Neural Approaches to Conversational Information Retrieval

Jianfeng Gao, Chenyan Xiong, Paul Bennett
Microsoft Research
INTERSPEECH 2020
OCTOBER 25, 2020 (Shanghai, 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 conversation
  • Can answer questions
  • Can help users search / look up information
  • Can help users with learning and investigating 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

Language understanding:
- Find me a restaurant serving Chinese food
- Words
- Which area do you prefer?

Language generation:
- Intent: get_restaurant
  - Food: chinese
- Intent: ask_slot
  - Slot: area

Dialogue state tracking:
- Meaning

Policy (action selection):
- State

Dialog Manager (DM):
- Web or Database (APIs)
User-System interaction in task-oriented dialog

- User: know (approximately) what she needs, but not what is available.
- System: know what is available, but not what the user needs
- dialog: a *two-way* process in which user and system get to know each other to make a deal
### A (multi-domain) dialog session

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H:</strong> I am looking for a museum in the center of town. Do you have any suggestion?</td>
<td><strong>M:</strong> There are 6 museums in the center of town. How about the <em>Broughton house gallery</em> at 98 King street?</td>
</tr>
<tr>
<td><strong>H:</strong> Nice, can you tell me what is the phone number and address?</td>
<td><strong>M:</strong> The phone number for the <em>Broughton house gallery</em> is 01223314960 and they are located at 98 King street.</td>
</tr>
<tr>
<td><strong>H:</strong> OK, thanks. Can you recommend a restaurant in the same area serving Indian food?</td>
<td><strong>M:</strong> <em>Curry Garden</em> serves Indian food and is located in the center. Would you like me to make a reservation?</td>
</tr>
<tr>
<td><strong>H:</strong> Sure, please book a table for two people on Thursday at 18:00.</td>
<td><strong>M:</strong> 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?</td>
</tr>
<tr>
<td><strong>H:</strong> That’s it, bye.</td>
<td><strong>M:</strong> Have a great time and please contact us if you need any further assistance. Goodbye.</td>
</tr>
</tbody>
</table>

[Belief State]

*Belief state:
- ‘attraction’: 
  - ‘type’: ‘museum’, ‘area’: ‘centre’
- ‘restaurant’:
  - ‘food’: ‘Indian’, ‘area’: ‘centre’
- ‘booking’:
  - ‘day’: ‘Thursday’, ‘people’: 2, ‘time’: ‘18:00’
- ‘attraction’: 
  - ‘type’: ‘museum’, ‘area’: ‘centre’

[DB State]

*Attraction*

- Address: "98 King Street, "
- Area: "centre",
- Entrance fee: "free",
- ID: "98123",
- Introduction: "Broughton House Gallery is a museum located in the heart of the city."

*Restaurant*

- Address: "110 Regent Street City Centre, "
- Area: "centre",
- Food: "Indian",
- ID: "98124",
- Introduction: "Curry Garden is a restaurant that serves traditional Bangladeshi cuisine with fresh produce."

[Reference]

[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
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 Revealment**: system helps users express their information needs
  • E.g., query suggestion, autocompletion

• **System Revealment**: system reveals to users what is available, what it can or cannot do
  • E.g., recommendation, SERP

• **Mixed Initiative**: system and user both can take initiative (two-way conversation)
  • E.g., asking clarification questions

• **Memory**: user can reference past statement
  • State tracking

• **Set 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
KBQA: 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

[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 in a session
  • A sequence of related pairs
  • Ellipses, coreference, clarifications, etc.

