When making your system conversational, …

Pengjie Ren

IRLab, Shandong University

renpengjie@sdu.edu.cn

Why making your system conversational?

- As human beings, our natural model of communication is through conversations.
- Conversations are more suitable for complex and exploratory information needs.
- Conversations are more friendly for some people and/or in some scenarios.
- Well, it looks more intelligent after all.

A sign from search engines

- **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/

Everything can be conversational.

- **Conversational search**
- **Conversational recommendation**
- **Conversational question answering**
- **Conversational machine reading comprehension**
- **Conversational summarization**
- **…**

What is different for <u>us</u>?

Mixed Initiative

- **User Initiative → Mixed Initiative**
- **Systems can ask clarifying questions.**
- **What to ask**
  - Dialogue Management
- **How to ask it**
  - Question Generation

Dialogue Management

- **Closed-domain**

act type	inform* / request* / select^{123} / recommend/^{123} / not found^{123} request booking info^{123} / offer booking^{1235} / inform booked^{1235} / decline booking^{1235} welcome* /greet* / bye* / reqmore*
slots	address* / postcode* / phone* / name^{1234} / no of choices^{1235} / area^{123} / pricerange^{123} / type^{123} / internet^{2} / parking^{2} / stars^{2} / open hours^{3} / departure^{45} destination^{45} / leave after^{45} / arrive by^{45} / no of people^{1235} / reference no.^{1235} / trainID^{5} / ticket price^{5} / travel time^{5} / department^{7} / day^{1235} / no of days^{123}

Paweł Budzianowski et al. MultiWOZ -- A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In EMNLP 2018

Dialogue Management

- **Closed-domain**

act type	inform* / request* / select^{123} / recommend/^{123} / not found^{123} request booking info^{123} / offer booking^{1235} / inform booked^{1235} / decline booking^{1235} welcome* /greet* / bye* / reqmore*
slots	address* / postcode* / phone* / name^{1234} / no of choices^{1235} / area^{123} / pricerange^{123} / type^{123} / internet^{2} / parking^{2} / stars^{2} / open hours^{3} / departure^{45} destination^{45} / leave after^{45} / arrive by^{45} / no of people^{1235} / reference no.^{1235} / trainID^{5} / ticket price^{5} / travel time^{5} / department^{7} / day^{1235} / no of days^{123}

New actions? New domains?

Paweł Budzianowski et al. MultiWOZ -- A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In EMNLP 2018

Dialogue Management

- **Open-domain**

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

Dialogue Management

- **Open-domain**

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

Dialogue Management

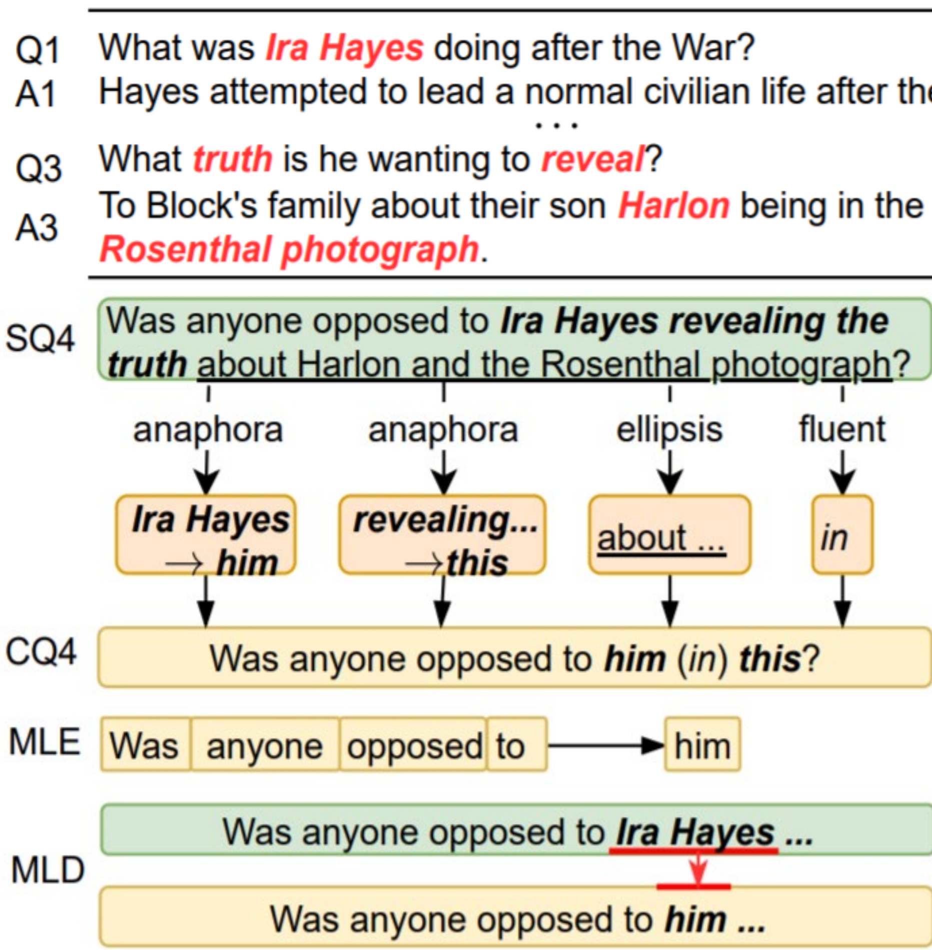
- **Open-domain**

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)	–

Fine granularity?

