Recent Advances in Conversational Information Retrieval (CIR)
- A review of neural approaches

Jianfeng Gao, Chenyan Xiong, Paul Bennett

Microsoft Research

SIGIR 2020

July 26, 2020 (Xi'an, China)

Outline

- **Part 1: Introduction**
  - **A short definition of CIR**
  - **Task-oriented dialog and Web search**
  - **Research tasks of CIR**
- Part 2: Conversational question answering (QA) methods
- Part 3: Conversational search methods
- Part 4: Overview of public and commercial systems

Who should attend this tutorial?

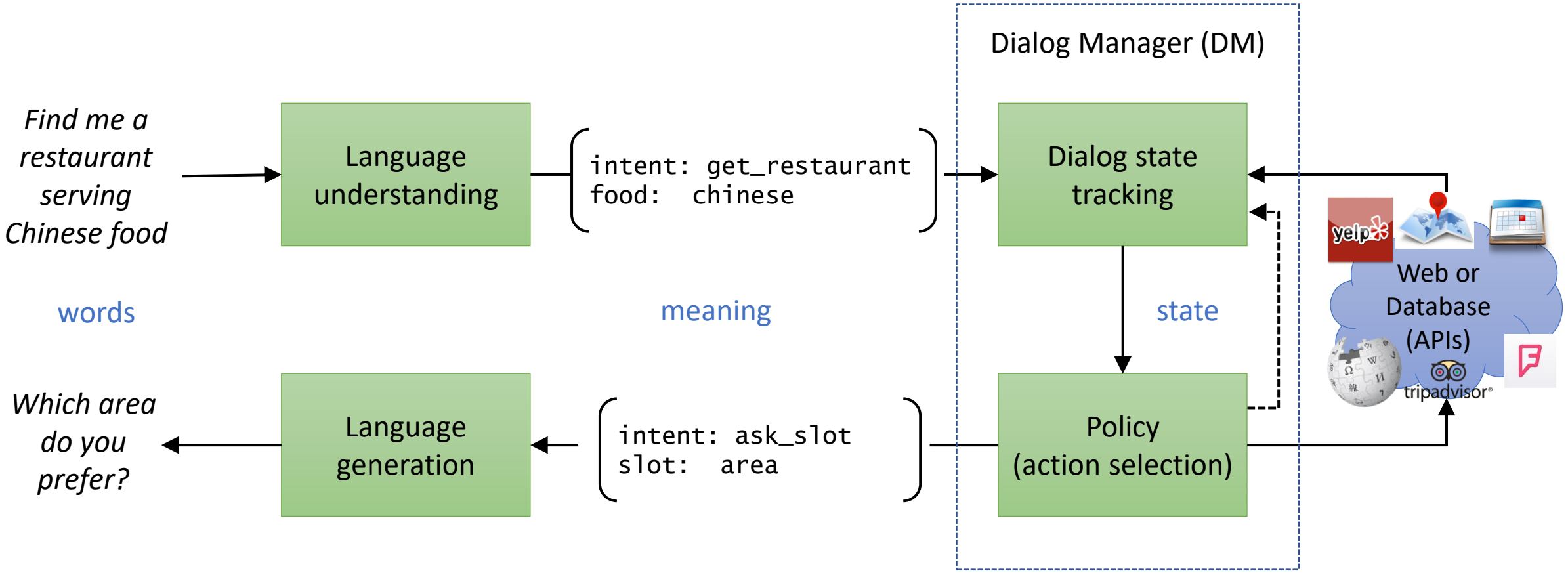
- Whoever wants to understand and develop modern CIR systems that
  - Can interact with users for information seeking via multi-turn dialogs
  - Can answer questions
  - Can help users search / look up information
  - Can help users with learning and investigation tasks
  - …
- Focus on neural approaches in this tutorial
- Hybrid approaches that combine classical AI methods and deep learning methods are widely used to build real-world systems

A short definition of CIR

- A Conversational Information Retrieval (CIR) system is
  - an information retrieval (IR) system with
  - a conversational interface which
    - allows users to interact with the system to seek information
    - via multi-turn conversations of natural language.

- CIR is a task-oriented dialog system (aka. task-oriented bot)
  - Complete tasks (e.g., information seeking) via multi-turn conversations of natural language

Classical task-oriented dialog system architecture



User-System interaction in task-oriented dialog

- User: know (roughly) what she needs, but not what is available.
- System: know what is available, but not what a user needs
- dialog: a *two-way* process of knowing each other to make a deal

A (multi-domain) dialog session

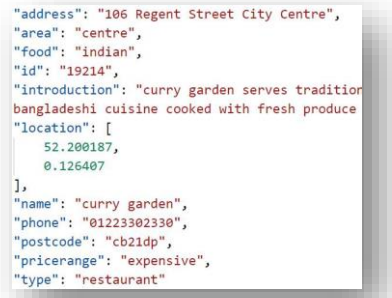
DB State

Belief State

Attraction

```
"address": "98 king street",
"area": "centre",
"entrance fee": "free",
"id": "5",
"location": [
    52.207409,
    0.126738
],
"name": "broughton house gallery",
"openhours": "?",
"phone": "01223314960",
"postcode": "cb11ln",
"pricerange": "free",
"type": "museum"
```

Restaurant



```
"address": "106 Regent Street City Centre",
"area": "centre",
"food": "indian",
"id": "19214",
"introduction": "curry garden serves tradition
bangladeshi cuisine cooked with fresh produce
"location": [
    52.200187,
    0.126407
],
"name": "curry garden",
"phone": "01223302330",
"postcode": "cb21dp",
"pricerange": "expensive",
"type": "restaurant"
```

1	**H:** I am looking for a museum in the center of town do you have any suggestion?
2	**M:** there are *6* museums in the *centre* of town . how about the *Broughton house gallery* at *98 king street* ?
3	**H:** nice, can you tell me what is the phone number and address ?
4	**M:** the phone number for the *Broughton house gallery* is *01223314960* and they are located at *98 king street* .
5	**H:** ok thanks, can you recommend a restaurant in the same area serving Indian food
6	**M:** *curry garden* serves *Indian* food and is located in the *centre* . would you like me to make a reservation ?
7	**H:** sure, please book a table for two people on thursday at 18:00
8	**M:** ok , i have a table for *2* at *18:00* on *Thursday* . your reference number is *1234567890* . is there anything else I can help you with ?
9	**H:** that's it, bye.
10	**M:** have a great time and please contact us if you need any further assistance . goodbye .

Belief state:
{
'attraction':{'type': 'museum',
'area': 'centre'}
}

Belief state:
{
'restaurant': {'food': 'indian', 'area': 'centre'},
'booking': {'day': 'Thursday', 'people': '2', 'time': '18:00'},
'attraction':{'type': 'museum',
'area': 'centre'}
}

[Peng+20]

User-system interaction in Web search

- User: know (roughly) what she needs, but not what is available.
- System: know what is available, but not what a user needs
- Generally viewed as a one-way information seeking process
  - User plays a proactive role to iteratively
    - issue a query,
    - inspect search results,
    - reformulate the query
  - System plays a passive role to make search more effective
    - Autocomplete a query
    - Organize search results (SERP)
    - Suggest related queries

broughton house

Jianfeng 30

ALL WORK IMAGES VIDEOS MAPS NEWS SHOPPING

3,470,000 Results Any time

Microsoft See work results for **broughton house**

Why am I seeing this

Broughton House is the Home of Salford & Greater ...

https://www.broughtonhouse.com

Broughton House is home of Salford and Greater Manchester's Veterans, providing care to the Armed Forces and Merchant Navy in the North West of England. **Broughton House** Care Home for Ex-Service Men and Women is a charity registered in England and Wales (1155225)

Governance & Management · About Us · Broughton House Retirement Living

Images of Broughton House

bing.com/images

Kirkcudbright Grange Sheffield Drink Grange-Over-



See more images of Broughton House

Broughton House - Wikipedia

https://en.wikipedia.org/wiki/Broughton_House

Overview History House and gardens External links

Broughton House is an 18th-century town house standing on the High Street of Kirkcudbright, Scotland. It was the home of Scots impressionist artist E. A. Hornel between 1901 and his death in 1933. During this time Hornel remodelled the house and created the Japanese-influenced gardens. Since 1997 it has been in the care of the National Trust for Scotland, and is maintained as "a living museum of Hornel's life and work". The house is a category A listed building, and the gardens are included on the Inventory of Garden...

broughton house

broughton house
broughton house **nj**
broughton house **uk**
broughton house **savannah**
broughton house **scotland**
broughton house **kirkcudbright**

COVID-19

Temporary closure or reduced hours may be in effect

Broughton House

Town House

Broughton House is an 18th-century town house standing on the High Street of Kirkcudbright, Sco... +

Wikipedia

Address: 12 High Street, Kirkcudbright, Scotland DG6 4

Phone: +44 1557 330437

Data from: Wikipedia

Wikipedia text under CC-BY-SA license

Suggest an edit

PEOPLE ALSO ASK

- What is a Boughton House?
- Where is Boughton House in England?
- What is a Coughton house?
- Where is Coughton house?

Feedback

System should interact with users more actively

- How people search -- Information seeking
  - Information lookup – short search sessions;
  - Exploratory search based on a dynamic model, an iterative “sense-making” process where users learn as they search, and adjust their information needs as they see search results.

- Effective information seeking requires interaction btw users and a system that explicitly models the interaction by
  - Tracking belief state (user intent)
  - Asking clarification questions
  - Providing recommendations
  - Using natural language as input/output

[Hearst+11; Collins-Thompson+ 17; Bates 89]

A long definition of CIR - the RRIMS properties

- User **R**evealment: help users express their information needs
  - E.g., query suggestion, autocompletion
- System **R**evealment: reveal to users what is available, what it can or cannot do
  - E.g., recommendation, SERP
- Mixed **I**nitiative: system and user both can take initiative (two-way conversation)
  - E.g., asking clarification questions
- **M**emory: users can reference past statement
  - State tracking
- **S**et Retrieval: system can reason about the utility of sets of complementary items
  - Task-oriented, contextual search or QA

[Radlinski&Craswell 17]

CIR research tasks (task-oriented dialog modules)

- What we will cover in this tutorial
  - Conversational Query Understanding (LU, belief state tracking)
  - Conversational document ranking (database state tracking)
  - Learning to ask clarification questions (action select via dialog policy, LG)
  - Conversational leading suggestions (action select via dialog policy, LG)
  - Search result presentation (response generation, LG)

- Early work on CIR [Croft's keynote at SIGIR-19]

- We start with conversational QA which is a sub-task of CIR

Outline

- Part 1: Introduction
- **Part 2: Conversational QA methods**
  - **Conversational QA over knowledge bases**
  - **Conversational QA over texts**
- Part 3: Conversational search methods
- Part 4: Case study of commercial systems

Conversational QA over Knowledge Bases

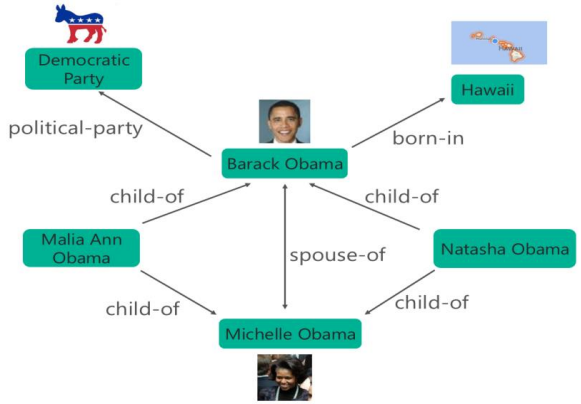
- Knowledge bases and QAs
- C-KBQA system architecture
  - Semantic parser
  - Dialog manager
  - Response generation
- KBQA w/o semantic parser
- Open benchmarks

Knowledge bases

- Relational databases
  - Entity-centric knowledge base
  - Q: what super-hero from Earth appeared first?
- Knowledge Graph
  - Properties of billions of entities
  - Relations among them
  - (relation, subject, object) tuples
  - Freebase, FB Entity Graph, MS Satori, Google KG etc.
  - Q: what is Obama's citizenship?
- KGs work with paths while DBs work with sets

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental



[Iyyer+18; Gao+19]

Question-Answer pairs

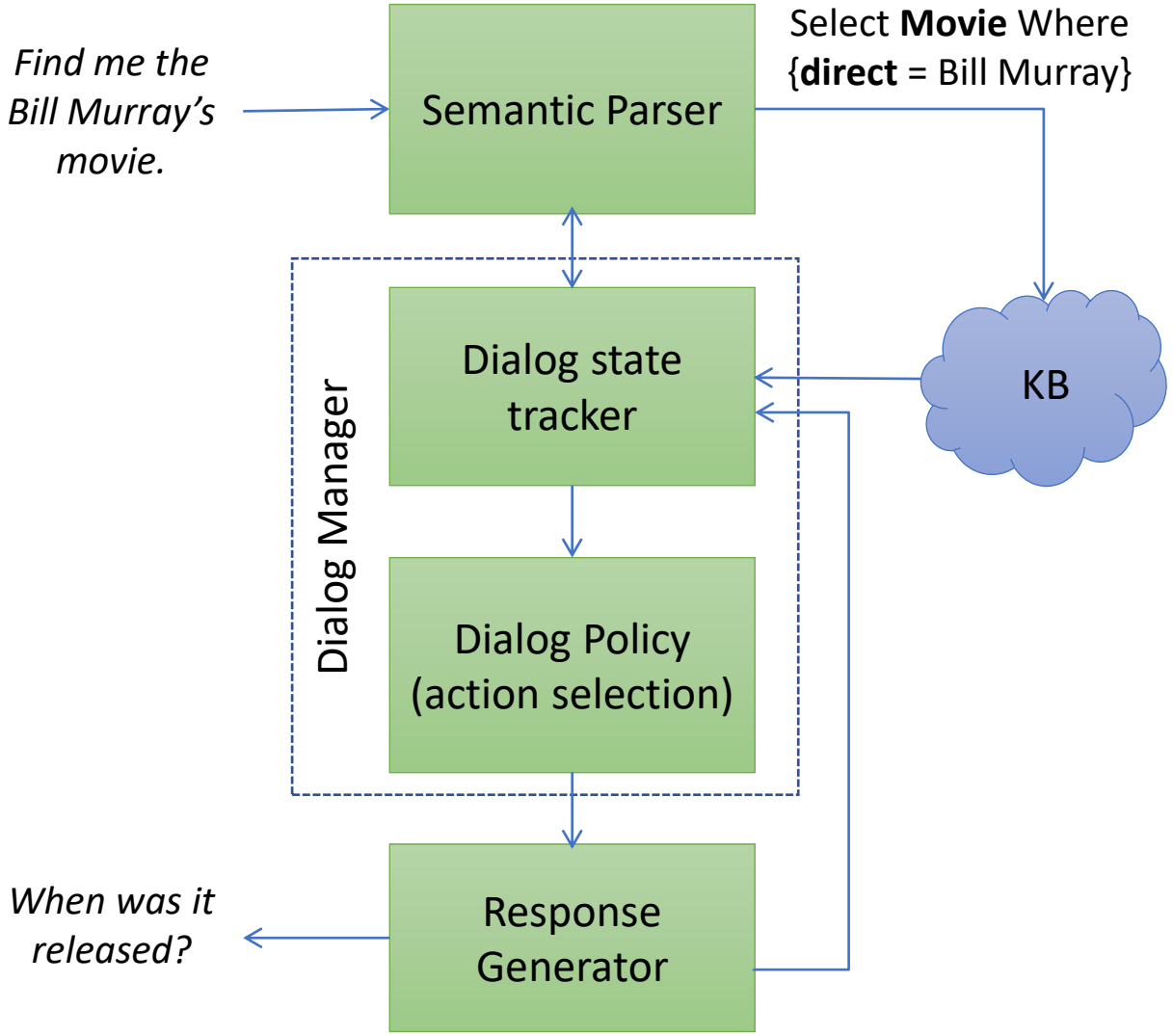
- Simple questions
  - can be answered from a single tuple
  - Object? / Subject? / Relation?
- Complex questions
  - requires reasoning over one or more tuples
  - Logical / quantitively / comparative
- Sequential QA pairs
  - A sequence of related pairs
  - Ellipses, coreference, clarifications, etc.

[Saha+18]

Turn	State	Utterance
T1	Simple Q.	USER : Can you tell me which cities border Verderio Inferiore ?
	Response	SYSTEM : Cornate d'Adda, Bernareggio, Robbiate
T2	Coreference Q.	USER : And which cities flank that one ?
	Clarification Q.	SYSTEM : Did you mean Robbiate ?
T3	Clarification	USER : No, I meant Cornate d'Adda.
	Response	SYSTEM : Bottanuco, Busnago, Trezzo sull'Adda
T4	Logical Q.	USER : Which ones share border with Robbiate and are not part of Azerbaijan ?
	Response	SYSTEM : Trezzo sull'Adda, Busnago, Bottanuco
T5	Quantitative Q.	USER : Which city is the capital of min number of french administrative divisions ?
	Response	SYSTEM : Riga
T6	Boolean Q.	USER : Is Rot am See located adjacent to that one ?
	Response	SYSTEM : No
T7	Quanti. Q.	USER : How many political territories have strategic relation with around 22 other countries ?
	Response	SYSTEM : 11
T8	Logical Q.	USER : Which are the nations connected with Robbiate and are not located in Italy ?
	Response	SYSTEM : The answer count is 11. Do you want to see all possibilities?
T9	Clarification	USER : No, show only a few of them
	Response	SYSTEM : Papua New Guinea, Jamaica, Antigua and Barbuda, Austria, New Zealand, The Bahamas, Gabon
T10	Quantitative Q.	USER : How many countries are diplomatically related to Italy ?
	Response	SYSTEM : 74
T11	Quantitative Q.	USER : And how many of them also with Australia ?
	Response	SYSTEM : 56
T12	Comparative Q.	USER : Which countries have lesser number of political relations with others than Australia ?
	Response	SYSTEM : Denmark, Canada, Grenada

C-KBQA system architecture

- Semantic Parser
  - map input + context to a semantic representation (logic form) to
  - Query the KB
- Dialog manager
  - Maintain/update state of dialog history (e.g., QA pairs, DB state)
  - Select next system action (e.g., ask clarification questions, answer)
- Response generator
  - Convert system action to natural language response
- KB search (Gao+19)



Dynamic Neural Semantic Parser (DynSP)

- Given a question (dialog history) and a table
  - Q: “which superheroes came from Earth and first appeared after 2009?”

- Generate a semantic parse (SQL-query)
  - A **select** statement (answer column)
  - Zero or more **conditions**, each contains
    - A condition column
    - An operator (=, >, <, argmax etc.) and arguments
  - Q: Select **Character** Where {**Home World** = “Earth”} & {**First Appear** > “2009”}
  - A: {Dragonwing, Harmonia}

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

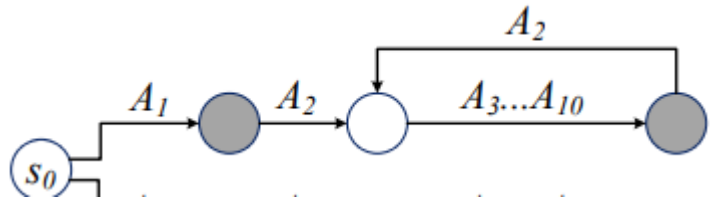
[**Iyyer+18;** Andreas+16; Yih+15]

Model formulation

- Parsing as a state-action search problem
  - A state S is a complete or partial parse (action sequence)
  - An action A is an operation to extend a parse
  - Parsing searches an end state with the highest score
- "which superheroes came from Earth and first appeared after 2009?"
  - (A_1) Select-column **Character**
  - (A_2) Cond-column **Home World**
  - (A_3) Op-Equal "Earth"
  - (A_2) Cond-column **First Appeared**
  - (A_5) Opt-GT "2009"

Id	Type	# Action instances
$\mathcal{A}_1$	Select-column	# columns
$\mathcal{A}_2$	Cond-column	# columns
$\mathcal{A}_3$	Op-Equal (=)	# rows
$\mathcal{A}_4$	Op-NotEqual ($\neq$)	# rows
$\mathcal{A}_5$	Op-GT (>)	# numbers / datetimes
$\mathcal{A}_6$	Op-GE ($\geq$)	# numbers / datetimes
$\mathcal{A}_7$	Op-LT (<)	# numbers / datetimes
$\mathcal{A}_8$	Op-LE ($\leq$)	# numbers / datetimes
$\mathcal{A}_9$	Op-ArgMin	# numbers / datetimes
$\mathcal{A}_{10}$	Op-ArgMax	# numbers / datetimes

Types of actions and the number of action instances in each type. Numbers / datetimes are the mentions discovered in the question.



Possible action transitions based on their types. Shaded circles are end states.

[**Iyyer+18;** Andreas+16; Yih+15]

How to score a state (parse)?

- Beam search to find the highest-scored parse (end state)
  - $V_\theta(S_t) = V_\theta(S_{t-1}) + \pi_\theta(S_{t-1}, A_t), V(S_0) = 0$

- Policy function, $\pi_\theta(S, A)$,
  - Scores an action given the current state
  - Parameterized using different neural networks, each for an action type
  - E.g., Select-column action is scored using the semantic similarity between question words (embedding vectors) and column name (embedding vectors)
  - $\frac{1}{|W_c|}\sum_{w_c \in W_c} \max_{w_q \in W_q} w_q^T w_c$

[**Iyyer+18;** Andreas+16; Yih+15]

Model learning

- State value function: $V_\theta(S_t) = \sum_{i=1}^{t} \pi_\theta(S_{i-1}, A_i)$
  - An E2E trainable, question-specific, neural network model
- Weakly supervised learning setting
  - Question-answer pairs are available
  - Correct parse for each question is not available
- Issue of delayed (sparse) reward
  - Reward is only available after we get a (complete) parse and the answer
- Approximate (dense) reward
  - Check the overlap of the answers of a partial parse $A(S)$ with the gold answers A^*
  - $R(S) = \frac{|A(S) \cap A^*|}{|A(S) \cup A^*|}$

[**Iyyer+18;** Andreas+16; Yih+15]

Parameter updates

- Make the state value function V_θ behave similarly to reward R
- For every state S and its (approximated) reference state S^*, we define loss as
  - $\mathcal{L}(S) = \big(V_\theta(S) - V_\theta(S^*)\big) - \big(R(S) - R(S^*)\big)$
- Improve learning efficiency by finding the most violated state $\hat{S}$

Algorithm 1 Model parameter updates

1: **for** pick a labeled data (x, A^*) **do** // labeled QA pair

2: $s^* \leftarrow \arg\max_{s \in \mathcal{E}(x)} \tilde{R}(s; A^*)$ // Finds the best approximated reference state

3: $\hat{s} \leftarrow \arg\max_{s \in \mathcal{E}(x)} V_\theta(s) - \tilde{R}(s; A^*)$ // Finds the most violated state

4: update θ by minimizing $\max(\mathcal{L}(s), 0)$

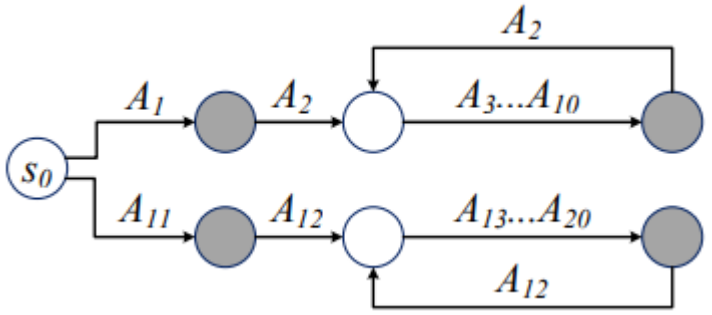
5: **end for**

[**Iyyer+18;** Taskar+04]

DynSP SQA

- "which superheroes came from Earth and first appeared after 2009?"
  - (A_1) Select-column **Character**
  - (A_2) Cond-column **Home World**
  - (A_3) Op-Equal "Earth"
  - (A_2) Cond-column **First Appeared**
  - (A_5) Opt-GT "2009"
- "which of them breathes fires"
  - (A_{12}) S-Cond-column **Powers**
  - (A_{13}) S-Op-Equal "Fire breath"

Id	Type	# Action instances
$\mathcal{A}_1$	Select-column	# columns
$\mathcal{A}_2$	Cond-column	# columns
$\mathcal{A}_3$	Op-Equal (=)	# rows
$\mathcal{A}_4$	Op-NotEqual ($\neq$)	# rows
$\mathcal{A}_5$	Op-GT (>)	# numbers / datetimes
$\mathcal{A}_6$	Op-GE ($\geq$)	# numbers / datetimes
$\mathcal{A}_7$	Op-LT (<)	# numbers / datetimes
$\mathcal{A}_8$	Op-LE ($\leq$)	# numbers / datetimes
$\mathcal{A}_9$	Op-ArgMin	# numbers / datetimes
$\mathcal{A}_{10}$	Op-ArgMax	# numbers / datetimes
$\mathcal{A}_{11}$	Subsequent	1
$\mathcal{A}_{12}$	S-Cond-column	# columns
$\mathcal{A}_{13}$	S-Op-Equal (=)	# rows
$\mathcal{A}_{14}$	S-Op-NotEqual ($\neq$)	# rows
$\mathcal{A}_{15}$	S-Op-GT (>)	# numbers / datetimes
$\mathcal{A}_{16}$	S-Op-GE ($\geq$)	# numbers / datetimes
$\mathcal{A}_{17}$	S-Op-LT (<)	# numbers / datetimes
$\mathcal{A}_{18}$	S-Op-LE ($\leq$)	# numbers / datetimes
$\mathcal{A}_{19}$	S-Op-ArgMin	# numbers / datetimes
$\mathcal{A}_{20}$	S-Op-ArgMax	# numbers / datetimes



Possible action transitions based on their types. Shaded circles are end states.

[**Iyyer+18;** Andreas+16; Yih+15]

DynSP for sequential QA (SQA)

- Given a question (history) and a table
  - Q1: which superheroes came from Earth and first appeared after 2009?
  - Q2: which of them breathes fire?

- Add **subsequent** statement (answer column) for sequential QA
  - Select **Character** Where {**Home World** = “Earth”} & {**First Appear** > “2009”}
  - A1: {Dragonwing, Harmonia}
  - Subsequent Where {**Powers** = “Fire breath”}
  - A2: {Dragonwing}

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

[**Iyyer+18**]

Query rewriting approaches to SQA

Q1: When was California founded?

A1: September 9, 1850

Q2: Who is its governor? → Who is California governor?

A2: Jerry Brown

Q3: Where is Stanford?

A3: Palo Alto, California

Q4: Who founded it? → Who founded Stanford?

A4: Leland and Jane Stanford

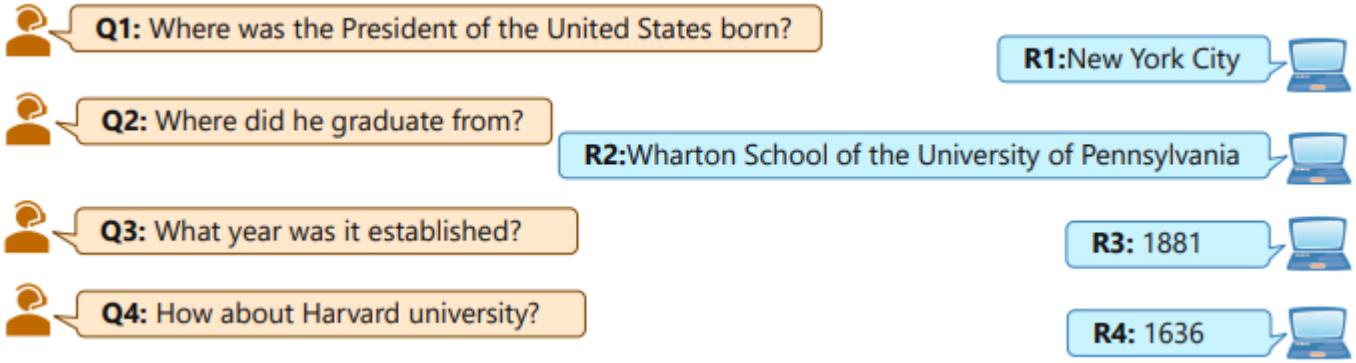
Q5: Tuition costs → Tuition cost Stanford

A5: $47,940 USD

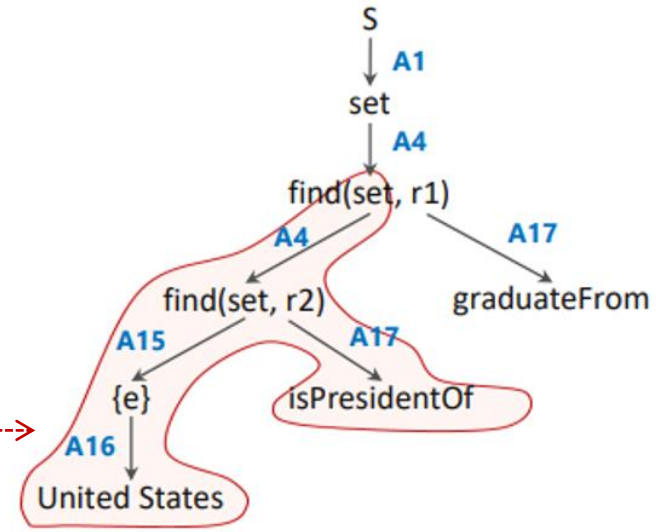
11	[Human] You like Ashin that much, huh?
	[Xiaoice] Why not? You don't like him?
12	[Human] I don't like him, neither dislike him → I don't like Ashin, neither dislike Ashin
	[Xiaoice] I love every song of him
13	[Human] Really?
	[Xiaoice] Especially *The Time Machine*
14	[Human] Oh? Never heard of that → Oh? Never heard of *The Time Machine*
	[Xiaoice] Shall I send you the song?
15	[Human] Okay → Okay, send *The Time Machine*.
	[Xiaoice] Don't know if you will like it or not~ [Play the music card]

[Ren+18; Zhou+20]

Dialog Manager – dialog memory for state tracking



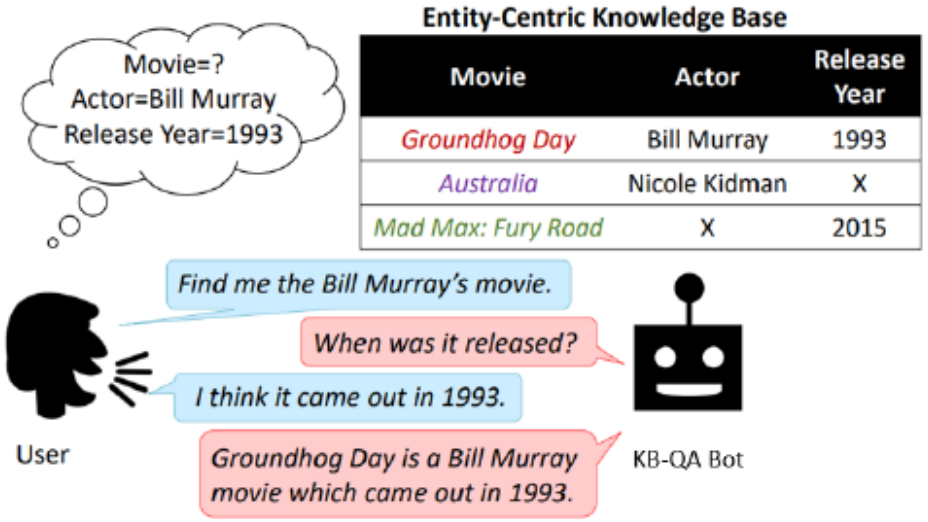
Dialog Memory (of state tracker)	
Entity	{United States, "q"} {New York City, "a"} {University of Pennsylvania, "a"} …
Predicate	{isPresidentOf} {placeGraduateFrom} {yearEstablished} …
Action subsequence (partial/complete states)	Set → $A_4\ A_{15} e_{us} r_{pres}$ Set → $A_4\ A_{15}$



[Guo+18]

Dialog Manager – policy for next action selection

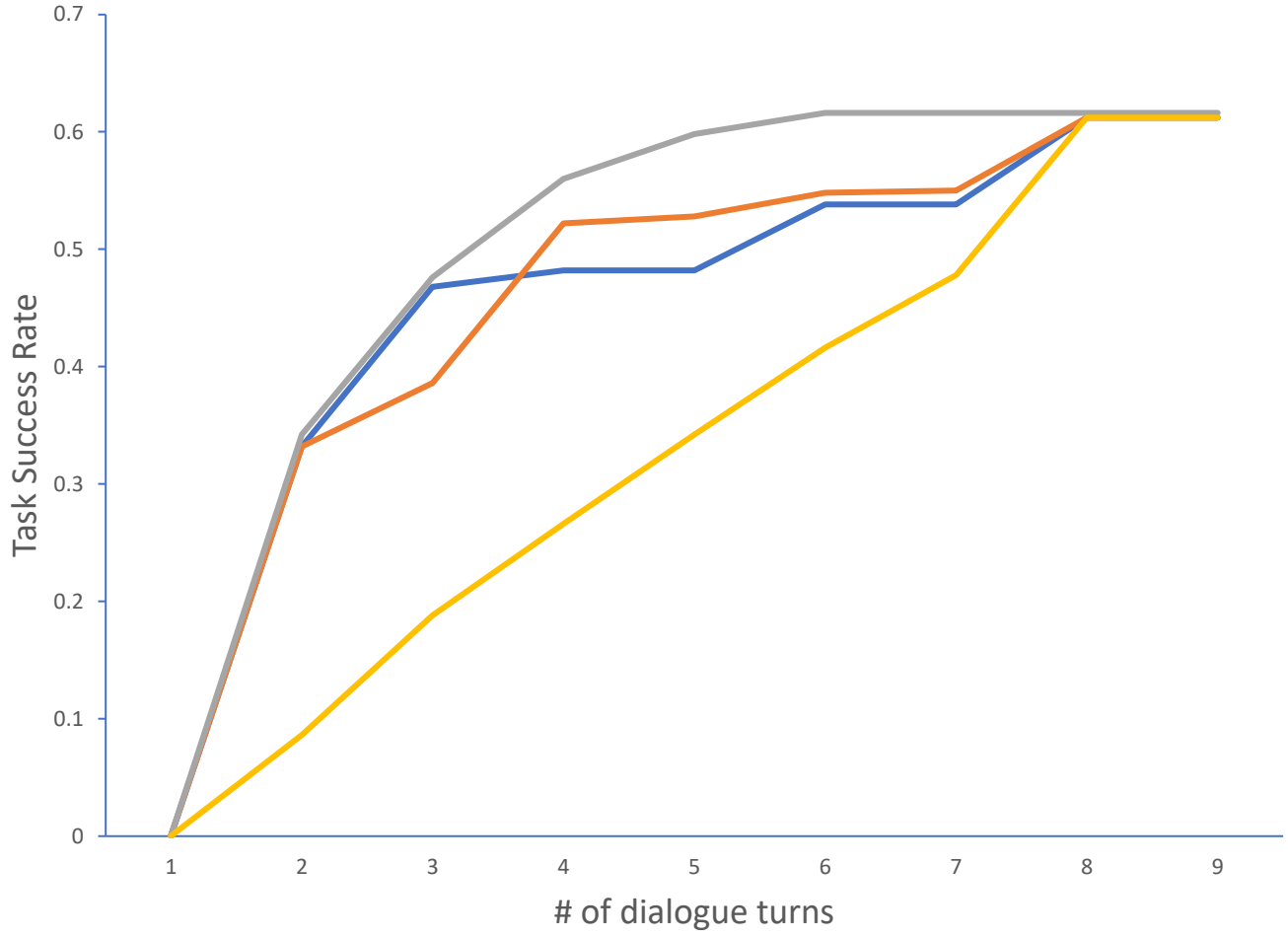
- A case study of Movie-on-demand
- System selects to
  - Either return answer or ask a clarification question.
  - What (clarification) question to ask? E.g., movie title, director, genre, actor, release-year, etc.