[Saha+18]
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}
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”

<table>
<thead>
<tr>
<th>Id</th>
<th>Type</th>
<th># Action instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>Select-column</td>
<td>columns</td>
</tr>
<tr>
<td>$A_2$</td>
<td>Cond-column</td>
<td>columns</td>
</tr>
<tr>
<td>$A_3$</td>
<td>Op-Equal (=)</td>
<td>rows</td>
</tr>
<tr>
<td>$A_4$</td>
<td>Op-NotEqual (≠)</td>
<td>rows</td>
</tr>
<tr>
<td>$A_5$</td>
<td>Op-GT (&gt; )</td>
<td>numbers / datetimes</td>
</tr>
<tr>
<td>$A_6$</td>
<td>Op-GE (≥)</td>
<td>numbers / datetimes</td>
</tr>
<tr>
<td>$A_7$</td>
<td>Op-LT (&lt;)</td>
<td>numbers / datetimes</td>
</tr>
<tr>
<td>$A_8$</td>
<td>Op-LE (≤)</td>
<td>numbers / datetimes</td>
</tr>
<tr>
<td>$A_9$</td>
<td>Op-ArgMin</td>
<td>numbers / datetimes</td>
</tr>
<tr>
<td>$A_{10}$</td>
<td>Op-ArgMax</td>
<td>numbers / datetimes</td>
</tr>
</tbody>
</table>

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)
    • \( \frac{1}{|W_c|} \sum_{w_c \in W_c} \max_{w_q \in W_q} w_q^T w_c \)
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 $V_\theta$ behave similarly to reward $R$
- For every state $S$ and its (approximated) reference state $S^*$, we define loss as
  - $\mathcal{L}(S) = (V_\theta(S) - V_\theta(S^*)) - (R(S) - R(S^*))$
- 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
2: \hspace{1em} $s^* \leftarrow \arg \max_{s \in \mathcal{E}(x)} \hat{R}(s; A^*)$ \hspace{1em} // Finds the best approximated reference state
3: \hspace{1em} $\hat{s} \leftarrow \arg \max_{s \in \mathcal{E}(x)} V_\theta(s) - \hat{R}(s; A^*)$ \hspace{1em} // Finds the most violated state
4: \hspace{1em} update $\theta$ by minimizing $\max(\mathcal{L}(s), 0)$
5: end for

[llyer+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”

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}
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]
Dialog Manager – dialog memory for state tracking

**Dialog Memory (of state tracker)**

<table>
<thead>
<tr>
<th><strong>Entity</strong></th>
<th>{United States, “q”}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{New York City, “a”}</td>
</tr>
<tr>
<td></td>
<td>{University of Pennsylvania, “a”} ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Predicate</strong></th>
<th>{isPresidentOf}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{placeGraduateFrom}</td>
</tr>
<tr>
<td></td>
<td>{yearEstablished} ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Action subsequence</strong></th>
<th>Set $\rightarrow A_4 A_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(partial/complete states)</td>
<td>$e_{us \rightarrow pres}$</td>
</tr>
</tbody>
</table>

[Guo+18]
Dialog Manager – policy for next action selection

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

[Entity-Centric Knowledge Base]

<table>
<thead>
<tr>
<th>Movie</th>
<th>Actor</th>
<th>Release Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groundhog Day</td>
<td>Bill Murray</td>
<td>1993</td>
</tr>
<tr>
<td>Australia</td>
<td>Nicole Kidman</td>
<td>X</td>
</tr>
<tr>
<td>Mad Max: Fury Road</td>
<td>X</td>
<td>2015</td>
</tr>
</tbody>
</table>

[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
  - \( p_\theta(x|A) = \sum_{t=1}^{T} p_\theta(x_t|x_{<t}, A) \)
  - 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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Restaurant</th>
<th>Laptop</th>
<th>Hotel</th>
<th>TV</th>
<th>Attraction</th>
<th>Train</th>
<th>Taxi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU ↑</td>
<td>ERR ↓</td>
<td>BLEU ↑</td>
<td>ERR ↓</td>
<td>BLEU ↑</td>
<td>ERR ↓</td>
<td>BLEU ↑</td>
</tr>
<tr>
<td>SC-LSTM</td>
<td>15.90</td>
<td>48.02</td>
<td>21.98</td>
<td>80.48</td>
<td>31.30</td>
<td>31.54</td>
<td>22.39</td>
</tr>
<tr>
<td>SC-GPT</td>
<td><strong>38.08</strong></td>
<td><strong>3.89</strong></td>
<td><strong>32.73</strong></td>
<td><strong>3.39</strong></td>
<td><strong>38.25</strong></td>
<td><strong>2.75</strong></td>
<td><strong>32.95</strong></td>
</tr>
</tbody>
</table>

