Conversational Passage Retrieval

...

Antonios Minas Krasakis

October 04, 2021

(SIKS course in Advances in Information Retrieval)

University of Amsterdam

Intro - Who am I?

3rd year PhD student @ UvA (IRLab)

Topic: Conversational Search

Previously:

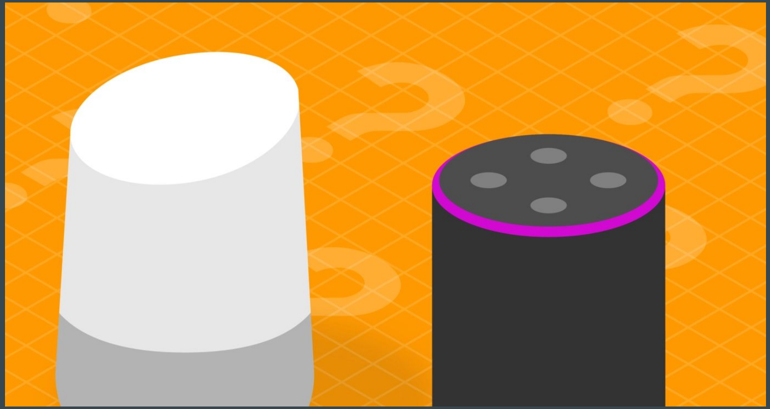
- Electrical & Computer Engineering (AuTh)
- MSc Data Science (UvA)
- Data Science @ bol.com

Conversational Search & Conv. Passage retrieval

Conversational Search

Many forms of conversations & Conv. Search:

- Conversational Question Answering [1]
- Conversations with Documents (Document-centered assistance) [2]
- Conversational Product Search [6]
- Conversational Passage Retrieval



Conversational Passage Retrieval

➔ **Given a conversation** (ie. user-system) **rank documents**/passages

Characteristics:

- Information-seeking
- Open domain
- Focus on **ranking** documents/passages
  (rather than ie. Question Answering)

Conversational Passage Retrieval

Different types (scenarios) of conversations:

- Clarifications

- Conversational Information-Seeking (CIS)

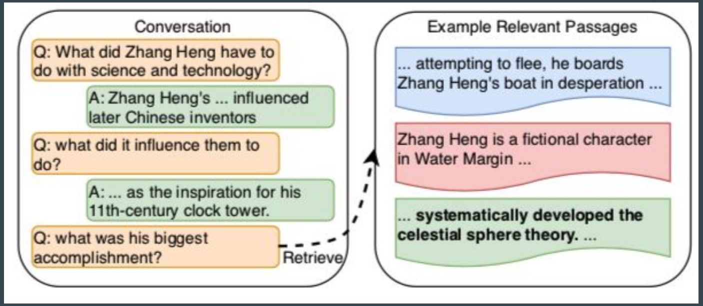
- More...

Clarification-based conversation

Conversational Passage Retrieval

Different types (scenarios) of conversations:

- Clarifications
- Conversational Information-Seeking (CIS)
- More...



Information-seeking conversation

Conversational Passage Retrieval

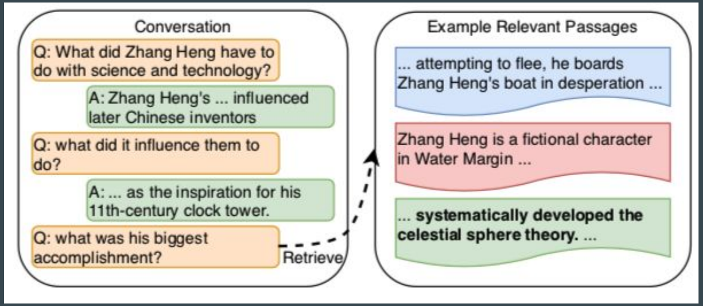
Different types (scenarios) of conversations:

- Clarifications
- Conversational Information-Seeking (CIS)
- More...