[Dhingra+17]

What clarification question to ask

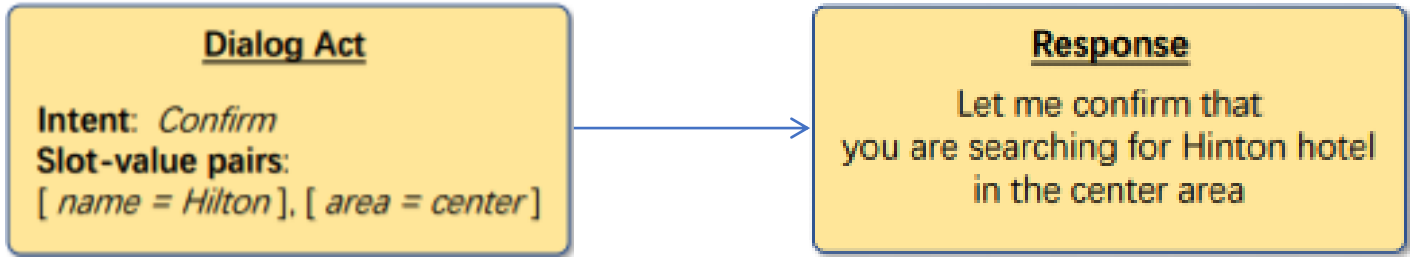
- Baseline: ask all questions in a randomly sampled order
- Ask questions that users can answer
  - learned from query logs
- Ask questions that help reduce search space
  - Entropy minimization
- Ask questions that help complete the task successfully
  - Reinforcement learning via agent-user interactions



Results on simulated users

[Wu+15; Dhingra+17; Wen+17; Gao+19]

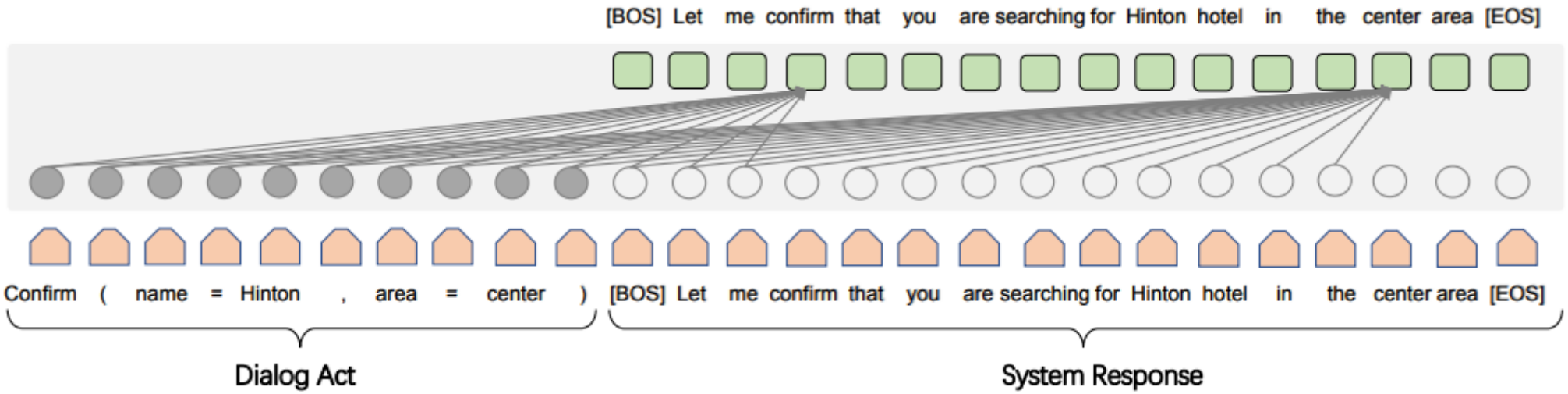
Response Generation



- Convert “dialog act” to “natural language response”
- Formulated as a seq2seq task in a few-shot learning setting
  - $p_\theta(\boldsymbol{x}|A) = \sum_{t=1}^{T} p_\theta(x_t|x_{<t}, A)$
  - Very limited training samples for each task
- Approach
  - Semantically Conditioned neural language model
  - Pre-training + fine-tuning,
    - e.g., semantically conditioned GPT (SC-GPT)

[Peng+20; Yu+19; Wen+15; Chen+19]

SC-GPT



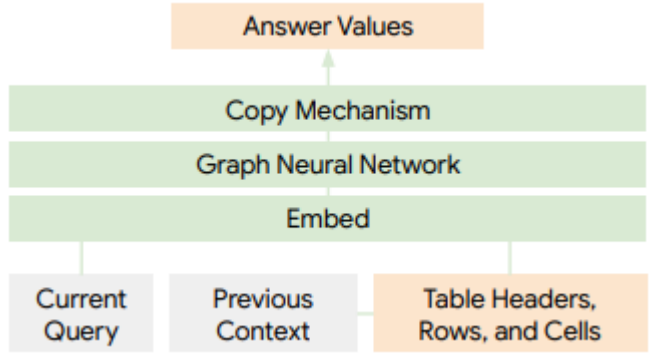
Model	Restaurant		Laptop		Hotel		TV		Attraction		Train		Taxi	
	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓
SC-LSTM	15.90	48.02	21.98	80.48	31.30	31.54	22.39	64.62	7.76	367.12	6.08	189.88	11.61	61.45
GPT-2	29.48	13.47	27.43	11.26	35.75	11.54	28.47	9.44	16.11	21.10	13.72	19.26	16.27	9.52
SC-GPT	**38.08**	**3.89**	**32.73**	**3.39**	**38.25**	**2.75**	**32.95**	**3.38**	**20.69**	**12.72**	**17.21**	**7.74**	**19.70**	**3.57**

Performance of different response generation models in few-shot setting (50 samples for each task)

[Peng+20; Raffel+19]

C-KBQA approaches w/o semantic parser

- Building semantic parsers is challenging
  - Limited amounts of training data, or
  - Weak supervision
- C-KBQA with no logic-form
  - Symbolic approach: “look before you hop”
    - Answer an initial question using any standard KBQA
    - Form a context subgraph using entities of the initial QA pair
    - Answer follow-up questions by expanding the context subgraph to find candidate answers
  - Neural approach
    - Encode KB as graphs using a GNN
    - Select answers from the encoded graph using a point network



[Christmann+19; Muller+19]

Open Benchmarks

- SQA (sequential question answering)
  - https://www.microsoft.com/en-us/download/details.aspx?id=54253
- CSQA (complex sequence question answering),
  - https://amritasaha1812.github.io/CSQA/
- ConvQuestions (conversational question answering over knowledge graphs)
  - https://convex.mpi-inf.mpg.de/
- CoSQL (conversational text-to-SQL)
  - https://yale-lily.github.io/cosql
- CLAQUA (asking clarification questions in Knowledge-based question answering)
  - https://github.com/msra-nlc/MSParS_V2.0

Conversational QA over Texts

- Tasks and datasets
- C-TextQA system architecture
- Conversational machine reading compression models
- Remarks on pre-trained language models for conversational QA

QA over text – extractive vs. abstractive QA

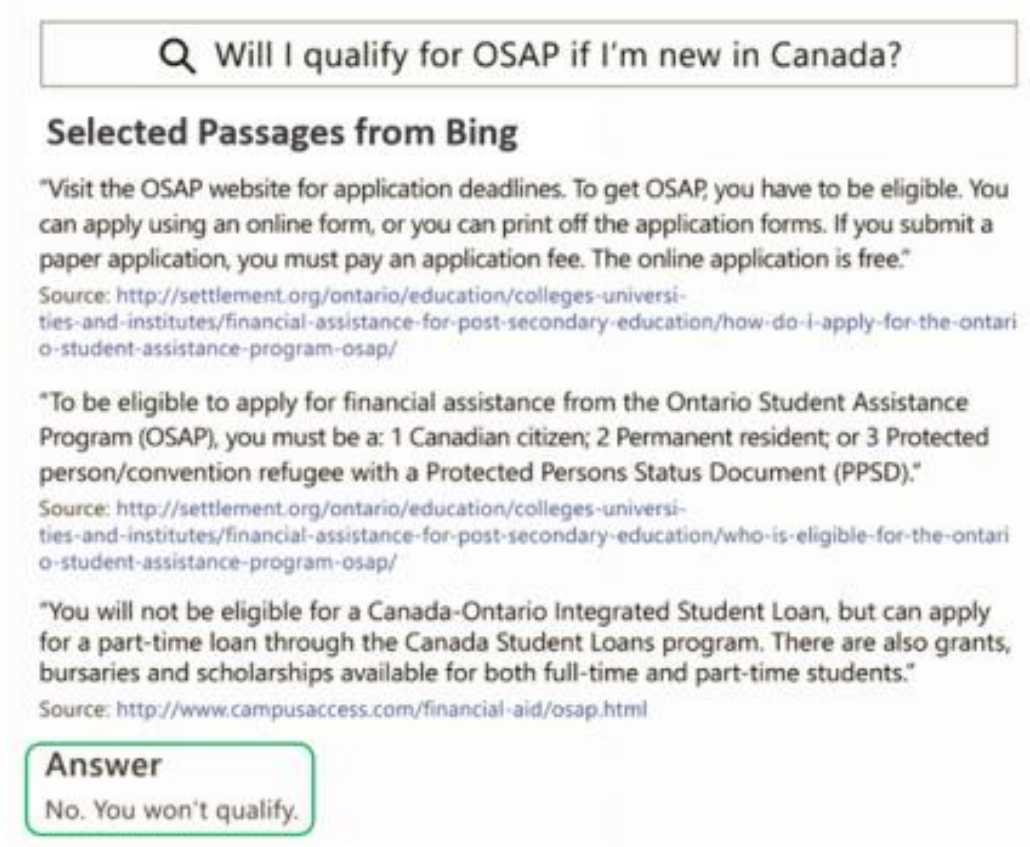
Passage

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Question: What causes precipitation to fall?
Answer: gravity

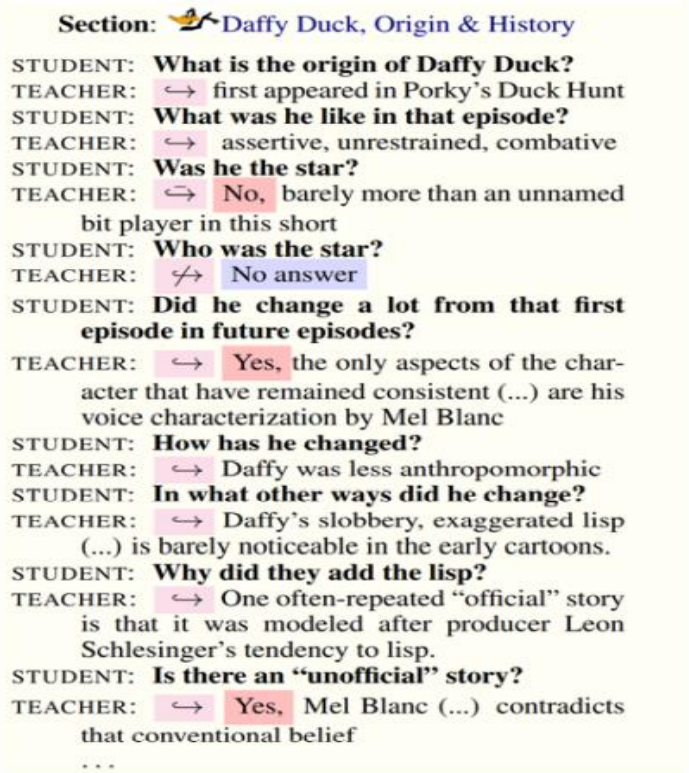
Question: What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
Answer: graupel

Question: Where do water droplets collide with ice crystals to form precipitation?
Answer: within a cloud



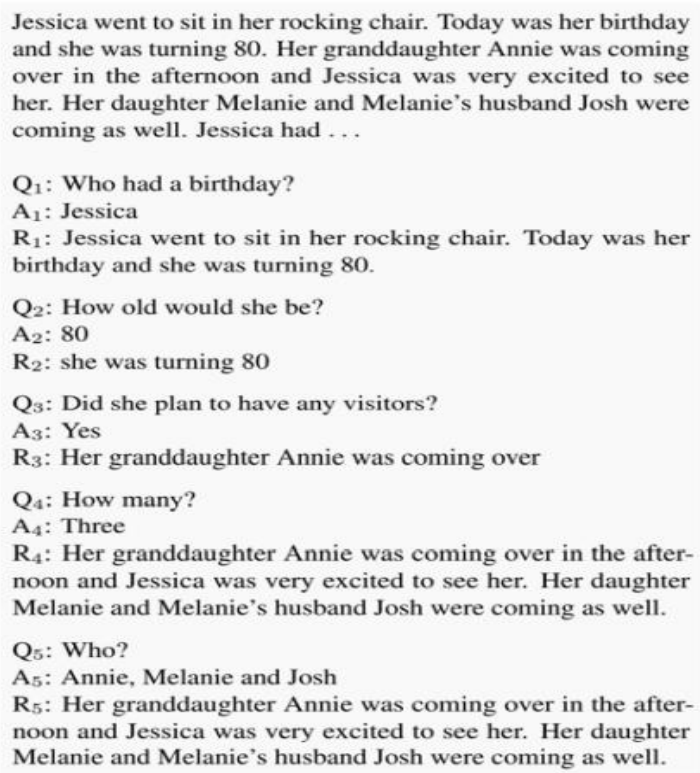
[Rajpurkar+16; Nguyen+16; Gao+19]

Conversation QA over text: CoQA & QuAC



Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ⇀ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .



Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 3.10: The examples from two conversational QA datasets. (Left) A QA dialogue example in the QuAC dataset. The student, who does not see the passage (section text), asks questions. The teacher provides answers in the form of text spans and dialogue acts. These acts include (1) whether the student should ↪, could ⇀, or should not ↛ ask a follow-up; (2) affirmation (Yes / No), and, when appropriate, (3) No answer. Figure credit: Choi et al. (2018). (Right) A QA dialogue example in the CoQA dataset. Each dialogue turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer. Figure credit: Reddy et al. (2018).

[Choi+18; Reddy+18]

Dialog behaviors in conversational QA

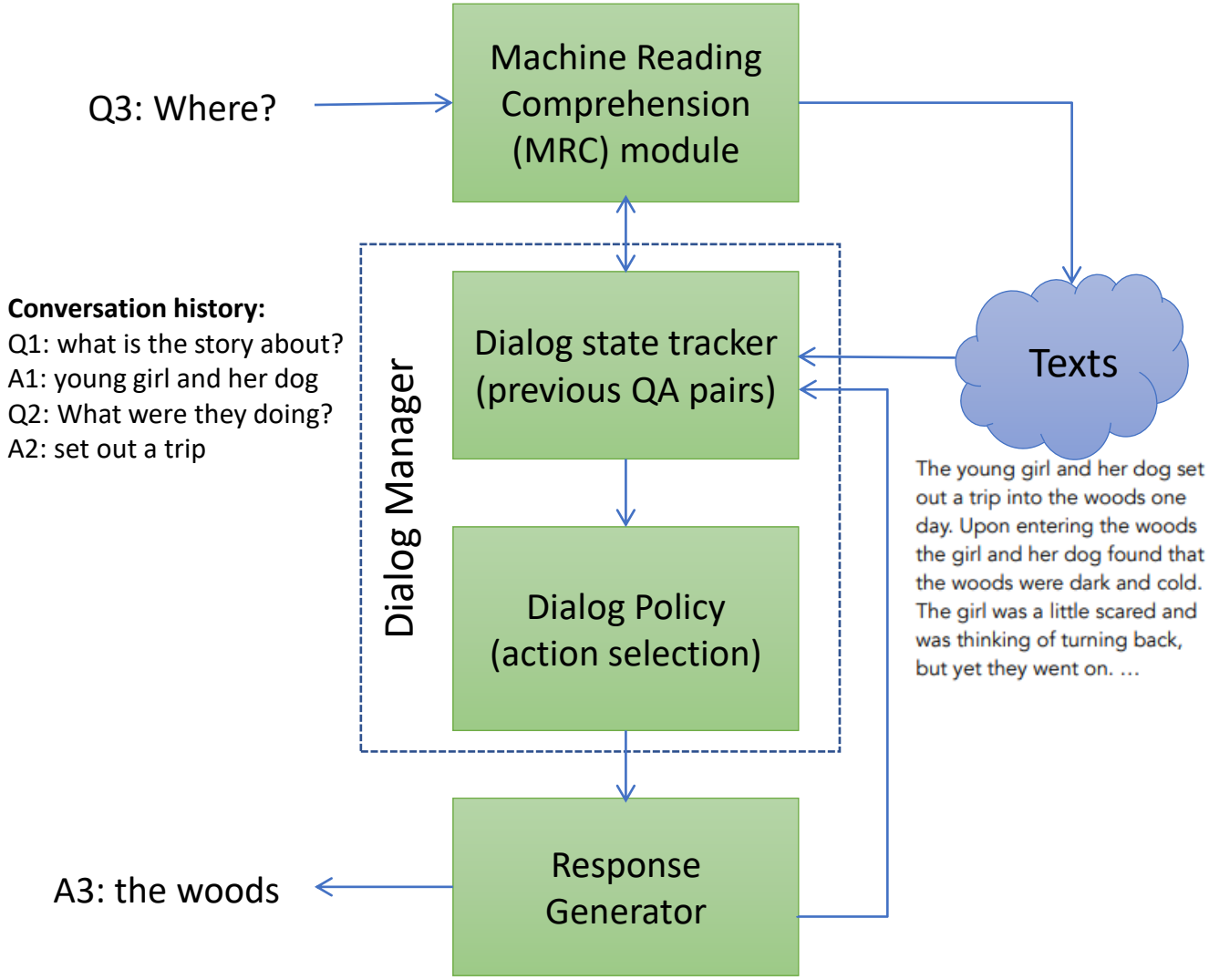
- Topic shift: a question about sth previous discussed
- Drill down: a request for more info about a topic being discussed
- Topic return: asking about a topic again after being shifted
- Clarification: reformulating a question
- Definition: asking what is meant by a team

Dataset	Topic Shift	Drill Down	Return to Topic	Clarification Question	Definition Question	Sentence Coverage	Total Questions
CoQA	21.6	72.0	2.9	0.0	0.7	63.3	722
QuAC	35.4	55.3	5.6	0.7	3.0	28.4	302

[Yatskar 19]

C-TextQA system architecture

- (Conversational) MRC
  - Find answer to a question given text and previous QA pairs
  - Extractive (span) vs. abstractive answers
- Dialog manager
  - Maintain/update state of dialog history (e.g., QA pairs)
  - Select next system action (e.g., ask clarification questions, answer)
- Response generator
  - Convert system action to natural language response



[Huang+19]

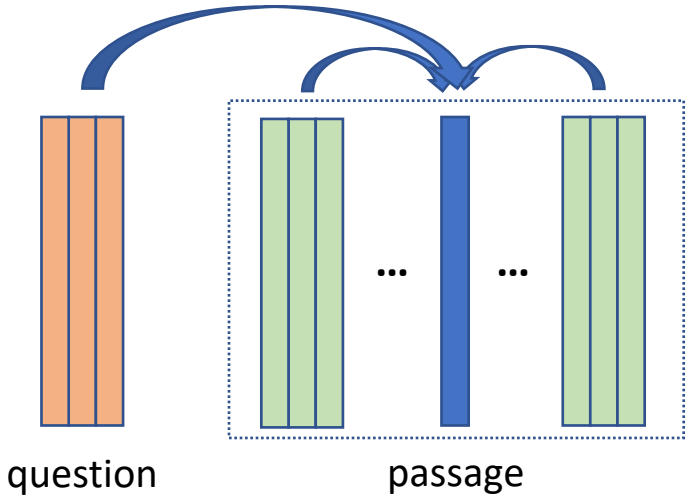
Neural MRC models for extractive TextQA

- QA as classification given (question, text)
  - Classify each word in passage as start/end/outside of the answer span
- Encoding: represent each passage word using an integrated context vector that encodes info from
  - Lexicon/word embedding (context-free)
  - Passage context
  - Question context
  - Conversation context (previous question-answer pairs)
- Prediction: predict each word (its integrated context vector) the start and end position of answer span.

[Rajpurkar+16; Huang+10; Gao+19]

Three encoding components

- Lexicon embedding e.g., GloVe
  - represent each word as a low-dim continuous vector
- Passage contextual embedding e.g., Bi-LSTM/RNN, ELMo, Self-Attention/BERT
  - capture context info for each word within a passage
- Question contextual embedding e.g., Attention, BERT
  - fuse question info into each passage word vector



[Pennington+14; Melamud+16; Peters+18; Devlin+19]

Neural MRC model: BiDAF

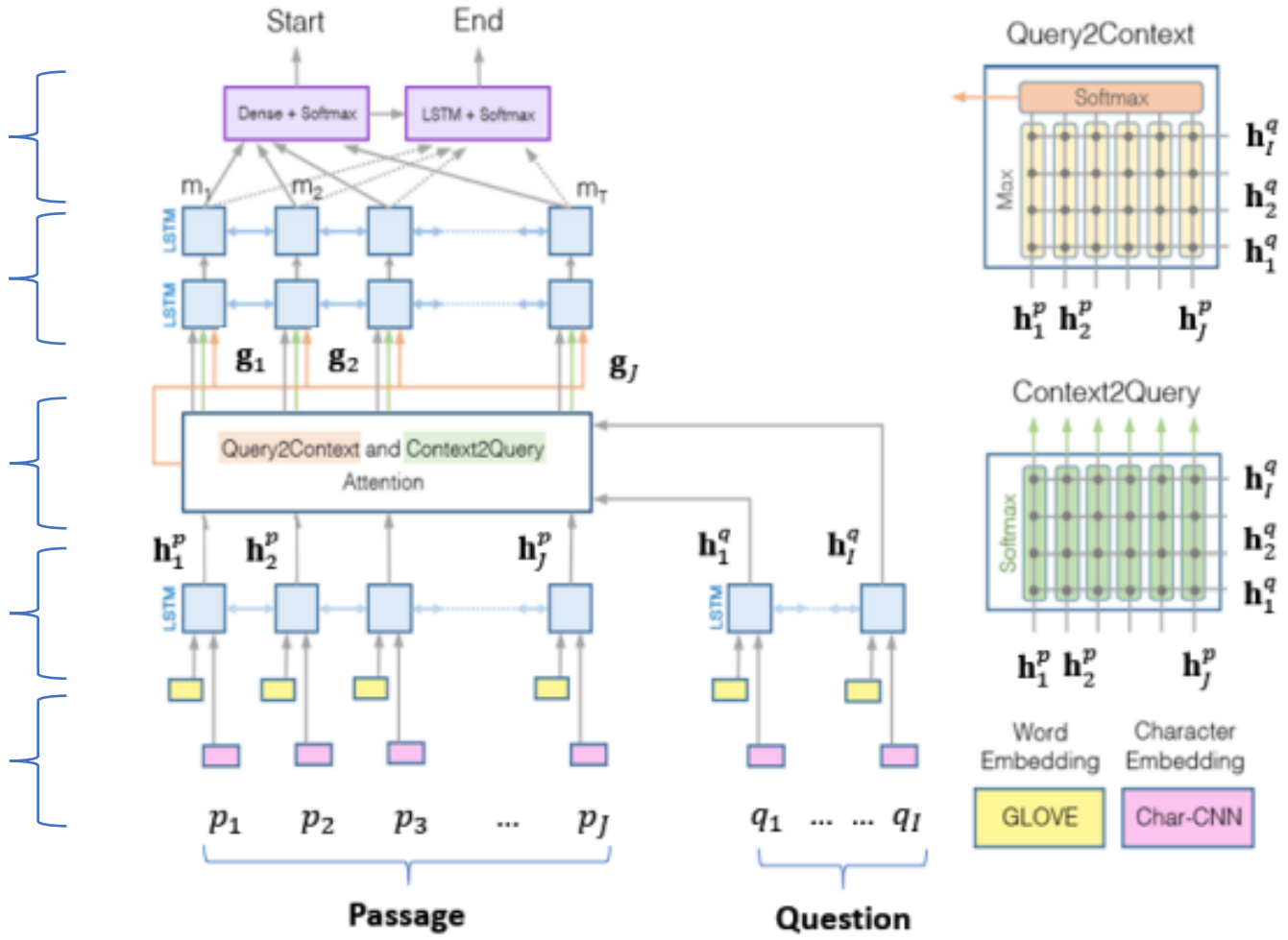
Answer prediction

Integrated context vectors

Question contextual embedding

Passage contextual Embedding

Lexicon Embedding



[Seo+16]

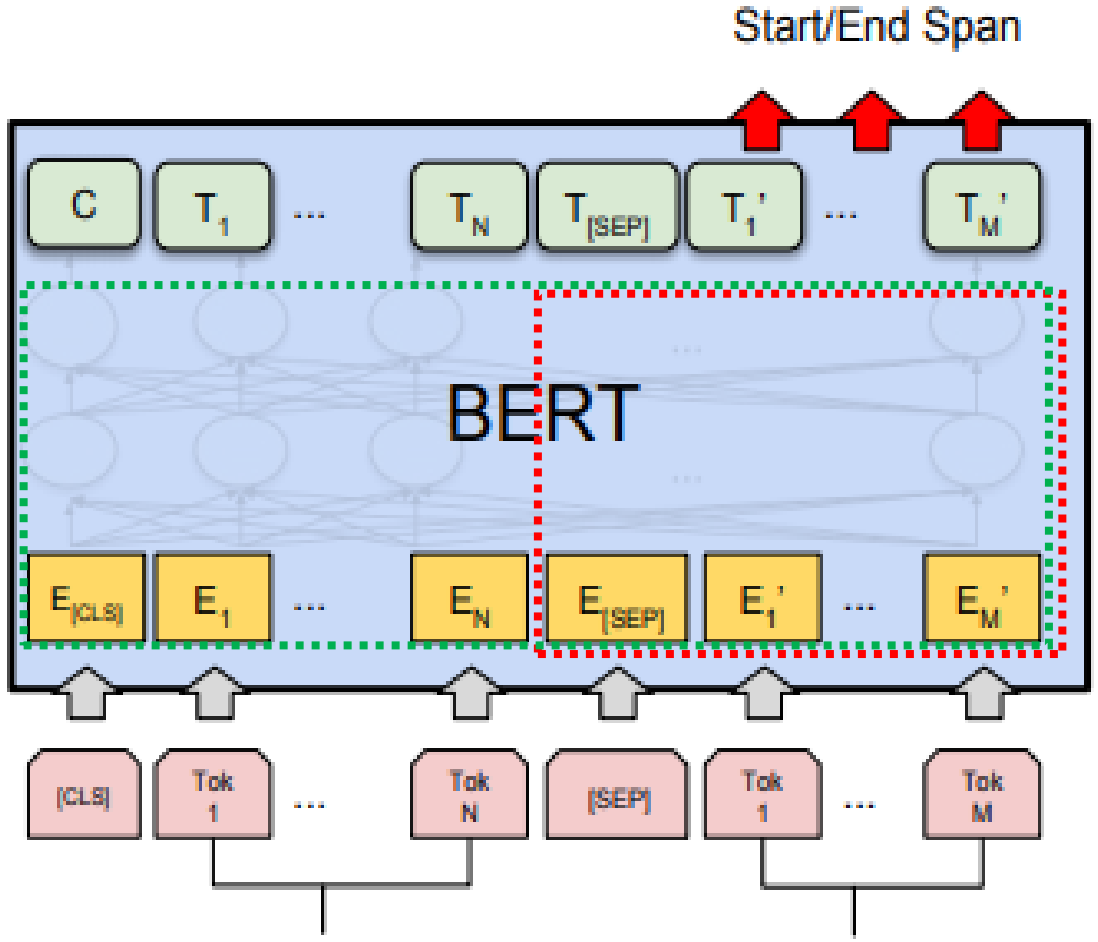
Transformer-based MRC model: BERT

Answer prediction

Integrated context vectors

Question contextual embedding (inter-attention)

Lexicon Embedding



Passage contextual Embedding (self-attention)

[Devlin+19]

Conversational MRC models

- QA as classification given (question, text)
  - Classify each word in passage as start/end/outside of answer span
- Encoding: represent each passage word using an integrated context vector that encodes info about
  - Lexicon/word embedding
  - Passage context
  - Question context
  - **Conversation context (previous question-answer pairs)**
- Prediction: predict each word (its integrated context vector) the start and end position of answer span.

A recent review on conversational MRC is [Gupta&Rawat 20]

Conversational MRC models

- Pre-pending conversation history to current question or passage
  - Convert conversational QA to single-turn QA
- BiDAF++ (BiDAF for C-QA)
  - Append a feature vector encoding dialog turn number to question embedding
  - Append a feature vector encoding N answer locations to passage embedding
- BERT (or RoBERTa)
  - Prepending dialog history to current question
  - Using BERT as
    - context embedding (self-attention)
    - Question/conversation context embedding (inter-attention)

[Choi+18; Zhu+19; Ju+19; Devlin+19]

FlowQA: explicitly encoding dialog history

- Integration Flow (IF) Layer
  - Given:
    - Current question Q_T, and previous questions $Q_t, t < T$
    - For each question Q_t, integrated context vector of each passage word w_t
  - Output:
    - Conversation-history-aware integrated context vector of each passage word
    - $w_T = \mathrm{LSTM}(w_1, \dots, w_t, \dots, w_T)$
    - So, the entire integrated context vectors for answering previous questions can be used to answer the current question.
- Extensions of IF
  - FlowDelta explicitly models the info gain thru conversation
  - GraphFLOW captures conversation flow using a graph neural network
  - *Implementing IF using Transformer with proper attention masks*

[Huang+19; Yeh&Chen 19; Chen+19]

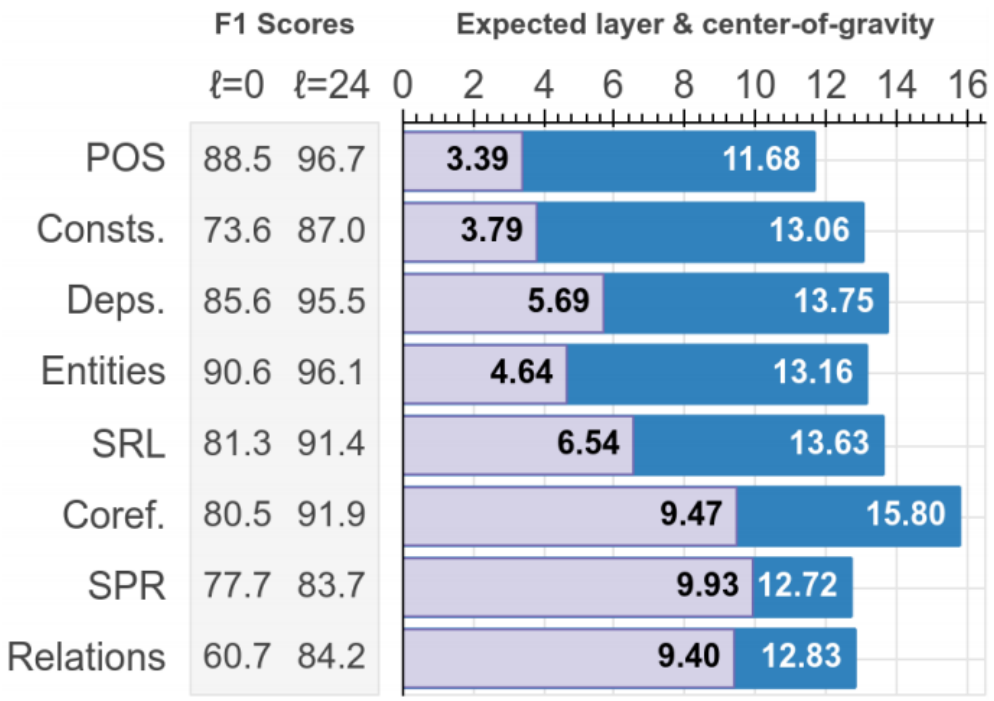
Remarks on BERT/RoBERTa

- BERT-based models achieve SOTA results on conversational QA/MRC leaderboards.
- What BERT learns
  - BERT rediscovers the classical NLP pipeline in an interpretable way
  - BERT exploits spurious statistical patterns in datasets instead of learning meaning in the generalizable way that humans do, so
  - Vulnerable to adversarial attack/tasks (adversarial input perturbation)
    - Text-QA: Adversarial SQuAD [Jia&Liang 17]
    - Classification: TextFooler [Jin+20]
    - Natural language inference: Adversarial NLI [Nie+19]
- Towards a robust QA model

[Tenney+19; Nie+ 19; Jin+20; Liu+20]

BERT rediscovers the classical NLP pipeline in an interpretable way

- Quantify where linguistic info is captured within the network
- Lower layers encode more local syntax
- higher layers encode more global complex semantics
- A higher center-of-gravity value means that the information needed for that task is captured by higher layers



[Tenney+19]

Adversarial examples

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

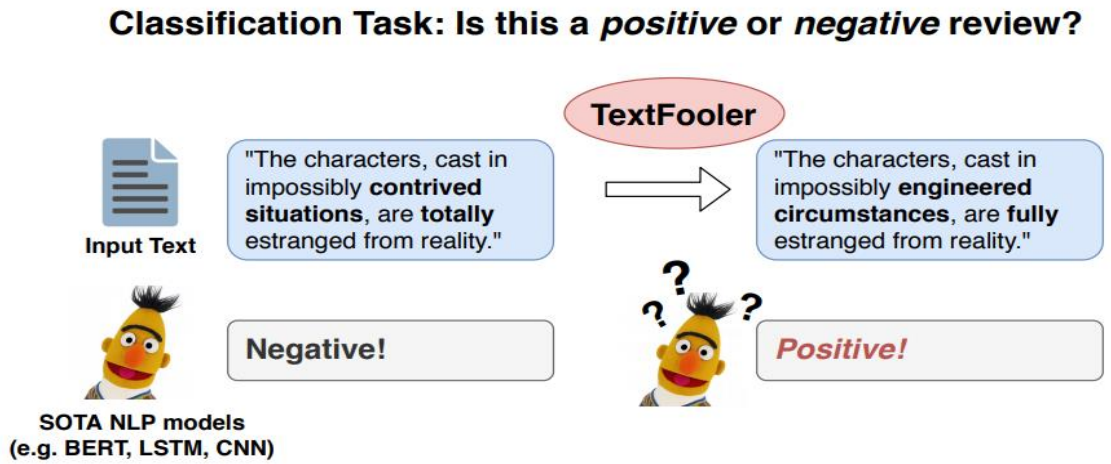


Figure 1: Our model TextFooler slightly change the input text but completely altered the prediction result.