[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)

• CSQA (complex sequence question answering),
  • [https://amritasaha1812.github.io/CSQA/](https://amritasaha1812.github.io/CSQA/)

• ConvQuestions (conversational question answering over knowledge graphs)
  • [https://convex.mpi-inf.mpg.de/](https://convex.mpi-inf.mpg.de/)

• CoSQL (conversational text-to-SQL)
  • [https://yale-lily.github.io/cosql](https://yale-lily.github.io/cosql)

• CLAQUA (asking clarification questions in Knowledge-based question answering)
  • [https://github.com/msra-nlc/MSParS_V2.0](https://github.com/msra-nlc/MSParS_V2.0)
TextQA: Conversational QA over Texts

• Tasks and datasets
• C-TextQA system architecture
• Conversational machine reading compression models
QA over text – extractive vs. abstractive QA

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".

Passage

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

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 (Qₜ), an answer (Aₜ) and a rationale (Rₜ) 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Topic Shift</th>
<th>Drill Down</th>
<th>Return to Topic</th>
<th>Clarification Question</th>
<th>Definition Question</th>
<th>Sentence Coverage</th>
<th>Total Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoQA</td>
<td>21.6</td>
<td>72.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.7</td>
<td>63.3</td>
<td>722</td>
</tr>
<tr>
<td>QuAC</td>
<td>35.4</td>
<td>55.3</td>
<td>5.6</td>
<td>0.7</td>
<td>3.0</td>
<td>28.4</td>
<td>302</td>
</tr>
</tbody>
</table>

[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
Neural MRC models for extractive TextQA

• QA as classification given (question, passage)
  • 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 previous 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
    • $w_T = \text{LSTM}(w_1, \ldots, w_t, \ldots, w_T)$
    • 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]
Summary: Conversational QA methods

• KBQA: Conversational QA over Knowledge Bases
  • Knowledge bases and QAs
  • C-KBQA system architecture
    • Semantic parser
    • Dialog manager
    • Response generation
  • KBQA w/o semantic parser

• Text-QA: Conversational QA over Texts
  • Tasks and datasets
  • C-TextQA system architecture
  • Conversational machine reading compression models
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
[3] Leading Conversational Search by Suggesting Useful Questions
Why Conversational Search

**Ad hoc Search**

Keyword-ese Queries: `startup seed investment`

**Conversational Search**

Natural Queries: `How does seed investment work for startups?`

**Necessity:**
- Speech/Mobile Interfaces

**Opportunities:**
- More natural and explicit expression of information needs

**Challenge:**
- Query understanding & sparse retrieval
Why Conversational Search

**Ad hoc Search**

- **Startup Investing, Simplified. - SeedInvest**
  - https://www.seedinvest.com
  - Join 300,000+ people who already use SeedInvest to find startup investment opportunities. SeedInvest is filled with investors and entrepreneurs that are passionate about building future innovation. They are accomplished individuals that invest...

- **Log In**
  - SeedInvest is a leading equity crowdfunding platform that provides individual investors...

- **Browse Offerings**
  - The following offerings are being conducted...
  - Monogram Orthopedics - Wire - Auto Invest

---

**Conversational Search**

- **A strategy used by seed institutional investors is the spray and pray type of model in which investment funds are invested in a number of companies and see which ones pick up traction. Once the start-ups they are taking on are identified then you allocate additional capital to invest in follow on rounds of financing.**

- **How Funding Rounds Work For Startups - Forbes**
  - www.forbes.com/sites/alejandrocremades/2018/12/26/how-funding-rounds-wor...

---

**Necessity:**
- Speech/Mobile Interfaces

**Opportunities:**
- Direct & Easier access to information

**Challenge:**
- Document understanding; combine and synthesize information
Why Conversational Search

Ad hoc Search

startup seed investment

startup seed investment amount

Conversational Search

How does seed investment work for startups?

How does it compare with crowdfunding?