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

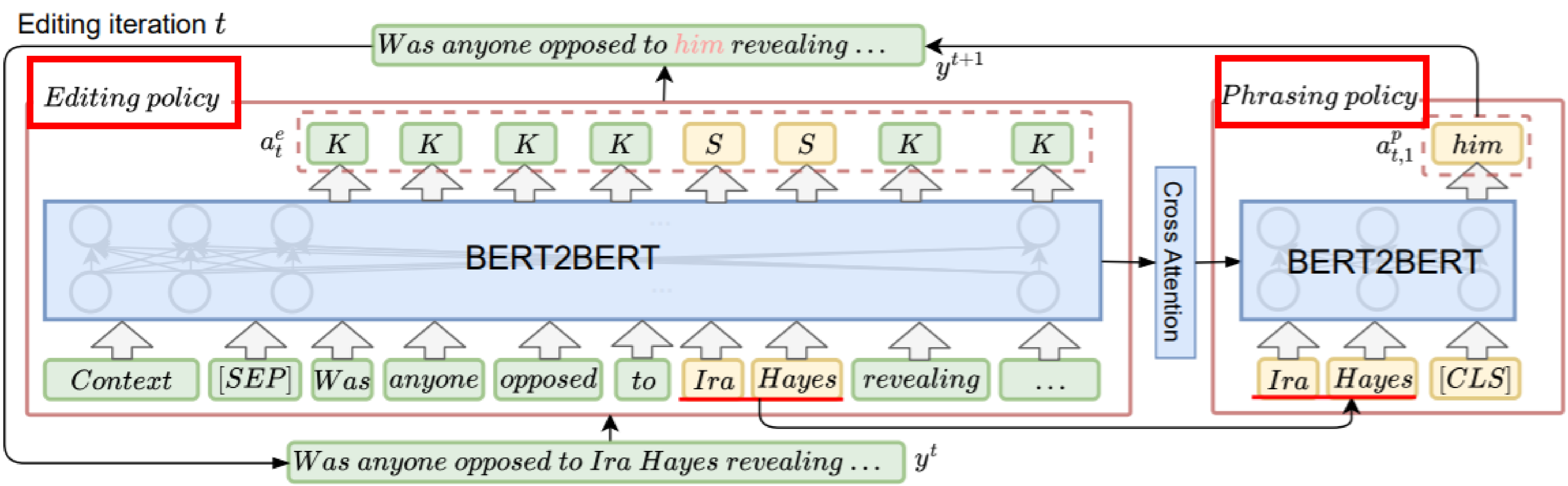
Question Generation



- ✓ 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

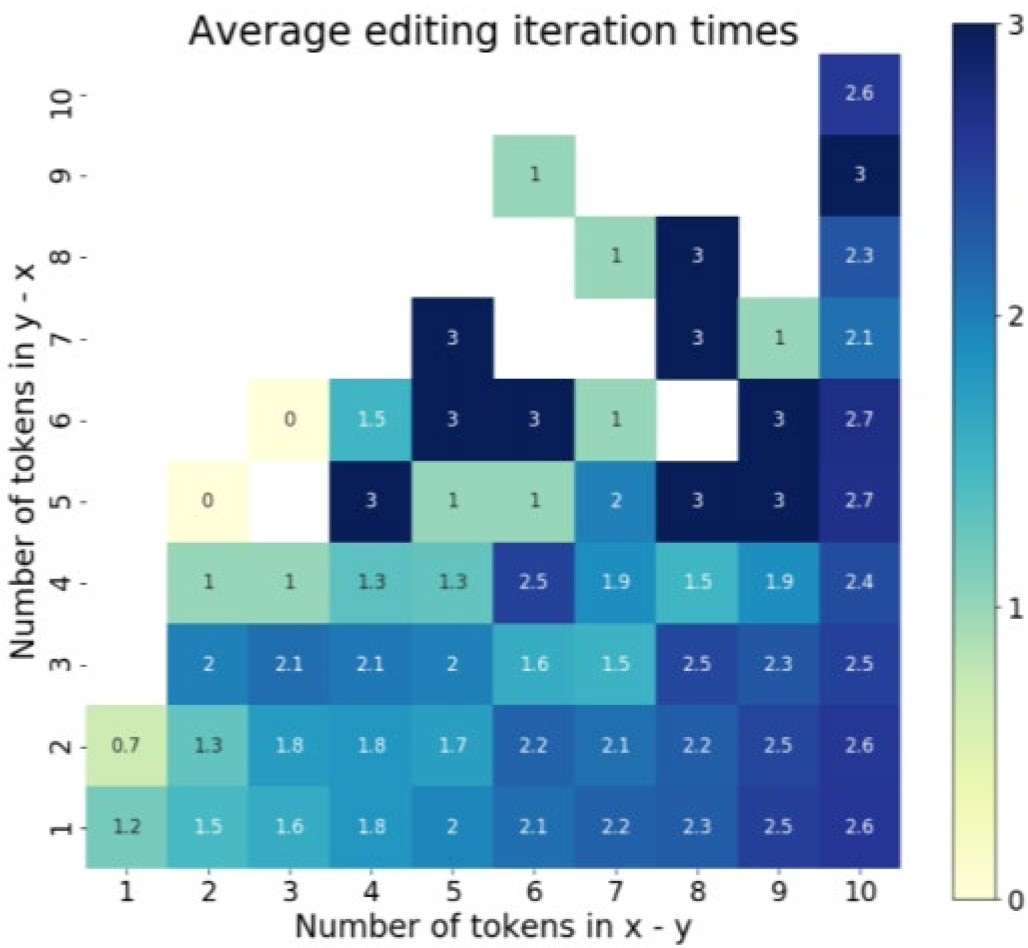
	CANARD (%)						CAsT (%) (unseen)					
Method	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**
Origin	54.7	47.0	40.6	35.3	70.9	3.460	75.9	69.2	62.9	57.6	85.0	5.946
Rule	55.0	47.0	40.2	34.8	70.5	3.420	78.0	71.4	65.3	60.0	86.1	6.220
Trans++	84.3	77.5	72.1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

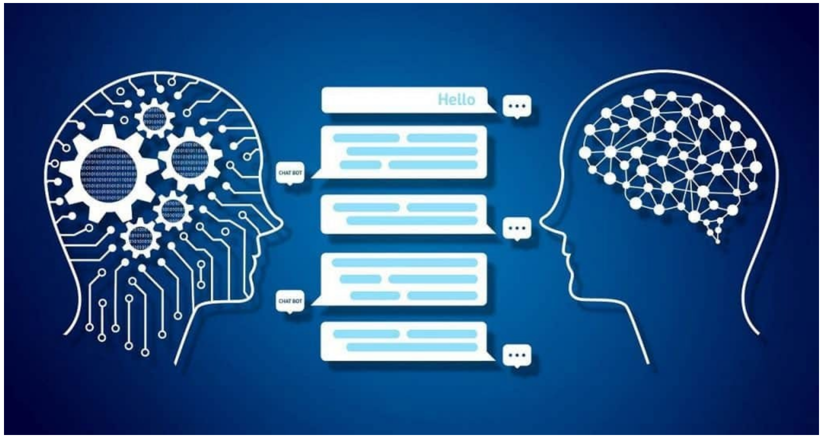
Zhongkun Liu et al. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In ACL 2021