Ranking with Information Seeking Conversations

Wizard setting:

1. User continuously asks (different) questions to a system
2. System responds with a passage



CIS task

Differences: CIS vs. ConvQA

	CIS	**Conv QA**
task	passage retrieval	answer selection / generation
models	retriever	(retriever) + **reader**
	open-domain	open- or closed-domain
evaluation	ranking metrics (NDCG, MRR, P, ...)	QA metrics (F1, EM, ...)

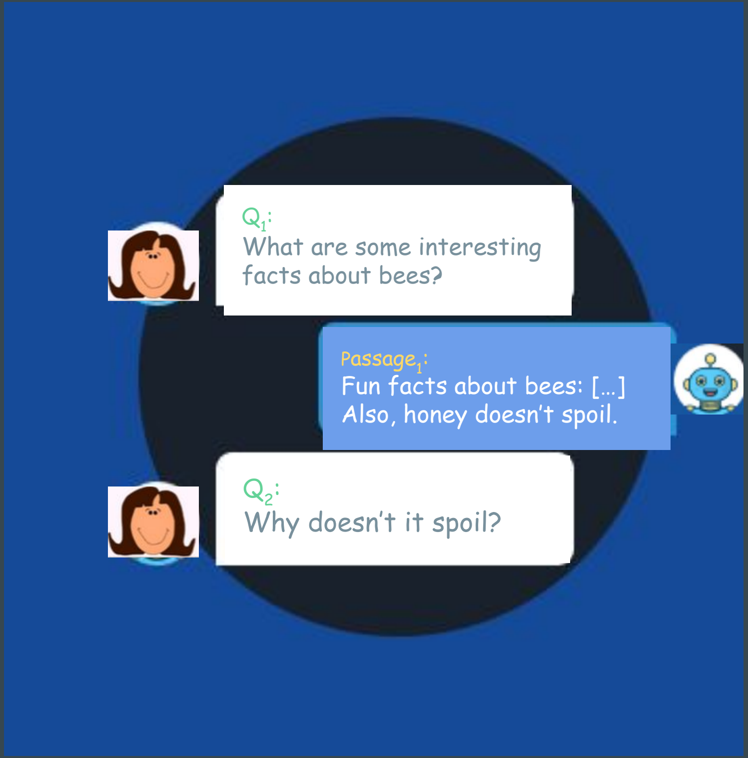
Comparison: CIS vs. Conv QA

Competitions/Datasets for CIS

1. Trec Conversational Assistant Track (CAsT) [2019-2021] [7]

	2019	2020/21
Next question depends on:	- Previous user questions (Q_1 , Q_2, … , Q_{X-1})	- Previous user questions (Q_1 , Q_2, … , Q_{X-1}) - Previous answer passages (P_1 , P_2, … , P_{X-1})

Task differences (per year)



CAsT 2020 example

Competitions/Datasets for CIS

2. OrQuaC

- **synthetic** data from QA pairs

- All relevant passages of a conversation come from 1 wikipedia article (dataset bias?)

