top-auto	0.267	0.715	0.436
	TREC-top-manual	0.405	0.879	0.589

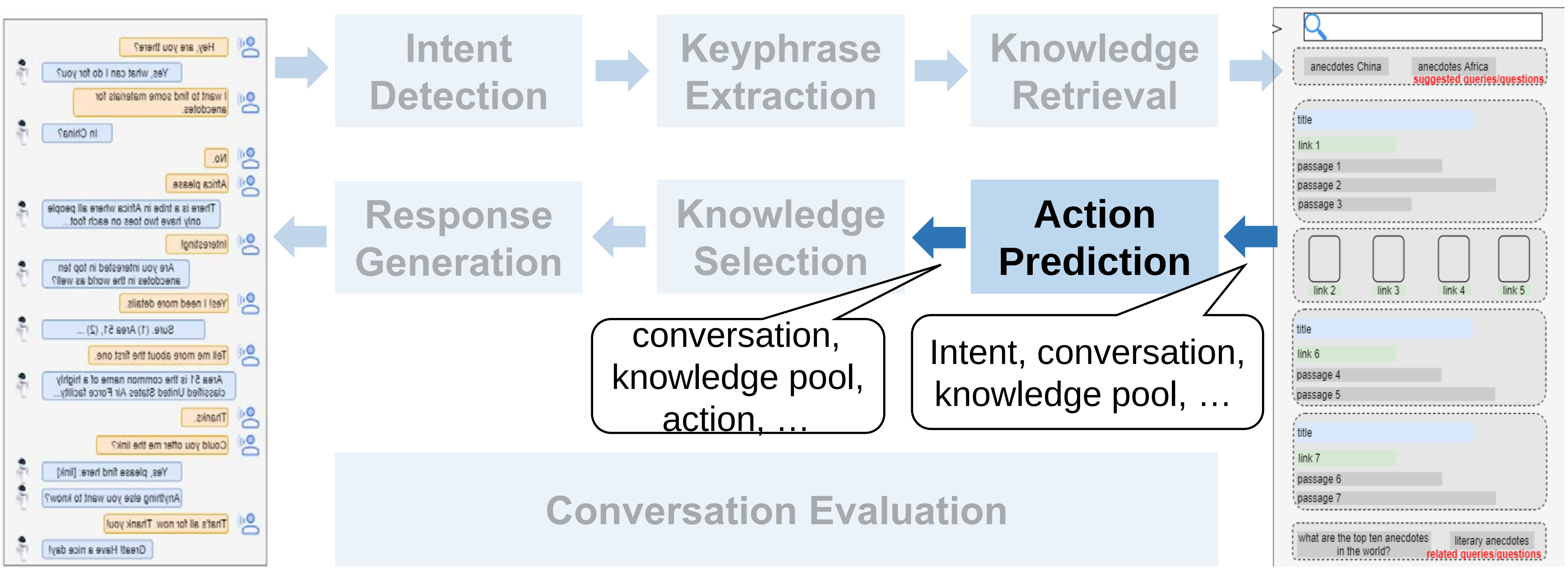
Extrinsic evaluation – reranking results on CAsT.

- ✓ The best model outperforms TRECtop-auto on all metrics.
- ✓ There is still plenty of room for improvement for reranking, which is a clear direction for future work.

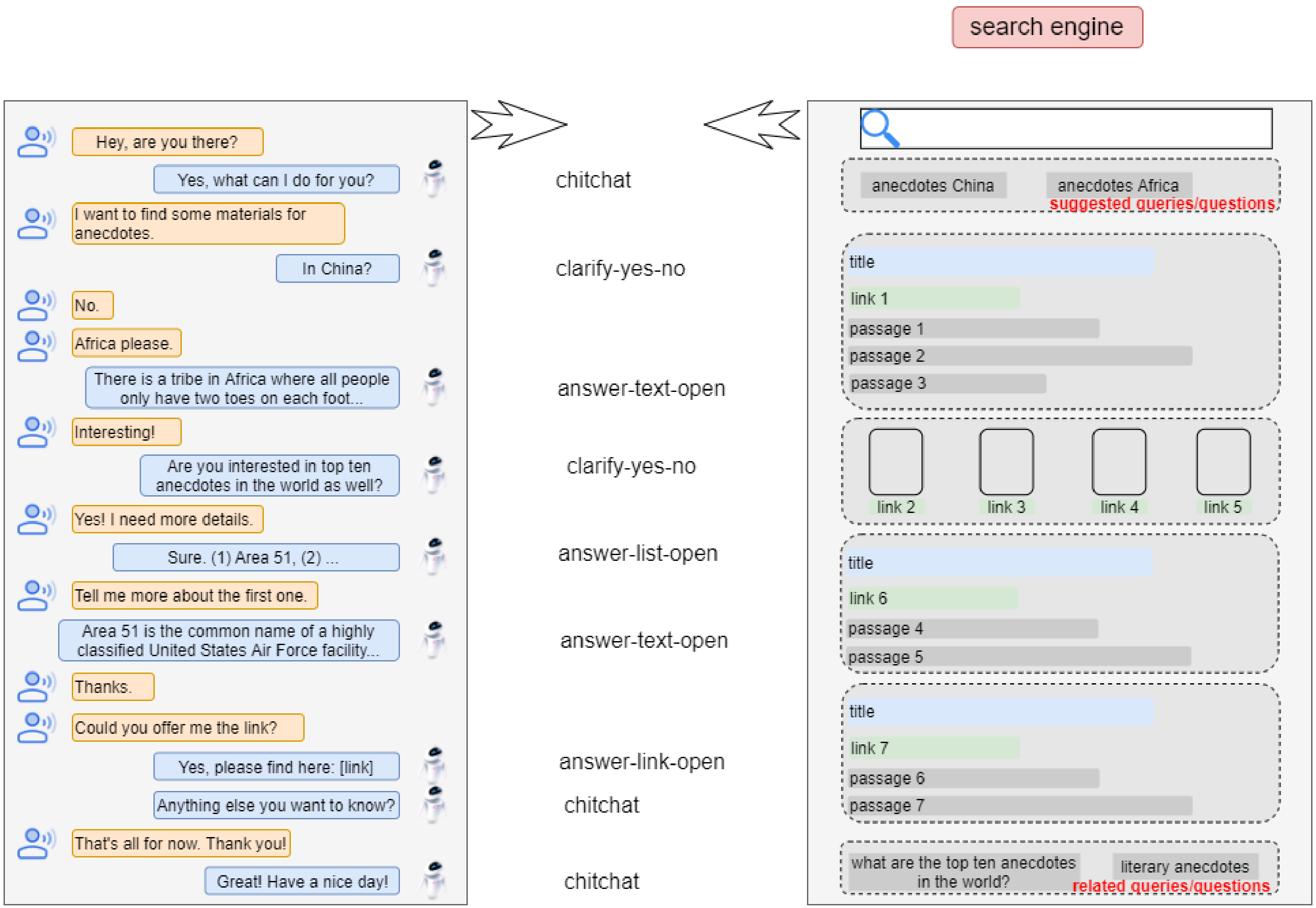
Conversations with Search Engines



Conversations with Search Engines



Action Prediction



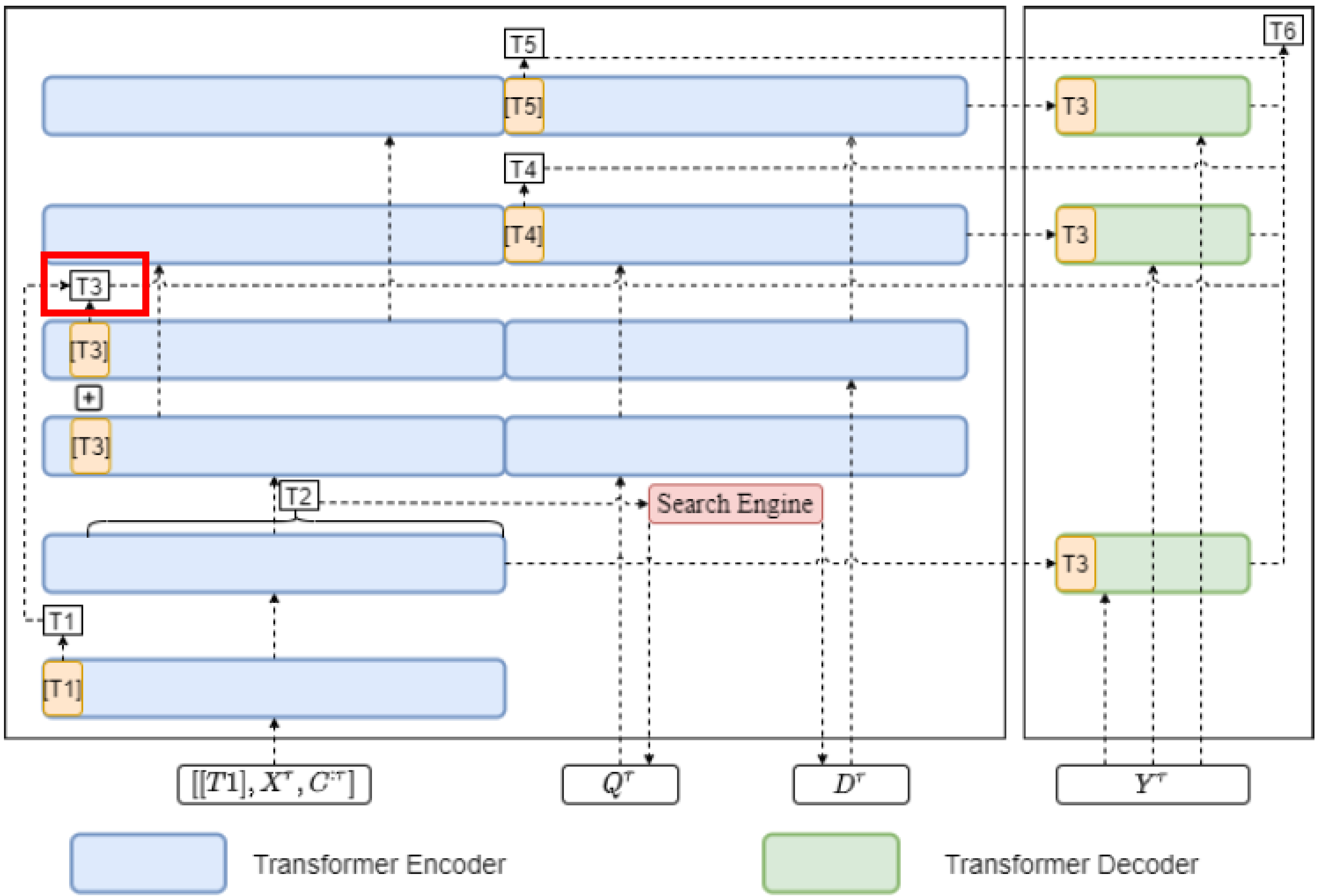
Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Action Prediction