Question Generation



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

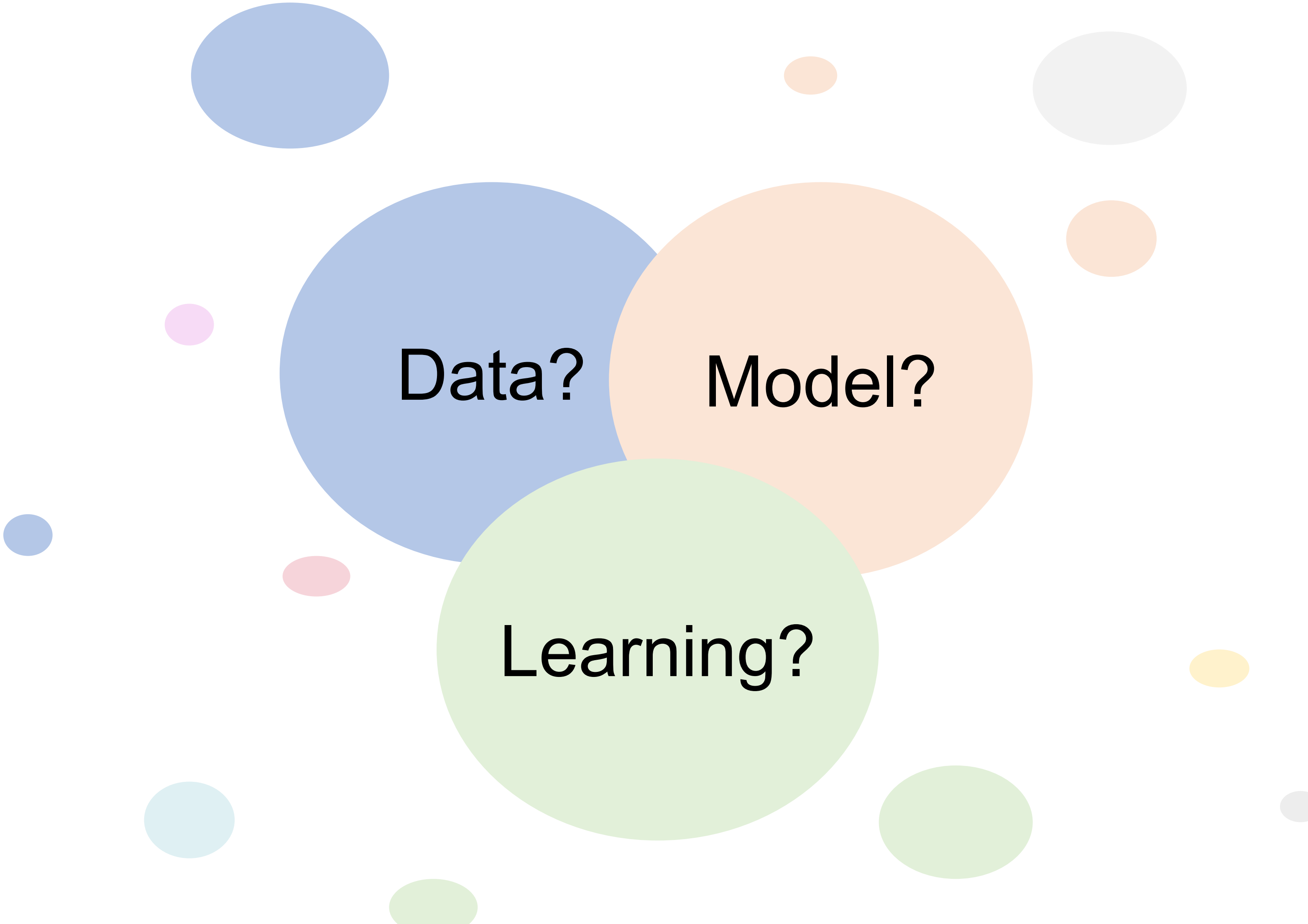
Supervision Signals



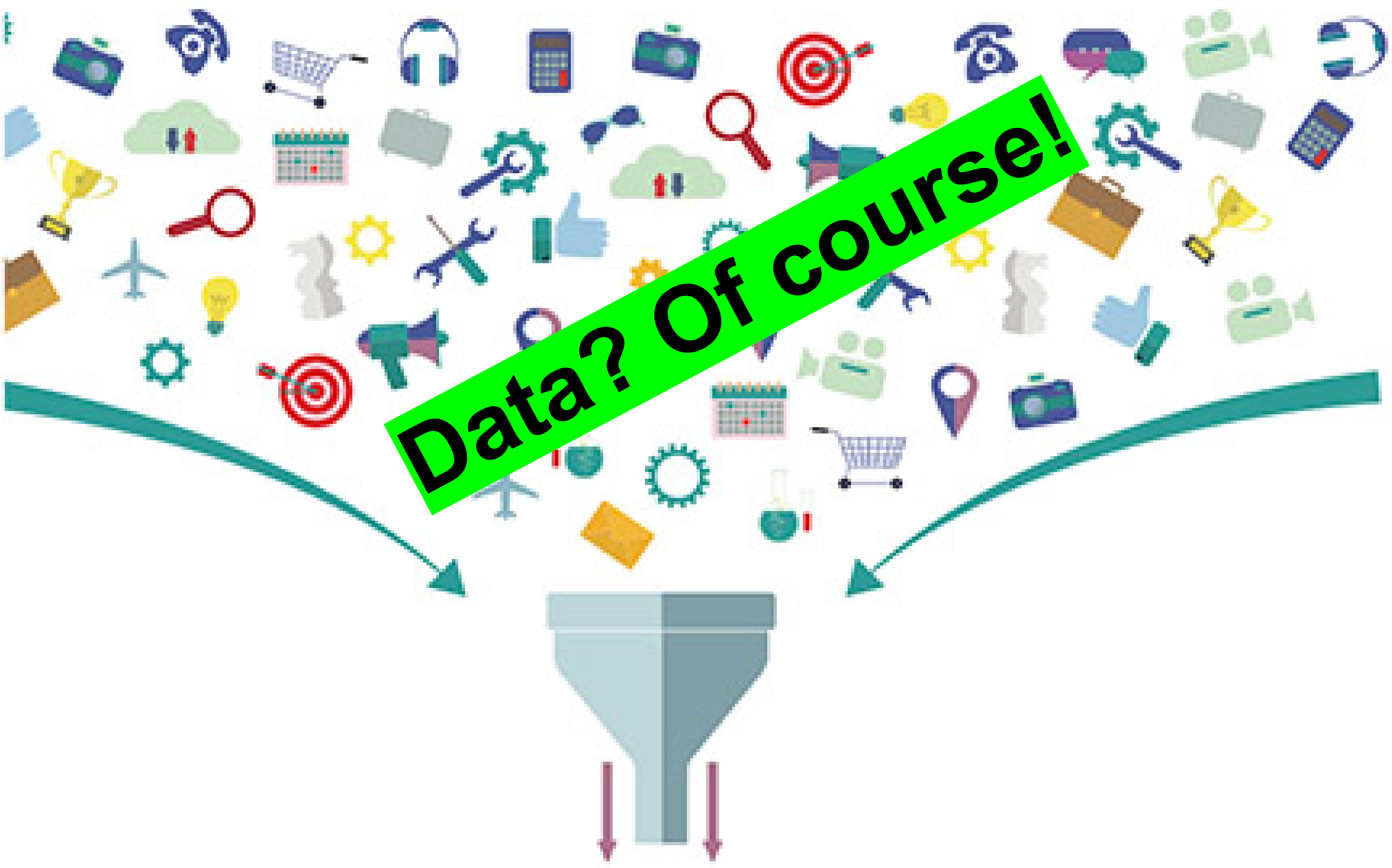
Human in the loop.

Lack of direct supervision signals.

What really matters for AI?



What really matters for AI? Data?



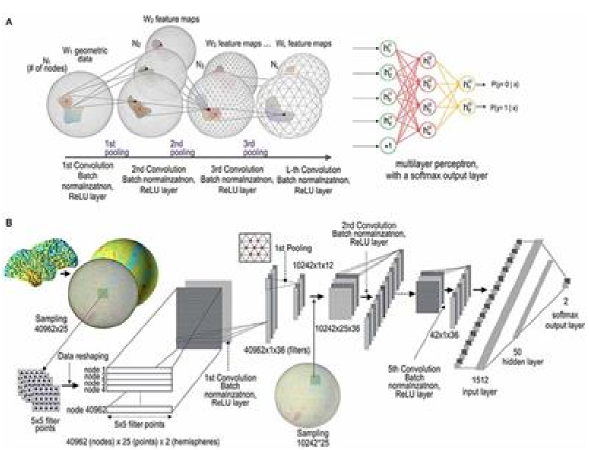
What really matters for AI? Data?

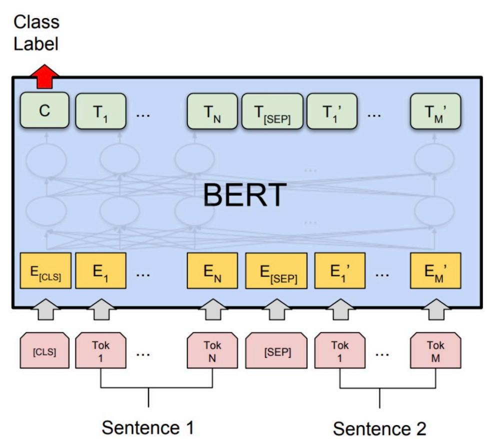
What really matters for AI? Model?