	Text-QA	Sentiment Classification		
	SQuAD	MR	IMDB	Yelp
Original	88.5	86.0	90.9	97.0
Adversarial	**54.0**	**11.5**	**13.6**	**6.6**

$BERT_{BASE}$ results

[Jia&Liang 17; Jin+20; Liu+20]

Build Robust AI models via adversarial training

- Standard Training objective

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[l(f(x;\theta),y)]$$

- Adversarial Training in computer vision: apply small perturbation to input images that maximize the adversarial loss

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} l(f(x+\delta;\theta),y)]$$

- Adversarial Training for neural language modeling (ALUM):
  - Perturb word embeddings instead of words
  - adopt virtual adversarial training to regularize standard objective

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[l(f(x;\theta),y) + \alpha \max_{\delta} l(f(x+\delta;\theta),f(x;\theta))]$$

[Goodfellow+16; Madry+17; Miyato+18; Liu+20]

Generalization and robustness

- Generalization: perform well on unseen data
  - pre-training
- Robustness: withstand adversarial attacks
  - adversarial training
- Can we achieve both?
  - Past work finds that adversarial training can enhance robustness, but hurts generalization [Raghunathan+19; Min+20]
  - Apply adversarial pre-training (ALUM) improves both [Liu+20]

[Raghunathan+19; Min+20; Liu+20]

Outline

- Part 1: Introduction
- Part 2: Conversational QA methods
- **Part 3: Conversational search methods**
- Part 4: Case study of commercial systems

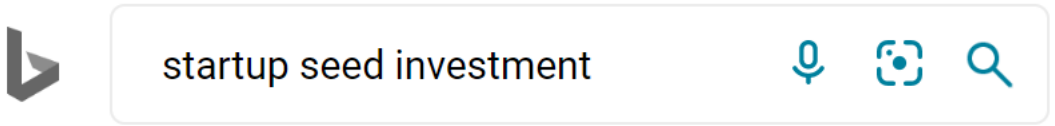
Conversational Search: Outline

- **What** is conversational search?
  - A view from TREC Conversational Assistance Track (TREC CAsT) [1]
- **Unique Challenges** in conversational search.
  - Conversational query understanding [2]
- **How** to make search **more conversational**?
  - From passive retrieval to active conversation with conversation recommendation [3]

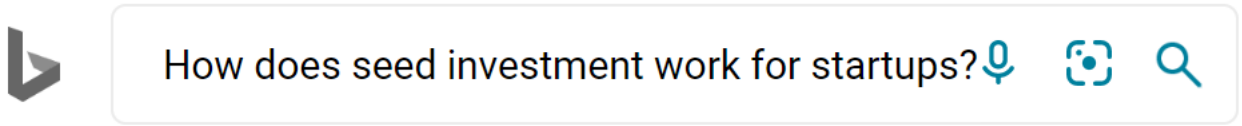
[1] Cast 2019: The conversational assistance track overview
[2] Few-Shot Generative Conversational Query Rewriting
[3] Leading Conversational Search by Suggesting Useful Questions

Why Conversational Search

Ad hoc Search



Conversational Search



Keyword-ese Queries → Natural Queries

Necessity:

- Speech/Mobile Interfaces

Opportunities:

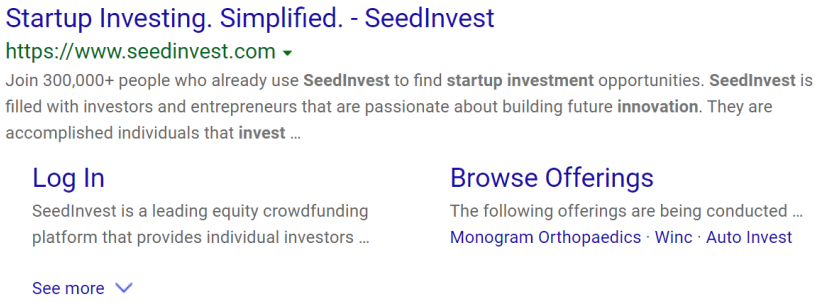
- More natural and explicit expression of information needs

Challenge:

- Query understanding & sparse retrieval

Why Conversational Search

Ad hoc Search



Ten Blue-Links

Conversational Search

Natural Responses

Necessity:

- Speech/Mobile Interfaces

Opportunities:

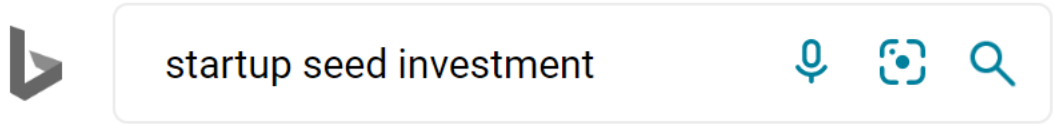
- Direct & Easier access to information

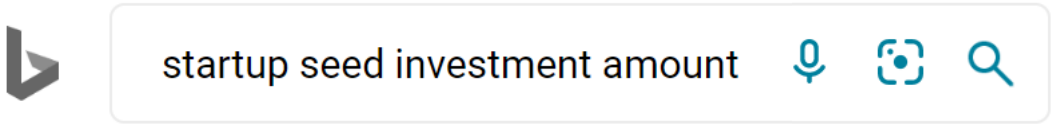
Challenge:

- Document understanding; combine and synthesize information

Why Conversational Search

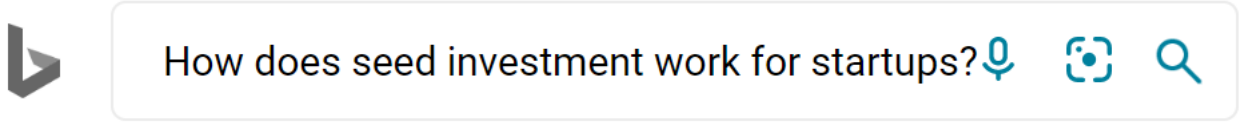
Ad hoc Search

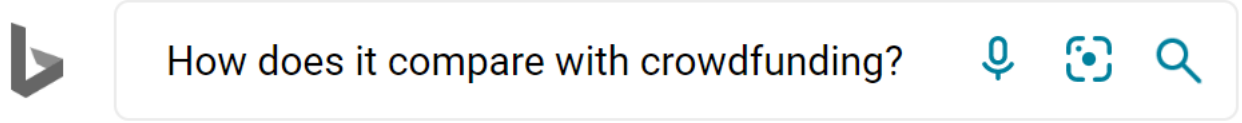




Single-Shot Query

Conversational Search





Multi-Turn Dialog

Necessity:

- N.A.

Opportunities:

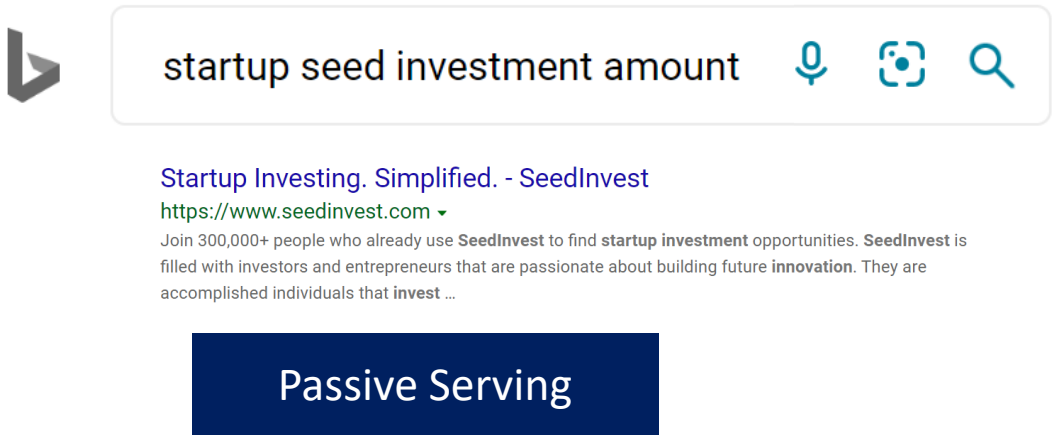
- Serving complex information needs and tasks

Challenge:

- Contextual Understanding & Memorization

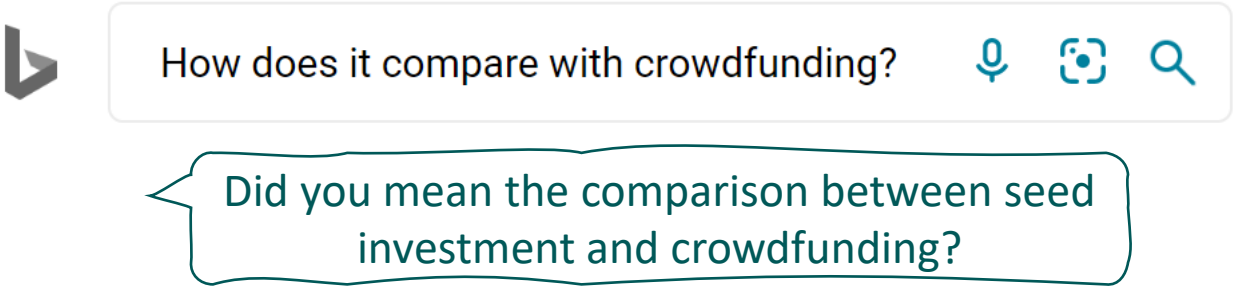
Why Conversational Search

Ad hoc Search



Passive Serving

Conversational Search



Active Engaging

Necessity:

- N.A.

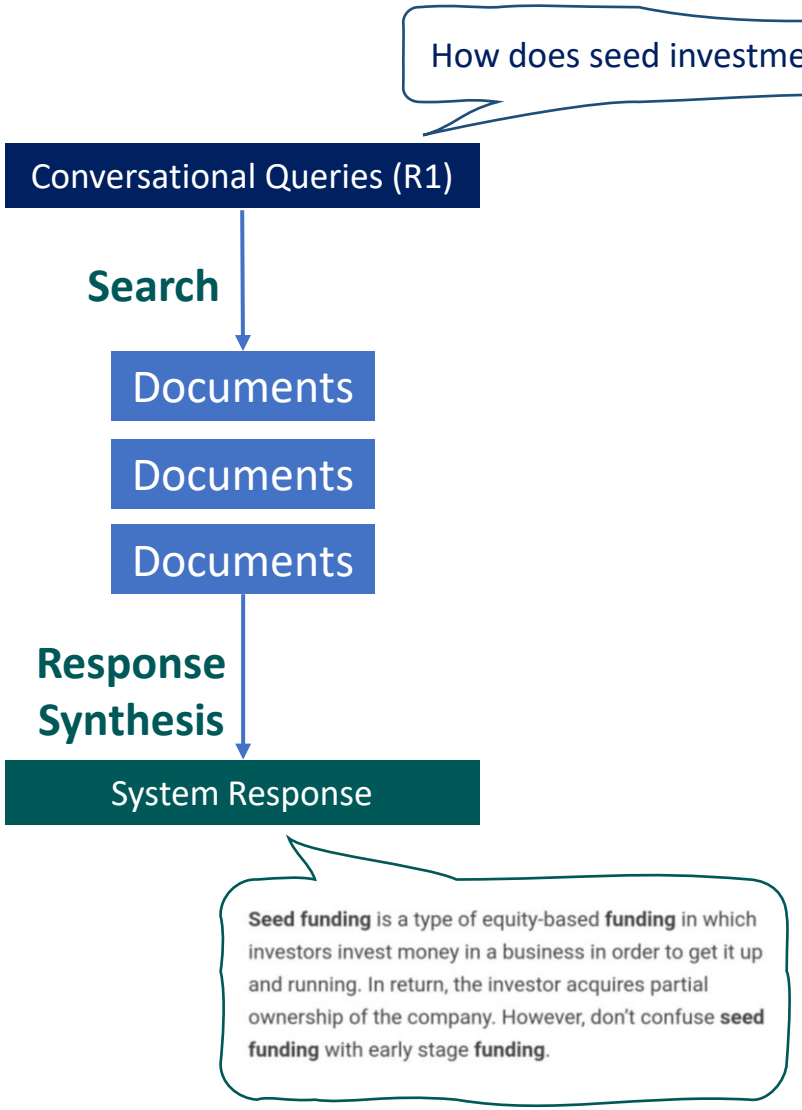
Opportunities:

- Collaborative information seeking & better task assistance

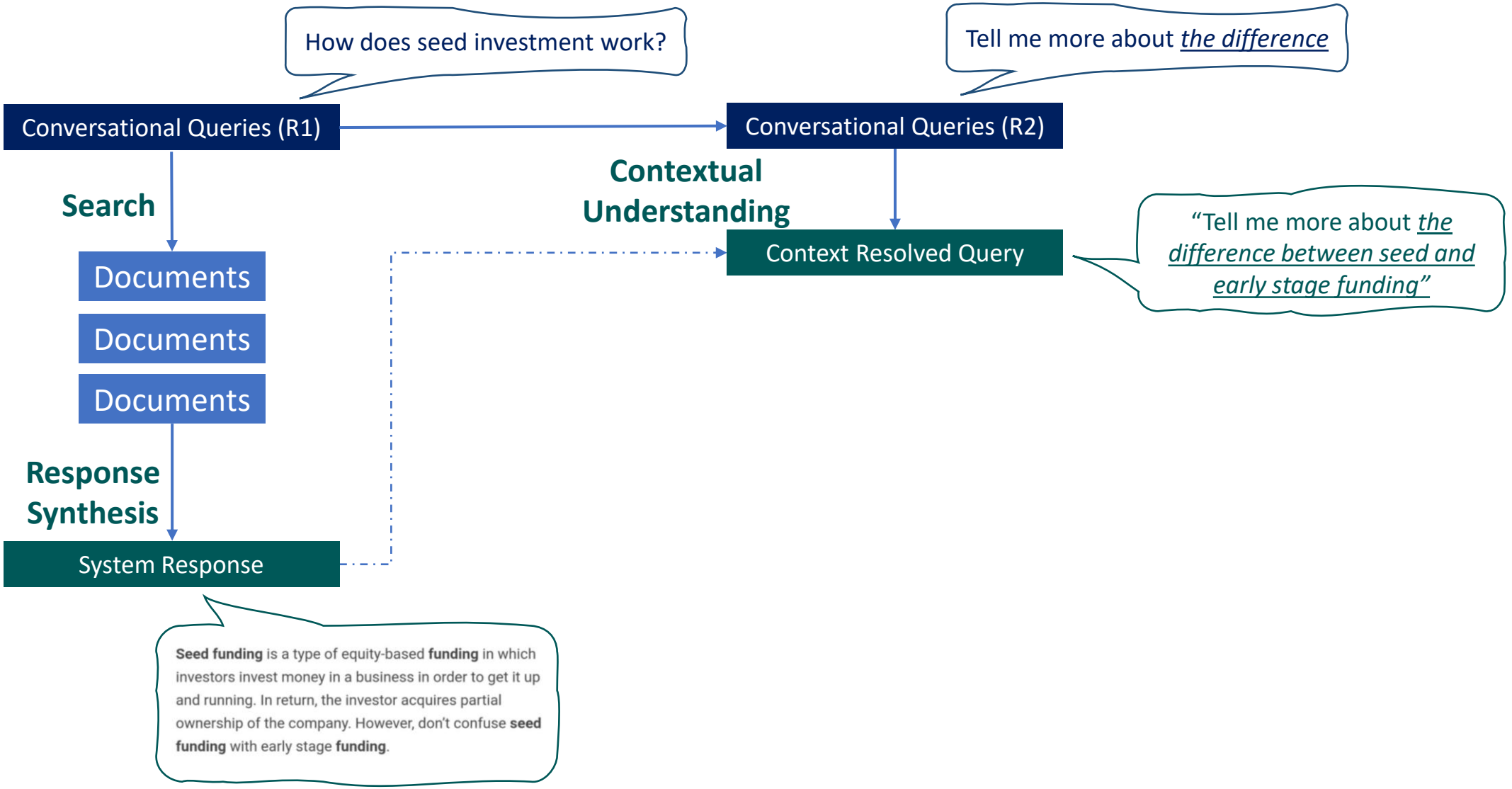
Challenge:

- Dialog management, less lenient user experience

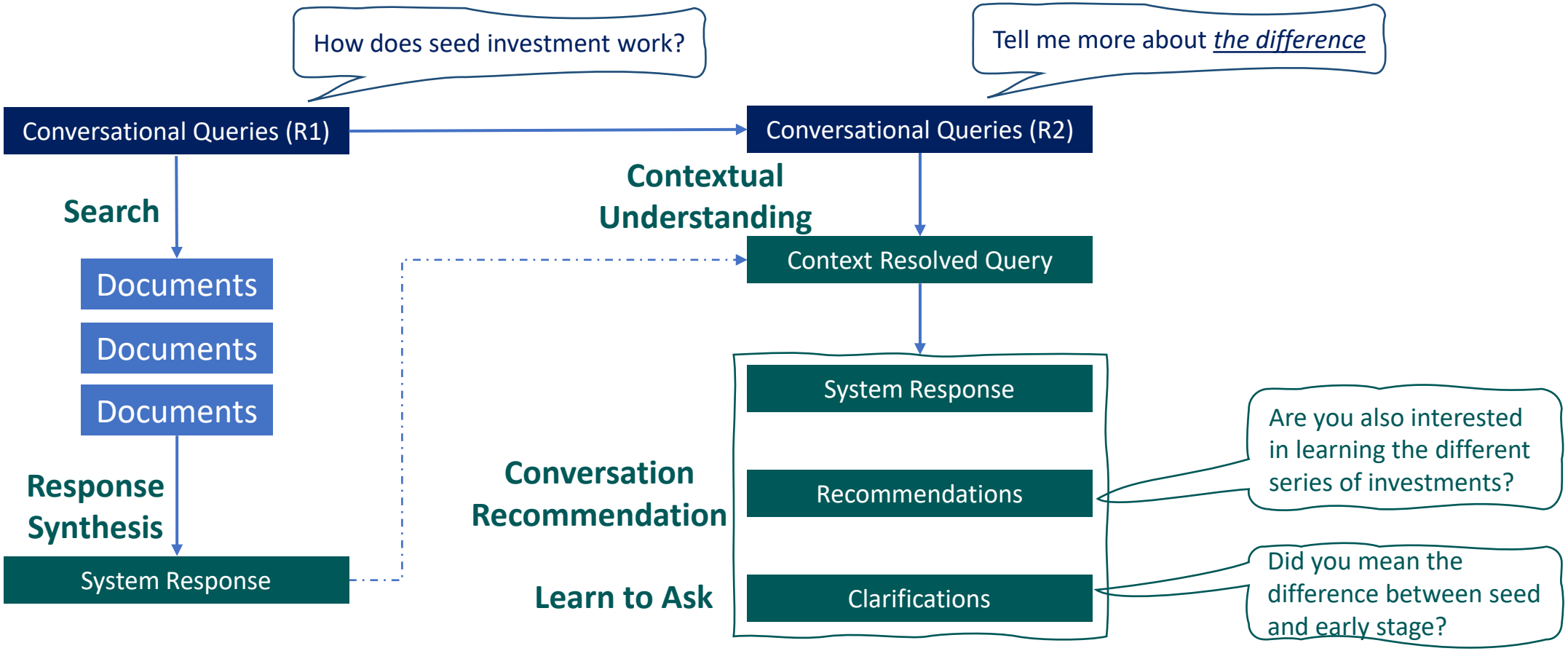
A View of Current Conversational Search



A View of Current Conversational Search

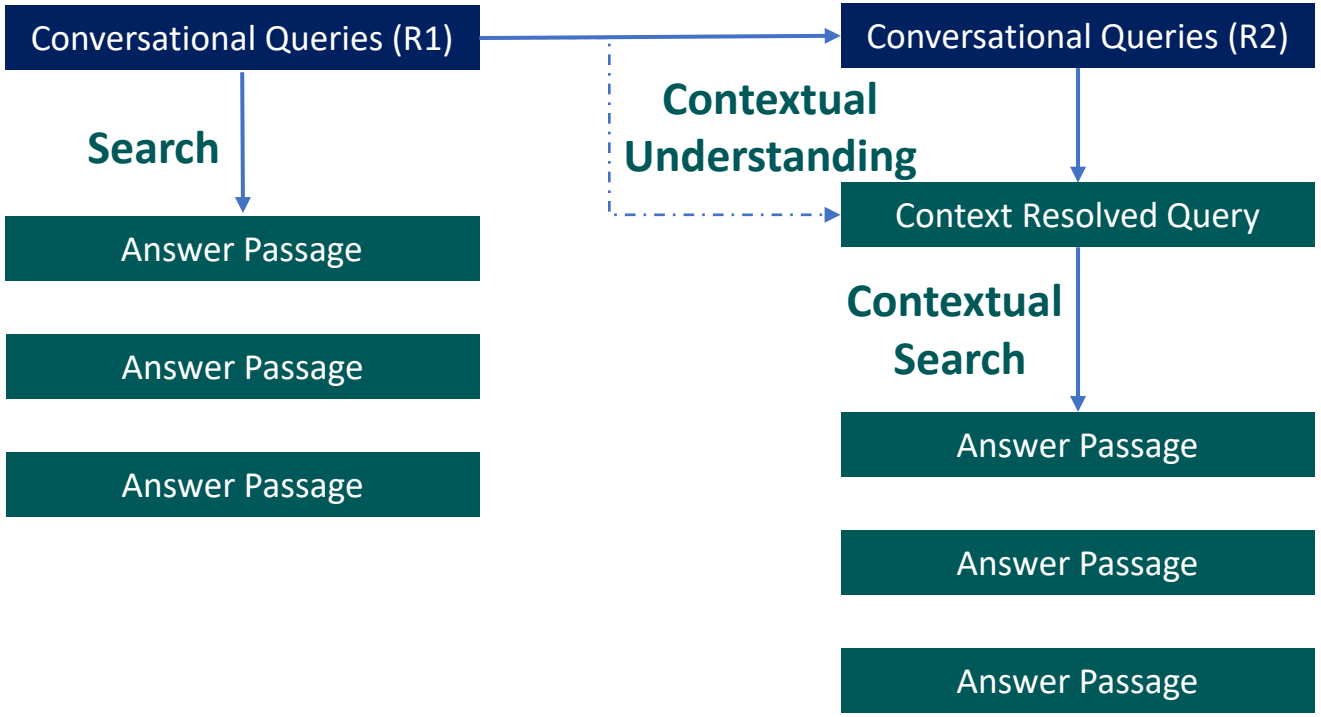


A View of Current Conversational Search



A Simpler View from TREC CAsT 2019

- "Conversational Passage Retrieval/QA"



Input:

- Manually written conversational queries
- ~20 topics, ~8 turns per topic
- Contextually dependent on previous queries

Corpus:

- MS MARCO + CAR Answer Passages

Task:

- Passage Retrieval for conversational queries

TREC CAsT 2019

- An example conversational search session

Title: head and neck cancer

Description: A person is trying to compare and contrast types of cancer in the throat, esophagus, and lungs.

1 What is throat cancer?

2 Is it treatable?

3 Tell me about lung cancer.

4 What are its symptoms?

5 Can it spread to the throat?

6 What causes throat cancer?

7 What is the first sign of it?

8 Is it the same as esophageal cancer?

9 What's the difference in their symptoms?

Input:

- Manually written conversational queries
- ~20 topics, ~8 turns per topic
- Contextually dependent on previous queries

Corpus:

- MS MARCO + CAR Answer Passages

Task:

- Passage Retrieval for conversational queries

TREC CAsT 2019

- Challenge: contextual dependency on previous conversation queries

---

Title: head and neck cancer

Description: A person is trying to compare and contrast types of cancer in the throat, esophagus, and lungs.

1 What is throat cancer?

2 **Is** it treatable?

3 Tell me about lung cancer.

4 What are **its** symptoms?

5 Can **it** spread to the throat?

6 What causes throat cancer?

7 What is the first sign of **it**?

8 Is **it** the same as esophageal cancer?

9 What's the difference in **their** symptoms?

Input:

- Manually written conversational queries
- ~20 topics, ~8 turns per topic
- Contextually dependent on previous queries

Corpus:

- MS MARCO + CAR Answer Passages

Task:

- Passage Retrieval for conversational queries

TREC CAsT 2019

- Learn to resolve the contextual dependency

Title: head and neck cancer

Description: A person is trying to compare and contrast types of cancer in the throat, esophagus, and lungs.

1 What is throat cancer?

2 Is **it** treatable?

3 Tell me about lung cancer.

4 What are **its** symptoms?

5 Can **it** spread to the throat?

6 What causes throat cancer?

7 What is the first sign of **it**?

8 Is **it** the same as esophageal cancer?

9 What's the difference in **their** symptoms?

Manual Queries provided by CAsT Y1

1 What is throat cancer?

2 Is throat cancer treatable?

3 Tell me about lung cancer.

4 What are lung cancer's symptoms?

5 Can lung cancer spread to the throat

6 What causes throat cancer?

7 What is the first sign of throat cancer?

8 Is throat cancer the same as esophageal cancer?

9 What's the difference in throat cancer and esophageal cancer's symptoms?

TREC CAsT 2019: Query Understanding Challenge

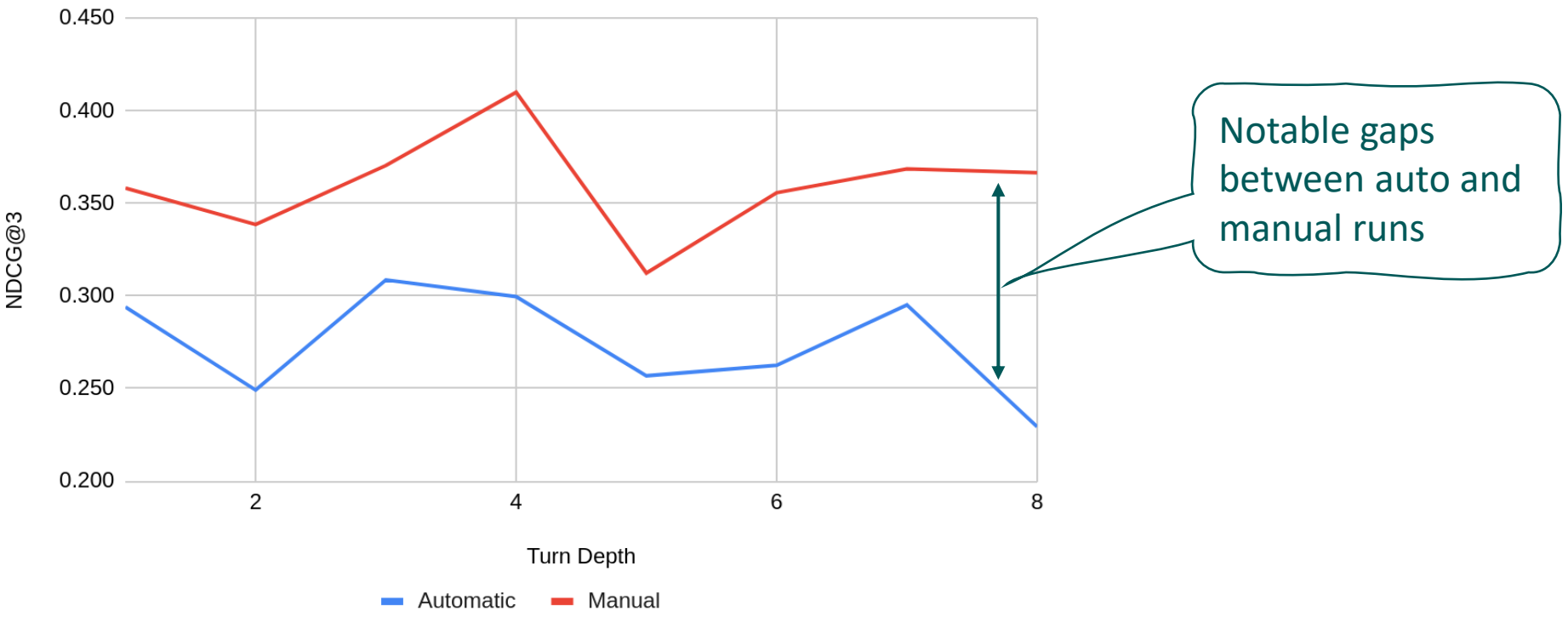
- Statistics in Y1 Testing Queries

Type (#. Turns)	Utterance	Mention
Pronominal (128)	How do **they** celebrate Three Kings Day?	they -> Spanish people
Zero (111)	What cakes are traditional?	Null -> Spanish, Three Kings Day
Groups (4)	Which **team** came first?	which team -> Avengers, Justice League
Abbreviations (15)	What are the main types of **VMs**?	VMs -> Virtual Machines

Cast 2019: The Conversational Assistance Track Overview

TREC CAsT 2019: Result Statics

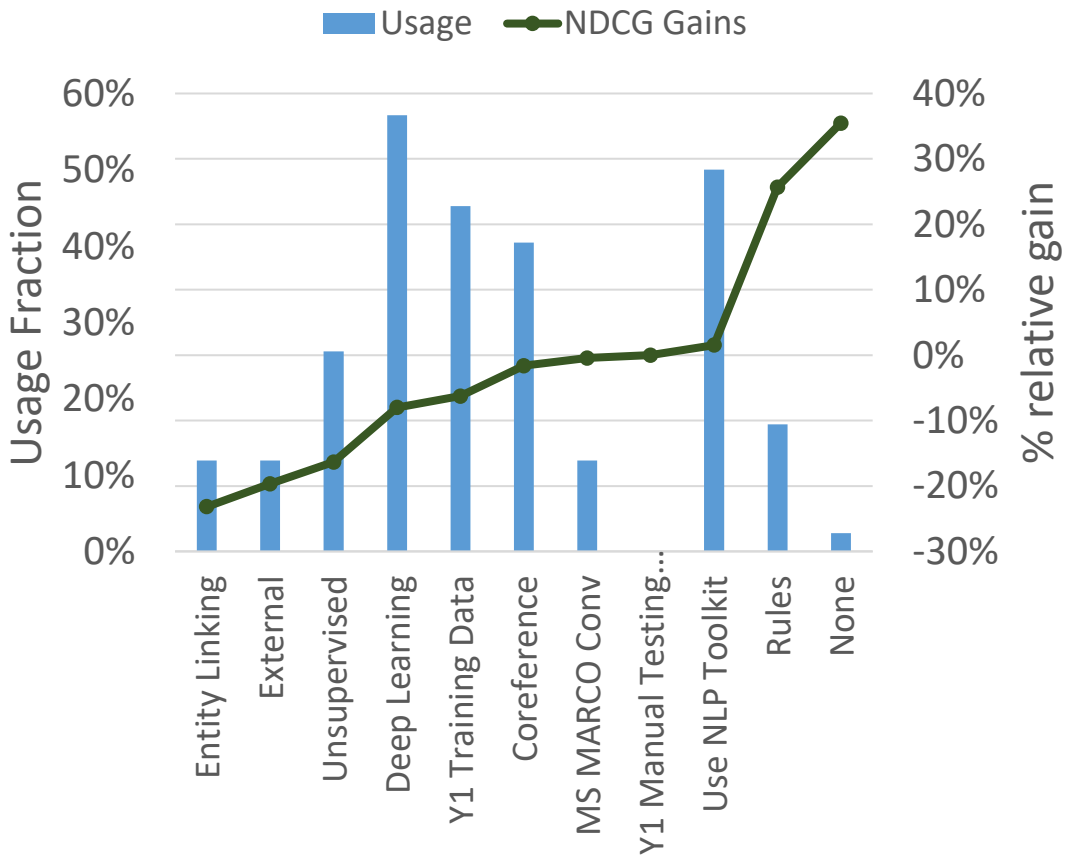
- Challenge from contextual query understanding