Pipeline methods for CIS

- CAsT-19
  - QuReTeC
  - Few-shot Query Rewriting
  - An overview

- CAsT-20
  - h2oloo submission

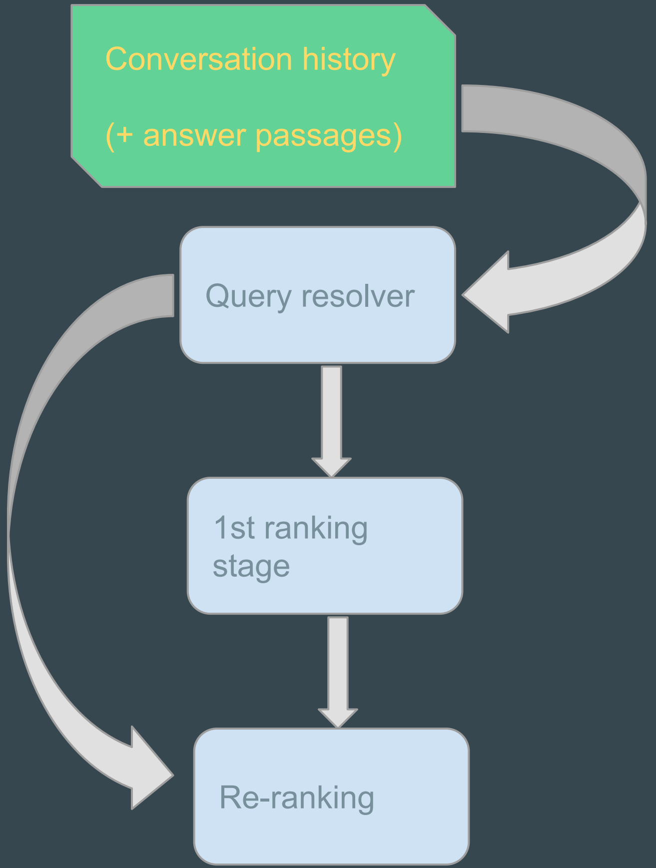
Pipeline methods

Focus on **query resolution/rewriting:**

- **Make the query** (last user utterance) → **self-contained**

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

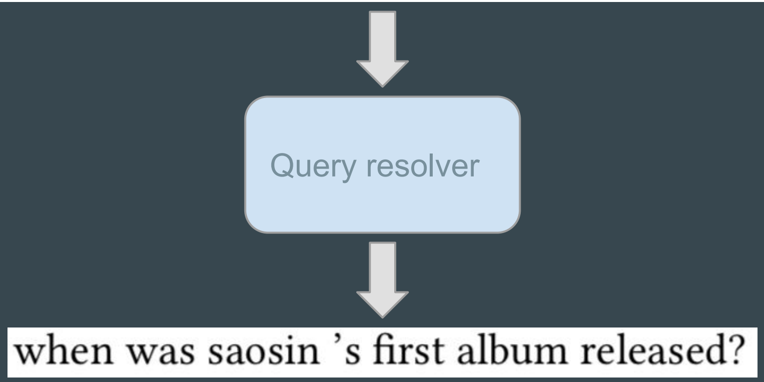
Cast-19 example



Conv. Passage Retrieval Pipeline (inference)

Query resolution / rewriting

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



Query resolution/rewriting

Inference pipeline (ranking)

Methods for query resolution:

I. Lexical methods (text)
   A. Query Resolution by Term Classification (QuReTeC) [8]
   B. Few-shot generative conversational query rewriting [9]
   C. h2oloo system (T5-based) [12]

II. Dense methods (embedding space)
   A. Few-shot conversational query rewriting [9]

Query Resolution by Term Classification (QuReTeC)

QuReTeC [8] approach

Key idea:

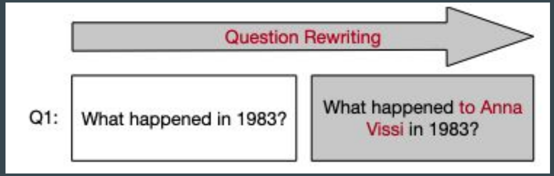
Query resolution as Term Classification
(of relevant history terms)

turn #	Question
$turn_1$	who formed **saosin**?
$turn_2$	when was the **band** founded?
$turn_3$	what was their **first** album?
$turn_4$	when was the album released?
resolution	when was the album released? **first saosin band**