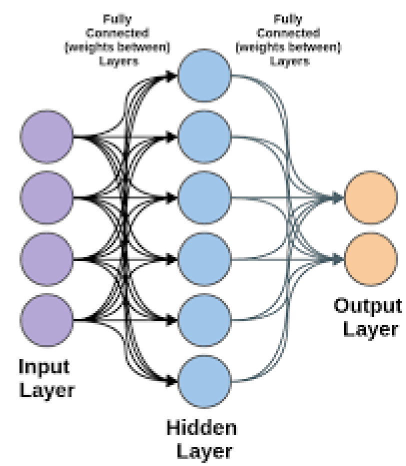
A lot to consider in modeling

Attention is all you need.

MLP is all you need.







Ashish Vaswani et al. Attention Is All You Need. NeurIPS 2017.
Ilya Tolstikhin et al. MLP-Mixer: An all-MLP Architecture for VisionMLP-Mixer: An all-MLP Architecture for Vision. arXiv 2021.
Luke Melas-Kyriazi. Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet. arXiv 2021.
Meng-Hao Guo et al. Beyond Self-attention: External Attention using Two Linear Layers for Visual Tasks. arXiv 2021.

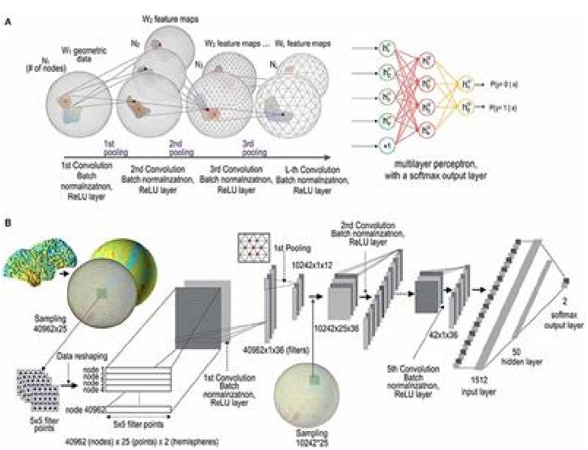
What really matters for AI? Model?

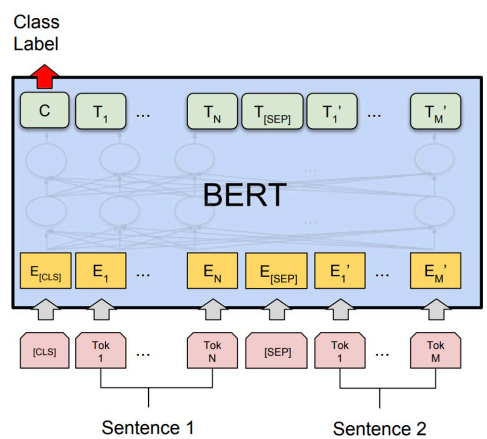
Model is getting simpler.

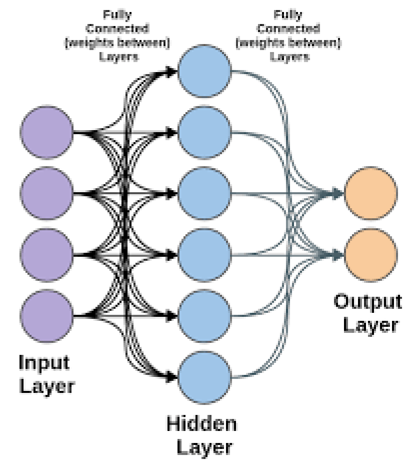
A lot to consider in modeling

Attention is all you need.

MLP is all you need.







Ashish Vaswani et al. Attention Is All You Need. NeurIPS 2017.
Ilya Tolstikhin et al. MLP-Mixer: An all-MLP Architecture for VisionMLP-Mixer: An all-MLP Architecture for Vision. arXiv 2021.
Luke Melas-Kyriazi. Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet. arXiv 2021.
Meng-Hao Guo et al. Beyond Self-attention: External Attention using Two Linear Layers for Visual Tasks. arXiv 2021.

What really matters for AI? Learning!

Do we have evidence?

MS MARCO Document Ranking Leaderboard

Search:

date		description	team	paper	code	type	MRR@100 (Dev)	MRR@100 (Eval)	tweet
2021/04/25	🏆	**PROP_step400K base + doc2query top1000(ensemble v0.2)**	Yingyan Li, Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xueqi Cheng - ICT, CAS	[paper]		full ranking	0.479	0.423	
2021/04/28		**Knowledge Retrieval**	HuaweiPoissonLab, RUCIR			full ranking	0.482	0.423	
2021/05/10		**Knowledge Retrieval**	HuaweiPoissonLab, RUCIR			full ranking	0.484	0.423	
2021/04/27		**ANCE BS+GL**	Jiajia Ding*, Chunyu Li* - PingAn			full ranking	0.489	0.421	
2021/04/18	🏆	**ANCE + LongP (ensemble)**	Soonhwan Kwon,Minyoung Lee, Samsung SDS AI Research			full ranking	0.481	0.420	

Do we have evidence?

MS MARCO Document Ranking Leaderboard

Search:

date		description	team	paper	code	type	MRR@100 (Dev)	MRR@100 (Eval)	tweet
2021/04/25		**PROP_step400K base + doc2query top1000(ensemble v0.2)**	Yingyan Li, Xinyu Ma, Jiafeng	[paper]		full	0.479	0.423	
2021/04/28		**Knowledge Retrieval**							
2021/05/10		**Knowledge Retrieval**							
2021/04/27		**ANCE BS+GL**							
2021/04/18		**ANCE + LongP (ensemble)**							

KeyPhrase Extraction(10/18/2019) ranked by F1 @3 on Eval

Rank	Model	Submission Date	F1 @1,@3,@5
1	**ETC-large** anonymous	May31 st, 2020	0.393, **0.420**, 0.360
2	**RoBERTa-JointKPE (Base)** Si Sun(1), Chenyan Xiong(2), Zhenghao Liu(3), Zhiyuan Liu(4), Jie Bao(5) - Tsinghua University(1,3,4,5), MSR AI(2)- [Sun et al '20] and [Code]	February 6th, 2020	0.364, **0.391**, 0.338
3	**RoBERTa-RankKPE (Base)** Si Sun(1), Chenyan Xiong(2), Zhenghao Liu(3), Zhiyuan Liu(4), Jie Bao(5) - Tsinghua University(1,3,4,5), MSR AI(2)- [Sun et al '20] and [Code]	February 6th, 2020	0.361, **0.390**, 0.337
4	**SpanBERT-JointKPE (Base)** Si Sun(1), Chenyan Xiong(2), Zhenghao Liu(3), Zhiyuan Liu(4), Jie Bao(5) - Tsinghua University(1,3,4,5), MSR AI(2)- [Sun et al '20] and [Code]	February 6th, 2020	0.359, **0.385**, 0.335

Do we have evidence?

MS MARCO Document Ranking Leaderboard

date	description
2021/04/25	PROP_step400K base + doc2query top1000(ensemble v0.2)
2021/04/28	Knowledge Retrieval
2021/05/10	Knowledge Retrieval
2021/04/27	ANCE BS+GL
2021/04/18	ANCE + LongP (ensemble)

KeyPhr

Rank	
1	
2	
3	
4	

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
Jan 27, 2021	RoR (Single model) *Anonymous*	**74.9**	**72.2**	**16.4**
2 Sep 3, 2020	EL-QA (Single model) *JD AI Research*	74.6	71.6	16.3
3 Jul 29, 2020	HistoryQA (single model) *PAII Inc.*	74.2	71.5	13.9
4 Dec 16, 2019	TR-MT (ensemble) *WeChat AI*	74.4	71.3	13.6
5 Nov 11, 2019	RoBERTa + DA (ensemble) *Microsoft Dynamics 365 AI*	74.0	70.7	13.1

Do we have evidence?