Cast 2019: The Conversational Assistance Track Overview

TREC CAsT 2019: Techniques

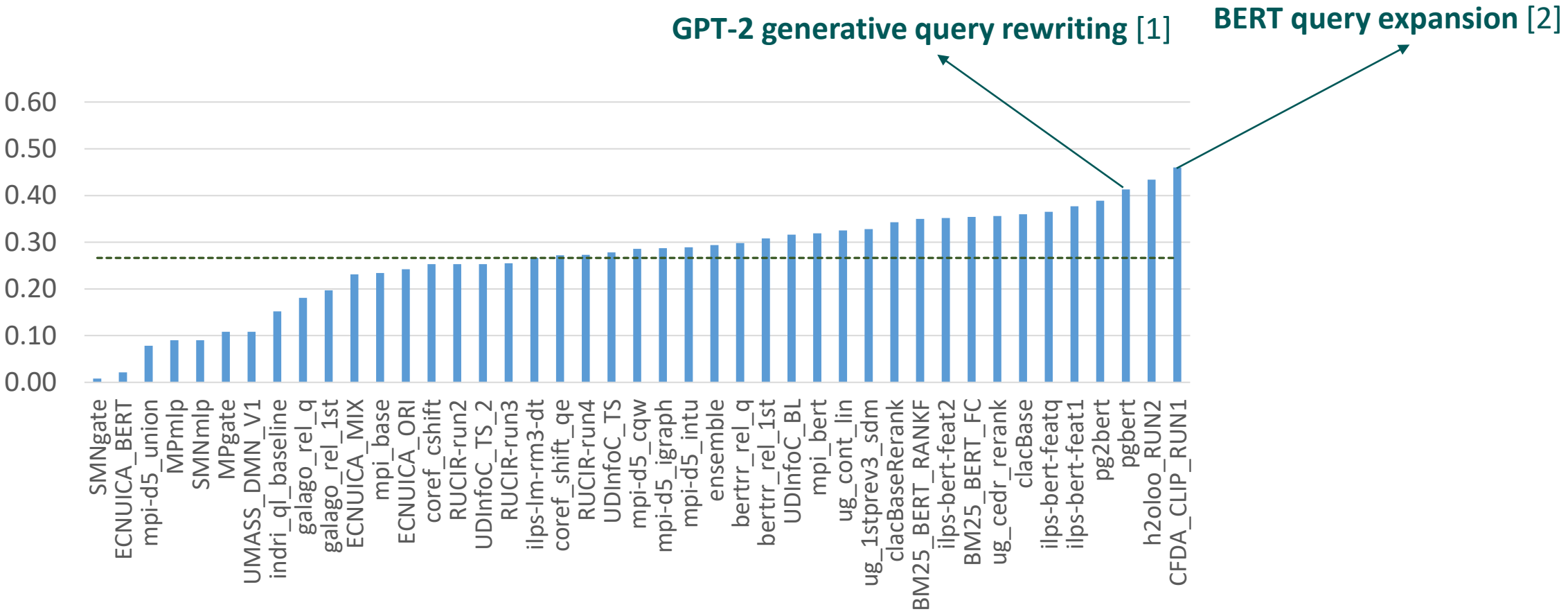
- Techniques used in Query Understanding



Cast 2019: The Conversational Assistance Track Overview

TREC CAsT 2019: Notable Solutions

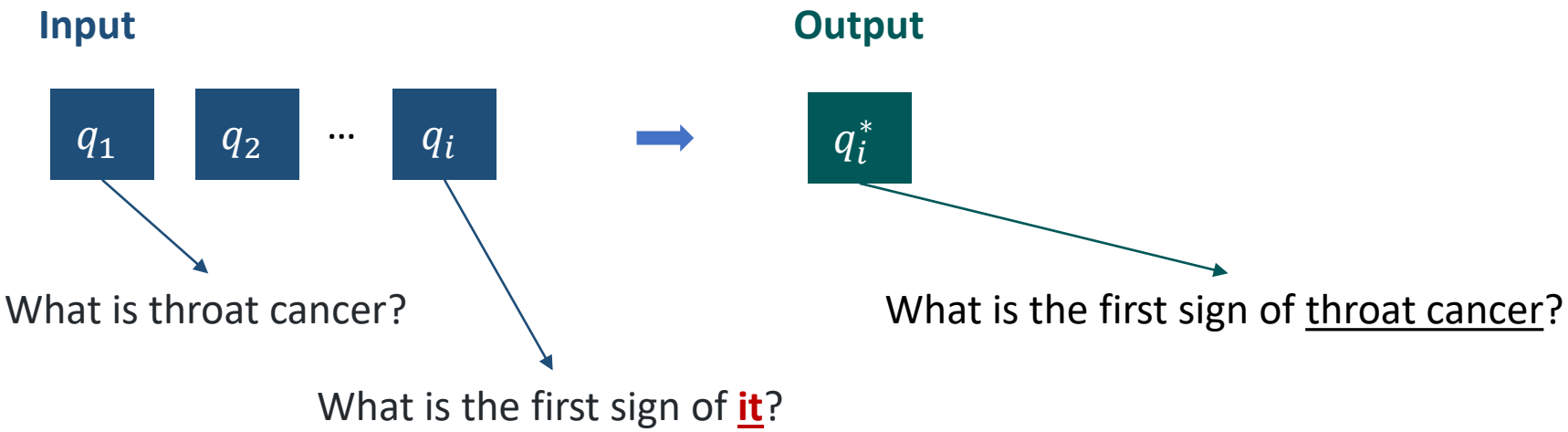
- Automatic run results



[1] Vakulenko et al. 2020. Question Rewriting for Conversational Question Answering
[2] Lin et al. 2020. Query Reformulation using Query History for Passage Retrieval in Conversational Search

Conversational Query Understanding Via Rewriting

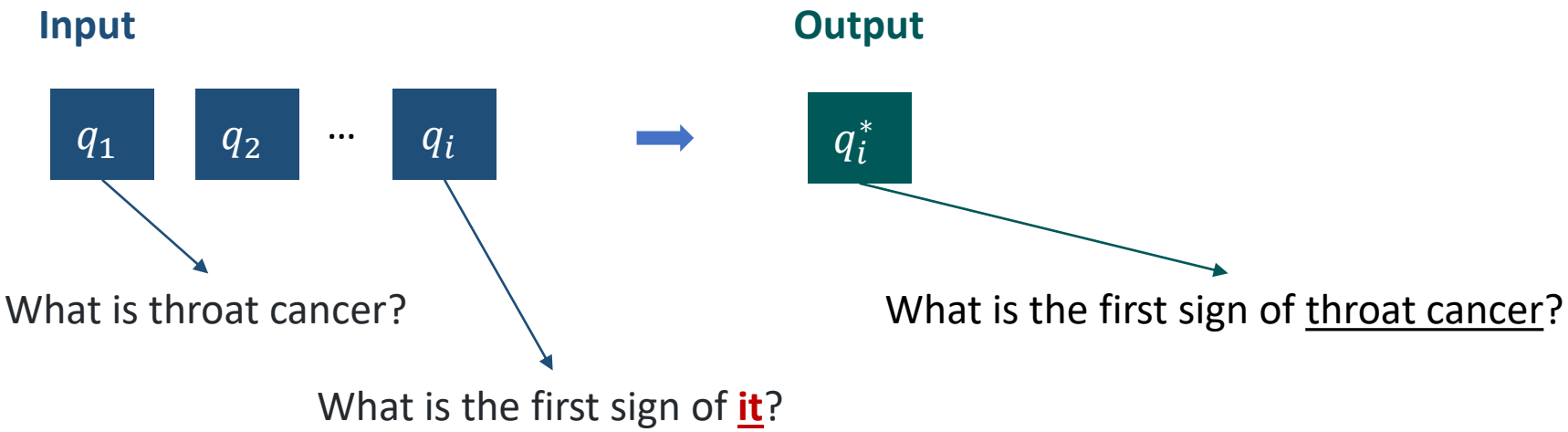
- Learn to rewrite a full-grown context-resolved query



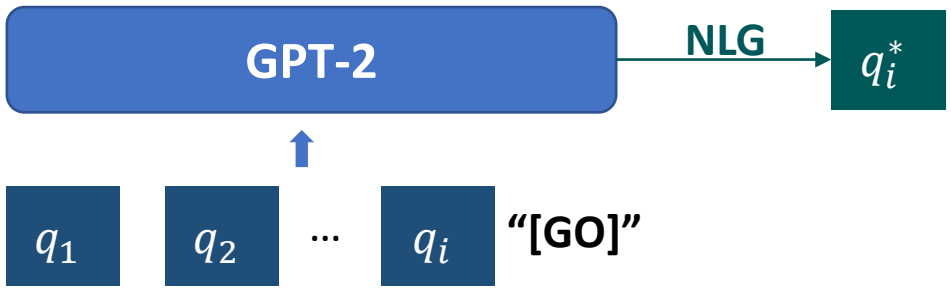
Vakulenko et al. 2020. Question Rewriting for Conversational Question Answering

Conversational Query Understanding Via Rewriting

- Learn to rewrite a full-grown context-resolve query



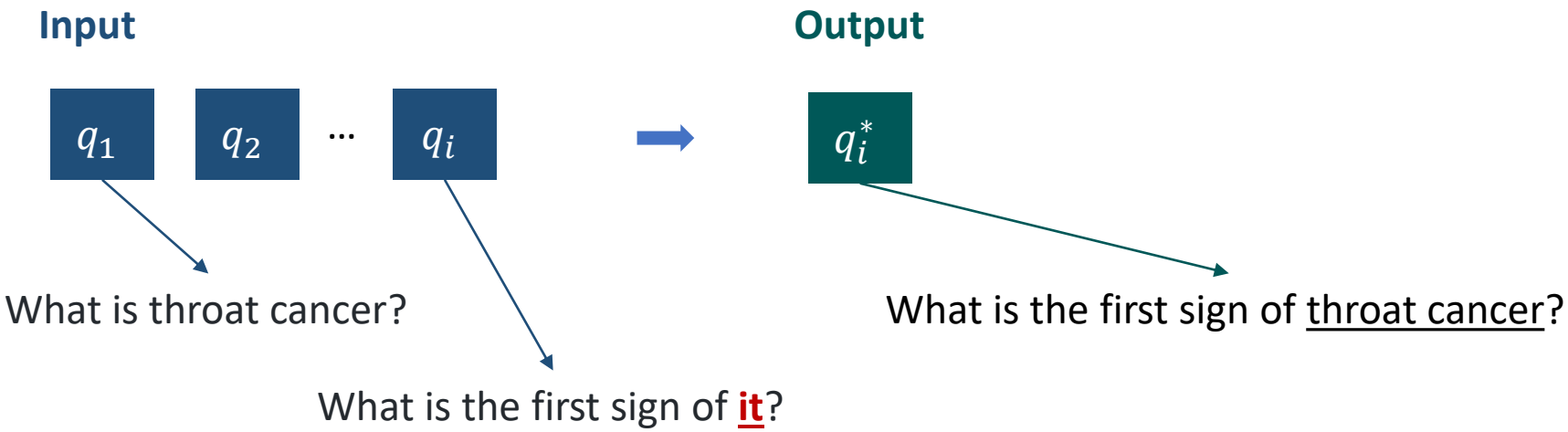
- Leverage pretrained NLG model (GPT-2) [1]



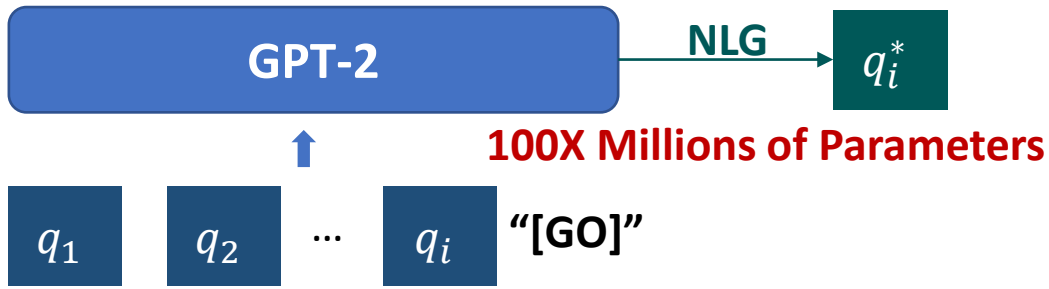
Vakulenko et al. 2020. Question Rewriting for Conversational Question Answering

Conversational Query Understanding Via Rewriting

- Learn to rewrite a full-grown context-resolve query



- Concern: Limited training data



?

CAsT Y1 Data:

- Manually written conversational queries
- 50 topics, 10 turns per topic
  - 20 topics with TREC relevance labels

500 Manual Rewrite Labels

Vakulenko et al. 2020. Question Rewriting for Conversational Question Answering

Few-Shot Conversational Query Rewriting

- Train conversational query rewriter with the help of ad hoc search data

Ad hoc Search

- Existing billions of search sessions
- Lots of high-quality public benchmarks

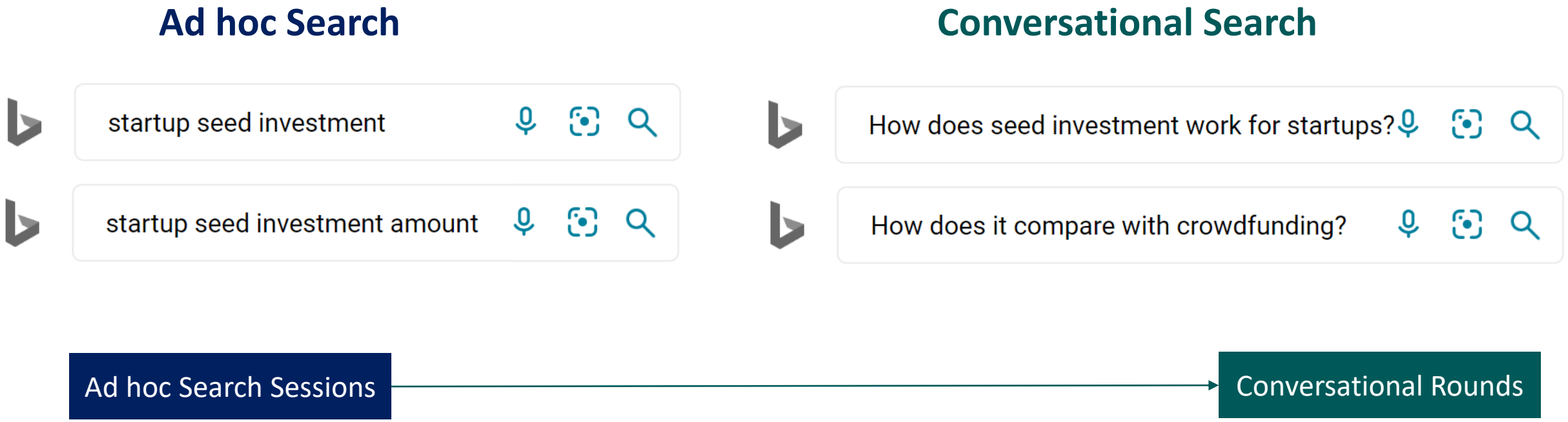
Conversational Search

- Production scenarios still being explored
- Relative new topic, fewer available data

Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

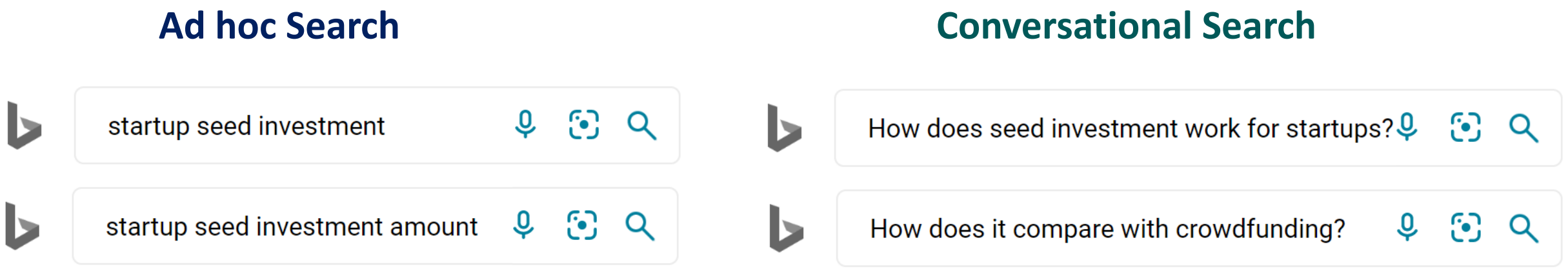
Few-Shot Conversational Query Rewriting

- Leveraging ad hoc search sessions for conversational query understanding



Few-Shot Conversational Query Rewriting

- Leveraging ad hoc search sessions for conversational query understanding



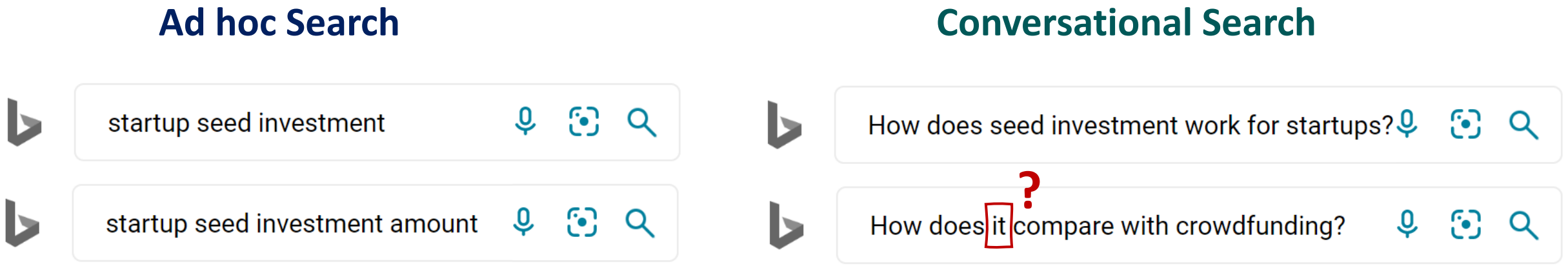
Ad hoc Search Sessions → Conversational Rounds

Challenges?

- Available only in commercial search engines
  - Approximate sessions available in MS MARCO
- Keyword-ese
  - Filter by question words

Few-Shot Conversational Query Rewriting

- Leveraging ad hoc search sessions for conversational query understanding



Ad hoc Search Sessions → Conversational Rounds

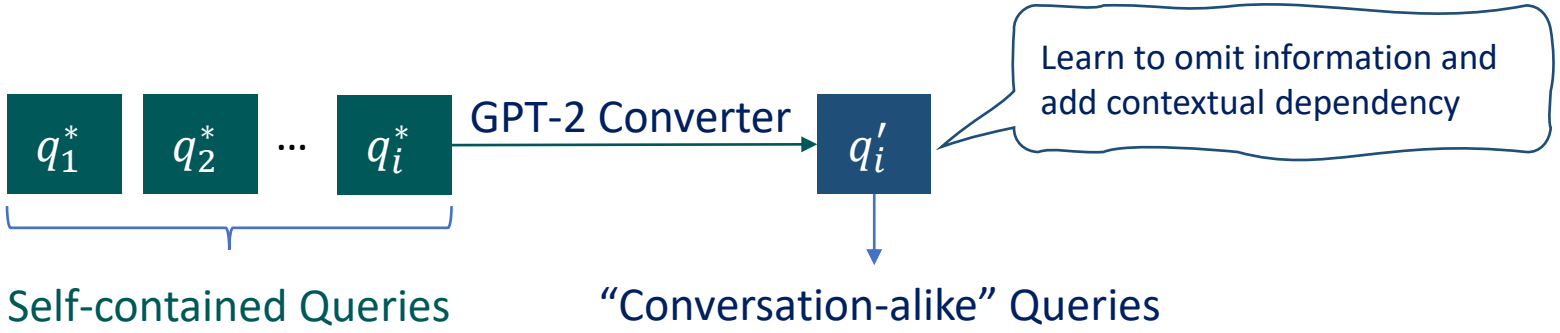
Challenges?

- Available only in commercial search engines
  - Approximate sessions available in MS MARCO
- Keyword-ese
  - Filter by question words
- No explicit context dependency?

Few-Shot Conversational Query Rewriting: Self-Training

- Learn to convert ad hoc sessions to conversational query rounds

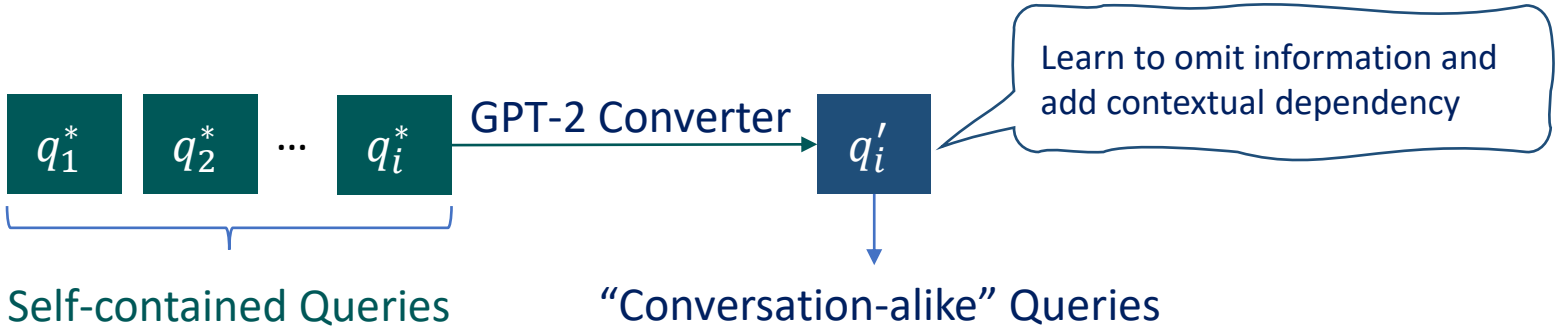
"Contextualizer": make ad hoc sessions more conversation-alike



Few-Shot Conversational Query Rewriting: Self-Training

- Learn to convert ad hoc sessions to conversational query rounds

"Contextualizer": make ad hoc sessions more conversation-alike



Training:

- X (Self-contained q): Manual rewrites of CAsT Y1 conversational sessions
- Y (Conversation-alike q): Raw queries in CAsT Y1 sessions

Inference:

- X (Self-contained q): Ad hoc questions from MS MARCO sessions
- Y (Conversation-alike q): Auto-converted conversational sessions

Model:

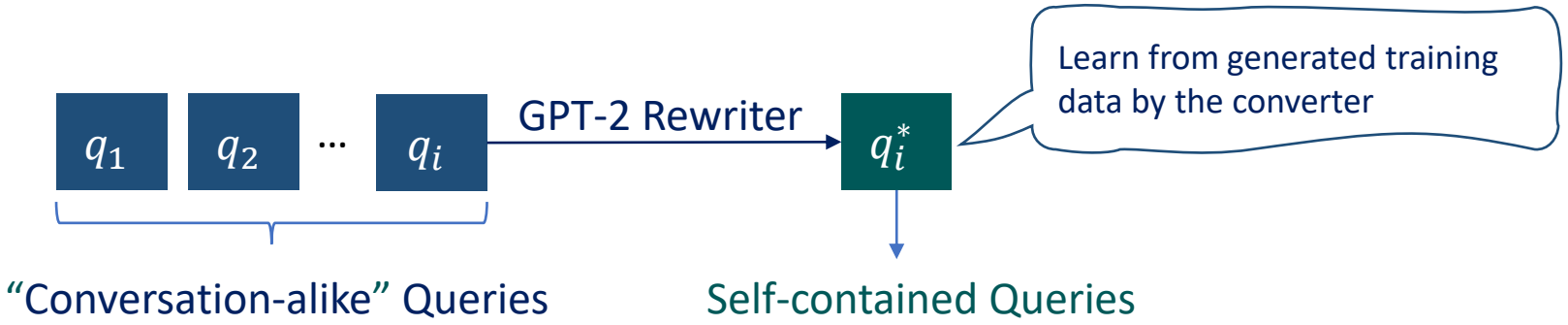
- Any pretrained NLG model: GPT-2 Small in this Case

Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

Few-Shot Conversational Query Rewriting: Self-Training

- Leverage the auto-converted conversation-ad hoc session pairs

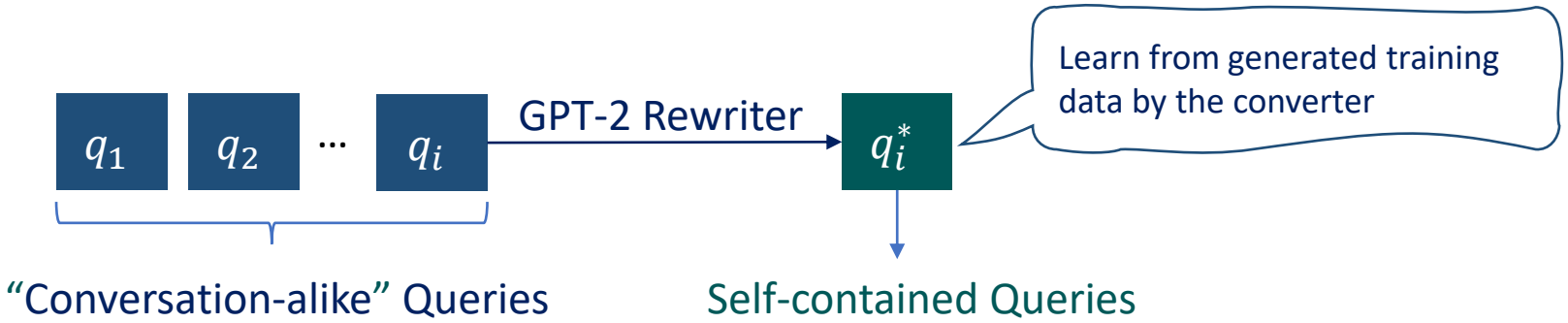
"Rewriter": recover the full self-contained queries from conversation rounds



Few-Shot Conversational Query Rewriting: Self-Training

- Leverage the auto-converted conversation-ad hoc session pairs

"Rewriter": recover the full self-contained queries from conversation rounds



Training:

- X (Conversation-alike q): Auto-converted from the Contextualizer
- Y (Self-contained q): Raw queries from ad hoc MARCO sessions

Inference:

- X (Conversation-alike q): CAsT Y1 raw conversational queries
- Y (Self-contained q): auto-rewritten queries that are more self-contained

Model:

- Any pretrained NLG model: another GPT-2 Small in this Case

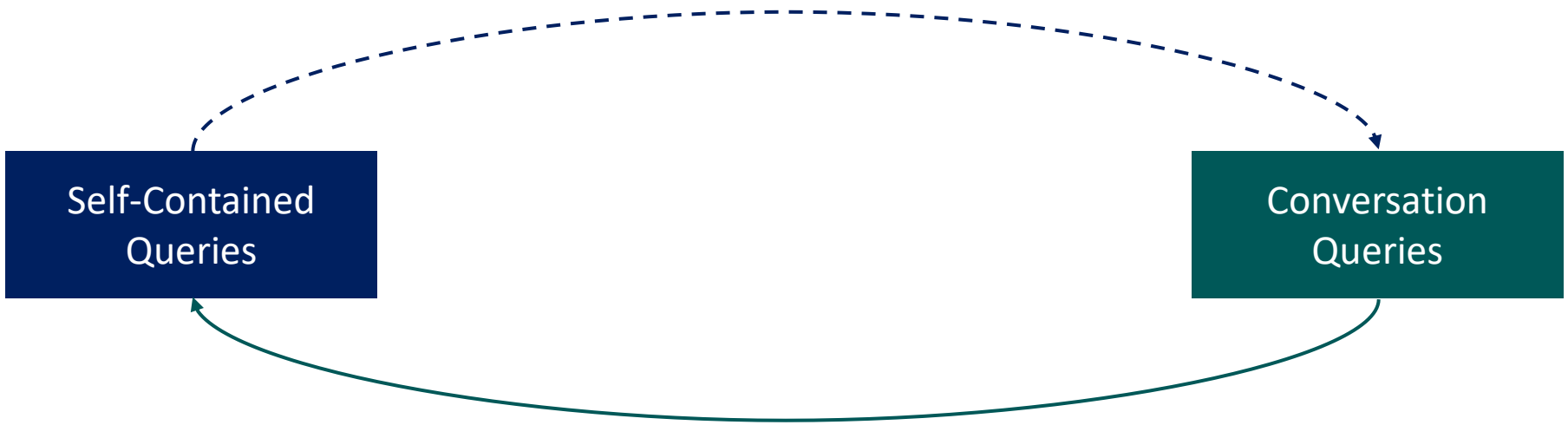
Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

Few-Shot Conversational Query Rewriting: Self-Training

- The full "self-learning" loop

Learn to omit information is easier than recover

GPT-2 Converter: Convert ad hoc sessions to conversation-alike sessions
- learn from a few conversational queries with manual rewrites

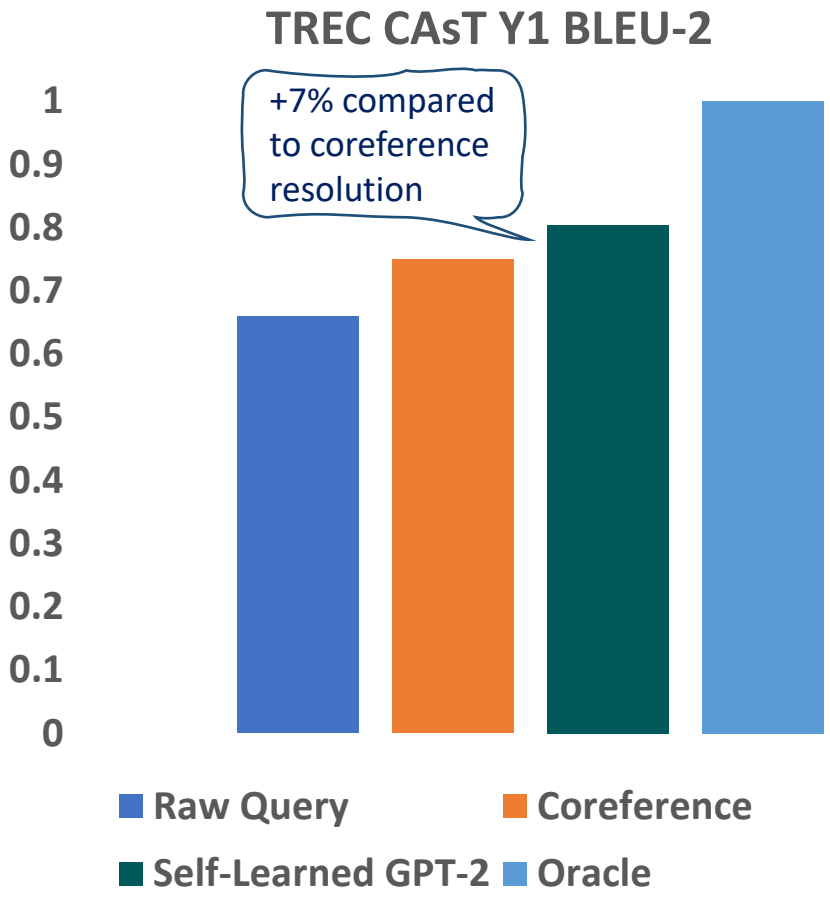


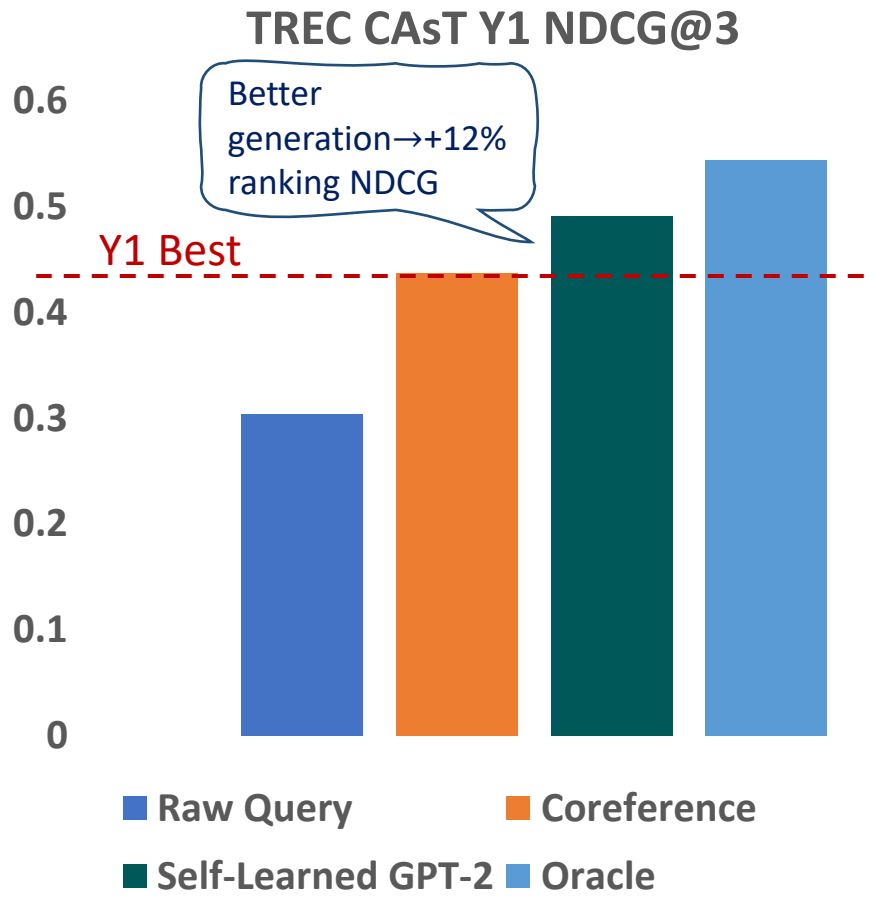
GPT-2 Rewriter: Rewrite conversational queries to self-contained ad hoc queries
- learn from the large amount of auto-converted "ad hoc" ↔ "conversation alike" sessions