Action		Explanation	Example	TSE operations
clarify	yes-no	Ask questions to clarify user intent when it is unclear or exploratory.	Do you want to the plot? (clarify-yes-no)	Suggest queries.
	choice		Do you want to know its plot, cast or director? (clarify-choice)	
	open		What information do you want to know? (clarify-open)	
answer-type	opinion	Give advice, ideas, suggestions, or instructions. The response is more subjective.	I recommend xxx, because … (answer-opinion)	Provide results.
	fact	Give a single, unambiguous answer. The response is objective and certain.	Her birthday is xxx. (answer-fact)	
	open	Give an answer to an open-ended question, or one with unconstrained depth. The response is objective but may be different depending on the perspectives.	One of the reasons of the earthquake is that… (answer-open)	
answer-form	free-text	Answer the user intent by providing information in the right form or when being asked to answer in a particular form.	The disadvantages of Laminate Flooring are that …… (answer_free_text)	
	list		Area 51. … (answer_list)	
	steps		1. Click on … 2. (answer_steps)	
	link		You can find the video here: [link]. (answer_link)	
no-answer		If there is no relevant information found, notice the user.	Sorry, I cannot find any relevant information. (no-answer)	No answer found.
request-rephrase		Ask the user to rephrase its question if it is unclear.	I didn't really get what you mean. (request-rephrase)	–
chitchat		Greetings or other content that are not related to the information need.	Hi. (chitchat) Yes, I am ready to answer your questions. (chitchat)	–

Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Action Prediction



Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Action Prediction

	AP (%)		
	P	R	F1
-ID	18.7	22.6	18.3
-KE	22.0	22.7	**19.1**
-AP	–	–	–
-QS	**22.7**	**23.2**	18.9
-PS	20.1	22.3	18.1
WISE	18.8	20.6	17.8

Results of joint learning.

	AP (%)		
	P	R	F1
test (unseen)	17.6	18.1	16.5
test (seen)	**19.9**	**24.2**	**19.0**
test	18.8	20.6	17.8

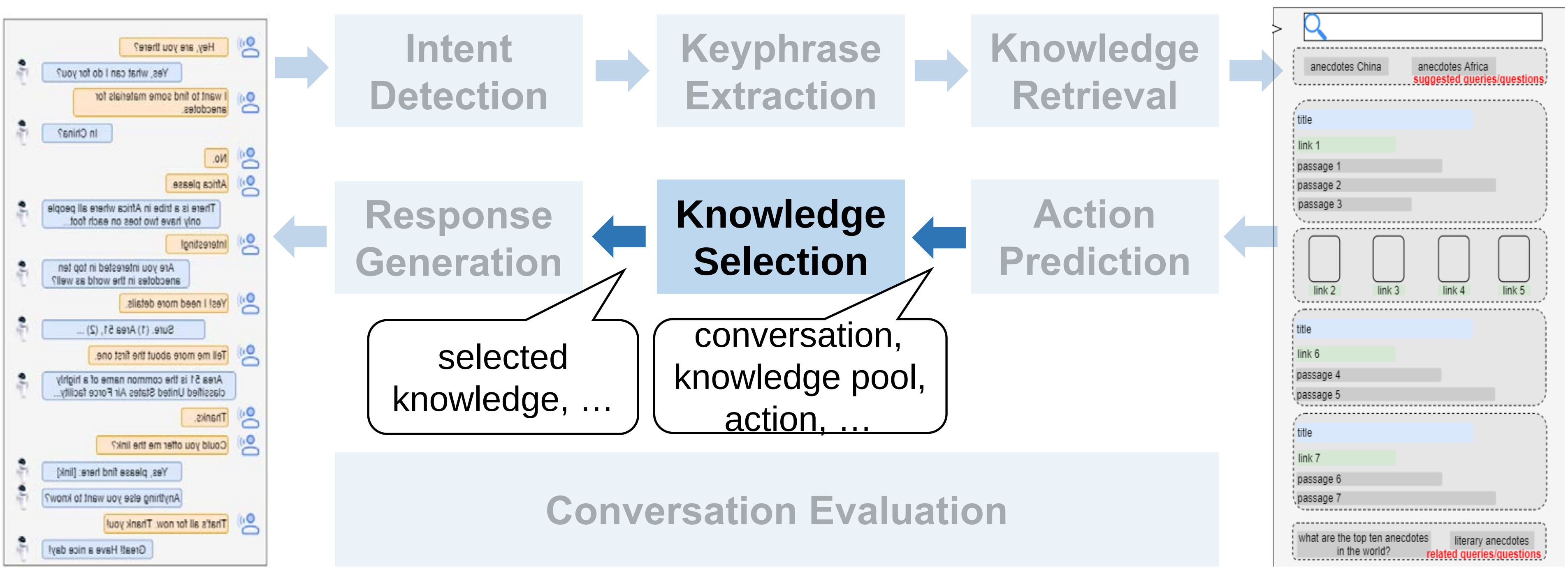
Results on seen/unseen data.

	AP (%)		
	P	R	F1
-DuReader	18.0	19.5	17.2
-KdConv	16.3	17.7	15.3
-DuConv	20.3	20.2	17.9
-WebQA	**20.9**	**20.9**	**18.8**
WISE	18.8	20.6	17.8

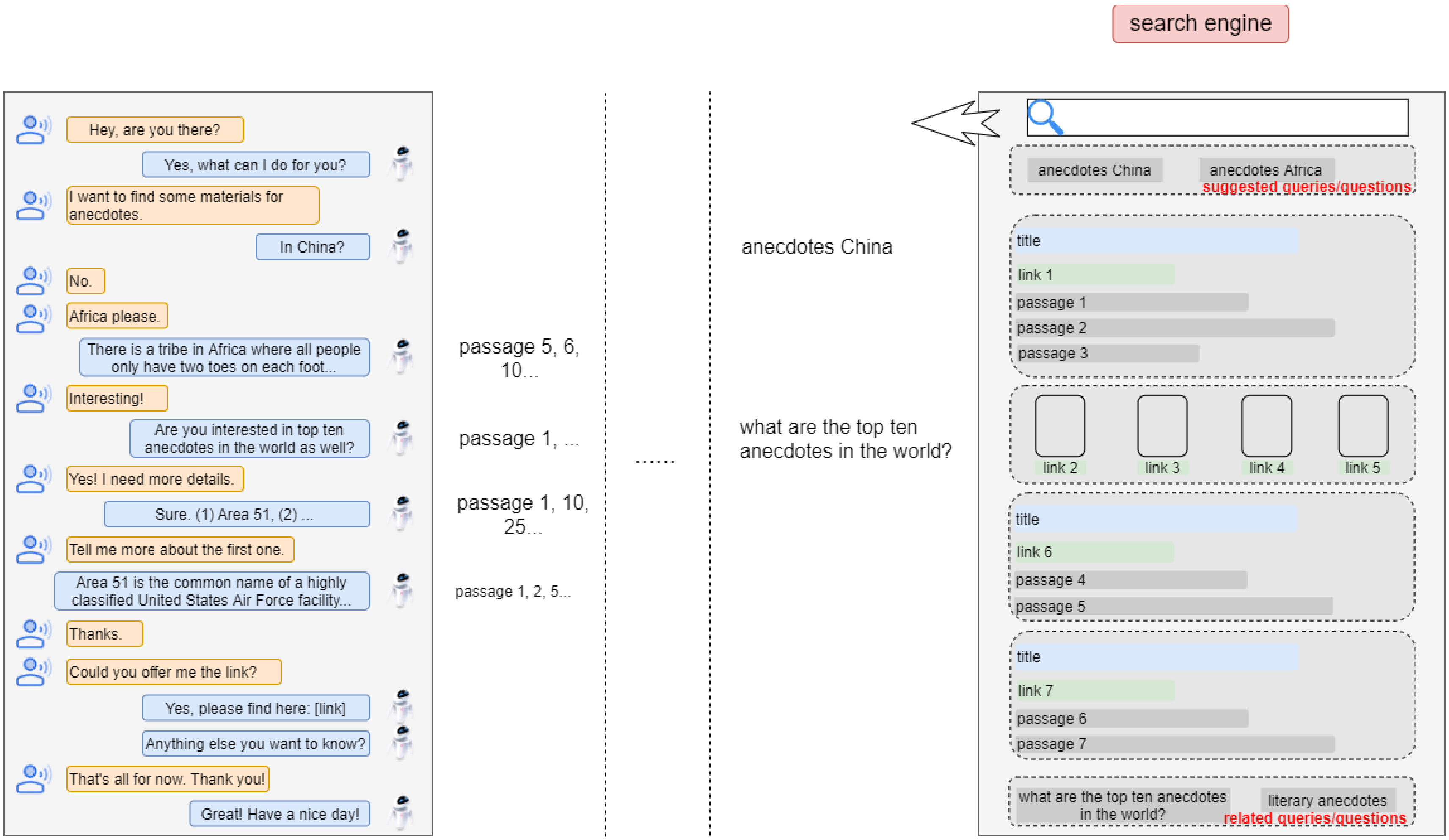
Results with different pretraining data.

- ✓ The joint learning tasks seem incompatible with the current architecture.
- ✓ Better performance on seen data.
- ✓ Not all pretraining data is helpful for ID performance.