MS MARCO Document Ranking Leaderboard

date		description
2021/04/25	🏆	PROP_step400K base + doc2query top1000(ensemble v0.2)
2021/04/28		Knowledge Retrieval
2021/05/10		Knowledge Retrieval
2021/04/27		ANCE BS+GL
2021/04/18	🏆	ANCE + LongP (ensemble)

KeyPhr

Rank	
1	
2	
3	
4	

@5

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
Jan 27, 2021	RoR (Single model) Anonymous	**74.9**	**72.2**	**16.4**

WMT 2014 EN-DE

Models are evaluated on the English-German dataset of the Ninth Workshop on Statistical Machine Translation (WMT 2014) based on BLEU.

Model	BLEU	Paper / Source
Transformer Big + BT (Edunov et al., 2018)	35.0	Understanding Back-Translation at Scale
DeepL	33.3	DeepL Press release
Admin (Liu et al., 2020)	30.1	Very Deep Transformers for Neural Machine Translation
MUSE (Zhao et al., 2019)	29.9	MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning
DynamicConv (Wu et al., 2019)	29.7	Pay Less Attention With Lightweight and Dynamic Convolutions

Do we have evidence?

MS MARCO Document Ranking Leaderboard

date	description
2021/04/25	PROP_step400K base + doc2query top1000(ensemble v0.2)
20	
20	
20	
20	

KeyPhr

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
	RoR (Single model) ous	**74.9**	**72.2**	**16.4**

,@5

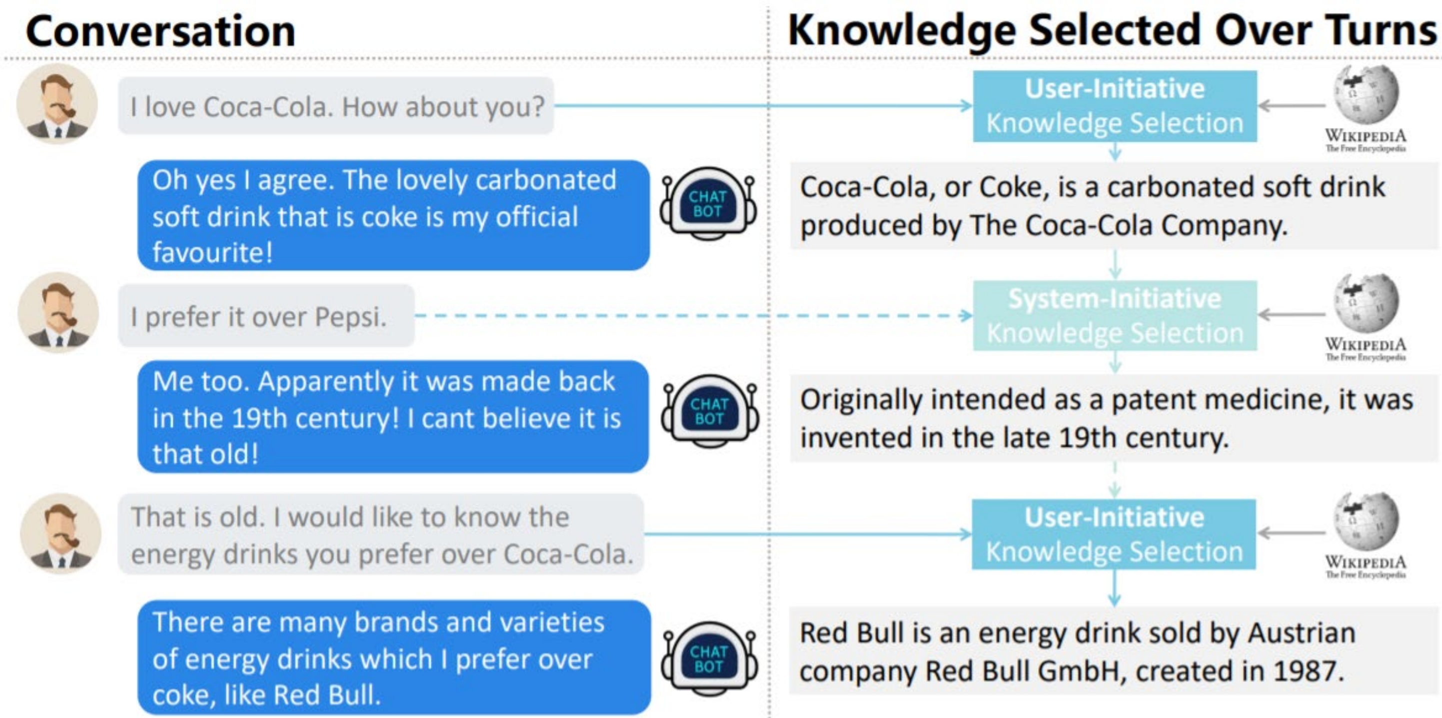
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

of the Ninth Workshop on Statistical

Source
nding Back-Translation at Scale
ess release
p Transformers for Neural Machine on
arallel Multi-Scale Attention for Sequence to e Learning
Attention With Lightweight and Dynamic ions