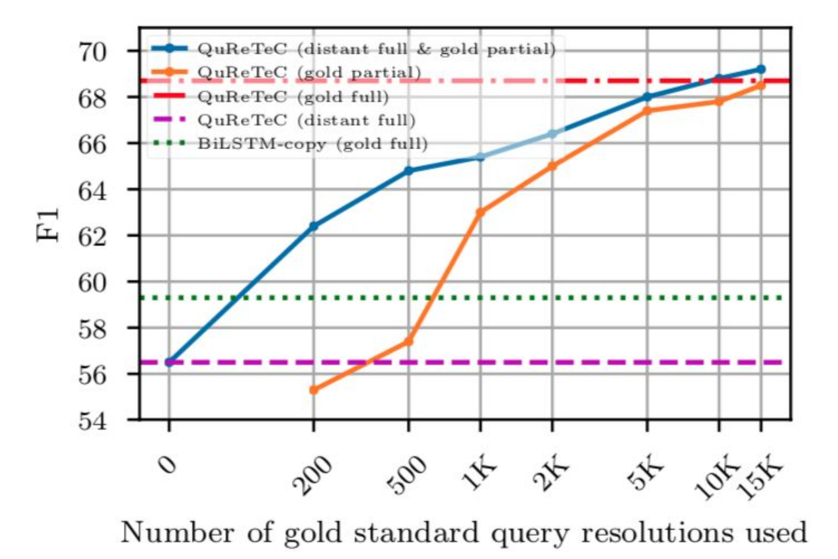
Query resolution by QuReTeC

Query Resolution by Term Classification (QuReTeC) [8]

training approaches:

Supervision	Labels	
Full	human resolutions (CAsT & CANARD)	
Weak	Distant supervision (from relevant passages)	Relevant terms are found in a relevant document

Query Resolution by Term Classification (QuReTeC) [8]



performance wrt. # of training examples

- ➔ **Good performance** with combination of

  human + distant supervision labels

- ➔ **Few human annotations needed**

Few-shot generative conversational query rewriting

Few-shot generative conversational query rewriting [9]

Key idea:

Try to address resolution by solving correferences & omissions

- ➢ Full rewriting (instead of adding salient terms)
- ➢ model: GPT-2

turn #	Question
$turn_1$	What causes throat cancer?
$turn_2$	What is the first sign of it?
$turn_3$	Is it the same as esophageal cancer?
$turn_4$	What's the difference in their symptoms?
resolution	What's the difference ***between*** throat cancer and esophageal cancer symptoms?

Query resolution by FSGCQR [9]

FS G CQR [9]: Training

Construct synthetic rewrite data

- ➔ Given a search session log:

  $\{query_1, query_2, ..., query_k\}$

- ➔ Use "QuerySimplifier" to simulate conversation (corrupt last query):

  $\{query_1, query_2, ..., query_k\} \rightarrow query_k^*$

- ➔ Teach rewriter to **reconstruct self-contained query**

  $\{query_1, query_2, ..., query_k^*\} \rightarrow query_k$

turn #	Session log
$turn_1$	who formed saosin?
$turn_2$	when was **the saosin band** founded?
$turn_2^*$	when was **their** band founded?

QuerySimplifier ($turn_2 \longrightarrow turn_2^*$)

FS G CQR [9]: "Conversational" Query corruption

QuerySimplifiers used:

A. **Rule-based**, from discourse phainomena:
   a. Correference *(he, it, …)*
   b. Omission

B. **Self-learn**
   a. few-shot GPT-2 corruptor from manual annotations

turn #	Question
$turn_1$	who formed saosin?
$turn_2$	when was the saosin band founded?
$turn_2^*$	when was their band founded?

Conversational Query simulation

FS G CQR [9]: Results

Method	BLEU-2	NDCG@3	QA-ROUGE
CAsT Queries			
Original	0.659	0.304	0.231
AllenNLP Coref w/o sw	–	0.314	–
AllenNLP Coref w/ sw	0.750	0.437	0.278
Oracle	1.000	0.544	0.314
Zero-Shot Rewriter			
GPT-2 Raw	0.112	0.124	0.196
MARCO Raw	0.380	0.172	0.183
Rule-Based	**0.755**	**0.437**	0.266
Few-Shot Rewriter			
Rule-Based + CV w/o PLM	0.178	0.065	0.151
Self-Learn	0.750	0.435	0.263
CV	0.793	0.467	0.280
Rule-Based + CV	**0.809**	**0.492**	**0.291**
Self-Learn + CV	0.804	0.491	**0.291**

Performance of zero- and few-shot rewriters [9]

- zero-shot: *(query corruption)*
  - substantial improvement over other zero-shot baselines
  - close to AllenNLP correference resolution (not zero-shot)
- few-shot *(query corruption + finetune)*
  - fine-tuning (CV) helps (+12%)
  - rule-based vs self-learn: on par
  - close to oracle (human resolutions)

Query Rewriters:
A comparison (so far)

Which resolution method works best ?