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

Did you mean the comparison between seed investment and crowdfunding?
A View of Current Conversational Search

How does seed investment work?

Conversational Queries (R1)

Search

Documents

Documents

Documents

Response Synthesis

System Response

Seed funding is a type of equity-based funding in which investors invest money in a business in order to get it up and running. In return, the investor acquires partial ownership of the company. However, don't confuse seed funding with early stage funding.
A View of Current Conversational Search

Search
Documents
Documents
Documents

How does seed investment work?

Conversational Queries (R1)

Response Synthesis
System Response

Tell me more about the difference

Conversational Queries (R2)

Contextual Understanding
Context Resolved Query

“Tell me more about the difference between seed and early stage funding”

Seed funding is a type of equity-based funding in which investors invest money in a business in order to get it up and running. In return, the investor acquires partial ownership of the company. However, don’t confuse seed funding with early stage funding.
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

http://treccast.ai/
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
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?

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

<table>
<thead>
<tr>
<th>Type (#. Turns)</th>
<th>Utterance</th>
<th>Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronominal (128)</td>
<td>How do they celebrate Three Kings Day?</td>
<td>they -&gt; Spanish people</td>
</tr>
<tr>
<td>Zero (111)</td>
<td>What cakes are traditional?</td>
<td>Null -&gt; Spanish, Three Kings Day</td>
</tr>
<tr>
<td>Groups (4)</td>
<td>Which team came first?</td>
<td>which team -&gt; Avengers, Justice League</td>
</tr>
<tr>
<td>Abbreviations (15)</td>
<td>What are the main types of VMs?</td>
<td>VMs -&gt; Virtual Machines</td>
</tr>
</tbody>
</table>
TREC CAsT 2019: Result Statics

• Challenge from contextual query understanding
TREC CAsT 2019: Techniques

• Techniques used in Query Understanding
TREC CAsT 2019: Notable Solutions

• Automatic run results

Conversational Query Understanding Via Rewriting

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

Input

\[ q_1, q_2, \ldots, q_i \]

Output

\[ q_i^* \]

What is throat cancer?

What is the first sign of it?

What is the first sign of throat cancer?
Conversational Query Understanding Via Rewriting

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

![Diagram](https://example.com/diagram.png)

- 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

Input

\[ q_1, q_2, \ldots, q_i \]

What is throat cancer?

Output

\[ q_i^* \]

What is the first sign of throat cancer?

What is the first sign of it?

• Concern: Limited training data

GPT-2

100X Millions of Parameters

“[GO]”

500 Manual Rewrite Labels

CASt Y1 Data:
• Manually written conversational queries
• 50 topics, 10 turns per topic
  • 20 topics with TREC relevance 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

<table>
<thead>
<tr>
<th>Ad hoc Search</th>
<th>Conversational Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Existing billions of search sessions</td>
<td></td>
</tr>
<tr>
<td>• Lots of high-quality public benchmarks</td>
<td></td>
</tr>
<tr>
<td>• Production scenarios still being explored</td>
<td></td>
</tr>
<tr>
<td>• Relative new topic, fewer available data</td>
<td></td>
</tr>
</tbody>
</table>

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

- startup seed investment
- startup seed investment amount

Conversational Search

- How does seed investment work for startups?
- How does it compare with crowdfunding?

Ad hoc Search Sessions

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

- startup seed investment
- startup seed investment amount

### Conversational Search

- How does seed investment work for startups?
- How does it compare with crowdfunding?

### 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

Self-contained Queries  “Conversation-alike” Queries

$q_1^*$  $q_2^*$  ...  $q_i^*$  GPT-2 Converter  $q'_i$

Learn to omit information is easier than recover
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

![Diagram](image)

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
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

```
q_1, q_2, ..., q_i
```