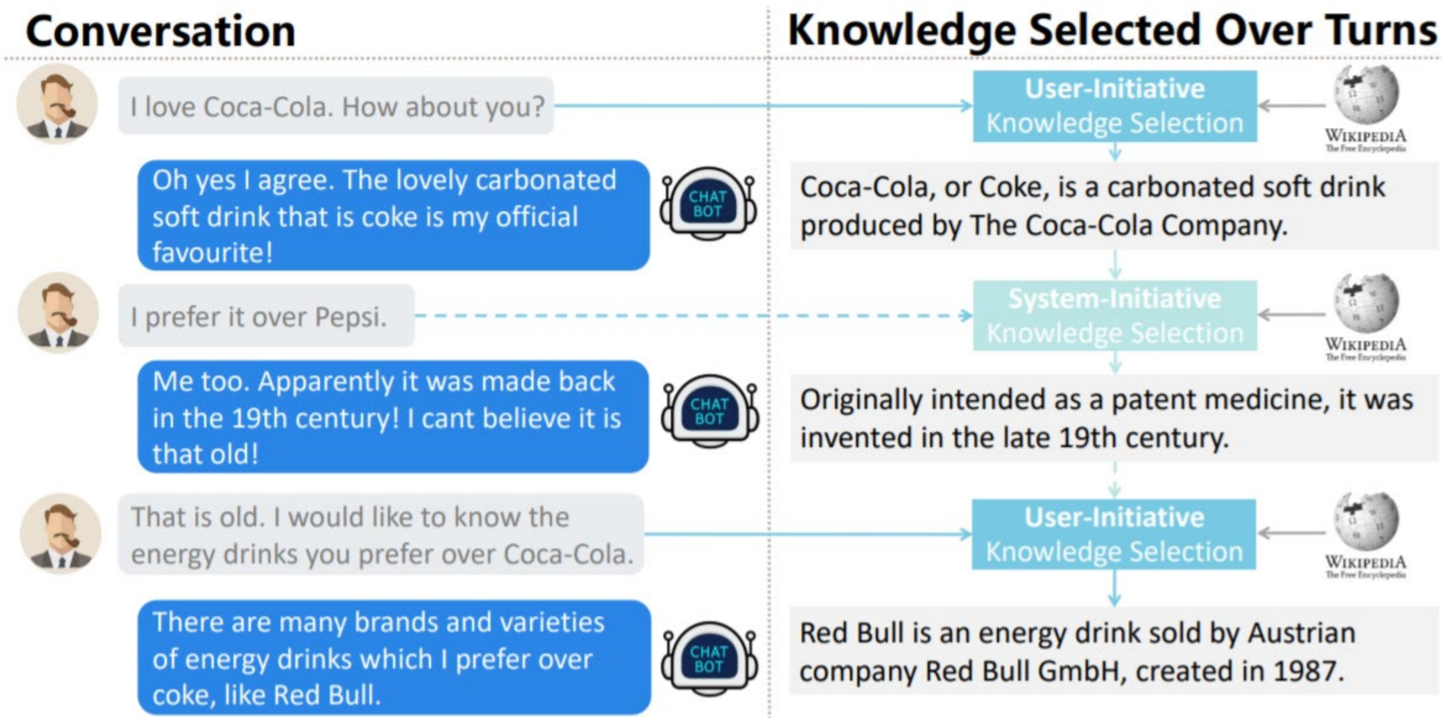
SSL for Knowledge Selection



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

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

SSL for Knowledge Selection



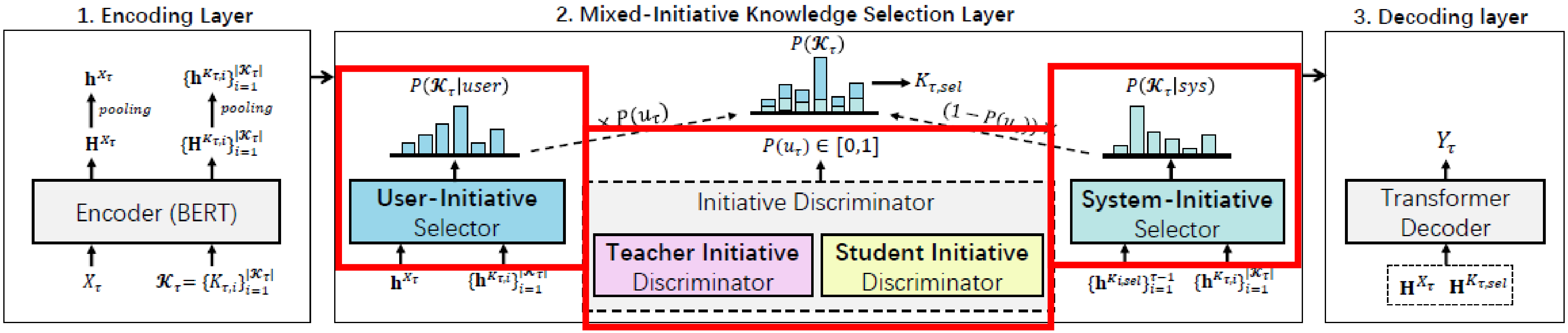
✓ Conversation is mixed initiative by nature.

✓ 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?

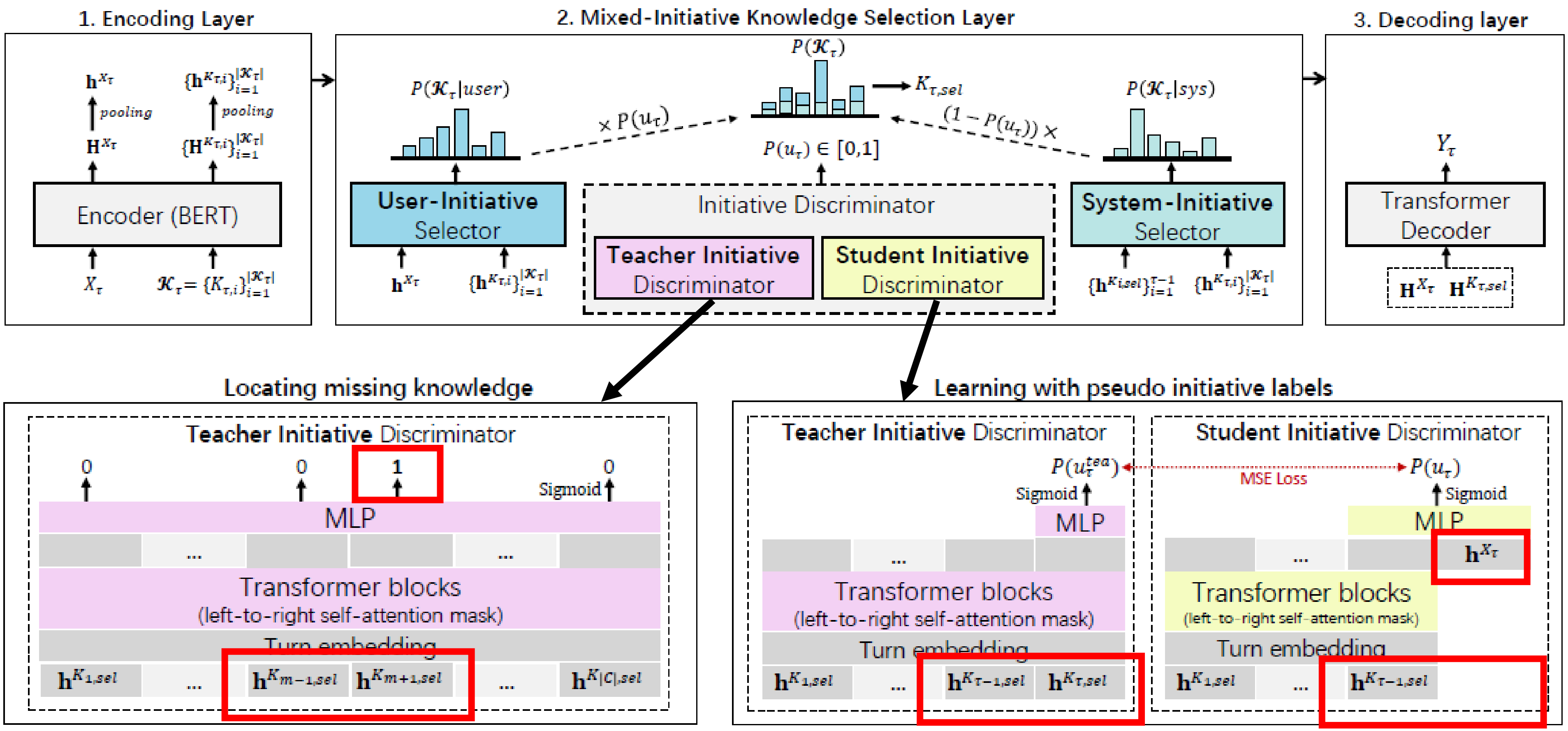
Chuan Meng et al. Initiative-Aware Self-Supervised learning for Knowledge-Grounded Conversations. In SIGIR 2021

SSL for Knowledge Selection



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

SSL for 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

SSL for 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.