- Classification *[QuReTeC]*
- Rewriting *[FS query rewriting / others]*

Query Rewriters: A comparison

QR Method	Recall@1000	NDCG@3		ROUGE-1		
	Initial	Initial	Reranked	P	R	F
Original	0.417	0.131	0.266	0.92	0.76	0.82
Transformer++ Q	0.743	0.265	**0.525**	**0.96**	0.88	**0.91**
Self-Learn Q	0.725	0.261	0.513	0.93	0.89	0.90
Rule-Based Q	0.717	0.248	0.487	0.94	0.89	0.91
QuReTeC Q	**0.768**	**0.296**	0.500	0.89	**0.90**	0.89

CAsT-19 results (from [9])

Common retrieval & rerank pipeline: **BM25 + BERT**

- @ retrieval:
  - best rewrite: QuReTeC
- @ re-ranking:
  - best rewrite: Transformer++

What is a good resolution?

- Should it be **human-readable**?
- **human resolution =? oracle**
- Rouge-**Precision** vs. Rouge-**Recall** ?
- Depends on pipeline component
  - ie. **Retrieval vs. Reranker**
  - ie. **BERT vs. T5**

Such issues affect **robustness of pipeline systems**

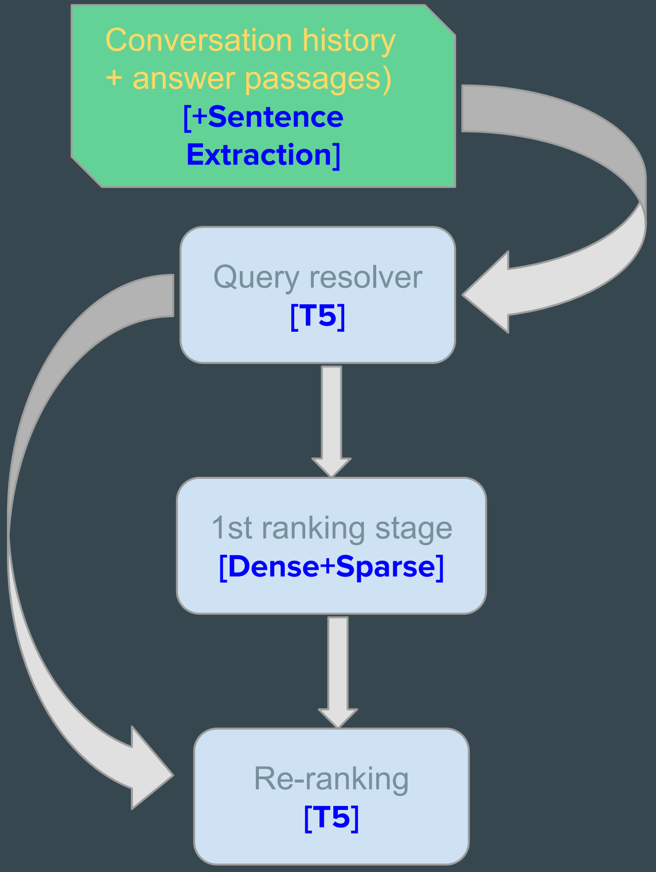
h2oloo trec-cast participation

h2oloo (Waterloo) system [12]

Key ideas:

1. use more **advanced models**
   a. **T5** *(rewriter + re-ranker)*
   b. **Hybrid Dense + Sparse** *retriever*