Conversations with Search Engines

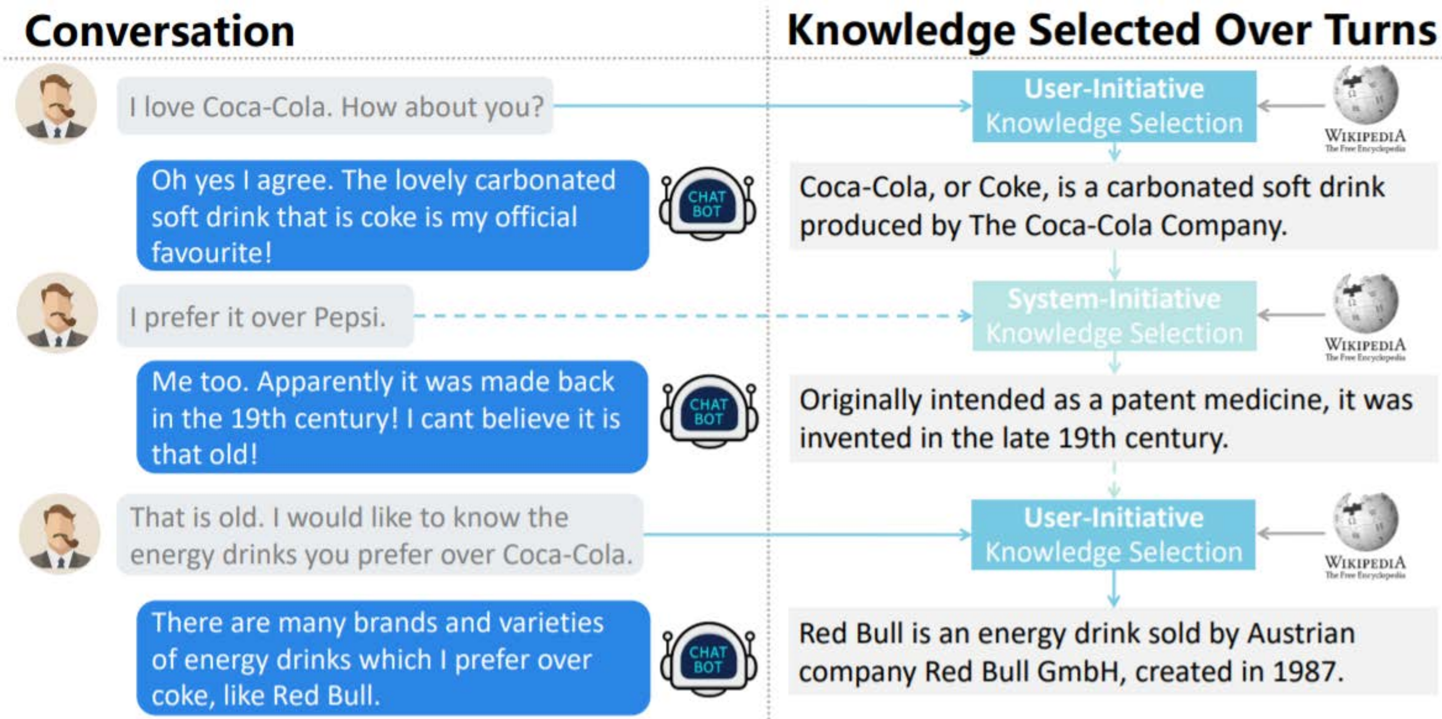


Knowledge Selection



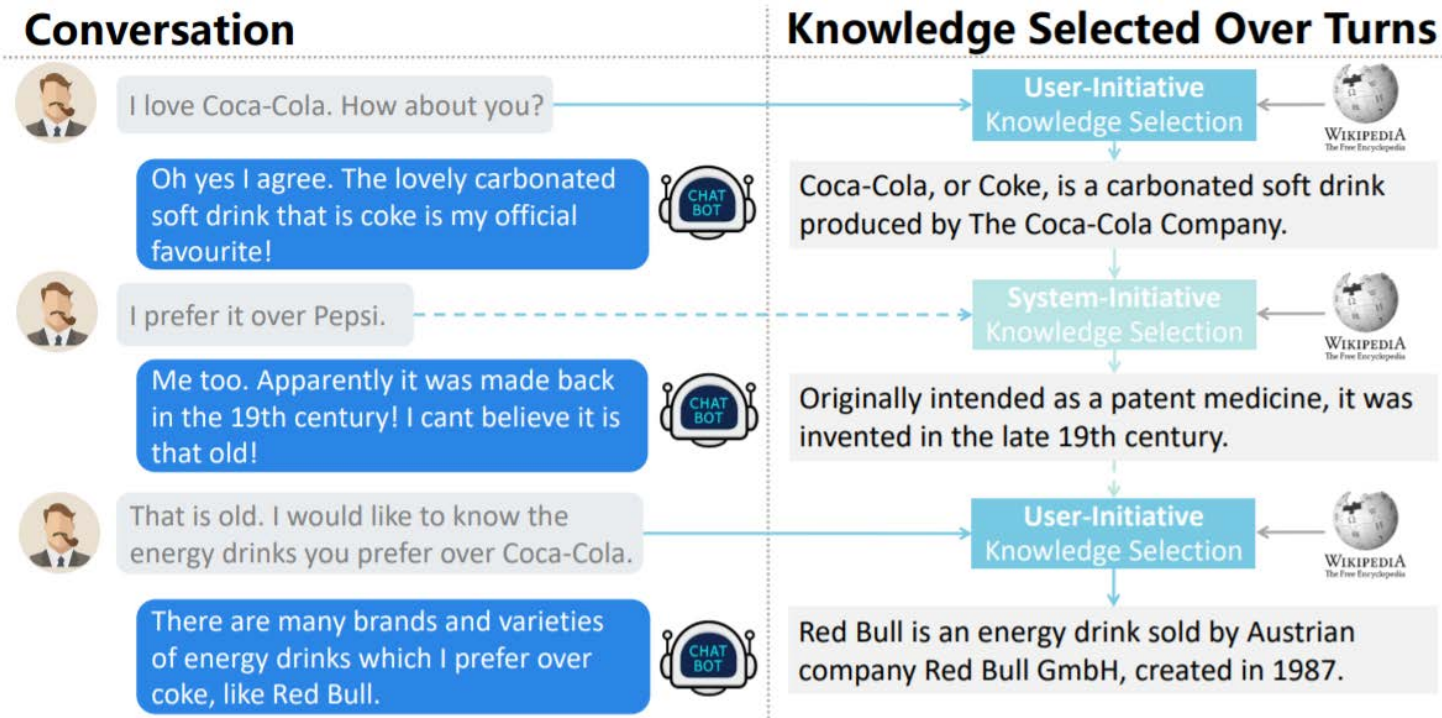
Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Knowledge Selection



- ✓ Conversation is mixed initiative by nature.
- ✓ Pretraining helps but not all conversation data has the required labels.

Knowledge Selection



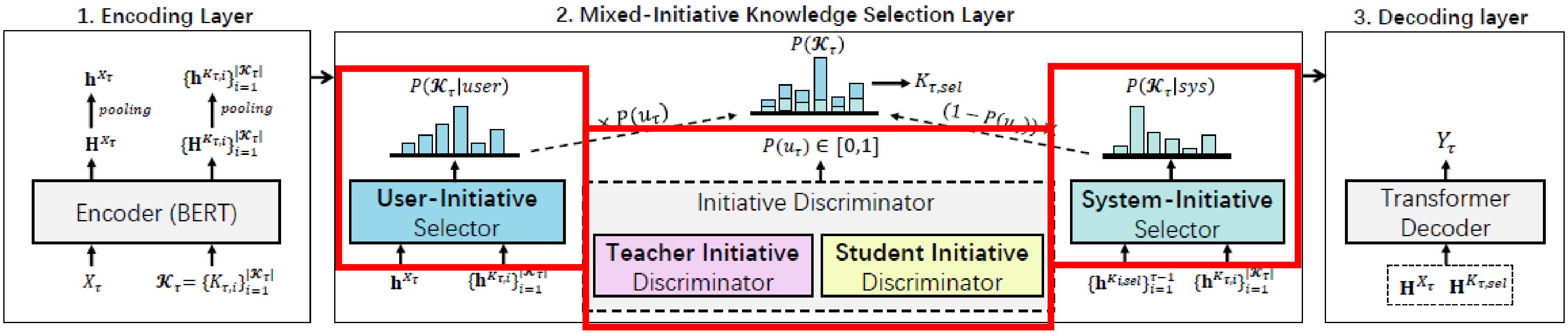
✓ Conversation is mixed initiative by nature.

✓ Pretraining helps but not all conversation data has the required labels.

So can we improve knowledge selection by leveraging the mixed initiative phenomenon without extra labelling required?

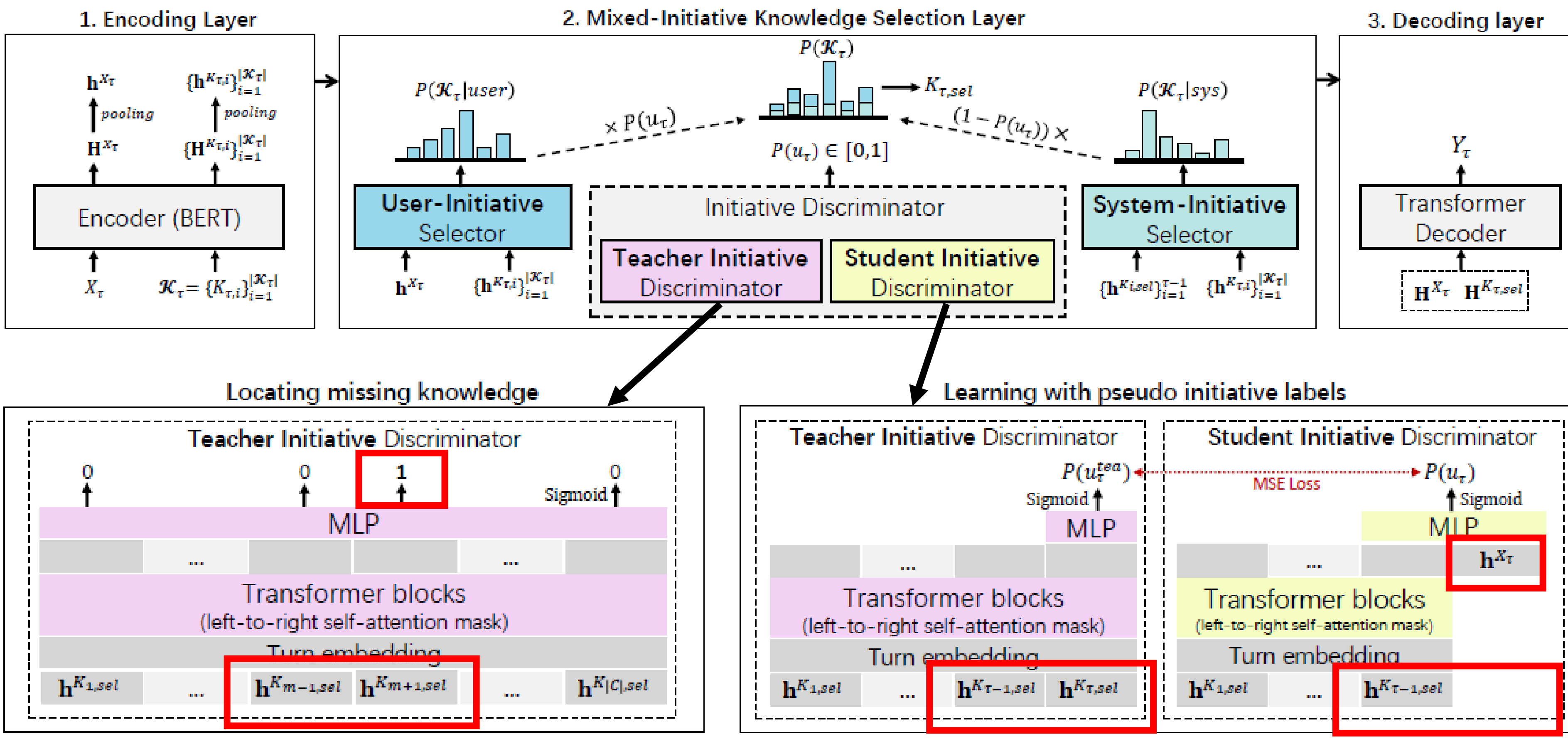
Chuan Meng et al. Initiative-Aware Self-Supervised learning for Knowledge-Grounded Conversations. In SIGIR 2021

Knowledge Selection



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

Knowledge Selection



Knowledge skipping

Assumption: Unsmooth knowledge shift is mostly because of user-initiative.