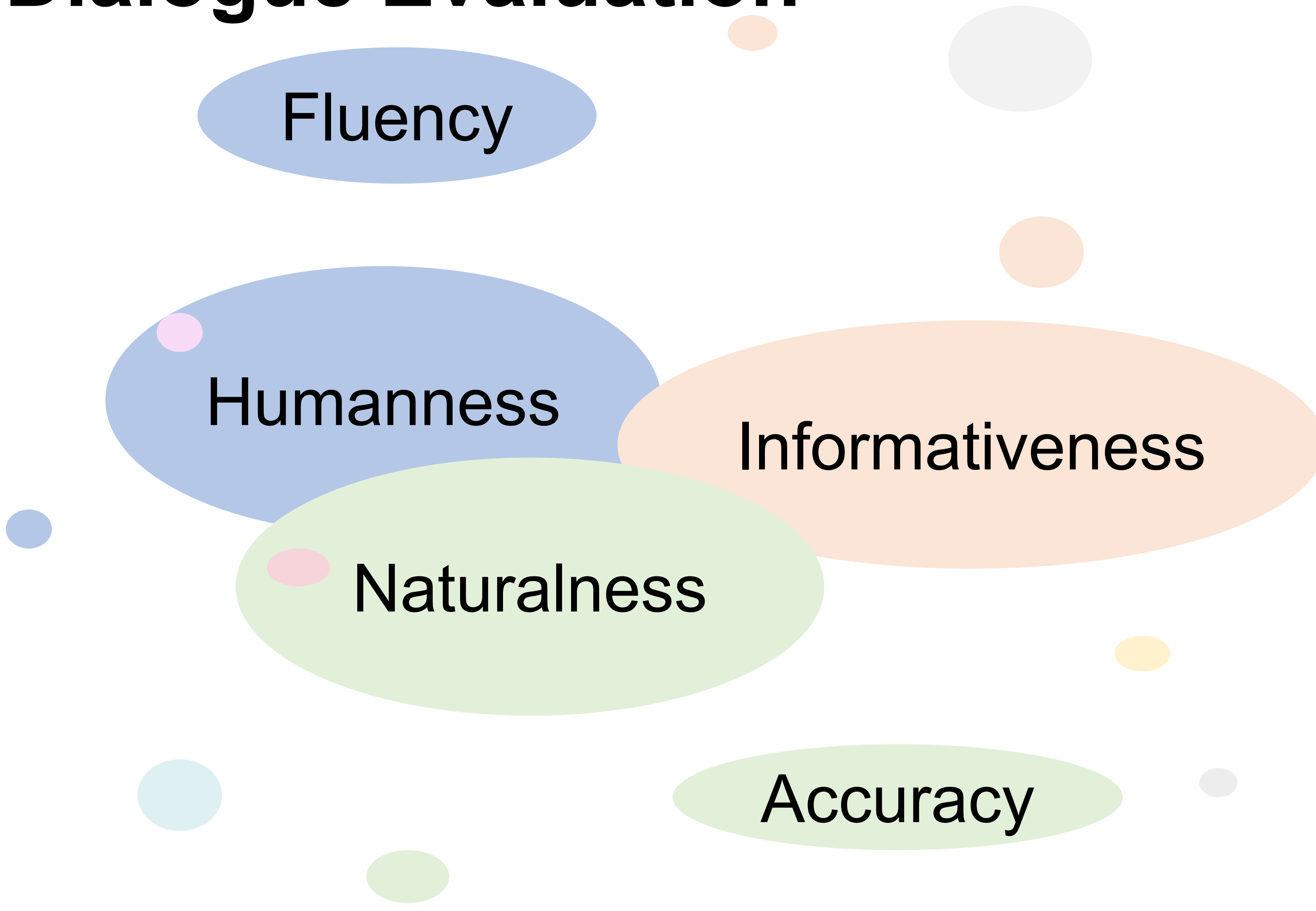
Chuan Meng et al. Initiative-Aware Self-Supervised learning for Knowledge-Grounded Conversations. In SIGIR 2021

SSL for Knowledge Selection

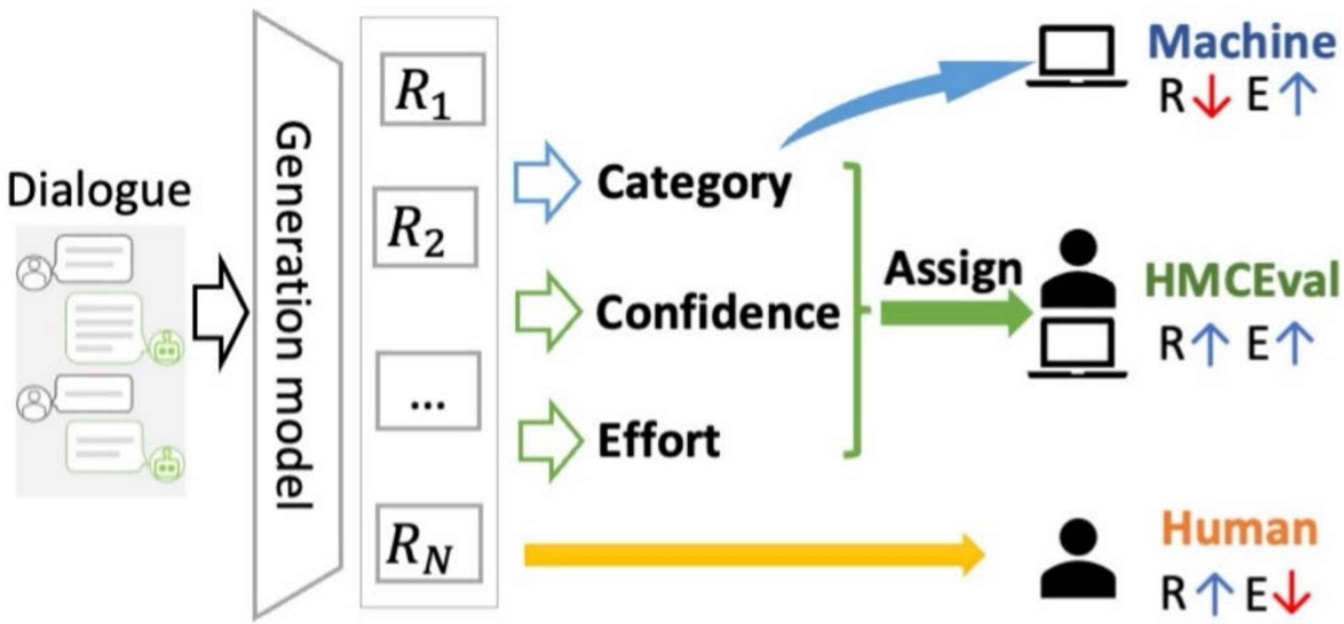
	Example 1 (Test seen)	Example 2 (Test unseen)
Knowledge pool	K_1: no knowledge used . K_2: while basketball is most often played as a team sport with five players on each side , two-on-two , and one-on-one competitions are also common . K_3: basketball is a limited contact sport played on a rectangular 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 .