2. **Incorporating previous answer passages : Sentence Extraction**
   a. long input causes **performance degradation** (language models)
   b. Time **efficiency**

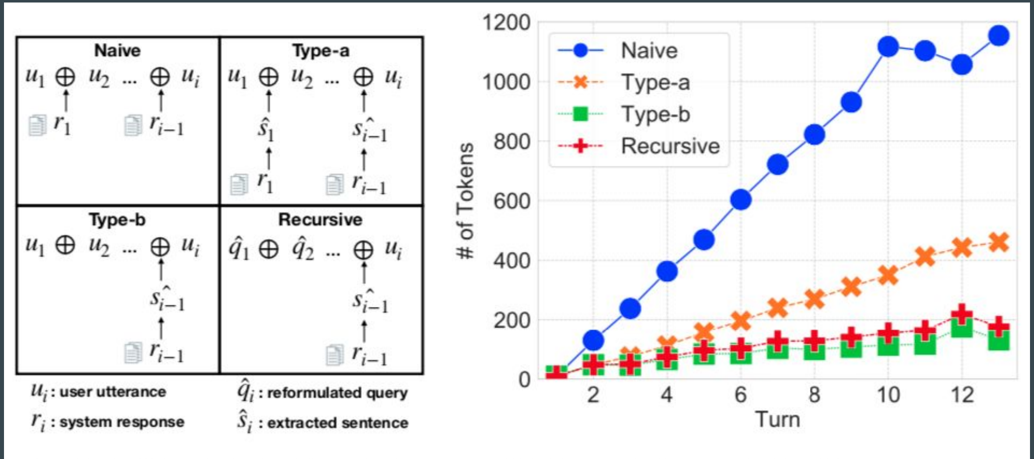


Conv. Passage Retrieval Pipeline
(inference)

h2oloo [12]: Handling previous answers

Extraction method	Canonical answer representation in query
Naive	Concatenate all previous answers
Type-a	Concatenate all extracted sentences
Type-b	Concatenate **last** extracted sentence
Recursive	Concatenate **previous query reformulations** + **last** extracted sentence

Query reformulation methods



(a) Query reformulation inference types (b) Average # of tokens input to T5 by turn depth
Figure 1: The comparison of different query reformulation methods

Query reformulation methods & lengths

h2oloo [12]: Experimental results

Query reformulation		Retrieval (dense+sparse)			Re-ranking (T5-3B)	
Model(T5)	Inference	R@1000	MAP	NDCG@3	MAP	NDCG@3
-	-	0.840	0.324	0.463	0.459	0.613
base	Query-only	0.668	0.225	0.343	0.330	0.452
base	Type-b	0.661	0.216	0.337	-	-
base	Recursive	0.684	0.220	0.328	-	-
large	Query-only	0.696	0.238	0.360	-	-
large	Type-a	0.708	0.239	0.364	-	-
large	Type-b	0.697	0.238	0.358	0.345	0.480
large	Recursive	**0.724**	**0.250**	**0.367**	**0.363**	**0.494**

Experimental results

- @ retrieval:
  - Reformulation roughly matters with T5-LARGE
- @ re-ranking:
  - T5-BASE + QUERY-ONLY: quite good already
  - [T5 BASE ⟶ T5 LARGE] + [QUERY-ONLY ⟶ TYPE-B]: +5%

h2oloo [12]: Experimental results

Query reformulation		Retrieval (dense+sparse)			Re-ranking (T5-3B)	
Model(T5)	Inference	R@1000	MAP	NDCG@3	MAP	NDCG@3
-	-	0.840	0.324	0.463	0.459	0.613
base	Query-only	0.668	0.225	0.343	0.330	0.452
base	Type-b	0.661	0.216	0.337	-	-
base	Recursive	0.684	0.220	0.328	-	-
large	Query-only	0.696	0.238	0.360	-	-
large	Type-a	0.708	0.239	0.364	-	-
large	Type-b	0.697	0.238	0.358	0.345	0.480
large	Recursive	**0.724**	**0.250**	**0.367**	**0.363**	**0.494**