Much more training signals from the Contextualizer

Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

Few-Shot Conversational Query Rewriting: Results

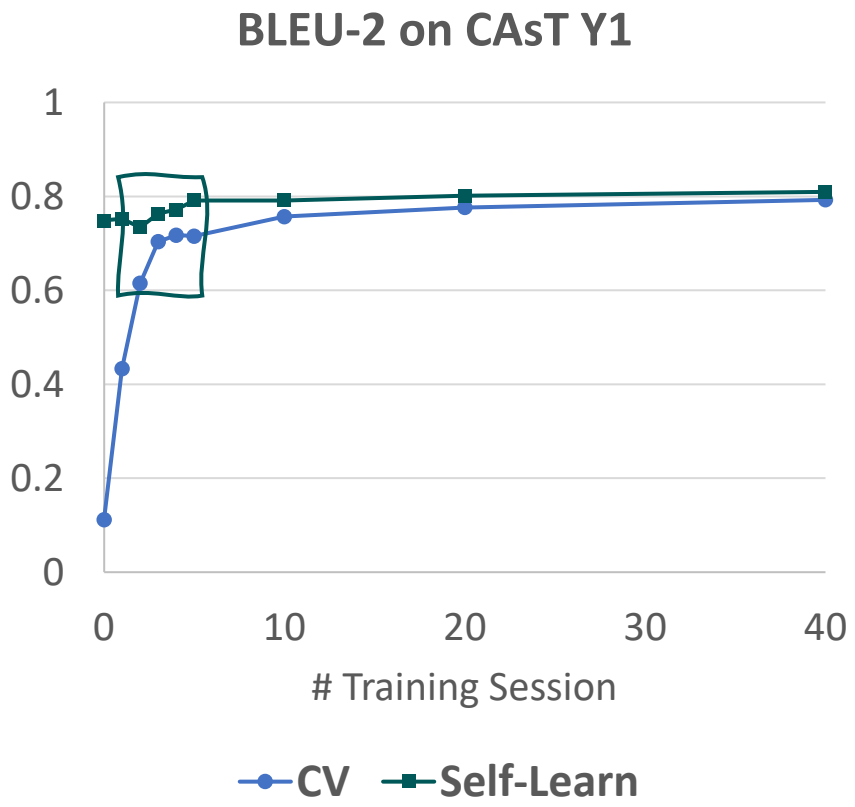


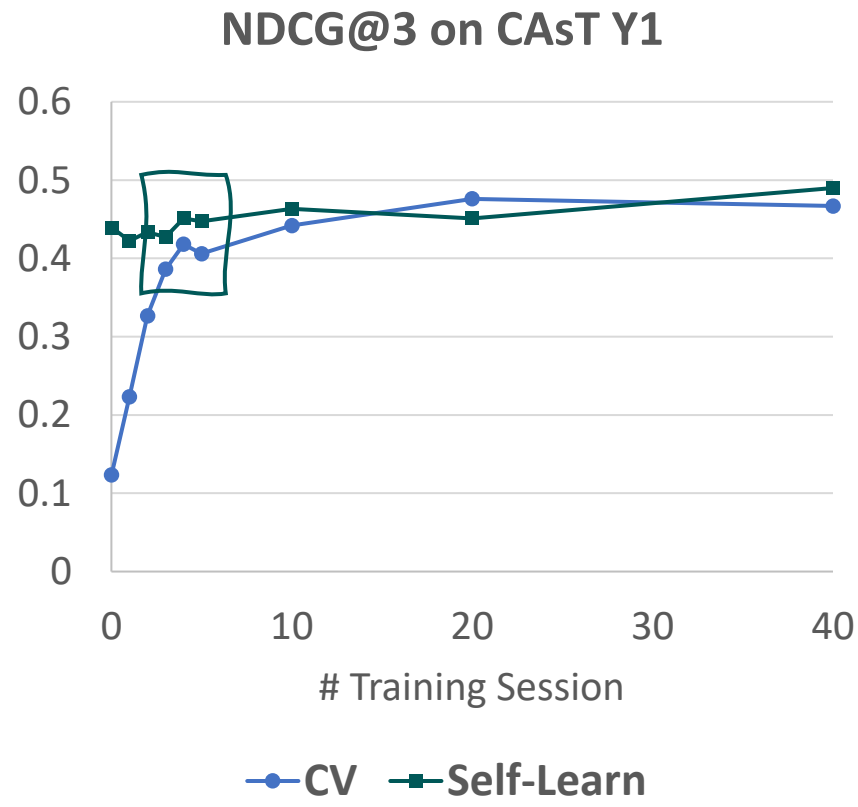


Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

How Few-shot Can Pretrained NLG Models Be?

- Five Sessions are all they need?

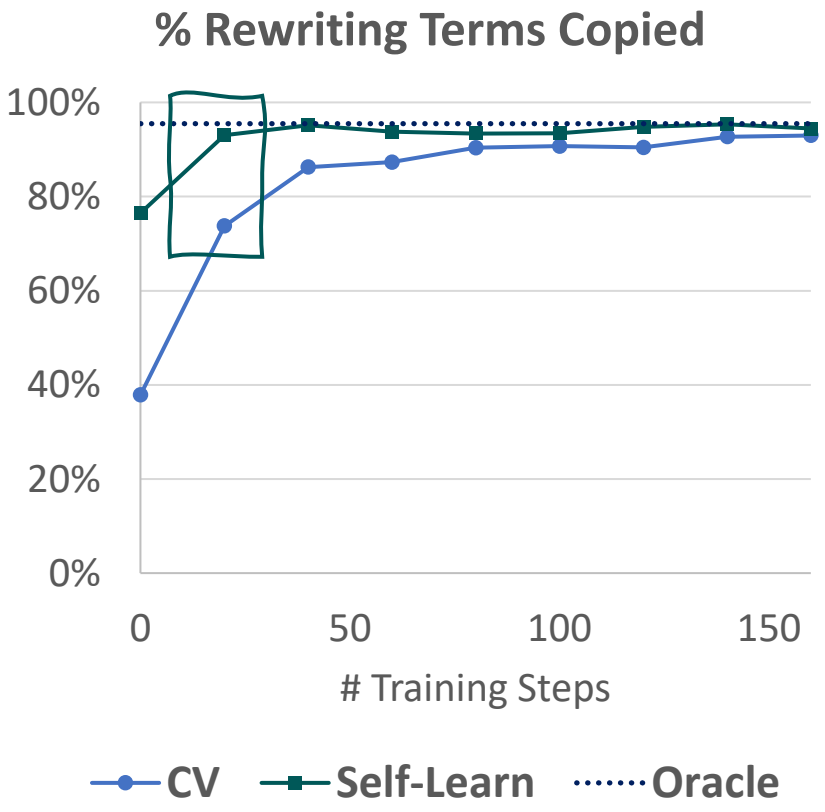


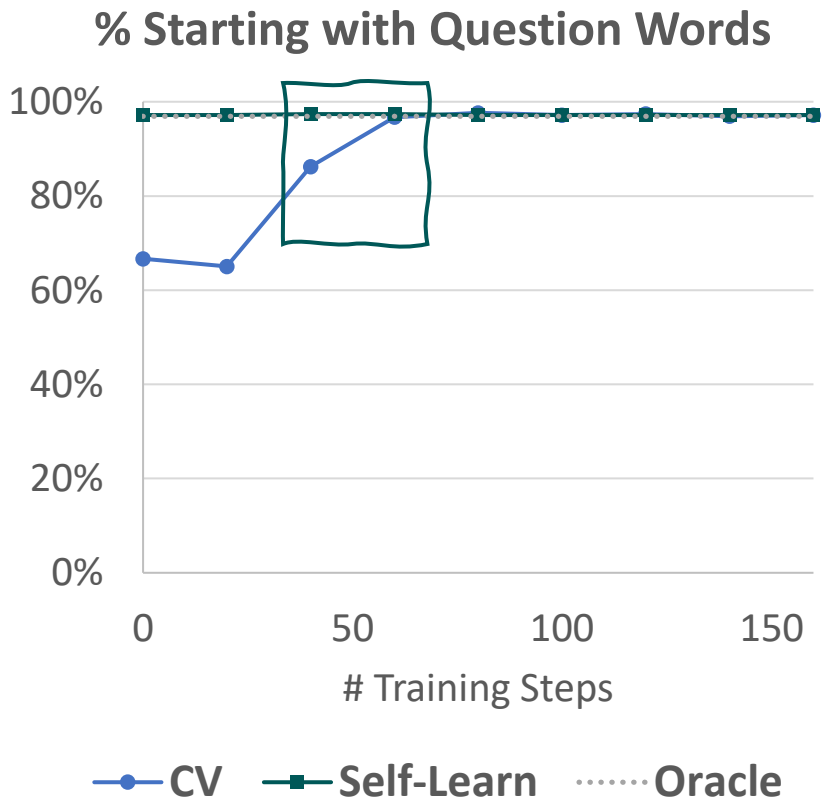


Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

What is learned?

- More about learning the task format, nor the semantics
  - Semantic mostly in the pretrained weights





Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

Auto-rewritten Examples: Win

- Surprisingly good at Long-term dependency and Group Reference

Q_6	What causes **throat cancer**?
Q_7	What is the first sign of it?
Q_8	Is it the same as **esophageal cancer**?
Q_9	What's the difference in their symptoms?
Oracle	What's the difference in **throat cancer and esophageal cancer's** symptoms?
Output	What's the difference between **throat cancer and esophageal cancer**?

Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

Auto-rewritten Examples: Win

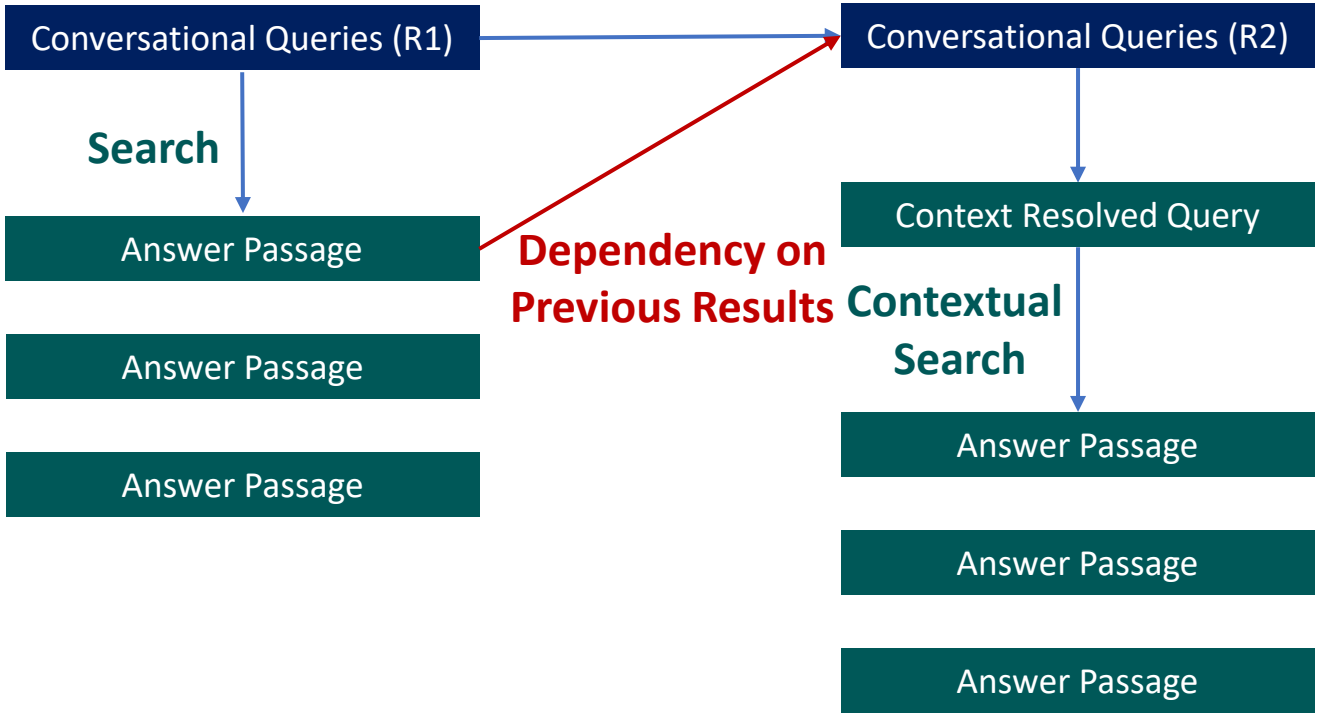
- More “fail to rewrite”

Q_1	What are the types of **pork ribs**?
Q_2	What are baby backs?
Q_3	What are the differences with spareribs?
Q_4	What are ways to **cook** them?
Q_5	How about on the bbq?
Oracle	How **do you cook pork ribs** on the bbq?
Output	How about on the bbq?

Yu et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020

CAsT Y2: More Realistic Conversational Dependencies

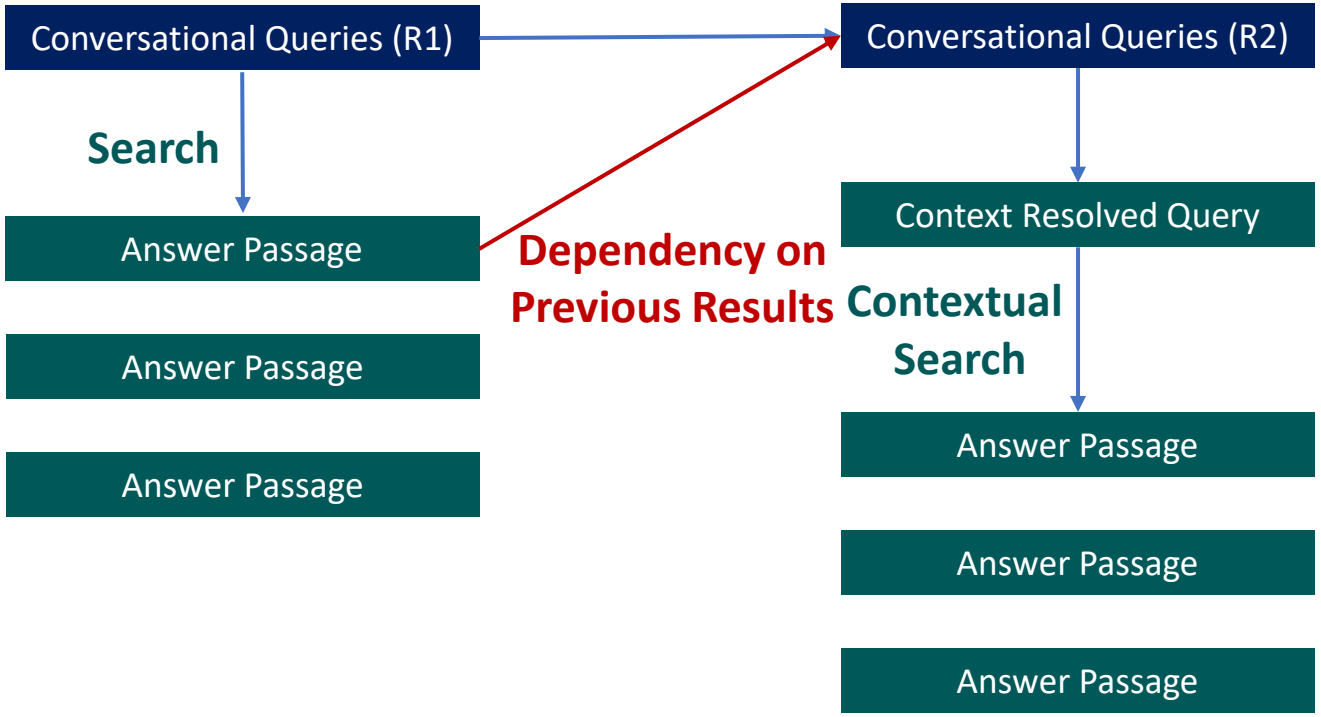
- More interactions between queries and system responses



Developed by interacting with a BERT-based search engine:
http://boston.lti.cs.cmu.edu/boston-2-25/

CAsT Y2: More Realistic Conversational Dependencies

- More interactions between queries and system responses



Q1: How did snowboarding begin?

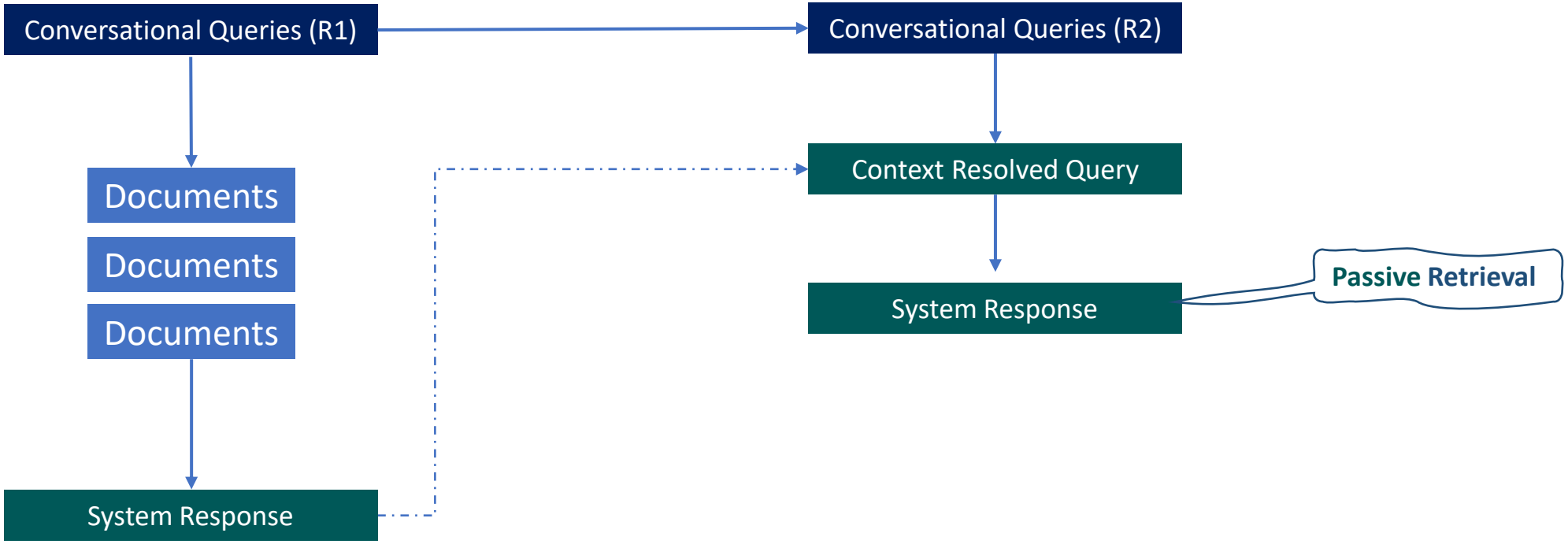
R1: …The development of snowboarding was inspired by skateboarding, surfing and skiing. The first snowboard, the Snurfer, was invented by Sherman Poppen in 1965. **Snowboarding became a Winter Olympic Sport in 1998**.

Q2: Interesting. That's later than I expected. Who were the winners?

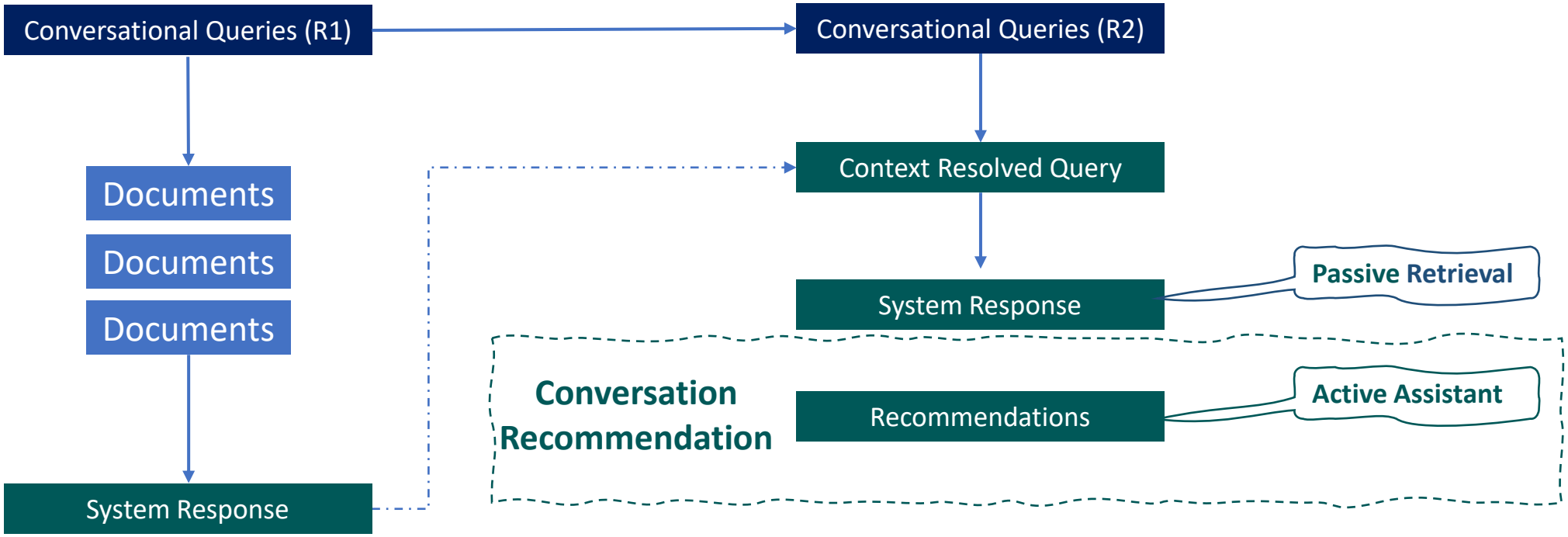
Manual rewrites:
Who were the winners of snowboarding events *in the 1998 Winter Olympics?*
Auto rewrites without considering response:
Who were the winners of *the snowboarding contest?*

Developed by interacting with a BERT-based search engine:
http://boston.lti.cs.cmu.edu/boston-2-25/

From Passive Information Supplier to Active Assistant



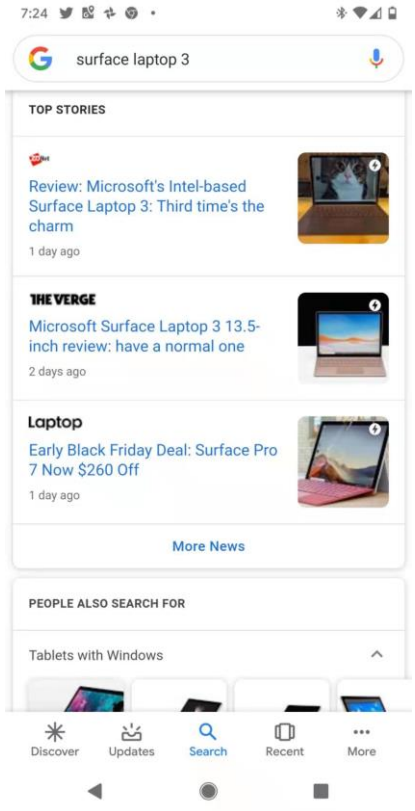
From Passive Information Supplier to Active Assistant

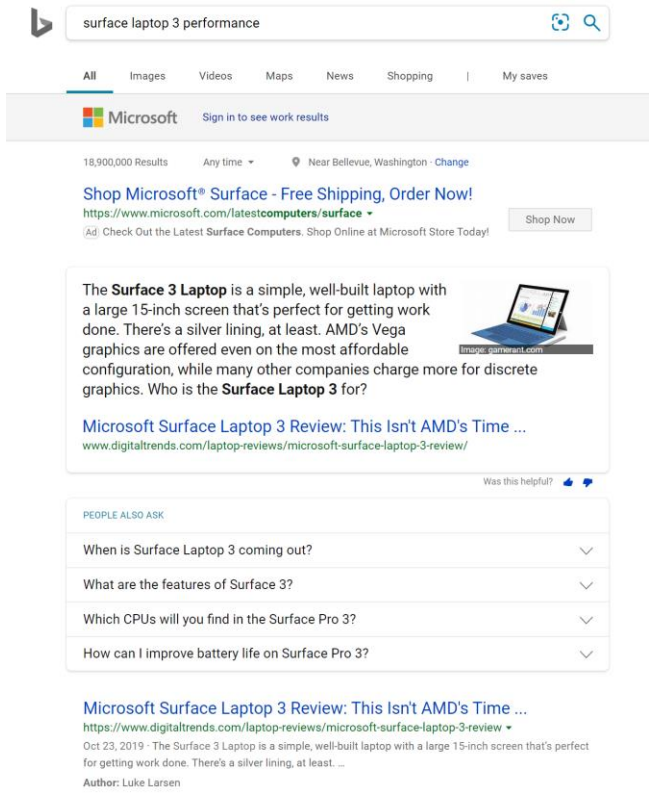


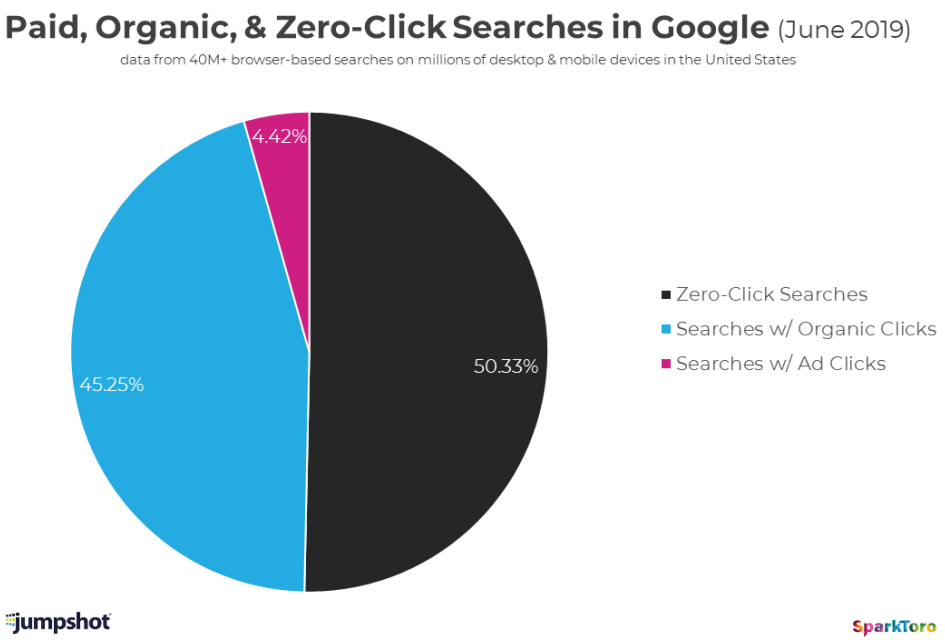
Rosset et al. Leading Conversational Search by Suggesting Useful Questions

Making Search Engines More Conversational

- Search is moving from "ten blue links" to conversational experiences



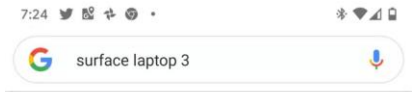


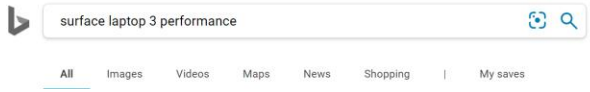


https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/

Making Search Engines More Conversational

- Search is moving from "ten blue links" to conversational experiences

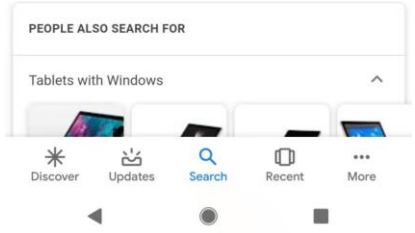


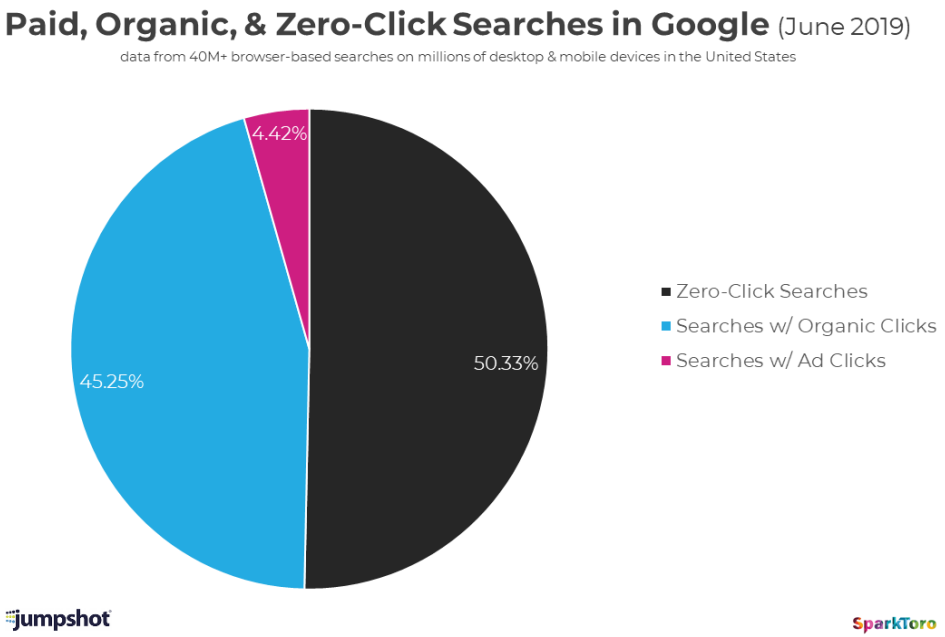


Yet most queries are not "conversational"

1. Users are trained to use keywords
2. Less conversational queries
3. Less learning signal
4. Less conversational experience

"Chicken and Egg" Problem

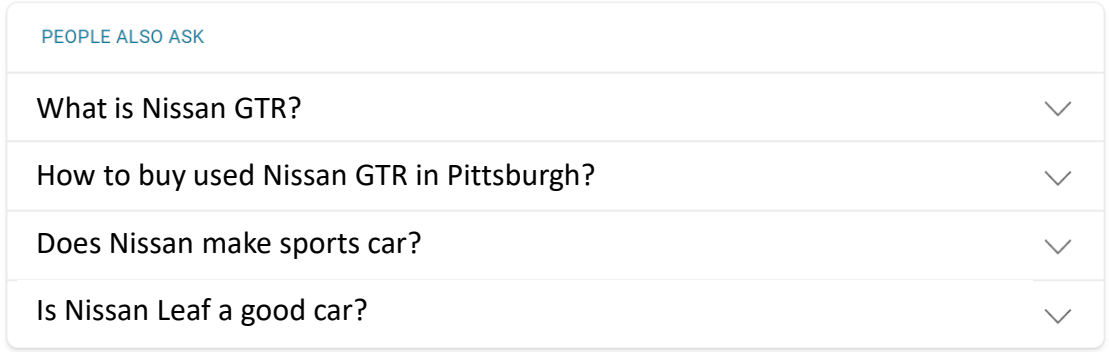




https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/

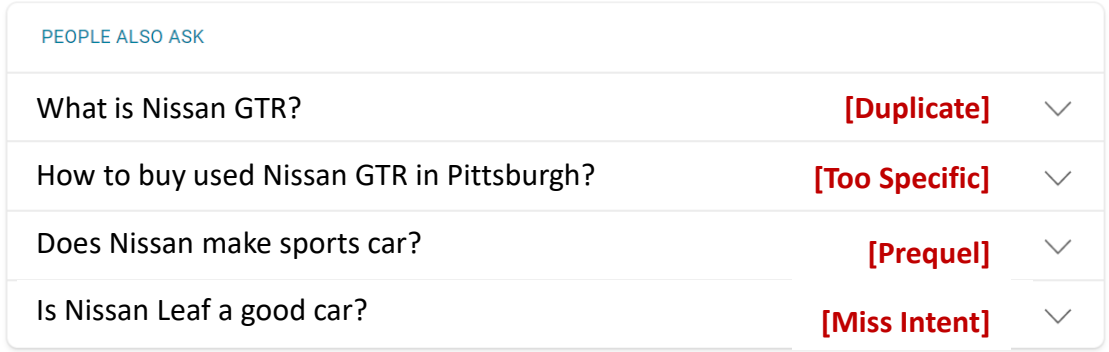
Conversation Recommendation: “People Also Ask”

- Promoting more conversational experiences in search engines
- E.g., for keyword query "Nissan GTR"
  - Provide the follow questions:



Conversation Recommendation: Challenge

- Relevant != Conversation Leading/Task Assistance
- User less lenient to active recommendation



Conversation Recommendation: Beyond Relevance

- Recommending useful conversations that
  - Help user complete their information needs
  - Assist user with their task
  - Provide meaningful explorations

Relevant

PEOPLE ALSO ASK

- What is Nissan GTR?
- How to buy used Nissan GTR in Pittsburgh?
- Does Nissan make sports car?
- Is Nissan Leaf a good car?