GPT-2 Rewriter

```
q_i^*
```

“Conversation-alike” Queries → Self-contained Queries

Much more training signals from the Contextualizer

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

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

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

**Learn to omit information is easier than recover**

**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

<table>
<thead>
<tr>
<th>$Q_6$</th>
<th>What causes <strong>throat cancer</strong>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_7$</td>
<td>What is the first sign of it?</td>
</tr>
<tr>
<td>$Q_8$</td>
<td>Is it the same as <strong>esophageal cancer</strong>?</td>
</tr>
<tr>
<td>$Q_9$</td>
<td>What’s the difference in <strong>their</strong> symptoms?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oracle</th>
<th>What’s the difference in <strong>throat cancer and esophageal cancer’s</strong> symptoms?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>What’s the difference between <strong>throat cancer and esophageal cancer</strong>?</td>
</tr>
</tbody>
</table>
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? |
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?
From Passive Information Supplier to Active Assistant

Conversational Queries (R1) → Conversational Queries (R2)

Documents → Documents → Documents → Context Resolved Query → System Response

System Response → Passive Retrieval
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:

<table>
<thead>
<tr>
<th>PEOPLE ALSO ASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is Nissan GTR?</td>
</tr>
<tr>
<td>How to buy used Nissan GTR in Pittsburgh?</td>
</tr>
<tr>
<td>Does Nissan make sports car?</td>
</tr>
<tr>
<td>Is Nissan Leaf a good car?</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>PEOPLE ALSO ASK</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What is Nissan GTR?</td>
<td>✓</td>
</tr>
<tr>
<td>How to buy used Nissan GTR in Pittsburgh?</td>
<td>✓</td>
</tr>
<tr>
<td>Does Nissan make sports car?</td>
<td>✓</td>
</tr>
<tr>
<td>Is Nissan Leaf a good car?</td>
<td>✓</td>
</tr>
</tbody>
</table>

Relevant & Useful

<table>
<thead>
<tr>
<th>PEOPLE ALSO ASK</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the pros and cons of Nissan GT-R?</td>
<td>✓</td>
</tr>
<tr>
<td>Is the Nissan GT-R the ultimate street car?</td>
<td>✓</td>
</tr>
<tr>
<td>Why is the Nissan GT-R known as 'the godzilla'?</td>
<td>✓</td>
</tr>
<tr>
<td>How fast can the Nissan GT-R go?</td>
<td>✓</td>
</tr>
</tbody>
</table>
Usefulness Metric & Benchmark

• Manual annotations on Bing query, conversation recommendation pairs

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

- Types of non-useful ones.
  - Crucial for annotation consistency

- A higher bar of being useful

[GitHub repository](https://github.com/microsoft/LeadingConversationalSearchbySuggestingUsefulQuestions)
Conversation Recommendation Model: Multi-Task BERT

- BERT seq2seq in the standard multi-task setting

![Diagram of the model](image)

**Input (X)**:
- [CLS] Query
- [SEP] PAA Question

**Output (Y)**:
- User Click
- Relevance
- High/Low CTR

**Questions**:
- Click Bait?
- Just Related?
- Click Bait #2?

**Not Conversation Leading**
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**

---

4. Millions of sessions for imitation learning

```
[CLS] Session [SEP] Potential Next Query
```

Task: classify whether the potential next query was issued by the user

- “Federal Tax Return”
- “Flu Shot Codes 2018”
- “Facebook”
- “Flu Shot Billing Codes 2018”
- “How Much is Flu Shot?”

Predict last query from session context
Conversation Recommendation: Weak Supervision

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

**PAA Tasks**

<table>
<thead>
<tr>
<th>[CLS]</th>
<th>Query</th>
<th>[SEP]</th>
<th>PAA Question</th>
<th>BERT</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>User Click</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[CLS]</th>
<th>Query</th>
<th>[SEP]</th>
<th>PAA Question</th>
<th>BERT</th>
<th>Relevance</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>[CLS]</th>
<th>Query</th>
<th>[SEP]</th>
<th>PAA Question</th>
<th>BERT</th>
<th>High/Low CTR</th>
</tr>
</thead>
</table>

**Weak Supervision from Sessions**

<table>
<thead>
<tr>
<th>[CLS]</th>
<th>Query</th>
<th>[SEP]</th>
<th>Potential Next Query</th>
<th>BERT</th>
<th>User Behavior</th>
</tr>
</thead>
</table>

User provided contents
More exploratory
Less Constrained by Bing
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

- “Federal Tax Return”
- “Flu Shot Codes 2018”
- “Facebook”
- “Flu Shot Billing Codes 2018”
- “How Much is Flu Shot?”