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

Knowledge Selection

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

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
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**
MIKe-ISLe	2.63	17.22	25.15	6.97	18.67	27.52	1.67	15.38	23.42	5.28	17.04	20.44
MIKe-ISLe-ID	2.48	17.28	24.90	6.64	18.24	26.58	1.46	14.70	22.87	5.16	16.36	19.35
MIKe-ISLe-ID-UIS	1.70	15.88	24.37	5.17	17.33	23.95	0.89	13.68	22.17	4.09	15.98	16.67
MIKe-ISLe-ID-SIS	1.68	15.76	24.33	5.08	17.21	23.88	0.87	13.44	22.01	3.88	15.79	15.99

Results on WoW.

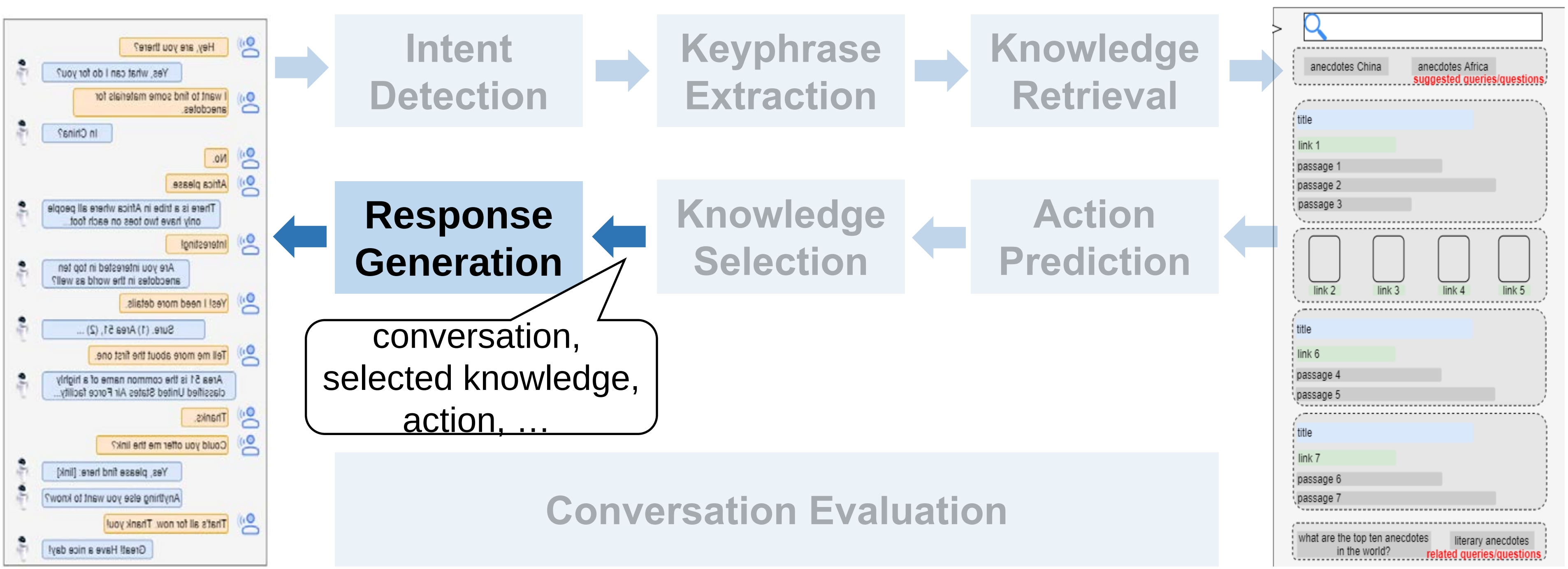
- ✓ MIKe outperforms other baselines in both knowledge selection and response generation.
- ✓ All components are beneficial for MIKe.

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

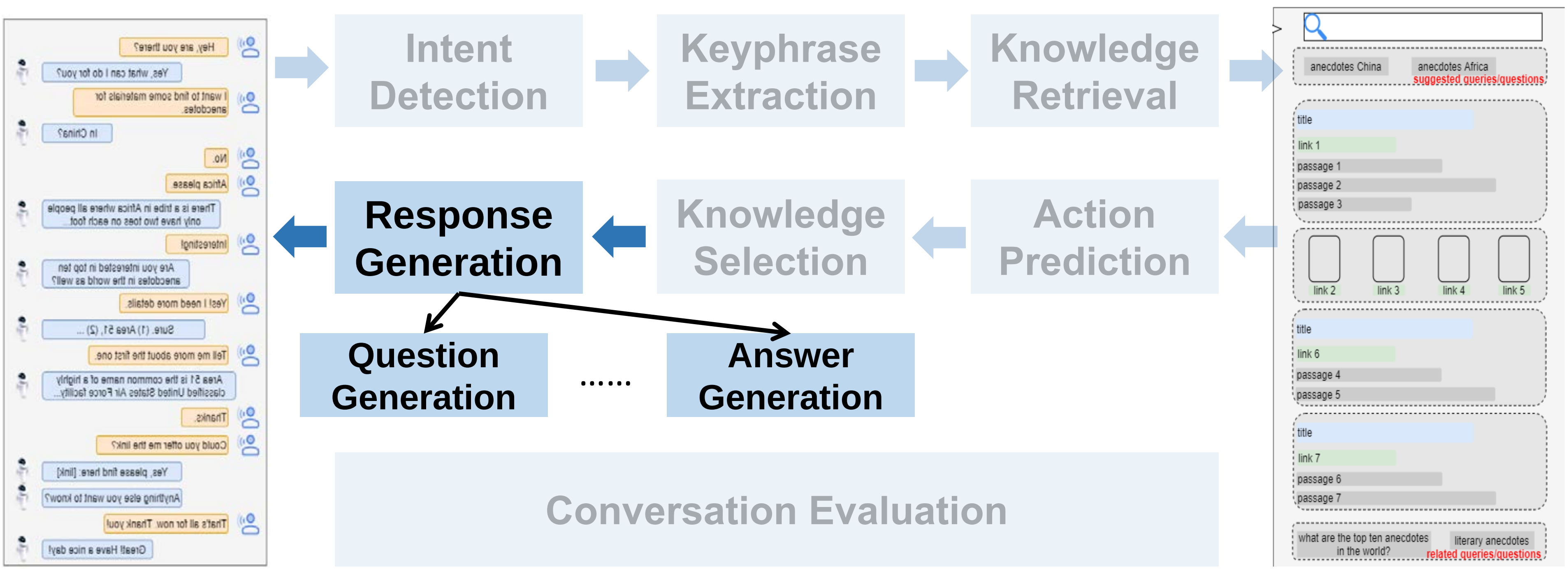
Knowledge Selection

	Example 1 (Test seen)	Example 2 (Test unseen)
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	K_1: no knowledge used . K_2: instagram is a mobile , desktop , and internet - based photo - sharing application and service that allows users to share pictures and videos either publicly , or privately to pre-approved followers . K_3: google photos is a photo sharing and storage service developed by google . K_4: instagram is owned by facebook . ...
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)	User: i hate to admit it but i spend way too much time on instagram ! System: (K_2) i use it for sharing photos and videos User: do you have a lot of followers ? (the current user utterance)
Initiative type	User-initiative KS	System-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 .	DiffKS + BERT: (K_3 ✗) i have a google DukeNet: (K_1 ✗) i have a lot of followers . SKT+PIPM+KDBTS: (K_1 ✗) i have not i have not . MIKe: (K_4 ✓) i have a lot of followers and i do know that it is owned by facebook .