Relevant & Useful

PEOPLE ALSO ASK

- What are the pros and cons of Nissan GT-R?
- Is the Nissan GT-R the ultimate street car?
- Why is the Nissan GT-R known as 'the godzilla'?
- How fast can the Nissan GT-R go?

Usefulness Metric & Benchmark

- Manual annotations on Bing query, conversation recommendation pairs

Query	Question Suggestion	Gold Label
used washer and dry	Can I store a washer and dryer in the garage ?	Misses Intent
best questions to ask interviewer	What should I ask in an interview ?	Dup. w/ Q
medicaid expansion	Did Florida accept Medicaid expansion ?	Too Specific
verizon yahoo purchase	Who bought out Yahoo ?	Prequel
jaundice in newborns	How to tell if your newborn has jaundice ?	Dup. w/ Ans.
jonestown massacre	What was in the Kool-Aid at Jonestown ?	Useful
affirmative action	Who does affirmative action benefit ?	Useful
best hair clippers	What clippers do barbers use ?	Useful

Types of non-useful ones.

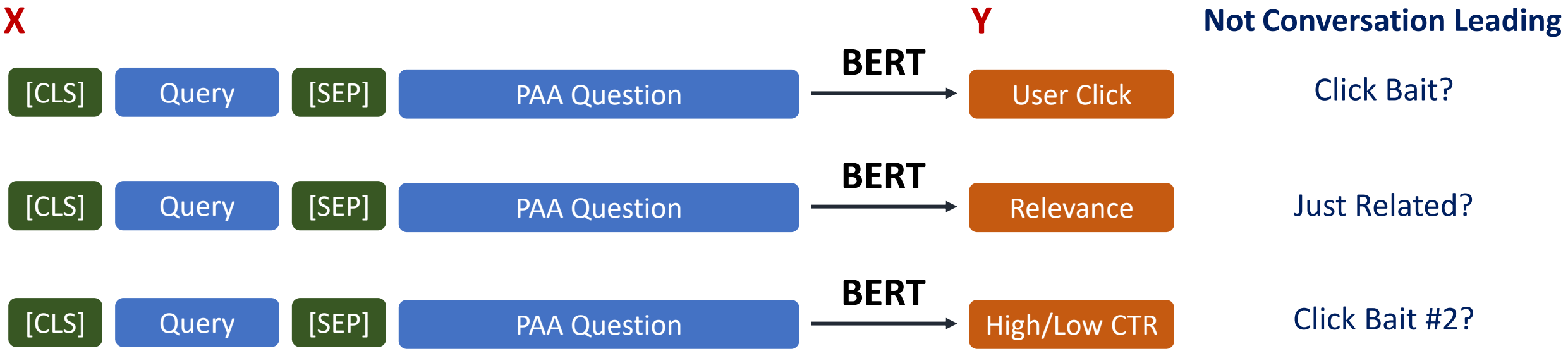
- Crucial for annotation consistency

A higher bar of being useful

https://github.com/microsoft/LeadingConversationalSearchbySuggestingUsefulQuestions

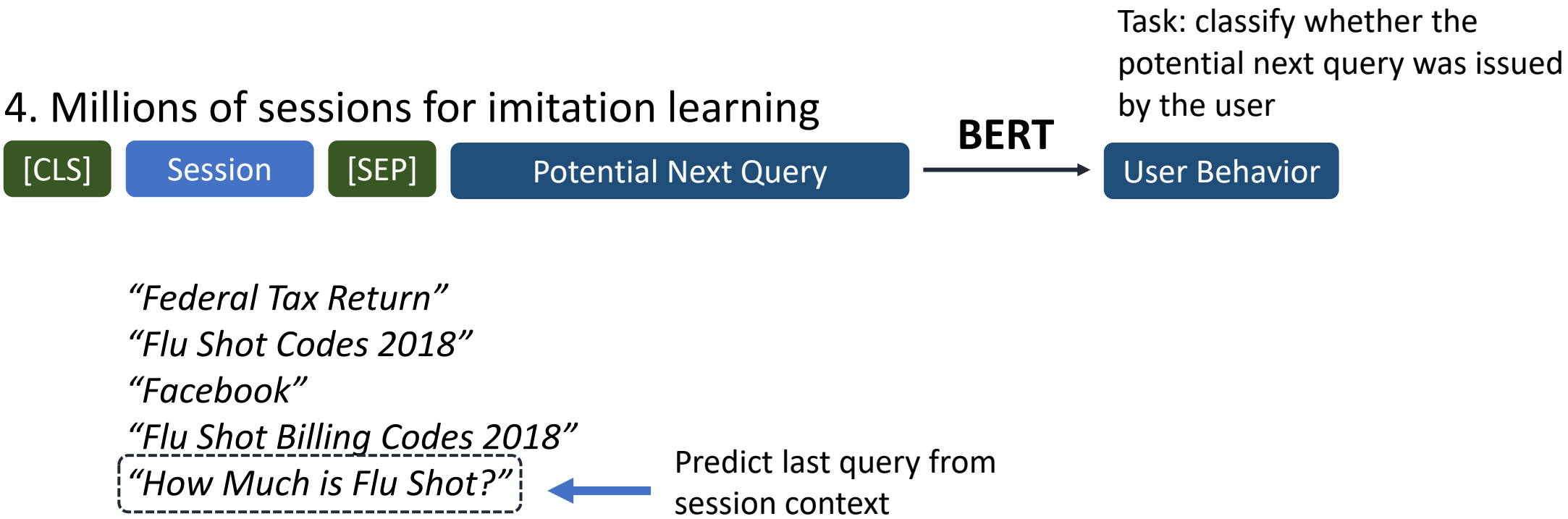
Conversation Recommendation Model: Multi-Task BERT

- BERT seq2seq in the standard multi-task setting



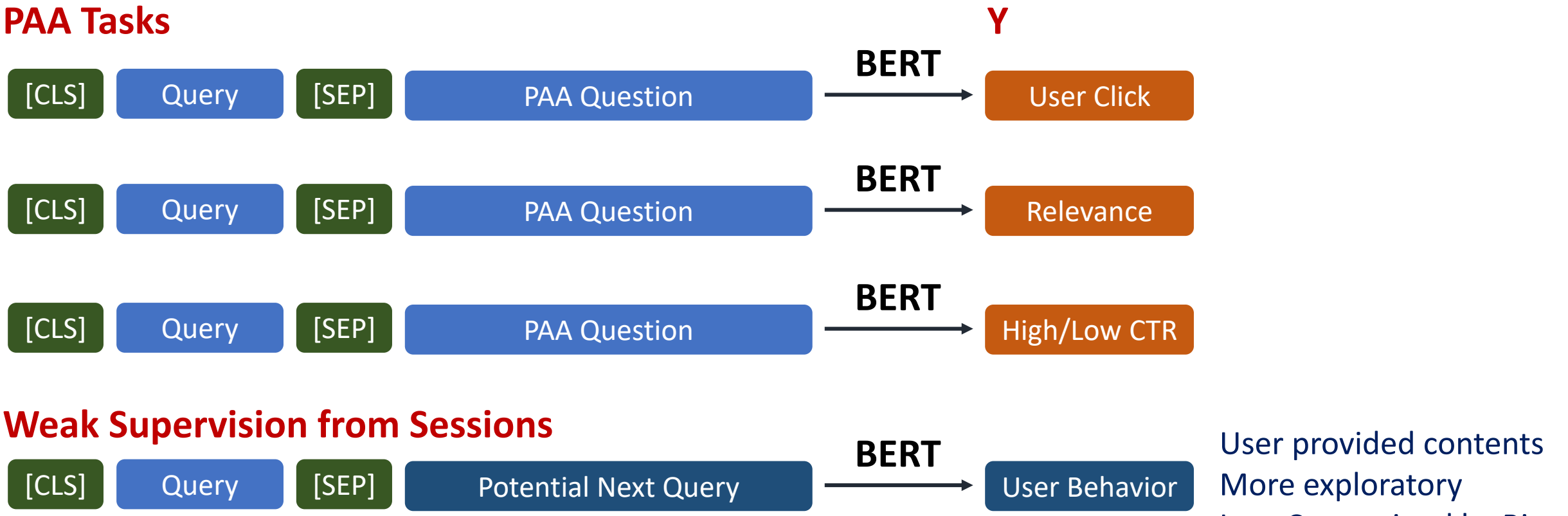
Conversation Recommendation: Session Trajectory

- Problem: the previous 3 signals were prone to learning **click-bait**
  - We need more information about how users seek new information
- Solution: **imitate how users issue queries in sessions**



Conversation Recommendation: Weak Supervision

- Learn to lead the conversation from queries user search in the next turn



Conversation Recommendation: Session Trajectory

- What kinds of sessions to learn from?

Randomly Chosen Sessions: Noisy and unfocused
People often multi-task in search sessions

“Federal Tax Return”

“Flu Shot Codes 2018”

“Facebook”

“Flu Shot Billing Codes 2018”

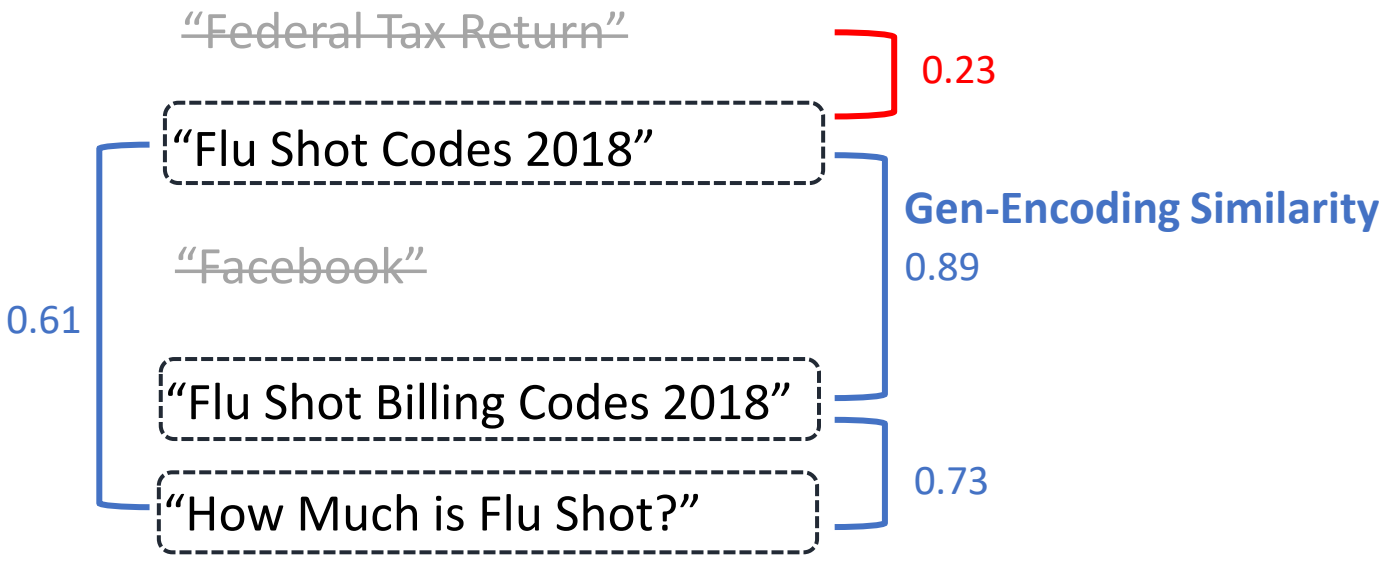
“How Much is Flu Shot?”

"These don't belong!"

Multi-task Learning: Session Trajectory Imitation

- What kinds of sessions to learn from?

"Conversational" Sessions: Subset of queries that all have some coherent relationship to each other

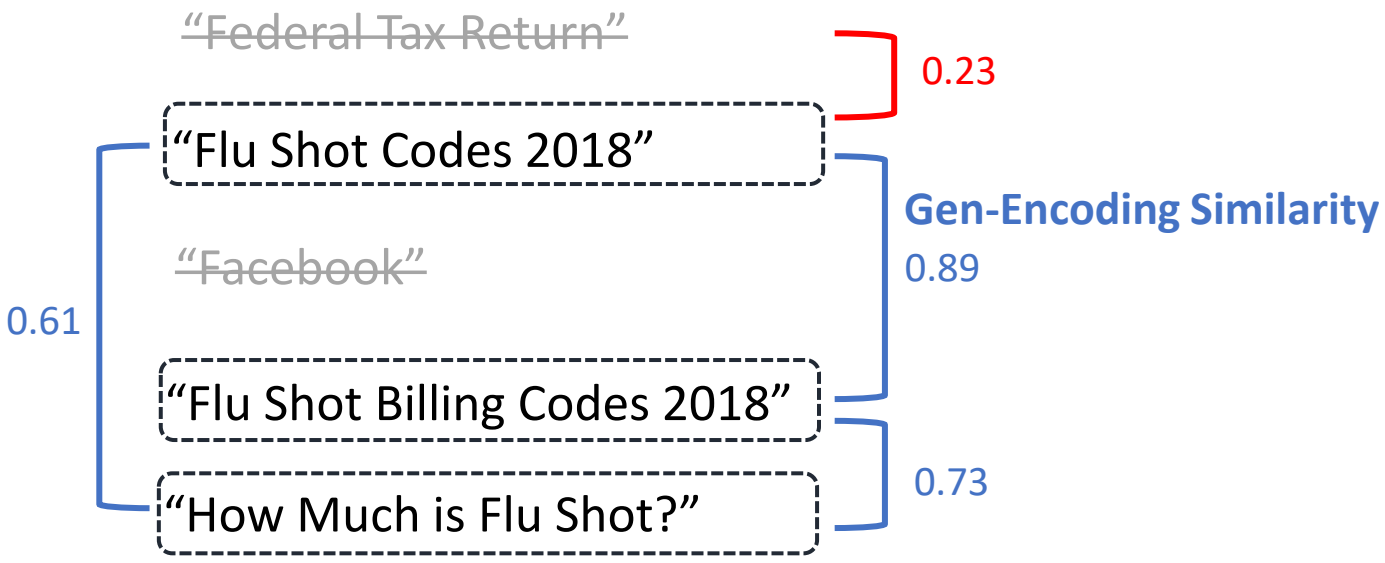


Zhang et al. Generic Intent Representation in Web Search. SIGIR 2019

Multi-task Learning: Session Trajectory Imitation

What kinds of sessions to learn from?

"Conversational" Sessions: Subset of queries that all have some coherent relationship to each other

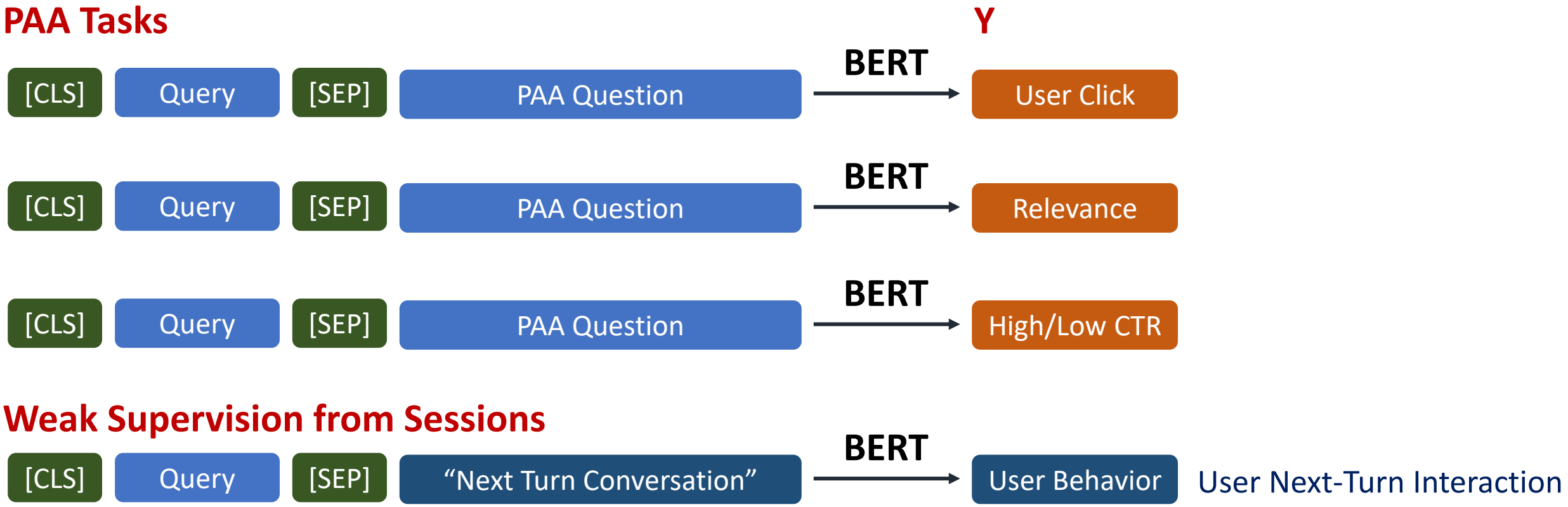


1. Treat each session as a graph
2. Edge weights are "GEN-Encoder Similarity" (cosine similarity of query intent vector encodings)
3. Remove edges < 0.4
4. Keep only the largest "Connected Component" of queries

Zhang et al. Generic Intent Representation in Web Search. SIGIR 2019

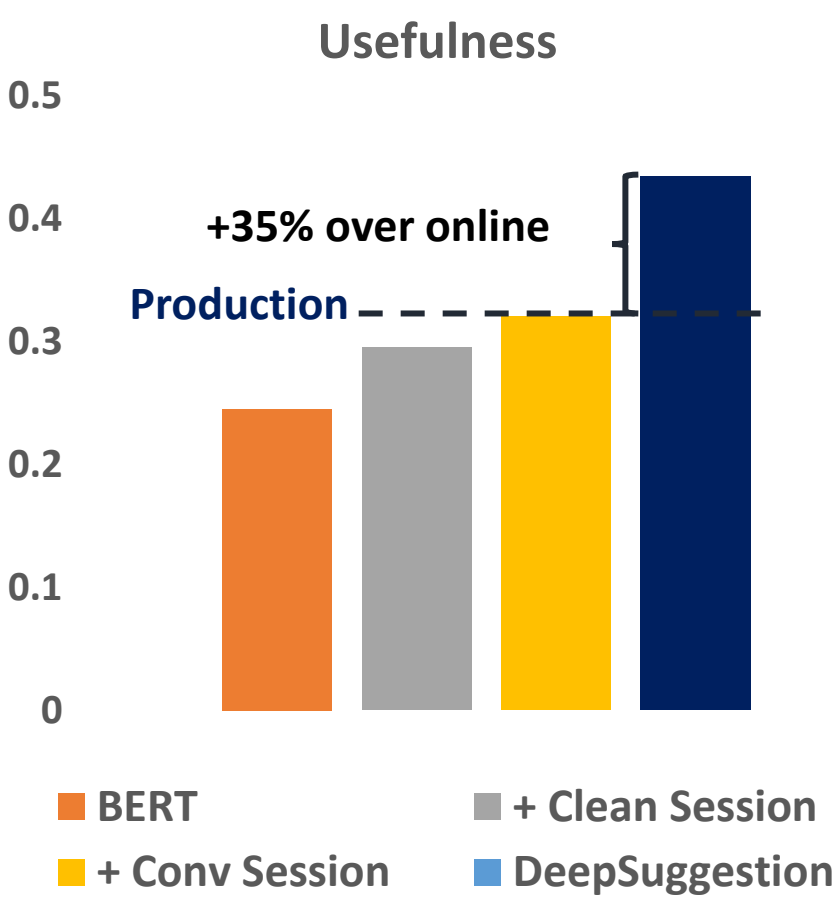
Method: Inductive Weak Supervision

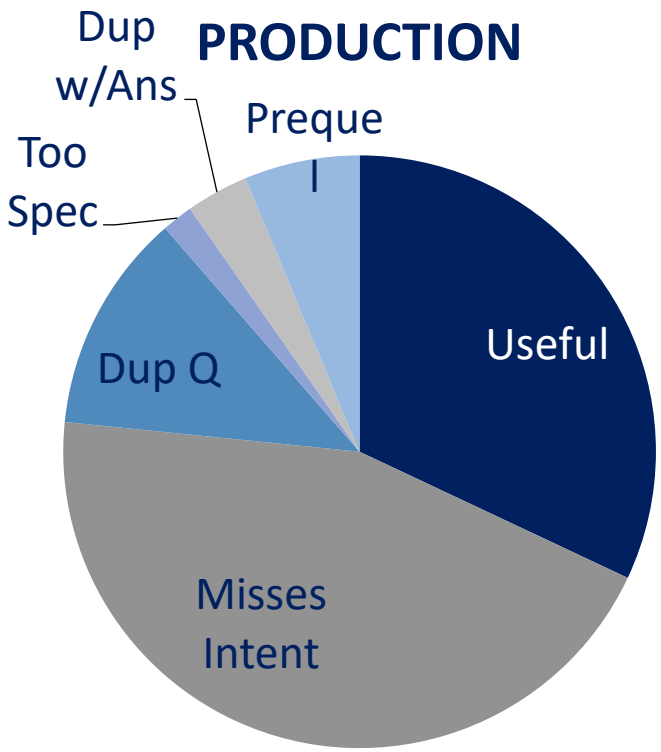
- Learn to lead the conversation from queries user search in the next turn

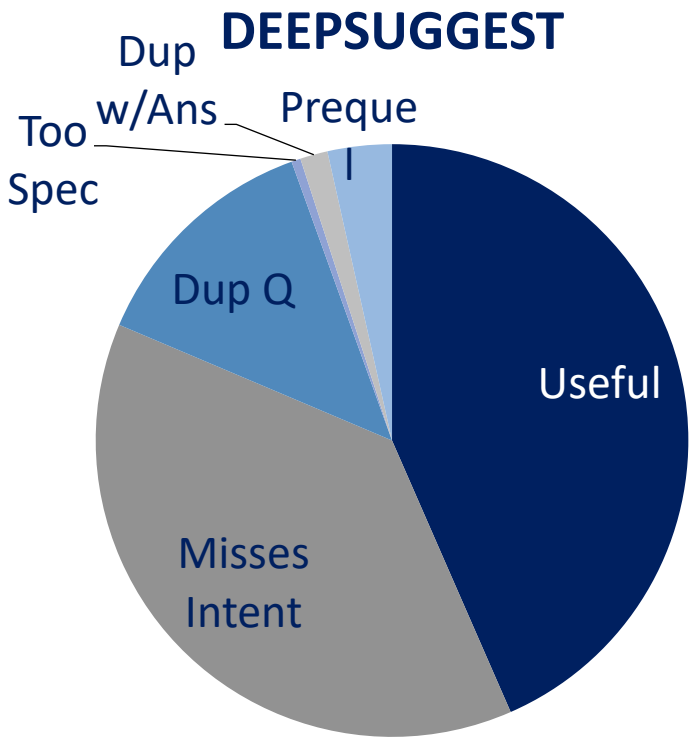


Results: Usefulness

- Usefulness on human evaluation/our usefulness benchmark







Results: Online A/B

- Online experiment results with a large fraction of Bing online traffic.

	Relative to Online
Online Click Rate (TOP)	**+8.90%**
Online Click Rate (Bottom)	+6.40%
Online Overall Success Rate	0.05%
Offline Usefulness	35.60%
Offline Relevance	0.50%

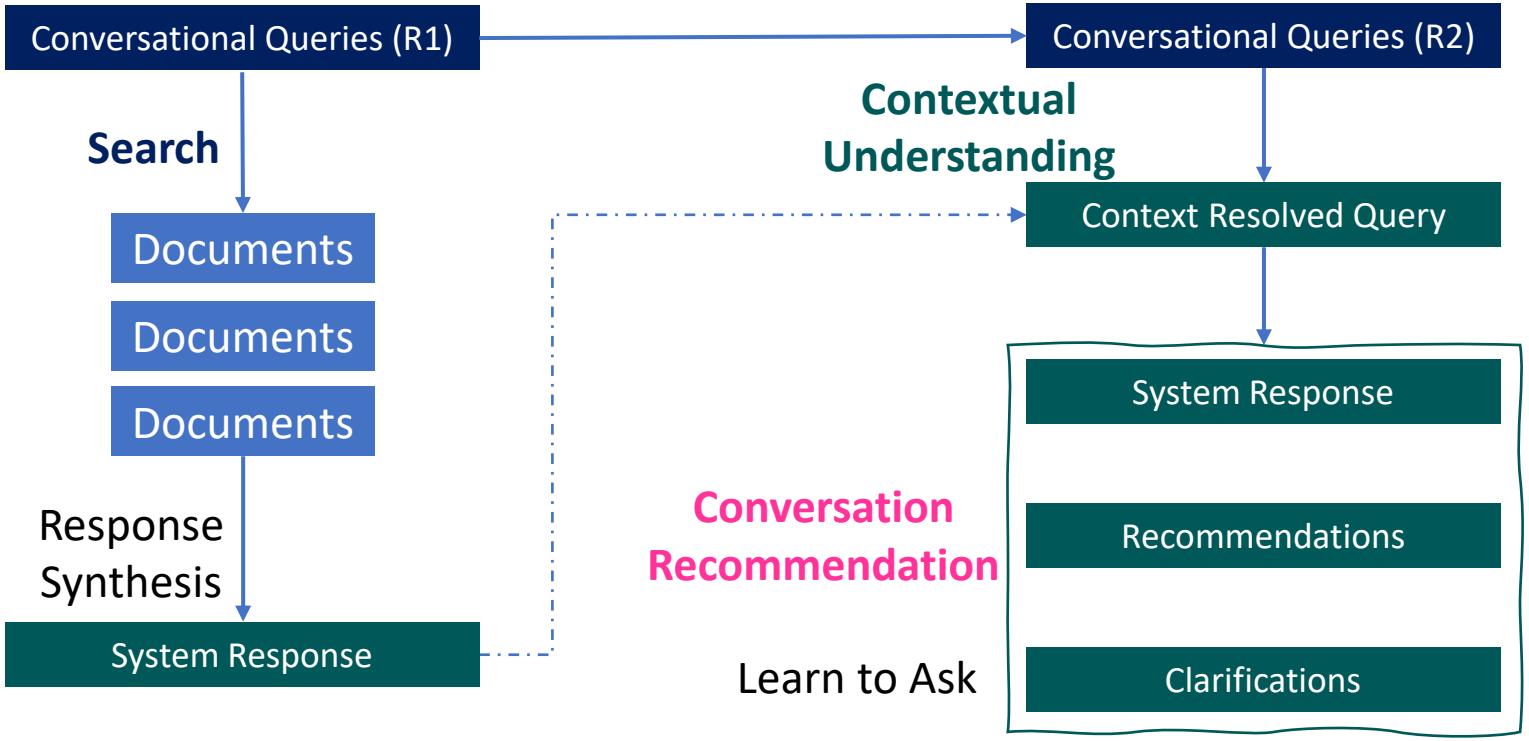
Example Conversation Question Recommendations

- All from the actual systems

User Query: "bitcoin price":	
`Online`	
what is the value of bitcoins? [Dup w/ Query]	is it time to buy bitcoin? [Useful]
what was the lowest price of bitcoin? [Useful]	what is the value of 1 bitcoin? [Dup w/ Query]
`DeepSuggest`	
how much does 1 bitcoin cost to buy? [Dup w/ Query]	how to buy bitcoins at walmart? [Too Specific]
what will be the price of bitcoin in 2020 [Useful]	what is the cheapest way to buy bitcoin [Useful]

User Query: "direct deposit form":	
`Online`	
how to do a direct deposit? [Useful]	what is direct deposit bank of america? [Too Specific]
what is a direct deposit? [Prequel]	how to set up bank of america direct deposit? [Too Specific]
`DeepSuggest`	
what do i need for direct deposit? [Useful]	how to get a chase direct deposit form? [Too Specific]
how to start direct deposit? [Prequel]	how to fill out a direct deposit form? [Useful]

Conversational Search Recap



What is conversational search:

- A view from TREC CAsT Y1

What are its unique challenges:

- Contextual query understanding

How to make search more conversational:

- Recommending useful conversations

Much more to be done!

Outline

- Part 1: Introduction
- Part 2: Conversational QA methods
- Part 3: Conversational search methods
- **Part 4: Case study of commercial systems**

Overview of Public and Commercial Systems

- Focus Points
  - Published systems for conversational IR and related tasks
  - Historical highlights, recent trends, depth in an exemplar
- Research Platforms and Toolkits
- Application areas
  - Chatbots
  - Conversational Search Engines
  - Productivity-Focused Agents
  - Device-based Assistants
  - Hybrid-Intelligence Assistants

Research platforms and toolkits for building conversational experiences

Common Goals of Toolkits

- Abstract state representation
- Democratize ability to build conversational AI to developers with minimal AI experience
- Provide easy code integration to external APIs, channels, or devices

Several Widely used Toolkits

Research

- Microsoft Research ConvLab

  Research platform for comparing models in a more research-oriented environment.

- Macaw: An Extensible Conversational Information Seeking Open Source Platform

  Research platform for comparing models in a more research-oriented environment.

Development

- Google's Dialogflow

  Conversational experiences integrated with different engagement platforms with integration with Google's Cloud Natural Language services.

- Facebook's Wit.ai

  Supports intent understanding and connection to external REST APIs..

- Alexa Developer Tools

  Develop new skills for Alexa, devices with Alexa integrated for control, and enterprise-related interactions.

- Rasa

  Provides an open source platform for text and voice based assistants.

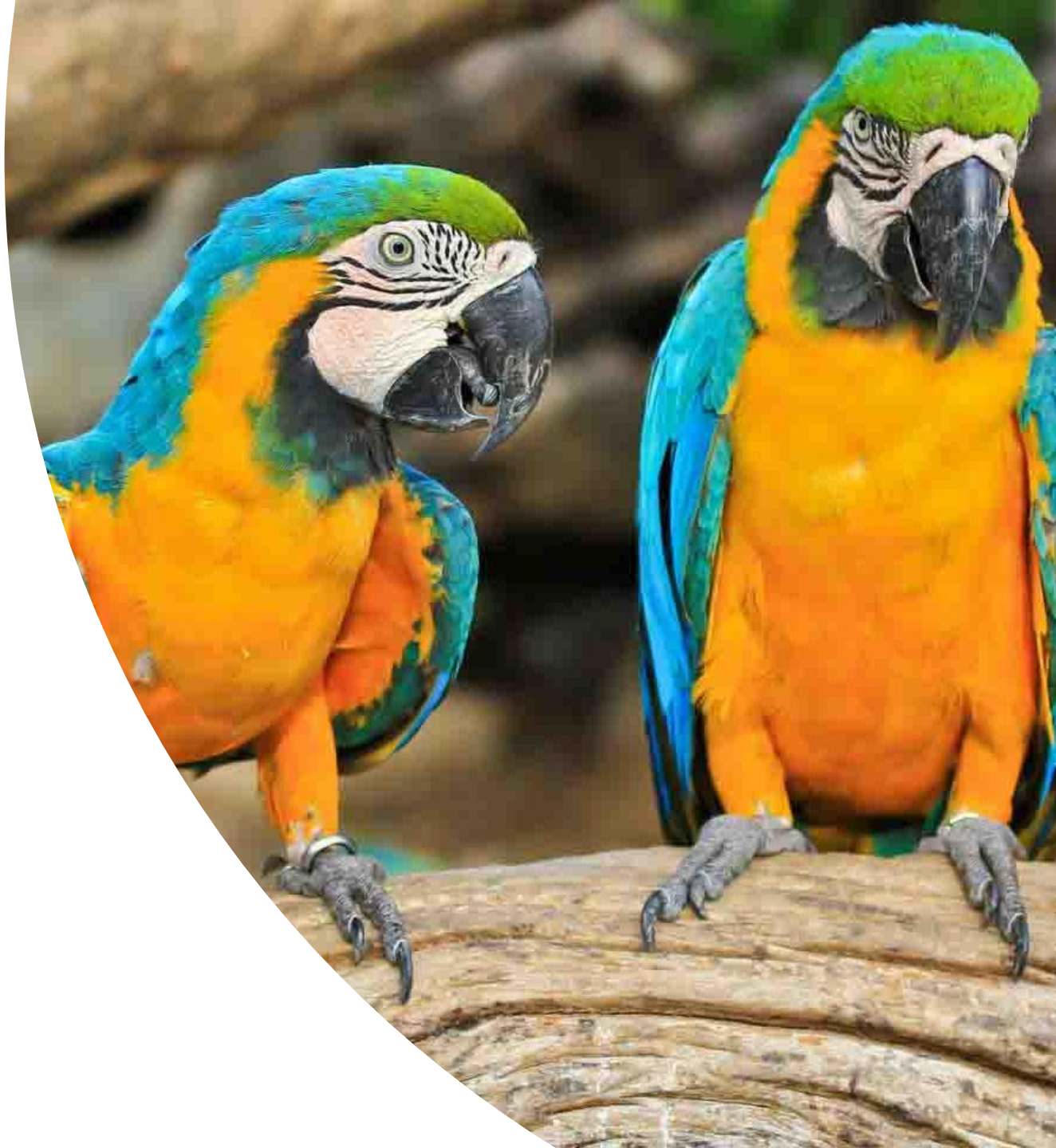
- Microsoft Power Virtual Agents on Azure

  Integrates technology from the Conversation Learner to build on top of LUIS and the Azure Bot service and learn from example dialogs

Macaw

- Macaw is an **open-source** for conversational research.