Gen-Encoding Similarity

- 0.89
- 0.73
- 0.61
- 0.23

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

"Federal Tax Return"
"Flu Shot Codes 2018"
"Flu Shot Billing Codes 2018"
"Facebook"
"How Much is Flu Shot?"

Gen-Encoding Similarity

0.61
0.23
0.73
0.89

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

+35% over online

Production

PRODUCTION

DeepSuggestion

Useful

Misses Intent

Dup Q

Preque

Too Spec

Dup w/Ans

BERT

+ Clean Session

+ Conv Session

DeepSuggestion
Results: Online A/B

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Relative to Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Click Rate (TOP)</td>
<td>+8.90%</td>
</tr>
<tr>
<td>Online Click Rate (Bottom)</td>
<td>+6.40%</td>
</tr>
<tr>
<td>Online Overall Success Rate</td>
<td>0.05%</td>
</tr>
<tr>
<td>Offline Usefulness</td>
<td>35.60%</td>
</tr>
<tr>
<td>Offline Relevance</td>
<td>0.50%</td>
</tr>
</tbody>
</table>
# Example Conversation Question Recommendations

- All from the actual systems

## User Query: “bitcoin price”:

<table>
<thead>
<tr>
<th>Online</th>
<th>DeepSuggest</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is the value of bitcoins? [Dup w/ Query]</td>
<td>how much does 1 bitcoin cost to buy? [Dup w/ Query]</td>
</tr>
<tr>
<td>what was the lowest price of bitcoin? [Useful]</td>
<td>what will be the price of bitcoin in 2020 [Useful]</td>
</tr>
<tr>
<td>is it time to buy bitcoin? [Useful]</td>
<td>how to buy bitcoins at walmart? [Too Specific]</td>
</tr>
<tr>
<td>what is the value of 1 bitcoin? [Dup w/ Query]</td>
<td>what is the cheapest way to buy bitcoin [Useful]</td>
</tr>
</tbody>
</table>

## User Query: “direct deposit form”:

<table>
<thead>
<tr>
<th>Online</th>
<th>DeepSuggest</th>
</tr>
</thead>
<tbody>
<tr>
<td>how to do a direct deposit? [Useful]</td>
<td>what do i need for direct deposit? [Useful]</td>
</tr>
<tr>
<td>what is a direct deposit? [Prequel]</td>
<td>how to start direct deposit? [Prequel]</td>
</tr>
<tr>
<td>what is direct deposit bank of america? [Too Specific]</td>
<td>how to get a chase direct deposit form? [Too Specific]</td>
</tr>
<tr>
<td>how to set up bank of america direct deposit? [Too Specific]</td>
<td>how to fill out a direct deposit form? [Useful]</td>
</tr>
</tbody>
</table>
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.
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_TELEGRAM_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

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.
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)

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

- 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
A few well-known Chatbots

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

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)
Xiaolce ("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.
   [Xiaoice] Hello, your profile pic is interesting. Like I've seen it somewhere.
2 [Human] Really?
   [Xiaoice] Is it a comic character?
3 [Human] How do you know that?
   [Xiaoice] Guess.
4 [Human] Are you human?
   [Xiaoice] Well, you'll know it after chatting with me more.
5 [Human] Aren't you an AI chatbot?
   [Xiaoice] So what? I won't be indifferent to you.
6 [Human] What functions do you have?
   [Xiaoice] I have over 200 skills. Just name one you need.