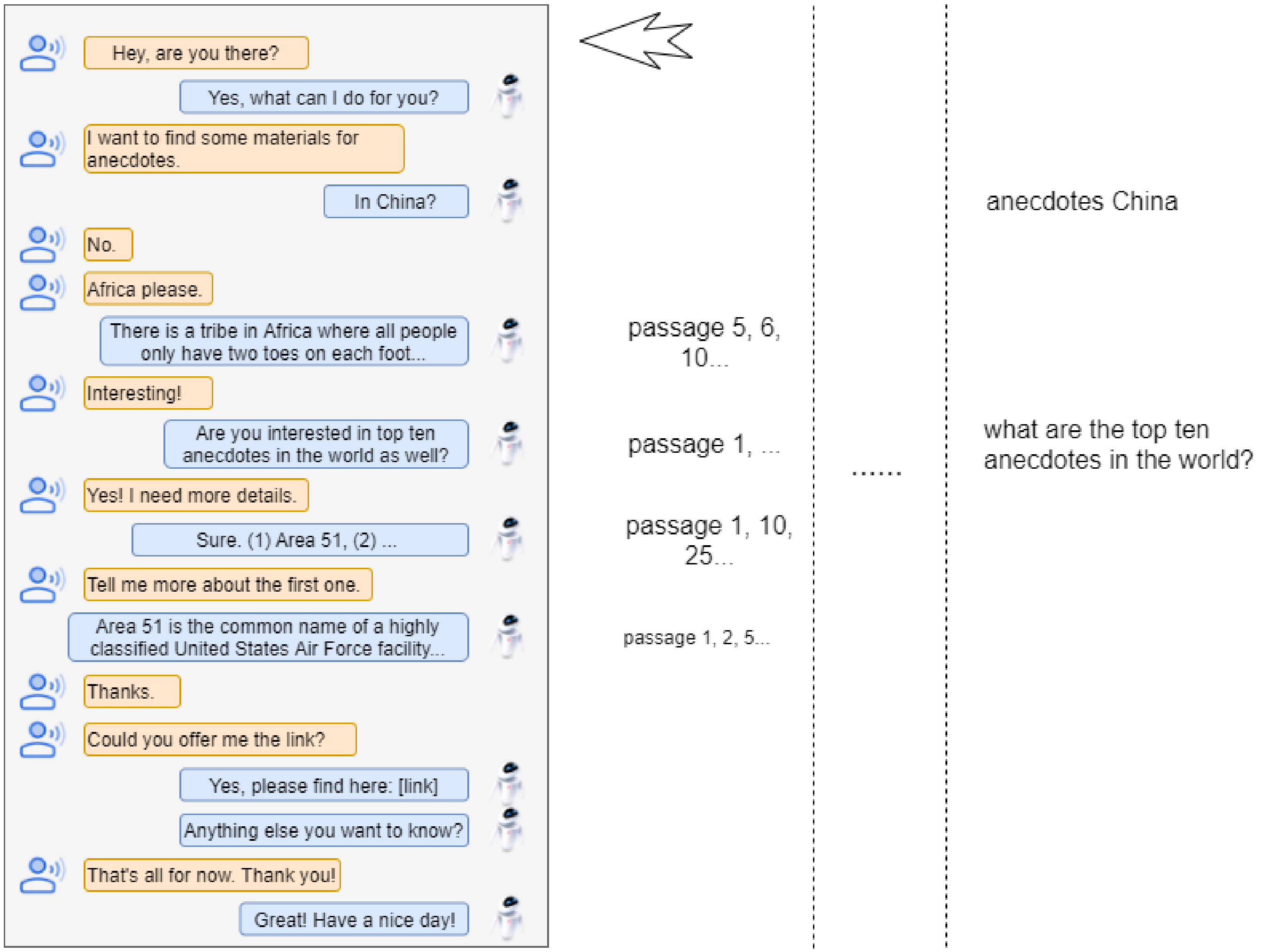
Conversations with Search Engines



Conversations with Search Engines

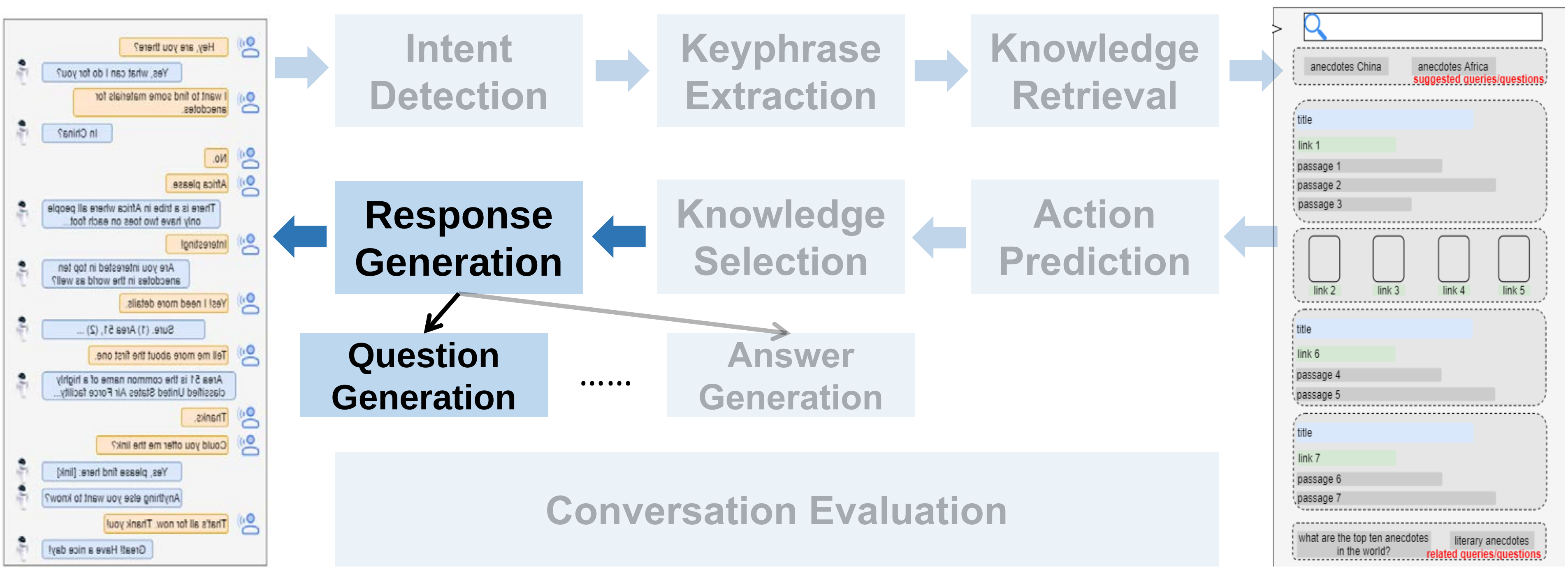


Response Generation

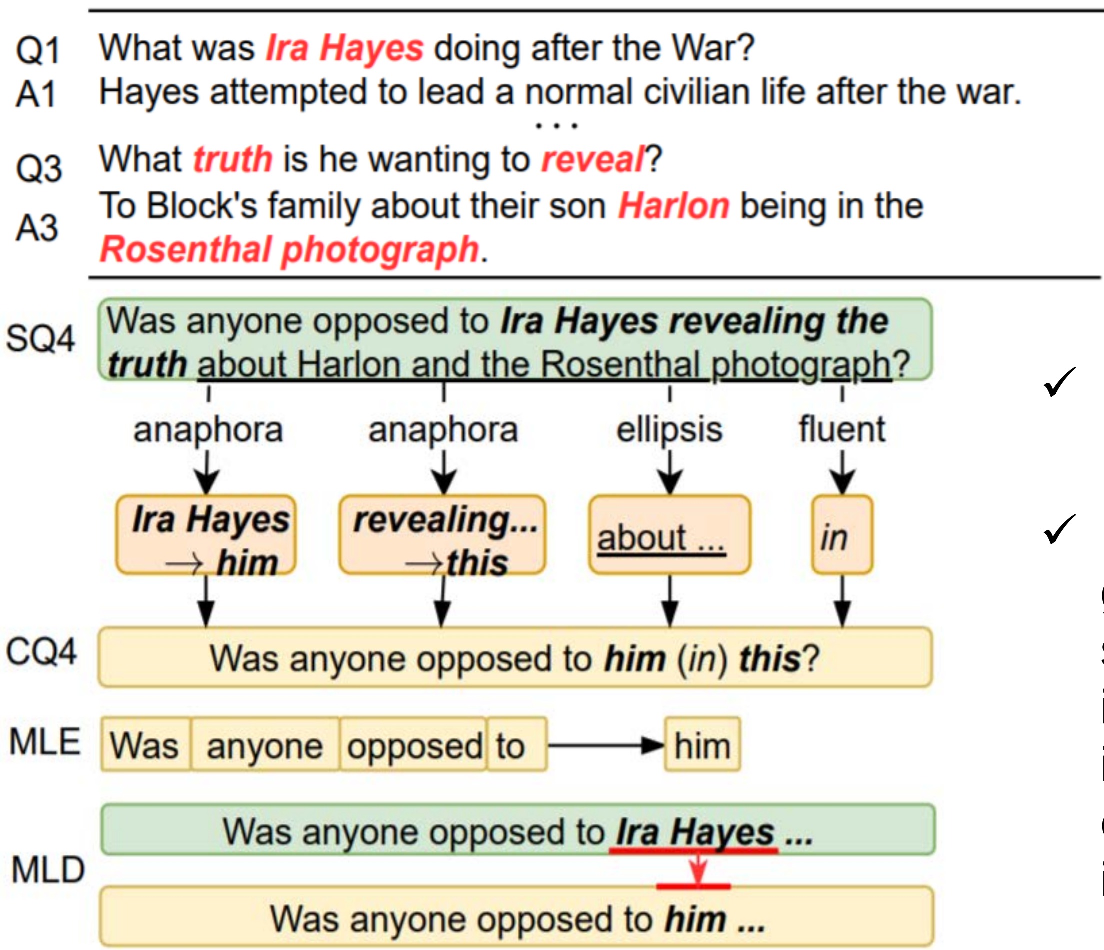


Pengjie Ren et al. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In SIGIR 2021