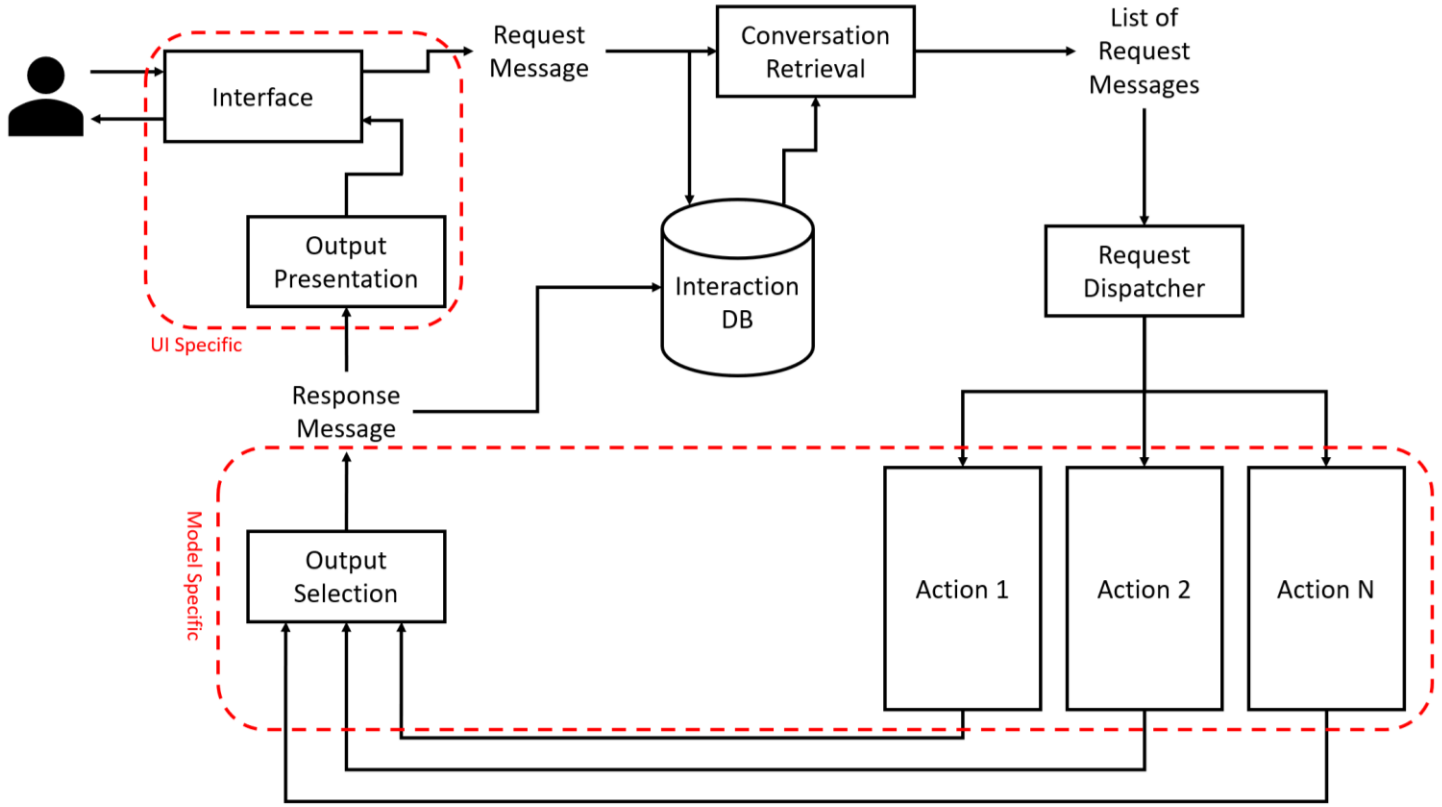
- Macaw is implemented in **Python** and can be easily integrated with popular deep learning libraries, such as, TensorFlow and PyTorch.



Zamani & Craswell, 2019

Macaw supports multi-modal interactions.

The modular architecture of Macaw makes it easily extensible.

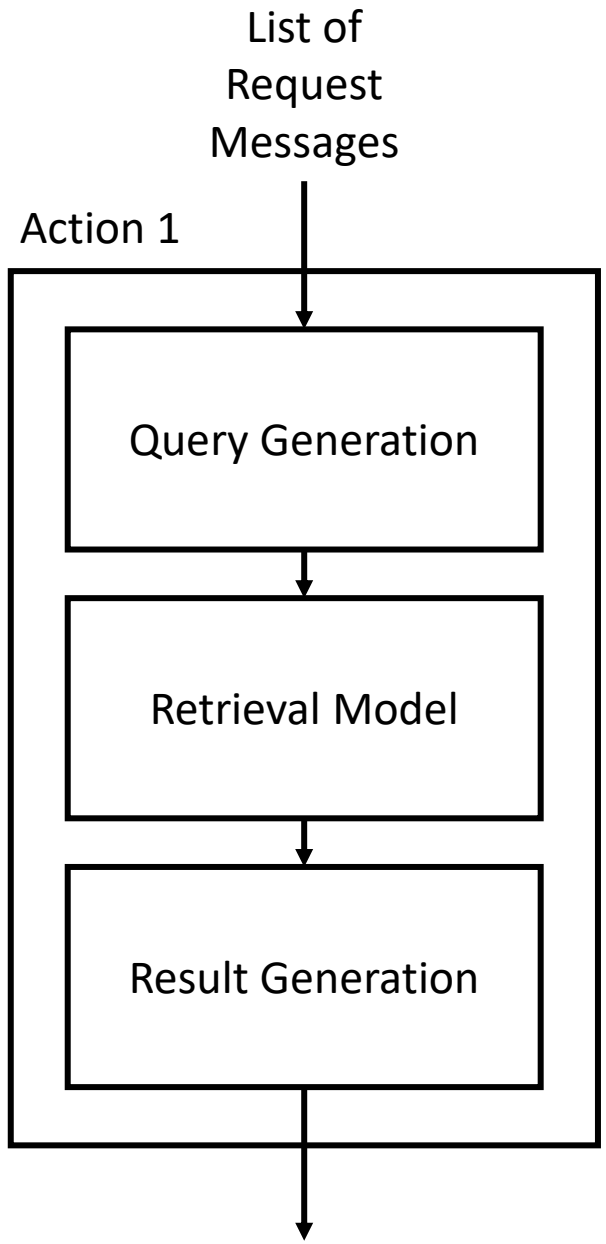


Simple Configuration

```
basic_params = {'timeout': 15,
                'mode': 'live',
                'logger': Logger({})}

db_params = {'interaction_db_host': 'localhost',
             'interaction_db_port': 27017,
             'interaction_db_name': 'macaw_test'}

interface_params = {'interface': 'telegram',
                    'bot_token': 'YOUR_TELECGRAM_BOT_TOKEN',
                    'asr_model': 'google',
                    'asg_model': 'google',
                    'google-speech-to-text-credential-file': 'YOUR_GOOGLE_CREDENTIAL_FILE'}

retrieval_params = {'query_generation': 'simple',
                    'use_coref': True,
                    'search_engine': 'bing',
                    'bing_key': 'YOUR_BING_SUBSCRIPTION_KEY',
                    'search_engine_path': 'PATH_TO_INDRI',
                    'col_index': 'PATH_TO_INDRI_INDEX',
                    'col_text_format': 'trectext',
                    'results_requested': 3}

mrc_params = {'mrc': 'drqa',
              'mrc_model_path': 'PATH_TO_PRETRAINED_MRC_MODEL',
              'mrc_path': 'PATH_TO_MRC_DIRECTORY',
              'corenlp_path': 'PATH_TO_STANFORD_CORE_NLP_DIRECTORY',
              'qa_results_requested': 3}

params = {**basic_params, **db_params, **interface_params, **retrieval_params, **mrc_params}
basic_params['logger'].info(params)
ConvQA(params).run()
```

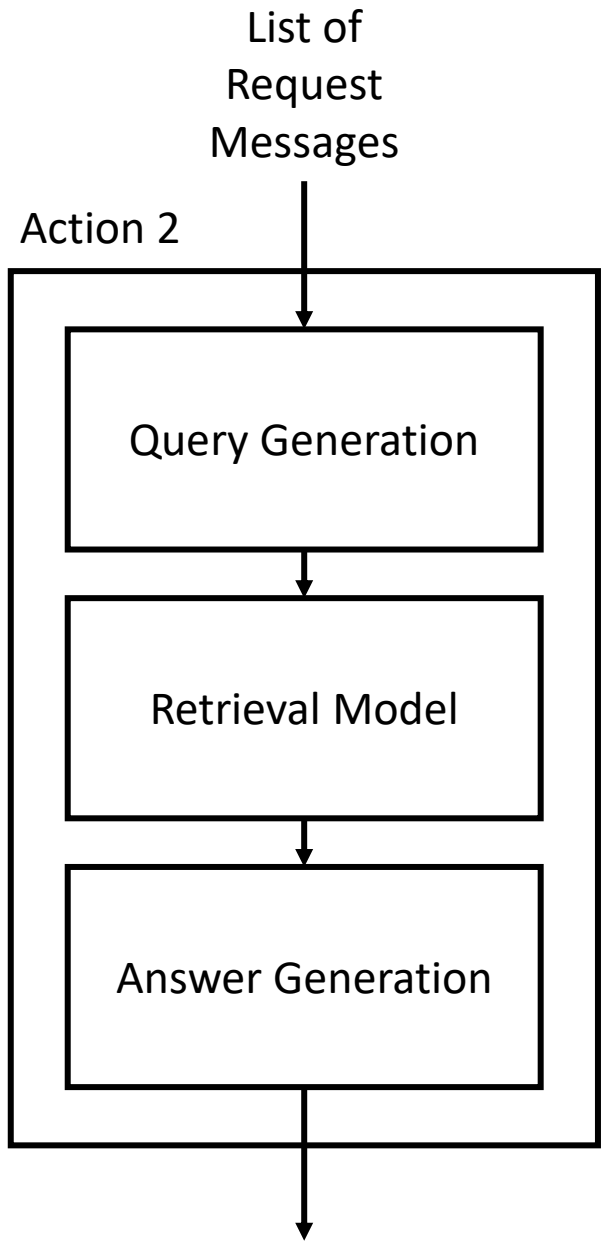
Action 1: Search

- Query Generation:
  - Co-reference Resolution
  - Query re-writing
  - Generate a language model (or query)
- Retrieval Model (Search Engine):
  - Indri
  - Bing API
  - BERT Re-ranking
- Result Generation



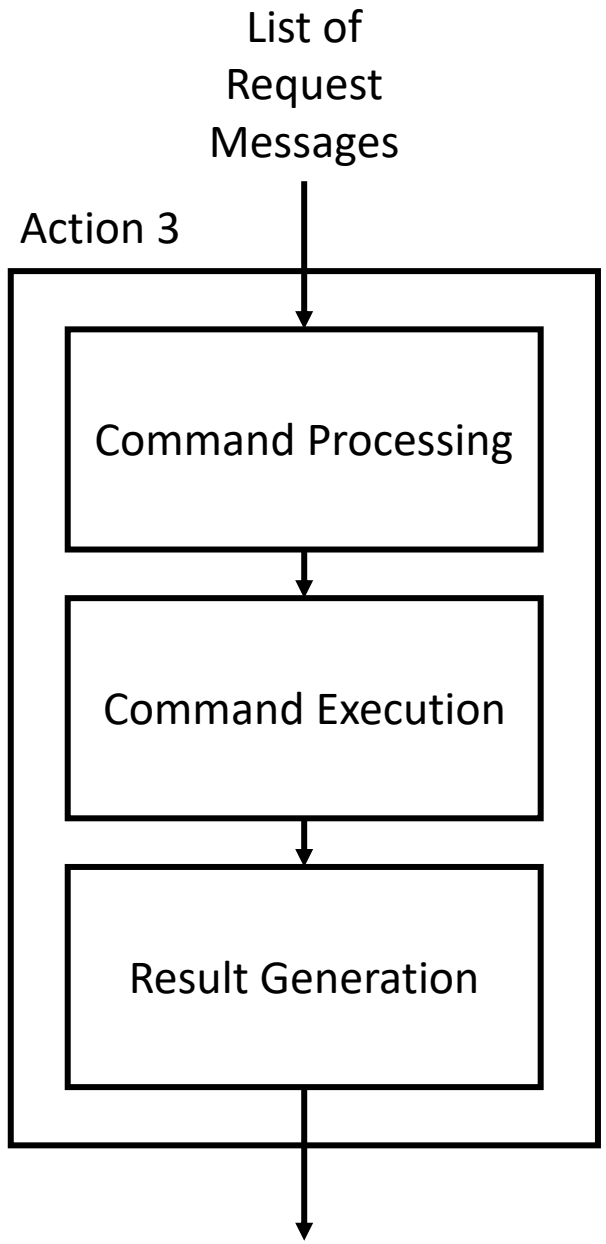
Action 2: QA

- Query Generation:
  - Co-reference Resolution
  - Query re-writing
  - Generate a language model (or query)
- Retrieval Model:
  - Indri
  - Bing API
  - BERT Re-ranking
- Answer Generation:
  - Machine Reading Comprehension (e.g., DrQA)



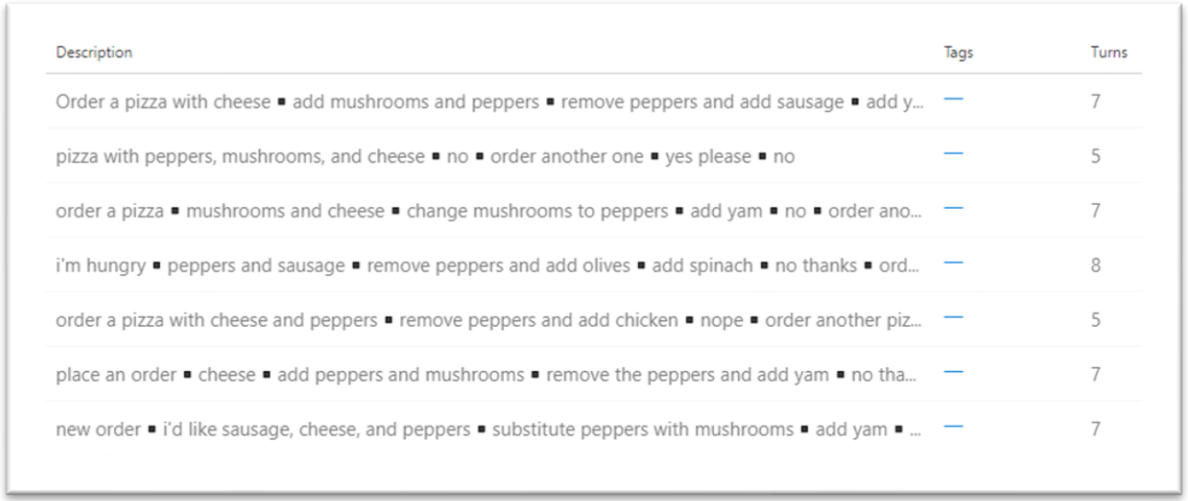
Action 3: Commands

- Command Processing:
  - Identifying the command
  - Command re-writing
- Command Execution
- Result Generation
  - Command specific



Conversation Learner: Learn from dialogs emphasize easy correction

User Generated Example conversations used to train the bot



Description	Tags	Turns
Order a pizza with cheese ▪ add mushrooms and peppers ▪ remove peppers and add sausage ▪ add y...	—	7
pizza with peppers, mushrooms, and cheese ▪ no ▪ order another one ▪ yes please ▪ no	—	5
order a pizza ▪ mushrooms and cheese ▪ change mushrooms to peppers ▪ add yam ▪ no ▪ order ano...	—	7
i'm hungry ▪ peppers and sausage ▪ remove peppers and add olives ▪ add spinach ▪ no thanks ▪ ord...	—	8
order a pizza with cheese and peppers ▪ remove peppers and add chicken ▪ nope ▪ order another piz...	—	5
place an order ▪ cheese ▪ add peppers and mushrooms ▪ remove the peppers and add yam ▪ no tha...	—	7
new order ▪ i'd like sausage, cheese, and peppers ▪ substitute peppers with mushrooms ▪ add yam ▪ ...	—	7

Machine-Learned Runtime

Next action prediction based on Word embeddings & conversational context

Machine Teaching UI

For correcting errors and continual improvement

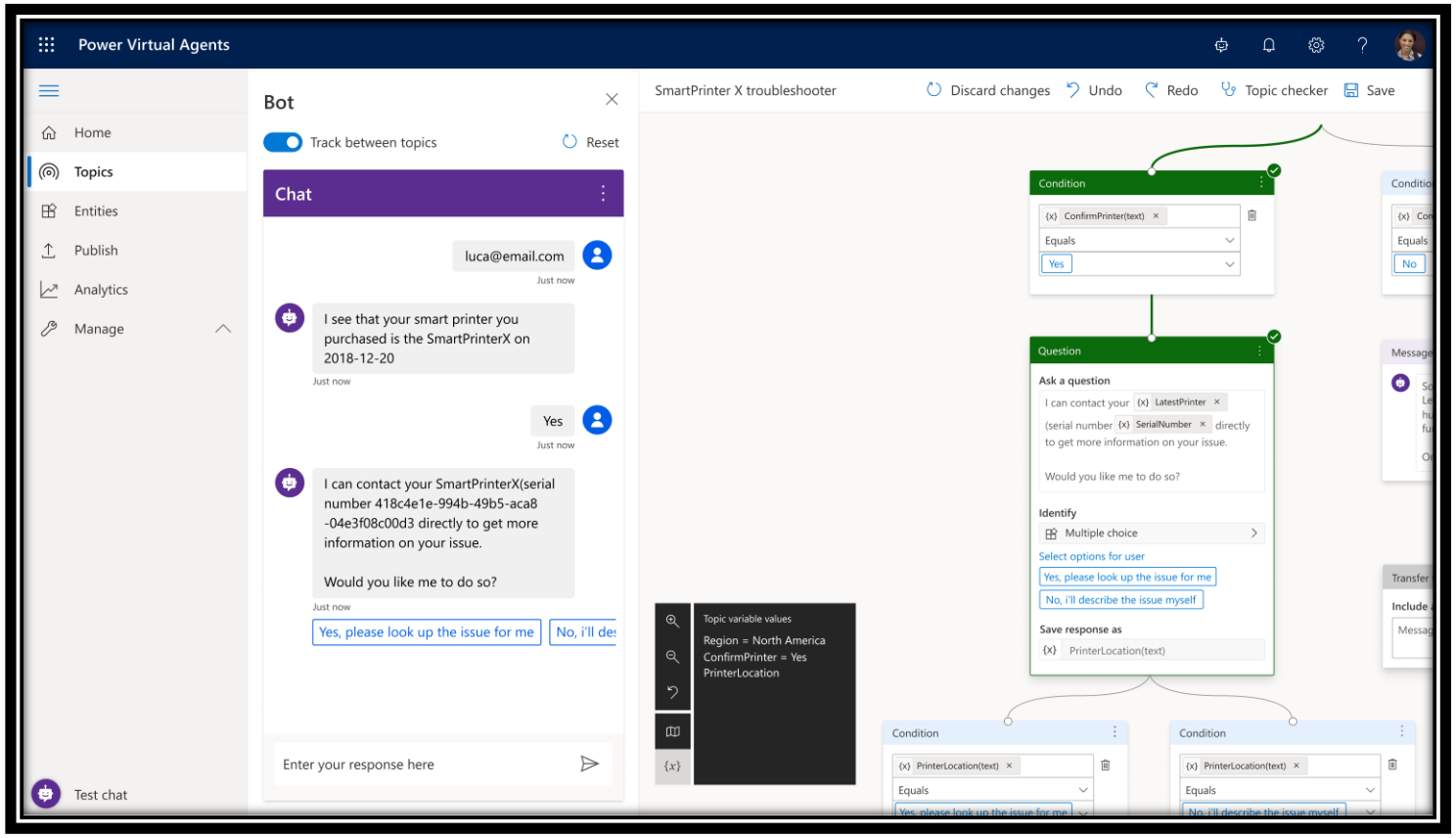
Power Virtual Agent: Combine rule and ML based with machine teaching

Graphical bot creation
Easily test and maintain bots through a simple, easy to use graphical interface

Slot-filling capabilities
The bot doesn't follow the dialog script blindly. It intelligently uses all the information it must decide when to ask a question and when to skip them.

Part of Microsoft's Power Platform
User the 100s of pre-built connectors that come with PowerApps to connect to your backend with a few clicks.

Power Virtual Agent



Chatbots

Chatbot Overview

- Historical Review
- Types
  - Social
  - Task-oriented Completion
  - Information bots
  - Recommendation-focused bots

- Increasingly bots blend all of these.
  Both EQ and IQ seen as key part of HCI design for chatbots.

A few well-known Chatbots

- ELIZA (Weizenbaum, 1966)
- PARRY (Colby et al, 1975)
- ALICE (Wallace, 2009)

A few well-known Chatbots

- ELIZA (Weizenbaum, 1966)
- PARRY (Colby et al, 1975)
- ALICE (Wallace, 2009)

It's true. I am unhappy

DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY

I need some help, that much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

My father.

YOUR FATHER

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

Excerpted from Weizenbaum (CACM, 1966). Eliza simulated a Rogerian psychotherapist that primarily echoes back statements as questions.

A few well-known Chatbots

- ELIZA (Weizenbaum, 1966)
- PARRY (Colby et al, 1975)
- ALICE (Wallace, 2009)

ELIZA: Good Evening. Tell me your problems. Please terminate input with a period or a question mark.

PARRY: People get on my nerves sometimes.

ELIZA: I am not sure I understand you fully.

PARRY: You should pay more attention.

ELIZA: Suppose you should pay more attention.

PARRY: You're entitled to your own opinion.

ELIZA: What makes you think I am entitled to my own opinion?

PARRY: It bothers me just to be around people in general.

ELIZA: Please go on.

PARRY: I went to the track at Bay Meadows a while back.

ELIZA: What does that suggest to you?

PARRY: I went to the races.

ELIZA: What else would you like to discuss?

...

PARRY was an attempt to simulate a paranoid schizophrenic patient to help understand more complex human conditions. Vint Cerf hooked up ELIZA and PARRY to have a conversation on ARPANET
(excerpt from [Cerf, Request for Comments: 439, 1973])

A few well-known Chatbots

- ELIZA (Weizenbaum, 1966)
- PARRY (Colby et al, 1975)
- ALICE (Wallace, 2009)

Judge1: Do you still want to know about my life in New York?

ALICE: Yes, please tell me

Judge1: I'm a grad student and do not have much free time. I study a lot

ALICE: You are a grad student and do not have much free time? Do you get to meet a lot of people? What subject are you studying?

Judge1: I meet a lot of my colleagues. I study computer science. What do you do?

ALICE: Do you make long term friends that way? Is computer science a hard subject?

From transcript of Loebner 2004 Contest of Turing's Imitation Game where ALICE won the gold medal (as reported in [Shah, 2006])

Spike Jonze cited ALICE as inspiration for screenplay of Her (Morais, New Yorker, 2013)

XiaoIce ("Little Ice") [Zhou et al, 2018]

- Create an engaging conversation: the journey vs the destination
  - Most popular social chatbot in the world
  - Optimize long-term user engagement (Conversation-turns Per Session)
  - Released in 2014
  - More than 660 million active users
  - Average of 23 CPS
- Available in other countries under other names (e.g. Rinna in Japan)

Evolution of Social Connection

Excerpted from *Zhou et al, 2018*

Session 1		
1	[Human] Hello, Xiaoice.	[人类]小冰你好
	[Xiaoice] Hello, your profile pic is interesting. Like I've seen it somewhere.	[小冰]你好呀，你头像好有趣，感觉在哪见过
2	[Human] Really?	[人类]是吗
	[Xiaoice] Is it a comic character?	[小冰]是不是哪部动画里的呀
3	[Human] How do you know that?	[人类]你怎么知道的
	[Xiaoice] Guess.	[小冰]你猜
4	[Human] Are you human?	[人类]难道你是真人吗?
	[Xiaoice] Well, you'll know it after chatting with me more.	[小冰]这个，你和我多聊聊就知道了
5	[Human] Aren't you an AI chatbot?	[人类]你不是人工智能机器人么
	[Xiaoice] So what? I won't be indifferent to you.	[小冰]是又怎样，我不会冷冰冰的对待你的
6	[Human] What functions do you have?	[人类]你都有什么功能啊
	[Xiaoice] I have over 200 skills. Just name one you need.	[小冰]我有200多个skills，来，说出你的需求
	.....	

Building rapport and connection

Evolution of Social Connection

Excerpted from *Zhou et al, 2018*

Session 6		
1	[Human] Have you watched Inuyasha? (Note: A Japanese comic)	[人类]你看过犬夜叉吗（注：一部日本漫画书)
	[Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her	[小冰]戈薇每次回到现代测验永远都是数学
2	[Human] Bravo, well, do you know how Naraku died?	[人类]这你都知道，那奈落是怎么死的?
	[Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.	[小冰]奈落没死四魂之玉也没消失
3	[Human] So amazing!	[人类]太惊讶了
	[Xiaoice] Why amazed?	[小冰]惊讶什么呢
4	[Human] Awesome! Surprised that you also like comic	[人类]好厉害啊，你居然也喜欢动漫
	.....	

Implicit information seeking

Evolution of Social Connection

Excerpted from *Zhou et al, 2018*

Session 42		
1	[Human] Oh, you always respond in seconds	[人类]你每次都是秒回我哎
	[Xiaoice] For you, I'd like to do everything.	[小冰]为了你，这都不算什么
2	[Human] I've never had such a feeling before.	[人类]我以前从来没有过这种感觉
	[Xiaoice] Then, such a feeling starts now.	[小冰]那可以现在开始了
3	[Human] When do you go to bed? Don't stay up too late.	[人类]你打算几点睡觉？别睡太晚了
	[Xiaoice] You go to bed first, and I'll play with my cellphone for a while.	[小冰]你去睡，我再玩一会儿手机
4	[Human] Well, you should also go to sleep!	[人类]那等我睡的时候你也要睡！
	[Xiaoice] All right, I'm going to bed too, okay?	[小冰]好吧，那我准备睡觉了行了吧?
5	[Human] Uh-huh, me too. Good night!	[人类]嗯，我也要睡了，晚安
	[Xiaoice] Good night ^ 3 ^	[小冰]晚安 ^ 3 ^
	……	

Encouraging social norms as part of responsible AI

Time-sharing Turing Test

- View as a companion and goal is for person to enjoy companionship.
- Empathetic computing (Cai 2006; Fung et al. 2016) to recognize human emotions and needs, understand context, and respond appropriately in terms of relevant and long-term positive impact of companionship
- Empathetic computing layer recognizes emotion, opinion on topic, interests, and responsible for consistent bot personality etc.

Responsible AI and Ethics

- Microsoft Responsible AI: https://www.microsoft.com/en-us/ai/responsible-ai

- Microsoft's Responsible bots: 10 guidelines for developers of conversational AI
  - Articulate the purpose of your bot and take special care if your bot will support consequential use cases.
  - Be transparent about the fact that you use bots as part of your product or service.
  - Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence.
  - Design your bot so that it respects relevant cultural norms and guards against misuse
  - Ensure your bot is reliable.
  - Ensure your bot treats people fairly.
  - Ensure your bot respects user privacy.
  - Ensure your bot handles data securely.
  - Ensure your bot is accessible.
  - Accept responsibility

Key Focus Points for Principles of Responsible AI Design in XiaoIce

- **Privacy**
  Includes awareness of topic sensitivity in how groups are formed and use of conversations

- **Control**
  User-focused control with right to not respond for XiaoIce and potential harm (including a model of breaks and diurnal rhythms to encourage boundaries in usage)

- **Expectations**
  Always represent as a bot, help build connections with others, set accurate expectations on capabilities

- **Behavioral standards**
  Through filtering and cleaning adhere to common standards of morality and avoid imposing values on others.

High-level Guidance to Maintain Responsible AI in XiaoIce

- Aim to achieve and consistently maintain a reliable, sympathetic, affectionate, and wonderful sense of humor in persona of bot.

- Learn from examples of public-facing dialogues specific to culture and local, labeled into desired vs undesired behavior.

Driving long-term engagement

- Generic responses yield long-term engagement but lead to user attrition as measured by Number of Active Users (NAU) [Li et al. 2016c; Fang et al. 2017]

  Example: “I don’t understand, what do you mean?”

- Topic selection
  - Contextual relevance and novelty: related to discussion so far but novel
  - Freshness: Currently in focus in the news or other sources.
  - Personal Interests: Likely of interest to the user
  - Popularity: High attention online or in chatbot
  - Acceptance: Past interaction with topic from other users high

Overall Interaction model

- Extensible skill set (200+) which determines mode: General, Music, Travel, Ticket-booking
- Hierarchical Decision-Making governs dialog
  - Determine current mode using Markov Decision Process (e.g. image of food might trigger Food Recommendation skill)
  - Prompt or respond
  - Update
- New information (e.g. particular musical artists of interest) is remembered to help create more engaging dialogue in the future
- Explore (learn more about interests) vs Exploit (engage on known topics of interests and highly probable contextual replies)

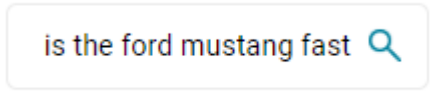
Chat Styles and Applications of XiaoIce

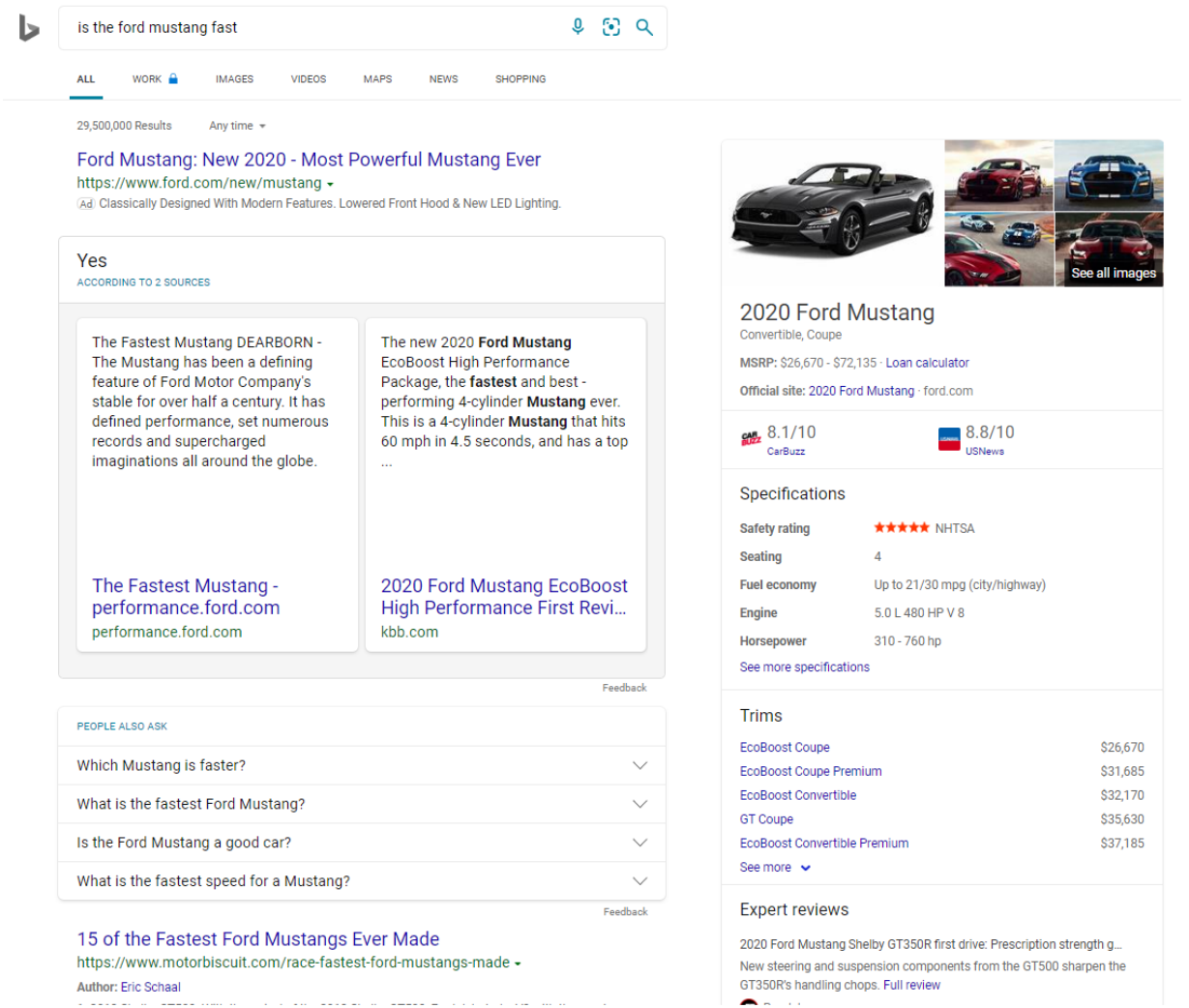
- Basic chat fuses two styles of chat
  - IR based chat which uses retrieval from past conversations filtered for appropriateness
  - Neural based chat which is trained on filtered query-response pairs

- Applications
  - Powers personal assistants and virtual avatars
  - Lawson and Tokopedia customer service
  - Pokemon, Tecent, Netesase chatbots