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

Dialogue Evaluation



Dialogue 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

Dialogue 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. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

Dialogue 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. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

Dialogue 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. 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

Dialogue 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. A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues. In ACL 2021

Dialogue Evaluation

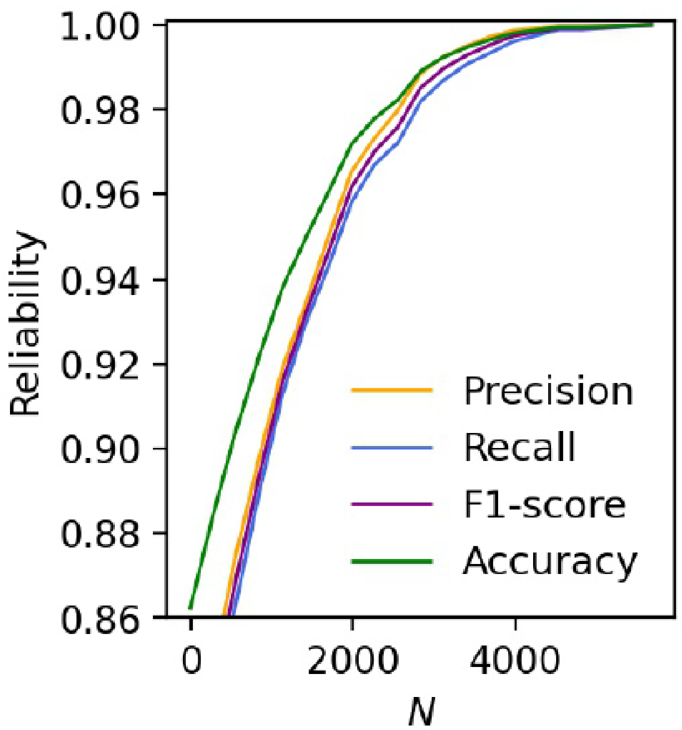
Metric	Machine	Human	HMCEval
Reliability			
Precision	0.818	1	0.983
Recall	0.803	1	0.976
F1-score	0.810	1	0.980
Accuracy	0.862	1	0.985
Efficiency			
Human ratio	0	1	0.500
Time cost	0	1	0.500

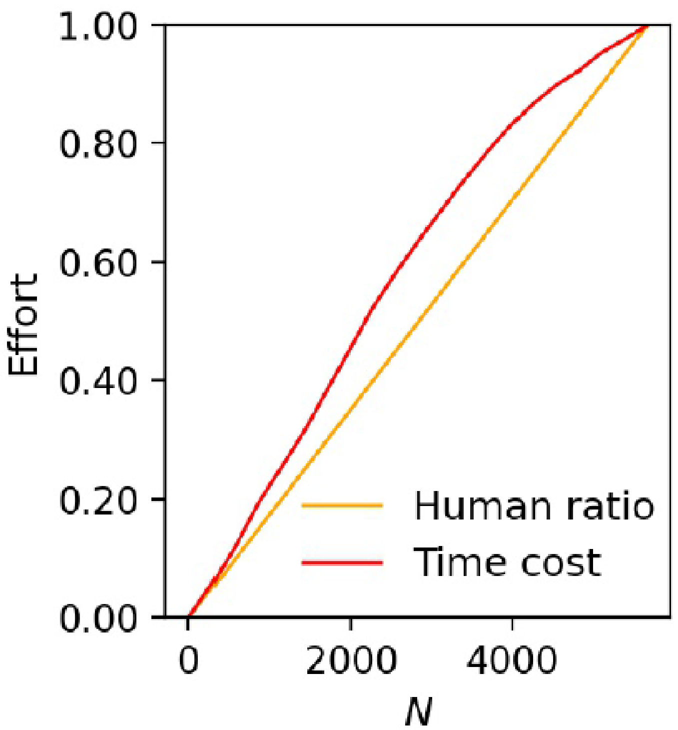
N/M=0.5

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

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

Dialogue Evaluation

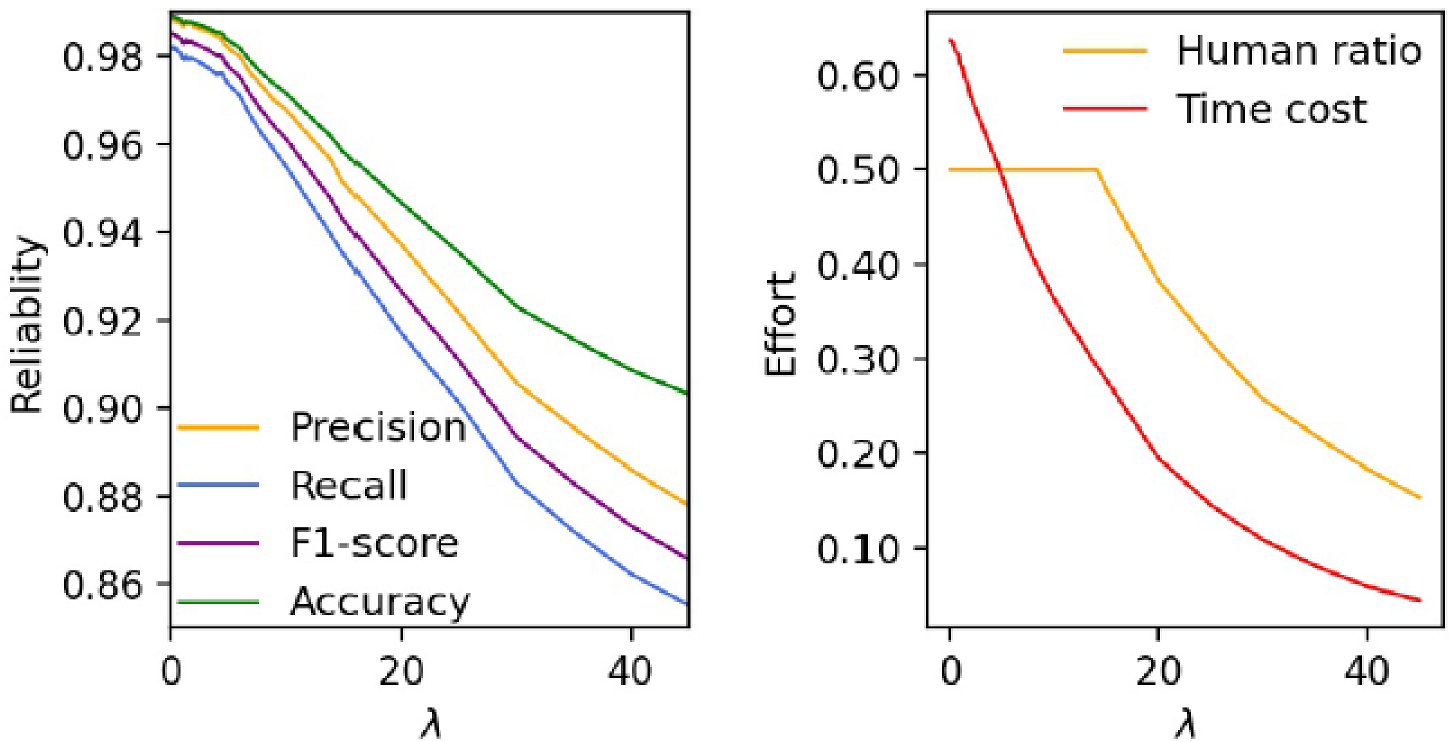




As N increases, HMCEval has better reliability, nevertheless the human effort increases.

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

Dialogue Evaluation



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

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

Yet there's more …

Future Directions

- Presentation form
  - ✓ Top n → Top 1
  - ✓ Summary, steps, list, link, ...
- Multi-modal conversations
  - ✓ Image, video, …
- Cross-/Multi-Lingual conversations
  - ✓ Leveraging available data better
- Ethics control
  - ✓ Safe AI

Thank you for your attention!

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

References

- Zhongkun Liu, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Maarten de Rijke and Ming Zhou. Learning to Ask Conversational Questions by 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.