Conversations with Search Engines



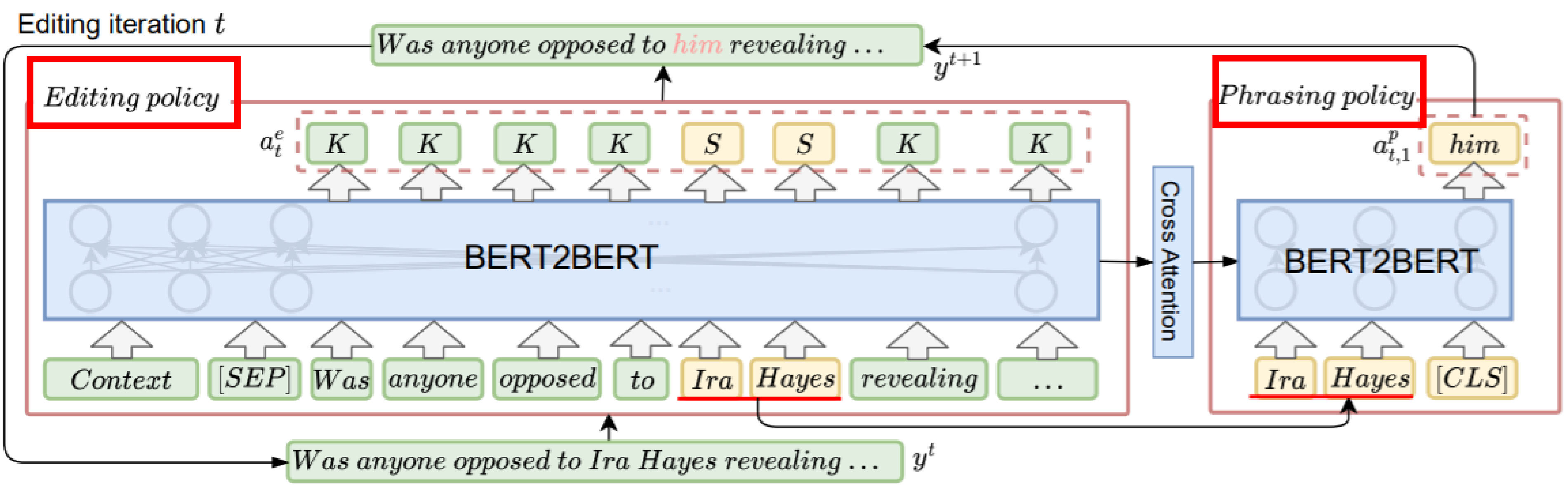
Question Generation



- ✓ 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

Question Generation



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

Question Generation

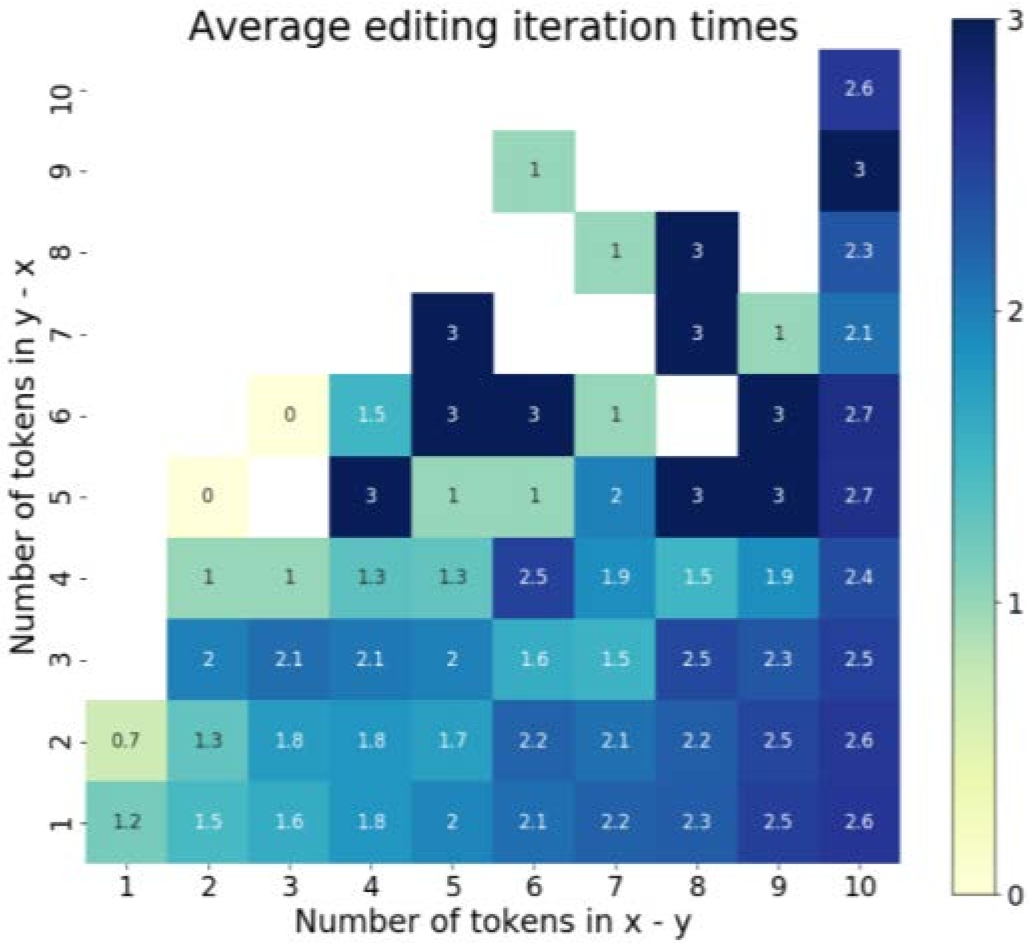
	CANARD (%)						CAsT (%) (unseen)					
Method	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**
Origin	54.7	47.0	40.6	35.3	70.9	3.460	75.9	69.2	62.9	57.6	85.0	5.946
Rule	55.0	47.0	40.2	34.8	70.5	3.420	78.0	71.4	65.3	60.0	86.1	6.220
Trans++	84.3	77.5	72.1	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.

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

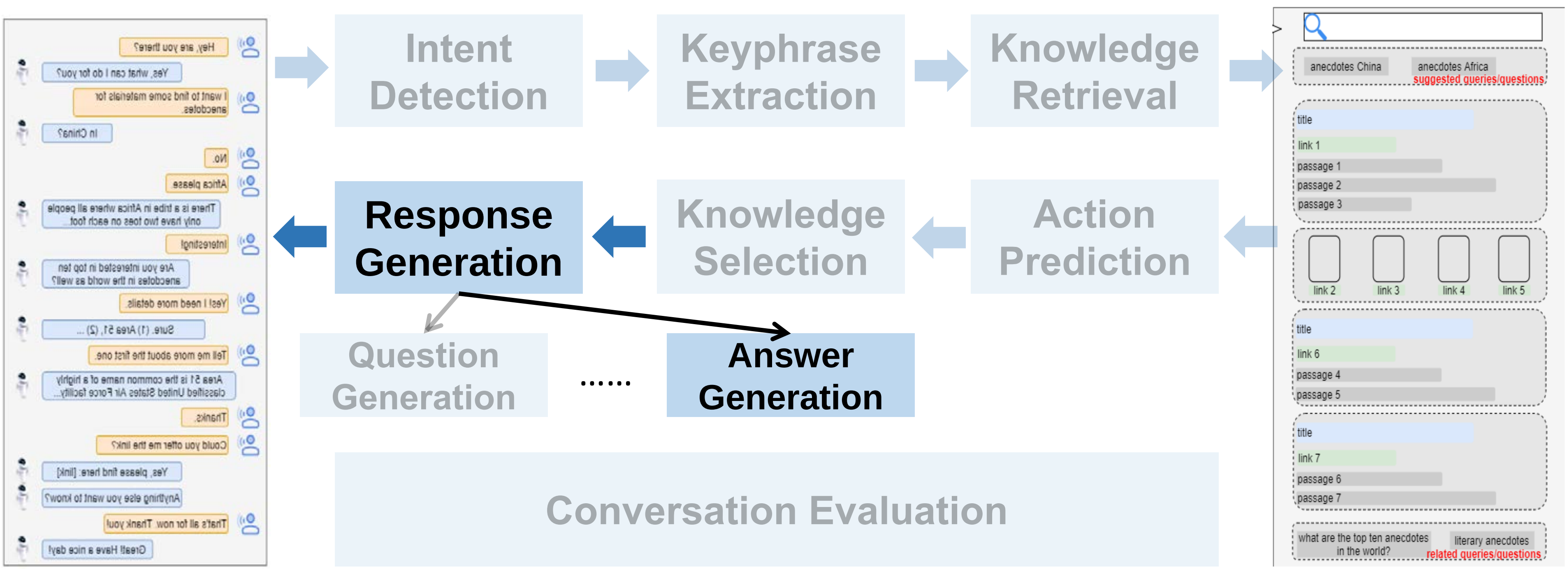
Question Generation



- ✓ 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

Conversations with Search Engines

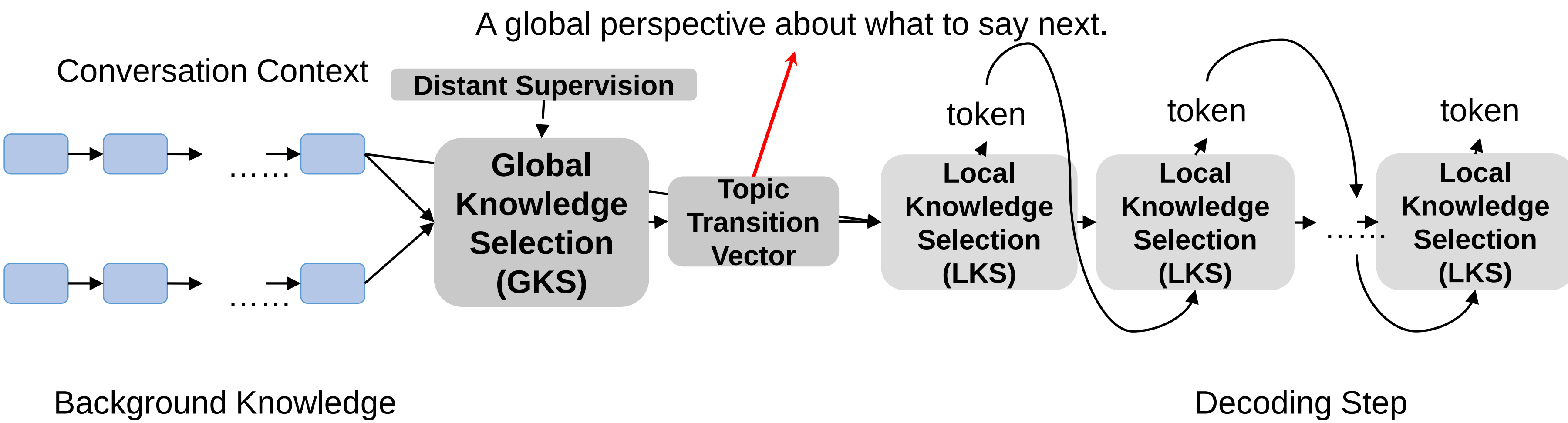


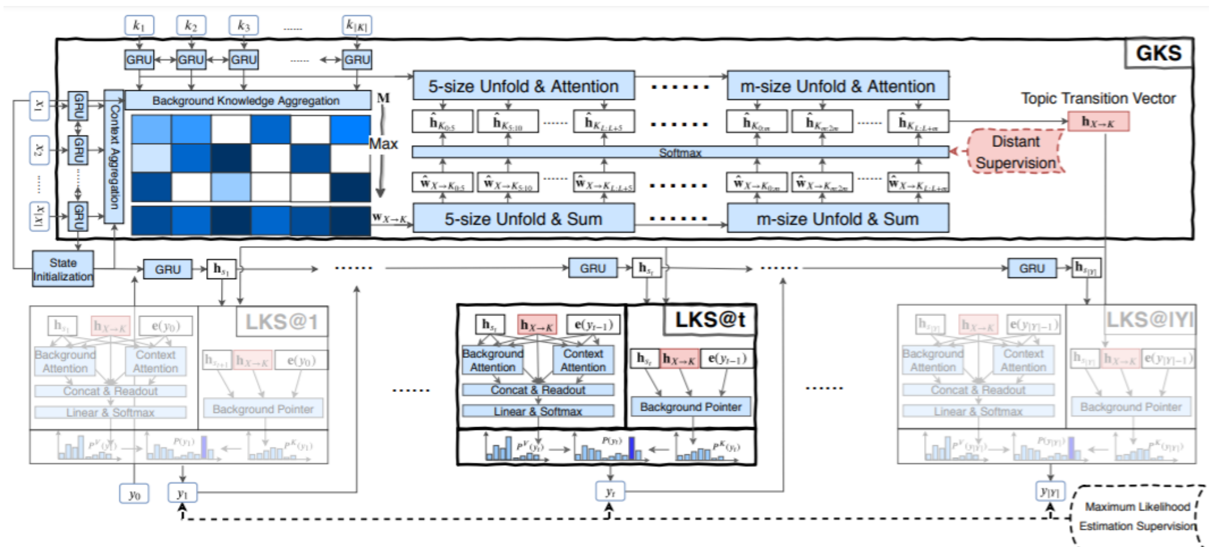
Answer Generation

- Selection methods (Community Question Answering)
  - ✓Not flexible
- Extraction methods (Reading Comprehension)
  - ✓Not fluent (not complete sentence for many cases)
- Abstraction methods (Conversational agents)
  - ✓Not precise (Not use knowledge properly)

Pengjie Ren et al. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. In AAAI 2020

Answer Generation





What to talk (be precise)

Thinking Globally

Acting Locally

How to talk it (be fluent)

Pengjie Ren et al. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. In AAAI 2020