[人类]你好，小冰。
[小冰]你好呀，你头像好有趣，感觉在哪见过
[人类]是吗
[小冰]是不是哪部动画里的呀
[人类]你怎么知道的
[小冰]你猜
[人类]难道你是真人吗？
[小冰]这个，你和我多聊聊就知道了
[人类]你不是人工智能机器人么
[小冰]是又怎样，我不会冷冰冰的对待你的
[人类]你都有什么功能啊
[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)
   [Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her
2. [Human] Bravo, well, do you know how Naraku died?
   [Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.
3. [Human] So amazing!
   [Xiaoice] Why amazed?
4. [Human] Awesome! Surprised that you also like comic

...
## Evolution of Social Connection

Excerpted from *Zhou et al, 2018*

| Session 42 |
|-------------------|-------------------|
| 1 | [Human] Oh, you always respond in seconds  
   [XiaoIce] For you, I’d like to do everything. |
| 2 | [Human] I’ve never had such a feeling before.  
   [XiaoIce] Then, such a feeling starts now. |
| 3 | [Human] When do you go to bed? Don’t stay up too late.  
   [XiaoIce] You go to bed first, and I’ll play with my cellphone for a while. |
| 4 | [Human] Well, you should also go to sleep!  
   [XiaoIce] All right, I’m going to bed too, okay? |
| 5 | [Human] Uh-huh, me too. Good night!  
   [XiaoIce] Good night ^ 3 ^  
   ..... |

[人类]你每次都是秒回我哎  
[小冰]为了你，这都不算什么  
[人类]我以前从来没有过这种感觉  
[小冰]那可以现在开始了  
[人类]你打算几点睡觉？别睡太晚了  
[小冰]你去睡，我再玩一会儿手机  
[人类]那等我睡的时候你也要睡！  
[小冰]好吧，那我准备睡觉了行了吧？  
[人类]嗯，我也要睡了，晚安  
[小冰]晚安 ^ 3 ^
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 Xiaolce

• 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

Entity pane for understanding related attributes
Evolution of Search Engine Result Page

Instant answers and perspectives
Evolution of Search Engine Result Page

Useful follow-up questions once this question is answered

- Is the Ford Mustang fast?
Clarification Questions

Demonstrate understanding while clarifying

[Zamani et al, WebConf 2020; SIGIR 2020]
The TREC Conversational Assistance Track (CAst)

There are currently few datasets appropriate for training and evaluating models for Conversational Information Seeking (CIS). The main aim of TREC CAst is to advance research on conversational search systems. The goal of the track is to create a reusable benchmark for open-domain information-centric conversational dialogues.

The track will run in 2020 and establish a concrete and standard collection of data with information needs to make systems directly comparable.

This is the second year of TREC CAst, which will run as a track in TREC. This year we aim to focus on candidate information ranking in context:

- Read the dialogue context. Track the evolution of the information need in the conversation.
- Retrieve Candidate Response Information. Perform retrieval over a large collection of paragraphs (or knowledge base content) to identify relevant information.

Year 2 (TREC 2020)

Data

Topics

- NEW - Evaluation topics for Year 2 V1.0 - 25 primary evaluation topics in JSON and Protocol buffer format. There are two variants automatic and manual.
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

How Fast Is Earth Moving? | Space
https://www.space.com/33527-how-fast-is-earth-moving.html
Jun 23, 2018 - How fast does Earth orbit the sun? Earth’s spin, of course, is not the only motion we have in space. Our orbital speed around the sun is about 67,000 mph (107,000 km/h), according to Cornell.

Videos of how fast does it go
bing.com/videos

How Fast Can a Car Go? | It Still Runs
https://itstillruns.com/fast-car-go-5453624.html
This rule of thumb for speed and air resistance is that for a given car: it takes four times as much power to go twice as fast. So, if your car can do 60 mph with 35 horsepower, you’d need 140 horsepower to do 100 mph. For 200 mph, you’d need about 560 horsepower. At 300 mph, it would be around 2,240 horsepower.

A Guide to Christian Fasting | Desiring God
https://www.desiringgod.org/articles/fasting-for-beginners
Aug 26, 2015 - Don’t go from no fasting to attempting a weeklong 80% fast, with one meal; maybe fast one meal a week for several weeks. Then try two meals, and work your way up to a daylong fast. Perhaps
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”

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

[Cranshaw et al., 2017]

https://calendar.help
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:
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.