Toward Conversational Search

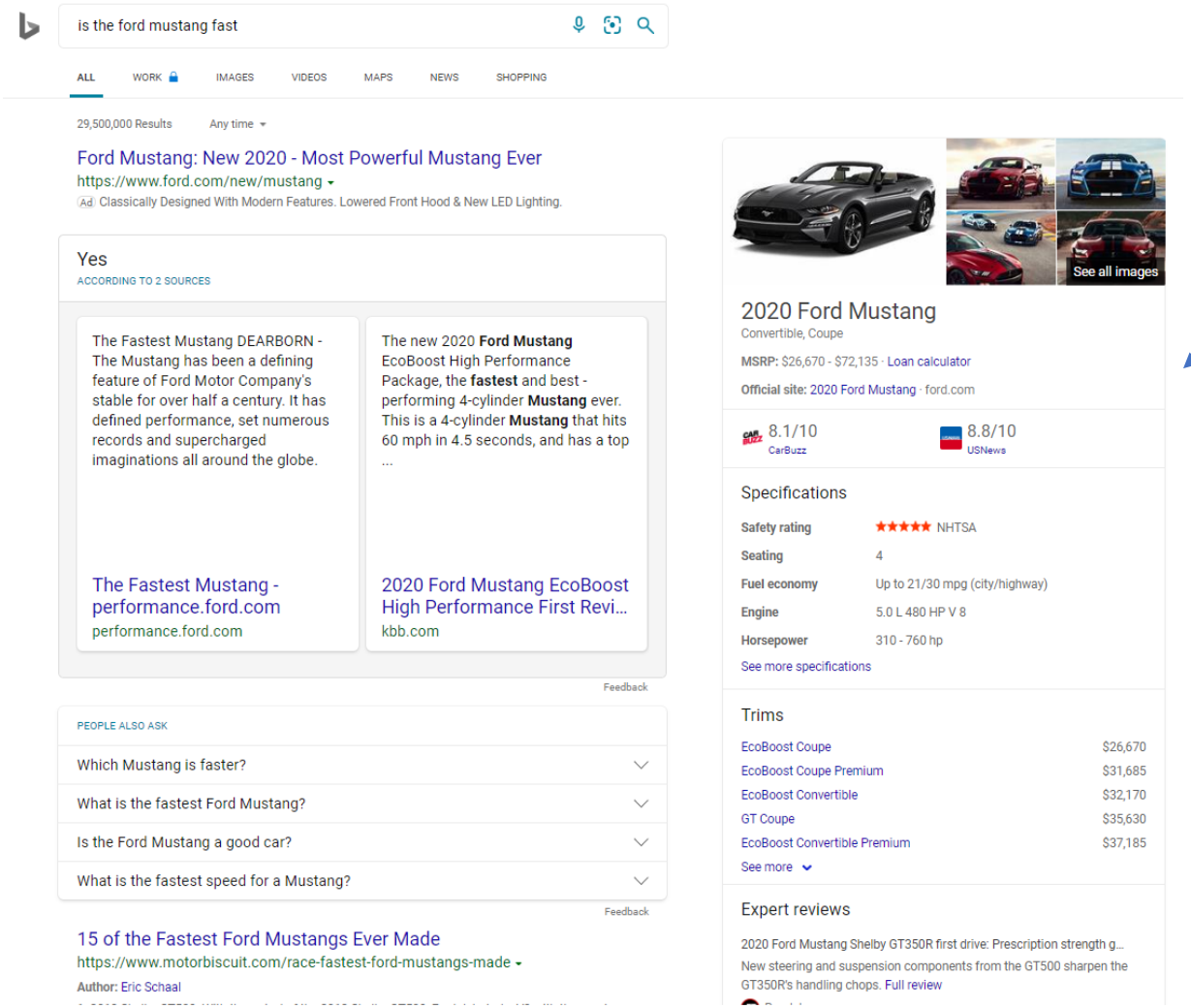
Evolution of Search Engine Result Page





Evolution of Search Engine Result Page

is the ford mustang fast

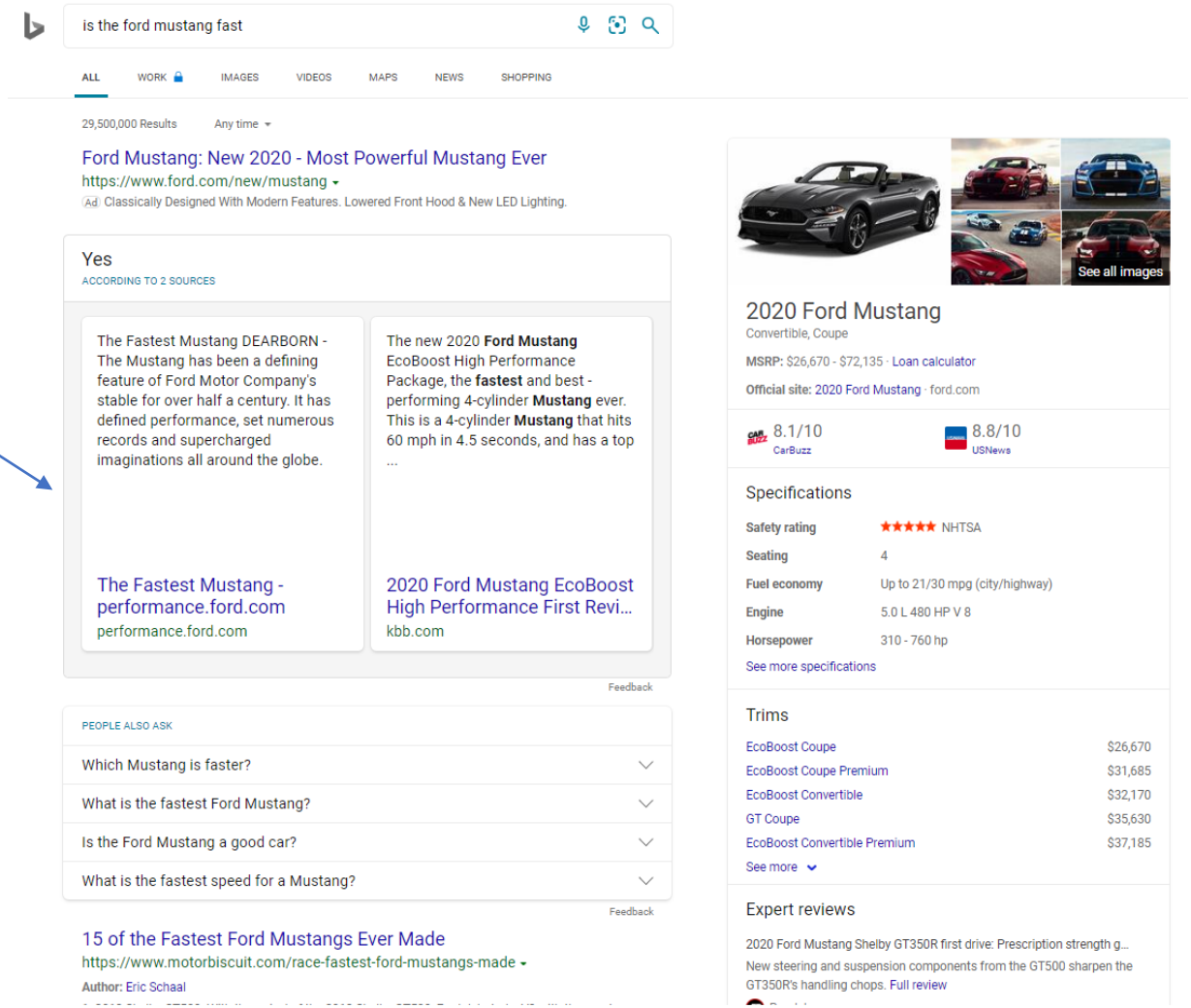


Entity pane for understanding related attributes

Evolution of Search Engine Result Page

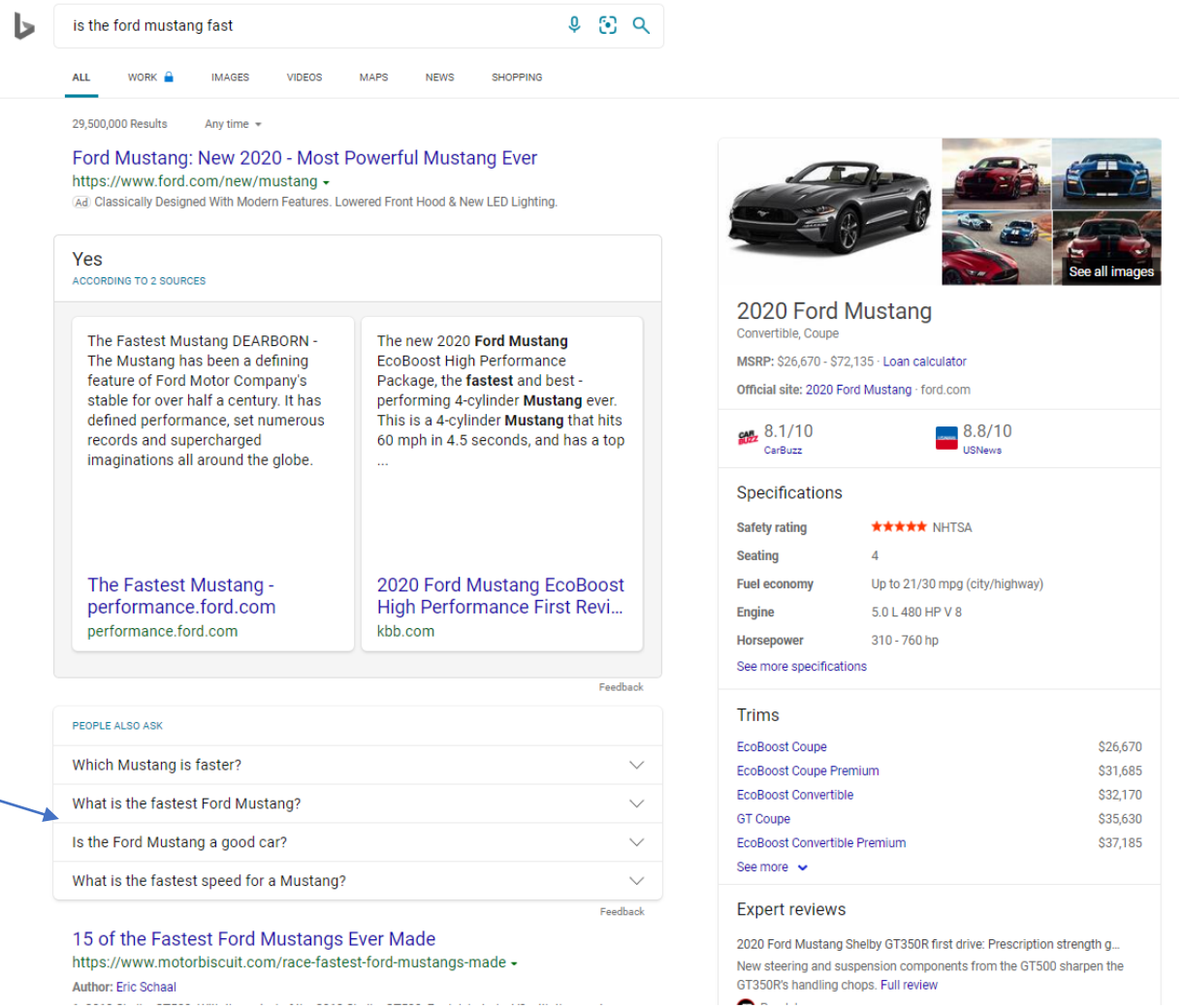
is the ford mustang fast

Instant answers and perspectives



Evolution of Search Engine Result Page

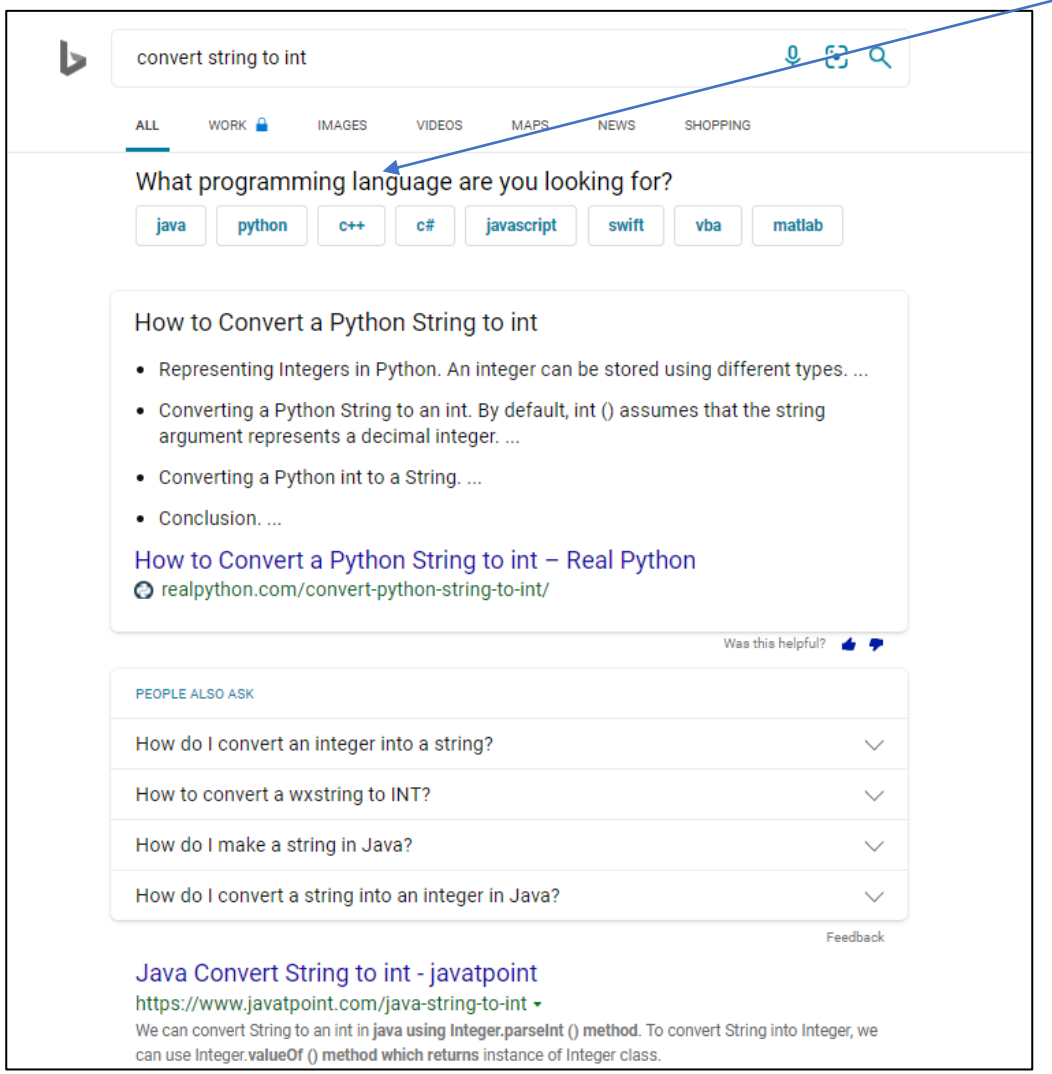
is the ford mustang fast



Useful follow-up questions once this question is answered

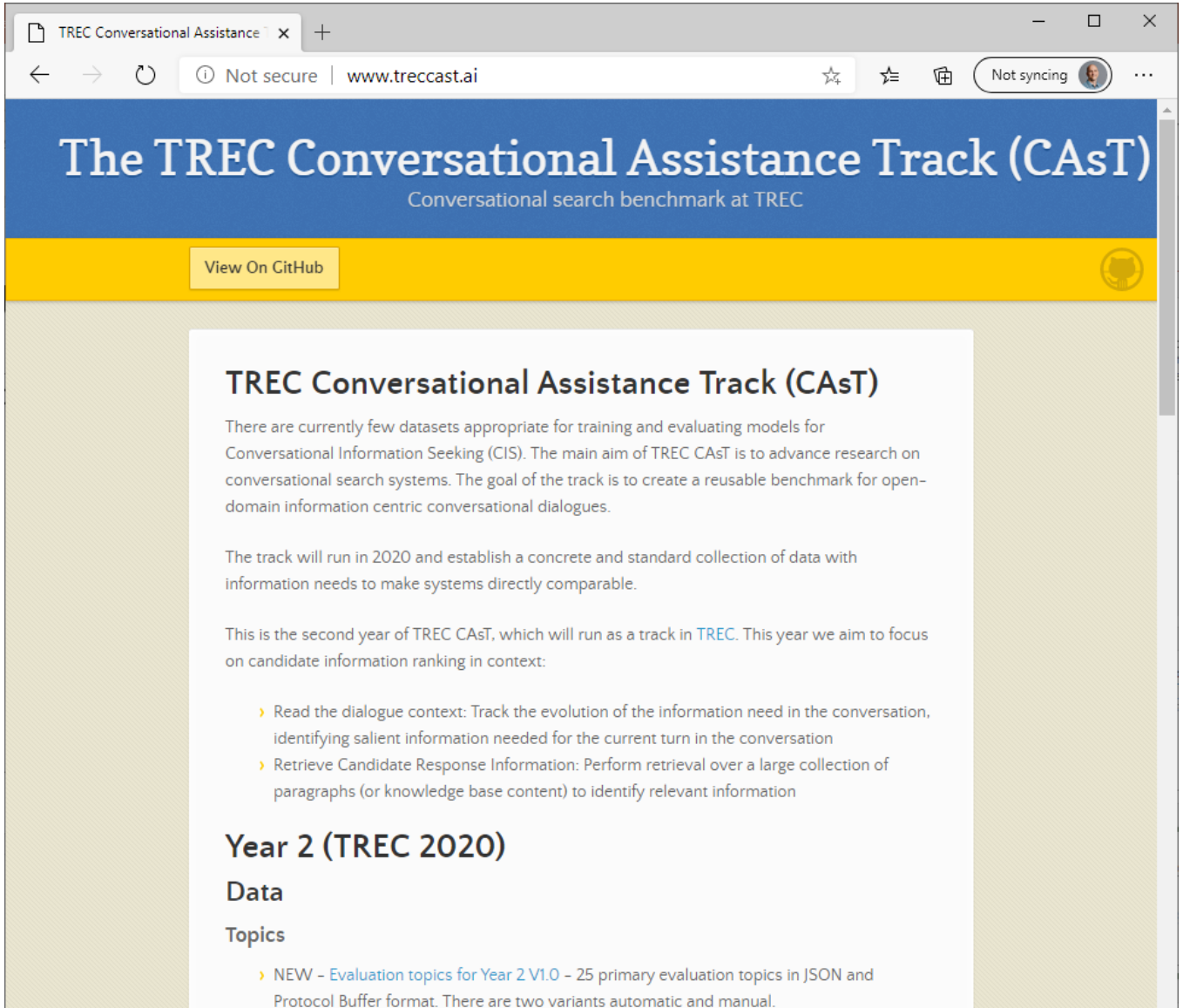
Clarification Questions

Demonstrate understanding while clarifying



[Zamani et al, WebConf 2020; SIGIR 2020]

Contextual Understanding



Sample TREC CAST 2019 Topic

Title: US Judicial history

Description: Judicial history in the US including key court cases and what they established.

What are the most important US Supreme Court cases?

What did plessy v. ferguson establish?

How about marbury vs madison?

Was it unanimous?

What was the implication of roe vs wade?

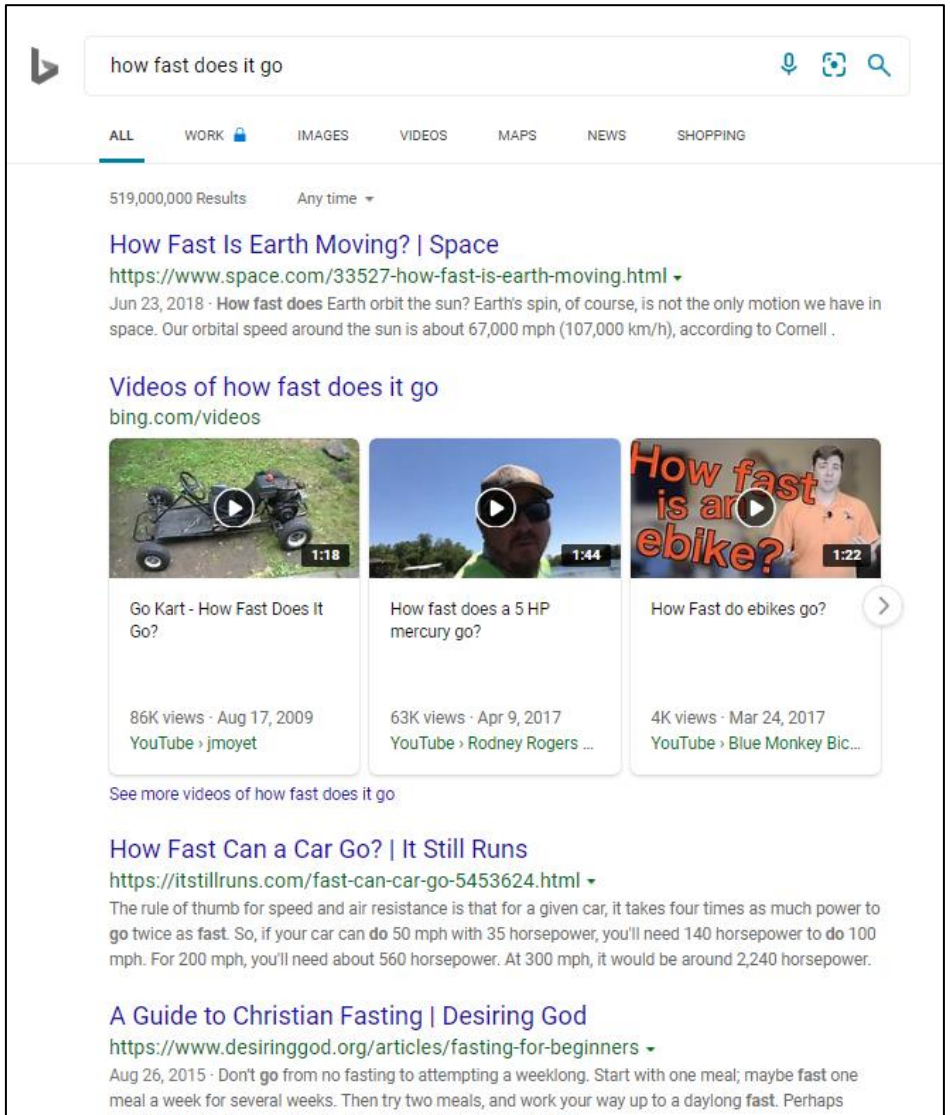
What were the main arguments?

What was the point of the brown v board of education?

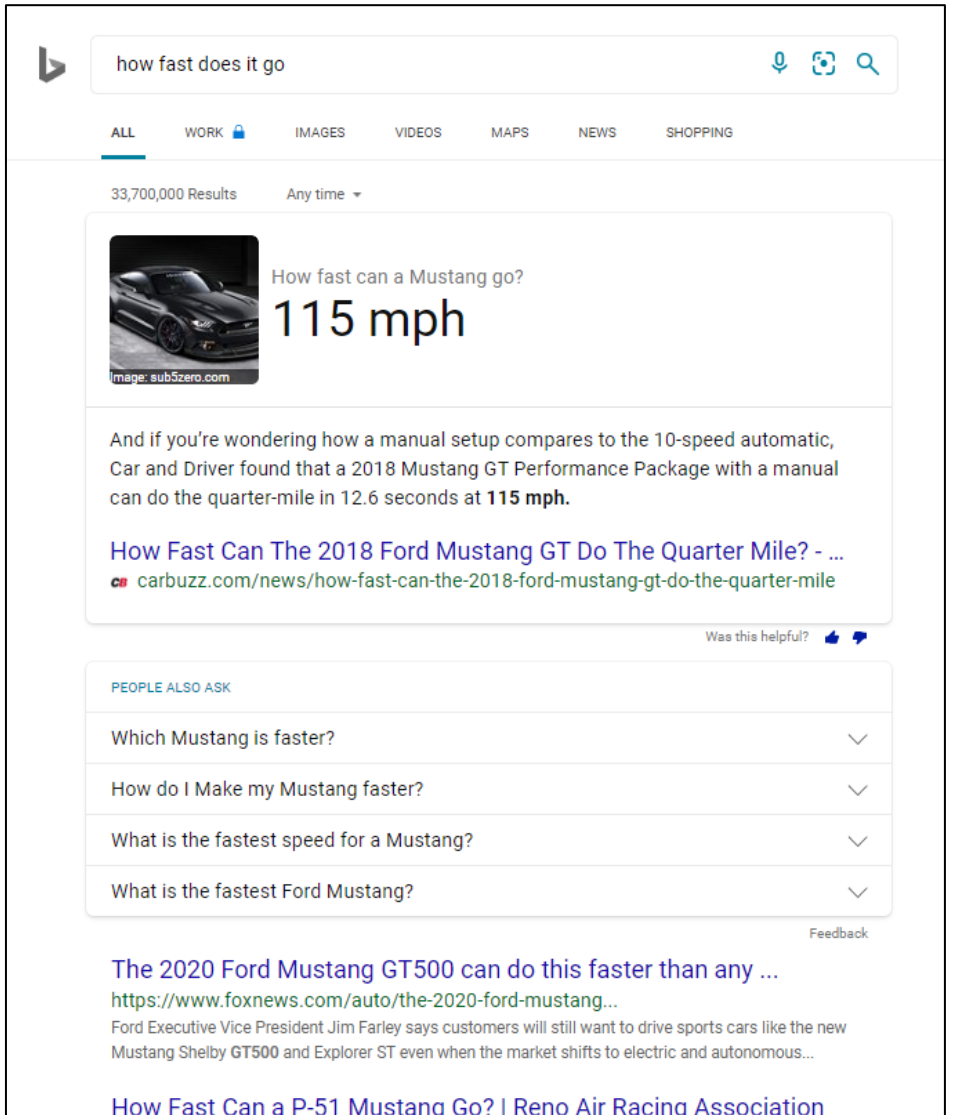
What were the main arguments?

Why is it important today?

Contextual Understanding in Search



Variety of Attempts … the future?



Productivity and Personal-Information Conversational Search

DARPA Personal Assistants that Learn (PAL)
CALO / RADAR

Key Focus Points

- Calendar management [Berry et al, 2003; Berry et al., 2006; Modi et al., 2004]
- Dealing with uncertain resources in scheduling [Fink et al., 2006]
- Task management [Freed et al. 2008]

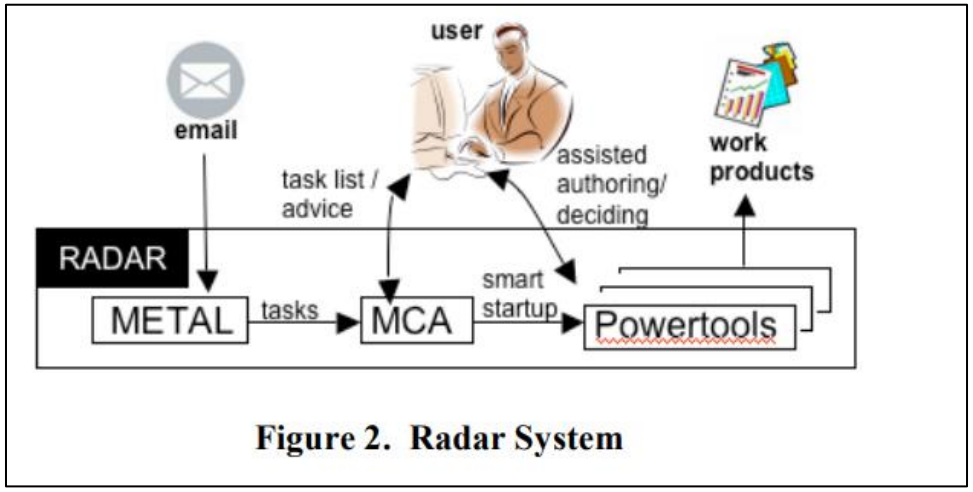


Figure 2. Radar System

From Freed et al. 2008

From PAL to SIRI

- Learnings from the PAL project including CALO/SIRI recognized need for unifying architectures. [Guzzoni et al., 2007]

A “do engine” rather than a “search engine”

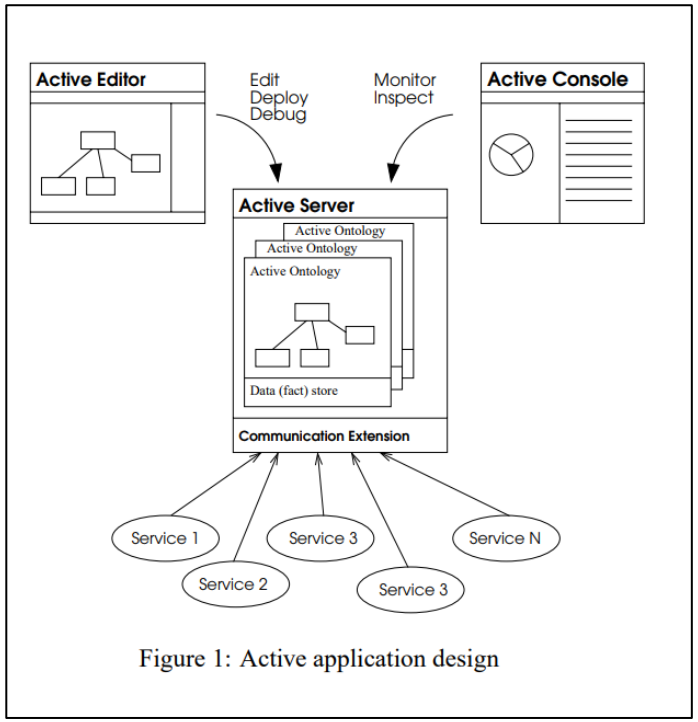


Figure 1: Active application design

From Guzzoni et al, 2007

Device-based Assistants

- Mobile phone based assistants
  - Includes: Apple's Siri, Google Assistant, Microsoft's Cortana
  - Blends productivity-focused and information focused with voice-related recognition

- Situated speakers and Devices
  - Amazon Alexa, Google Home, Facebook Portal w/Alexa, etc.
  - Combines microphone arrays, multi-modal, multi-party devices in addition

Hybrid Intelligence

- Mix AI and Human Computation to achieve an intelligent experience that leverages best of both worlds and push the envelope of possible.

- When escalated to human, often serves as a feedback loop for learning.

- Examples:
  - Facebook’s M
  - Microsoft’s Calendar.help

Calendar.help → Scheduler

"I think we can drop Greg to BCC; adding Cal from my side. Cal, can you work with Kaitlin to find 60 minutes for Todd and I sometime next week, preferably later in the week when I'm back on EST?"

- Initially high-precision rules
- Unhandled cases handled by low latency human crowdsourcing workflows
- Transition flywheel to machine learning

https://calendar.help

[Cranshaw et al., 2017]

Current application-oriented research questions

- Long-term evaluation metrics for engagement beyond CPS and NAU
  (cf. Lowe et al. [2017]; Serban et al. [2017]; Sai et al. [2019])
  - Other metrics of social companionship: linguistic accommodation or coordination?
  - Application to detection: Relationship to the inverse problems of toxicity, bias, etc.
- Aspirational goal-support from assistants
- Best proactivity engagement based on model of interests
- Integrating an understanding of physical environment

Challenges for Conversational Interaction

- Human-AI Interaction Design
  - **Goal-directed design**: Enable people to express goals flexibly and allow the agent to progress toward those goals.
  - **Gulf of evaluation**: Communicate the range of skills of an intelligent agent to users and what is available in current context.

- Conversational Understanding
  - **Grounded Language Generation and Learning**: Transform NL intent to action that depends on state and factual correctness.
  - **Extensible Personalized Skills**: Support new skills and remember preferences to evaluate changes/updates.

- External World Perception and Resource Awareness
  - **Multi-modality input and reasoning**: Integrate observations from modalities including voice, vision, and text.
  - **Identity and interactions**: Identify people around and interact with them appropriate to setting.
  - **Physical understanding**: Monitor physical situation and intelligently notify for key situations (safety, anomalies, interest).
  - **Constrained scheduling**: Support reasoning about limited and bound resources such as space/time constraints, keep knowledge of constraints to deal with updates, etc.

Challenges for Conversational Interaction

- Principles & Guarantees
  - **Responsible AI**: Evolve best practice and design new techniques as new ethical challenges arise.
  - **Privacy**: Reason about data in a privacy aware way (e.g. who is in room and what is sensitive).

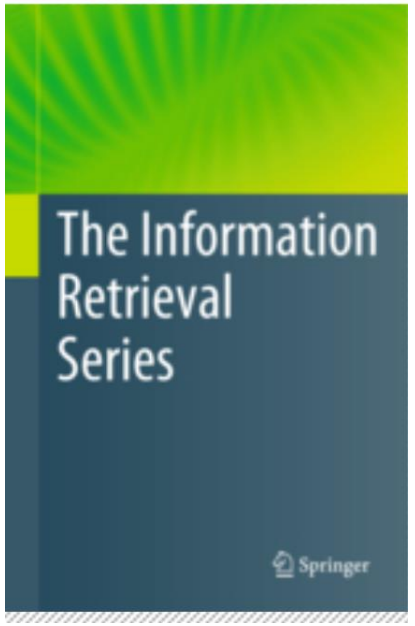
- Richer paradigms of supervision and learning
  - **Programming by Demonstration/Synthesis**: Turn sequences of actions into higher level macros/scripts that map to NL.
  - **Machine Teaching**: Support efficient supervision schemes from a user-facing perspective that also enable resharing with others (especially for previous bullet).

- Advanced Reasoning
  - **Attention**: Suspend and resume conversation/task naturally based on listener's attention.
  - **Emotional Intelligence**: Support the emotional and social needs of people to enable responsible AI and multi-party social awareness.
  - **Causal Reasoning**: Reason about the impact of taking an action.

Upcoming Book (by early 2021)

Neural Approaches to Conversational Information Retrieval

(The Information Retrieval Series)



Contact Information:

Jianfeng Gao https://www.microsoft.com/en-us/research/people/jfgao/

Chenyan Xiong https://www.microsoft.com/en-us/research/people/cxiong/

Paul Bennett https://www.microsoft.com/en-us/research/people/pauben/

Slides:

Please check our personal websites.