Answer Generation

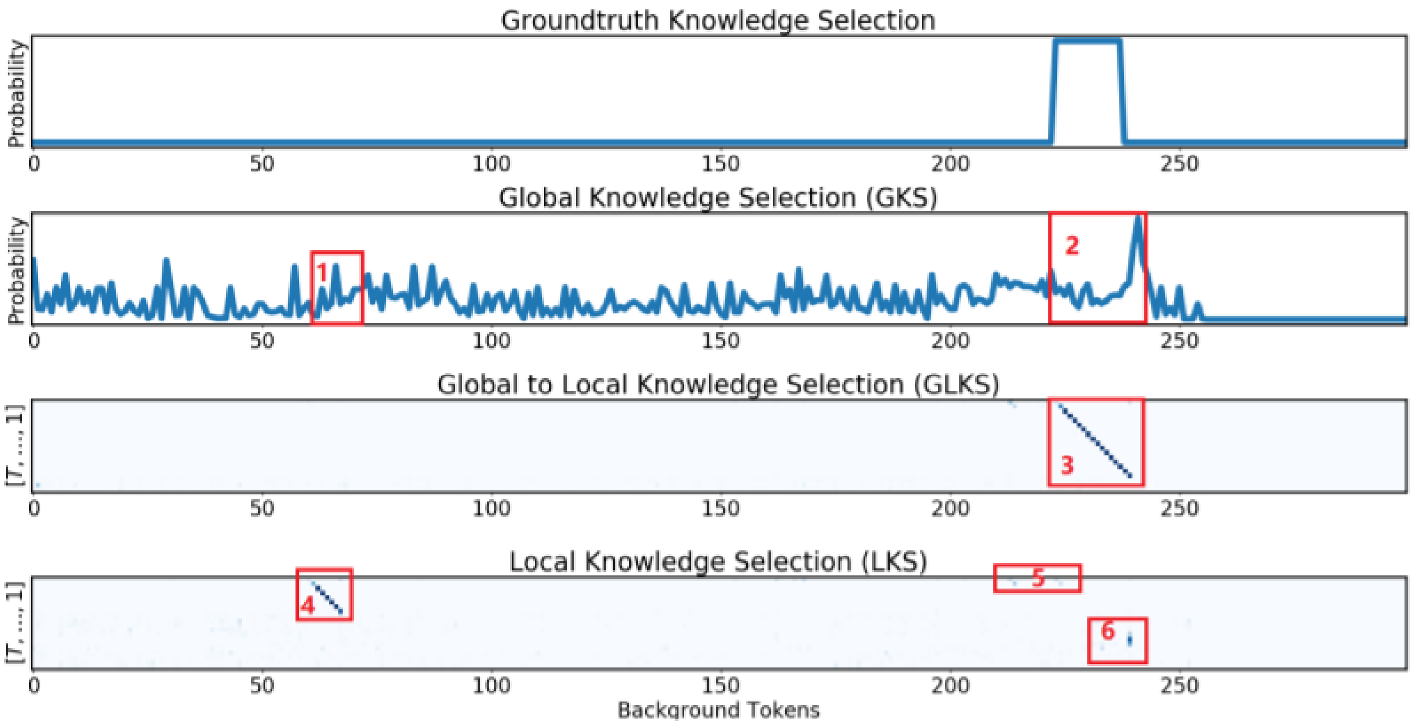
	ROUGE-1		ROUGE-2		ROUGE-L	
	SR	MR	SR	MR	SR	MR
no background						
S2S	27.15	30.91	09.56	11.85	21.48	24.81
HRED	24.55	25.38	07.61	08.35	18.87	19.67
oracle background (256-word)						
S2SA	27.97	32.65	14.50	18.22	23.23	27.55
GTTP	29.82	35.08	17.33	22.00	25.08	30.06
CaKe	42.82	48.65	30.37	36.54	37.48	43.21
RefNet	42.87	49.64	30.73	38.15	37.11	43.77
GLKS	**43.75***	**50.67***	**31.54***	**39.20***	**38.69***	**45.64***
mixed-short background (256-word)						
S2SA	26.36	30.76	13.36	16.69	21.96	25.99
GTTP	30.77	36.06	18.72	23.70	25.67	30.69
CaKe	41.26	45.81	29.43	34.00	36.01	40.79
RefNet	41.33	47.00	31.08	36.50	36.17	41.72
AKGCM	–	–	29.29	–	34.72	–
GLKS	**44.52***	**50.06***	**33.05***	38.87*	**39.63***	**45.12***
mixed-long background (1,200-word)						
S2SA	21.90	24.90	5.63	7.00	17.02	19.65
GTTP	23.64	28.81	10.11	14.34	17.60	22.04
RefNet	34.90	42.08	**22.12**	**29.74**	29.64	36.65
GLKS	**35.30**	**42.31**	21.86	29.35	**30.36**	**37.30**

Results on Holl-E.

- ✓ GLKS is much better at leveraging and locating the right knowledge to generate responses.
- ✓ Knowledge selection becomes much more difficult when the knowledge becomes longer and larger.

Pengjie Ren et al. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. In AAAI 2020

Answer Generation

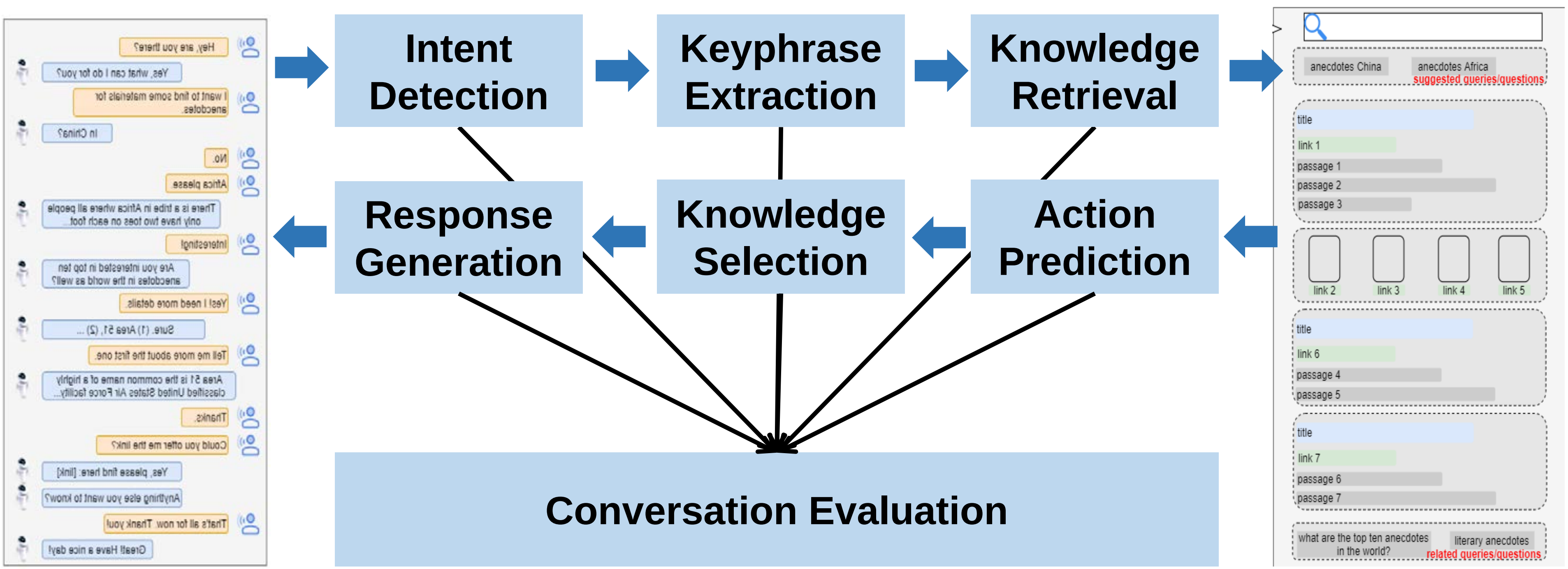


✓ GKS offers a more precise guide in knowledge selection.

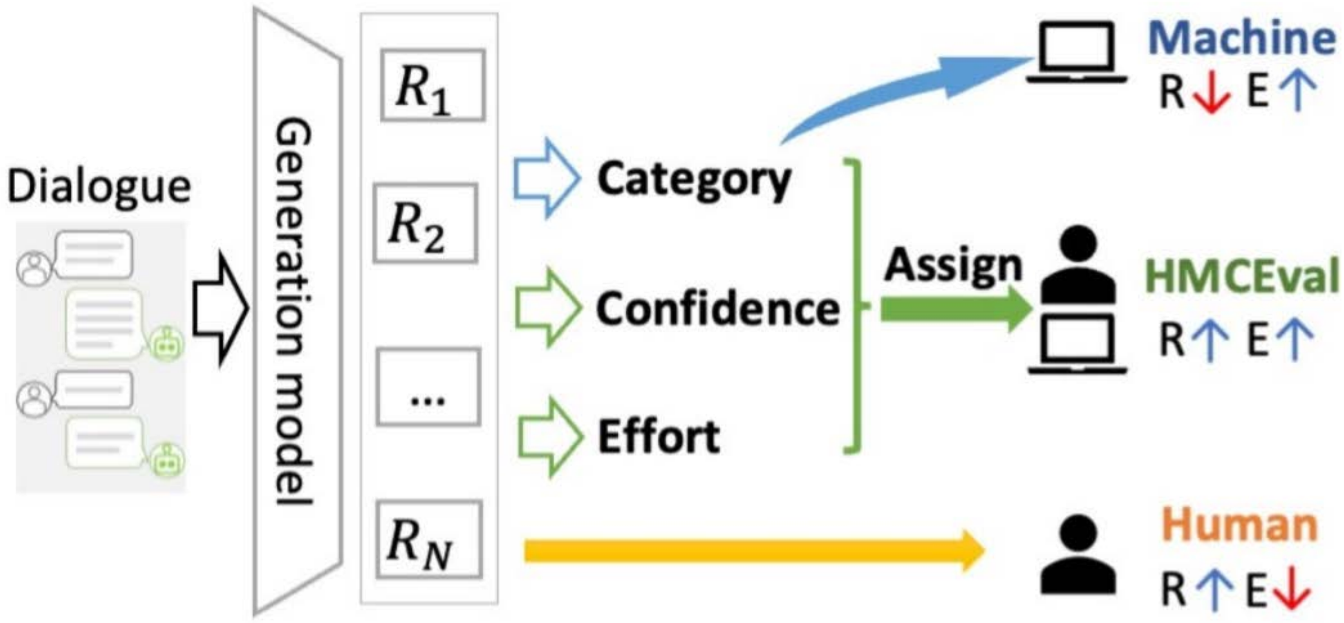
✓ LKS can focus better on response fluency.

Pengjie Ren et al. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. In AAAI 2020

Conversations with Search Engines



Conversation Evaluation



✓ Automatic Evaluation: Efficient but not reliable usually.

✓ Human Evaluation: Mostly reliable but not efficient.

Yangjun Zhang et al. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Conversation Evaluation

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.

Yangjun Zhang et al. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Conversation Evaluation

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. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Conversation Evaluation

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. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. 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

Conversation Evaluation

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…

Yangjun Zhang et al. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Conversation Evaluation

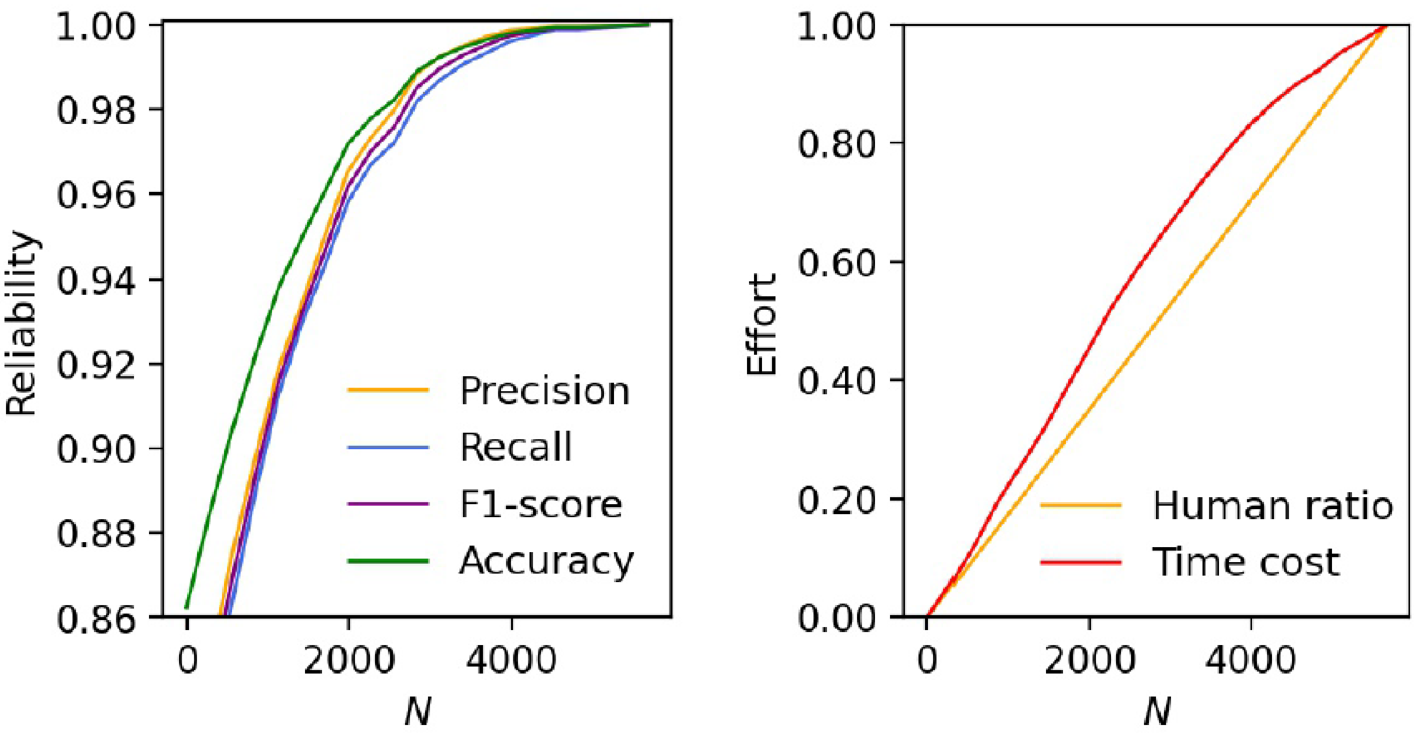
Metric	Machine	Human	HMCEval
Reliability			
Precision	0.818	1	0.983
Recall	0.803	1	0.976
F1-score	0.810	1	0.980
Accuracy	0.862	1	0.985
Efficiency			
Human ratio	0	1	0.500
Time cost	0	1	0.500

N/M=0.5

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

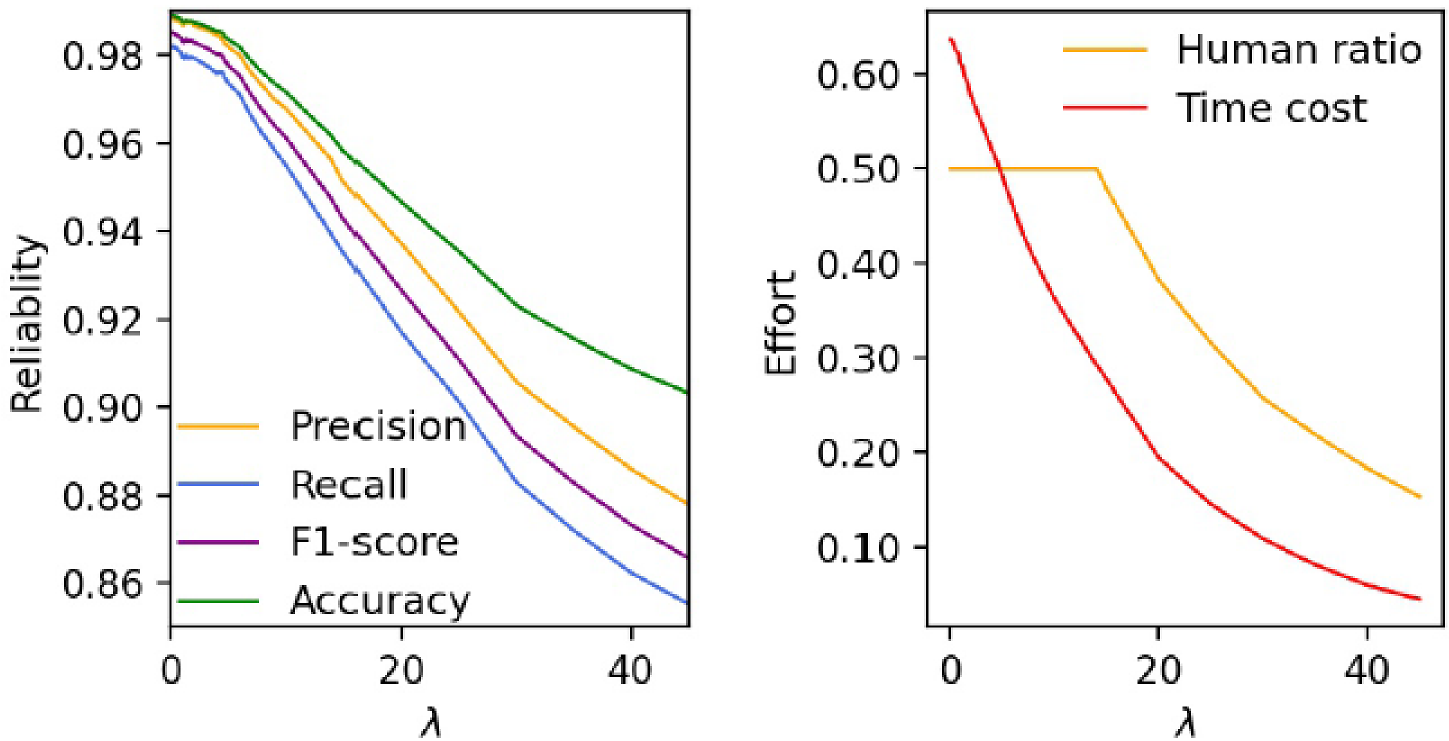
Yangjun Zhang et al. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Conversation Evaluation



As N increases, HMCEval has better reliability, nevertheless the human effort increases.

Conversation Evaluation



As λ increases, HMCEval gets more efficient, while the reliability gets worse.

Yangjun Zhang et al. Learning to Ask Conversational Questionsby Optimizing Levenshtein Distance. In ACL 2021

Yet there's more …

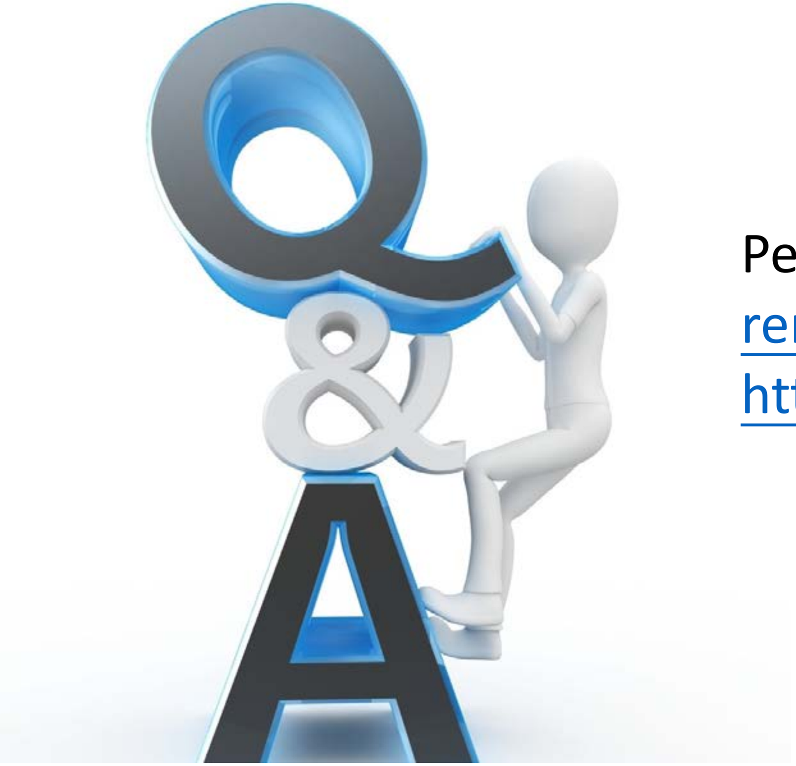
Future directions

- Feedback mining
  - ✓ Less clicks→ Conversations
- Intents/actions increase
  - ✓ Out-of-domain intents/actions
  - ✓ Varying intent/action space
- Response presentation form
  - ✓ Top n → Top 1
  - ✓ Summary, steps, list, link, ...
- Multi-modal conversations
  - ✓ Image, video, …
- Cross-/Multi-Lingual conversations
  - ✓ Leveraging available data better
- More data, more supervision
  - ✓ Building conversations → Labor intensive
- Ethics control
  - ✓ Safe AI

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.
- 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.
- 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.
- 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. Transactions on Information Systems (TOIS), 2021.
- Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas and Maarten de Rijke. Query Resolution for Conversational Search with Limited Supervision. The 43th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2020.
- Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, Maarten de Rijke. Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation. The 34th AAAI Conference on Artificial Intelligence (AAAI), 2020.

Thank you for your attention!



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