Experimental results

- model **size** matters
- sentence extraction **?**

- @ retrieval:
  - Reformulation roughly matters with T5-LARGE
- @ re-ranking:
  - T5-BASE + QUERY-ONLY: quite good already
  - [T5 BASE ⟶ T5 LARGE] + [QUERY-ONLY ⟶ TYPE-B]: +5%

Neural methods for CIS

Few-shot Conversational Dense Passage Retrieval [11]

First e2e Dense Retrieval architecture for Conv Passage retrieval:

Query resolution + retrieval: directly on the dense space

- backbone ranker: ANCE

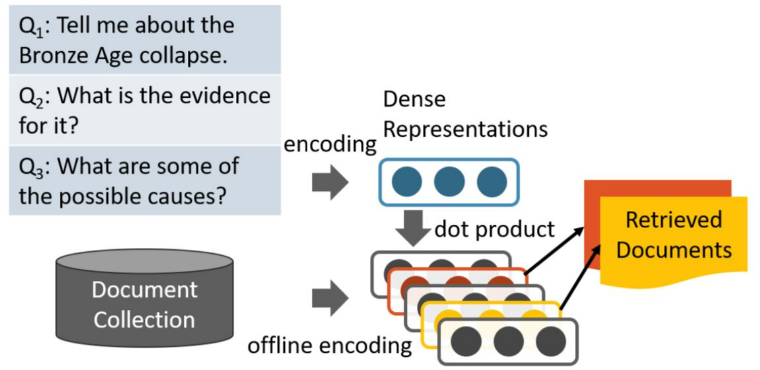


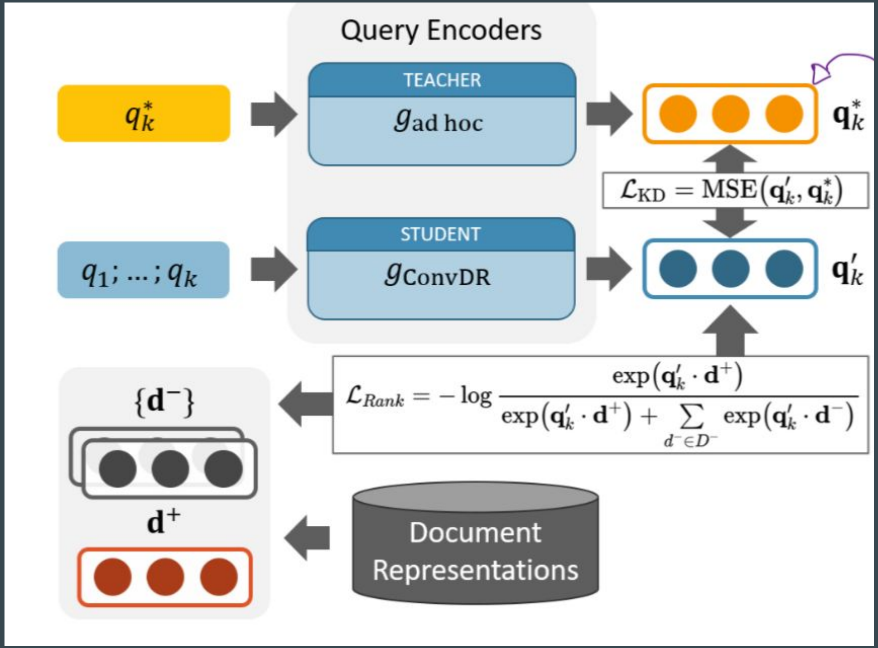
Figure 1: An illustration of Conversational Dense Retrieval. The conversational queries are contextually embedded and matched with documents via dense retrieval.

Inference (FS Conv DPR [11])

F.S. Conv.DPR [11]: Resolution in dense space

To resolve the query in the dense space:

➔ **Teacher-student network** that "pushes" the representation of the conversation to be close to the resolved query

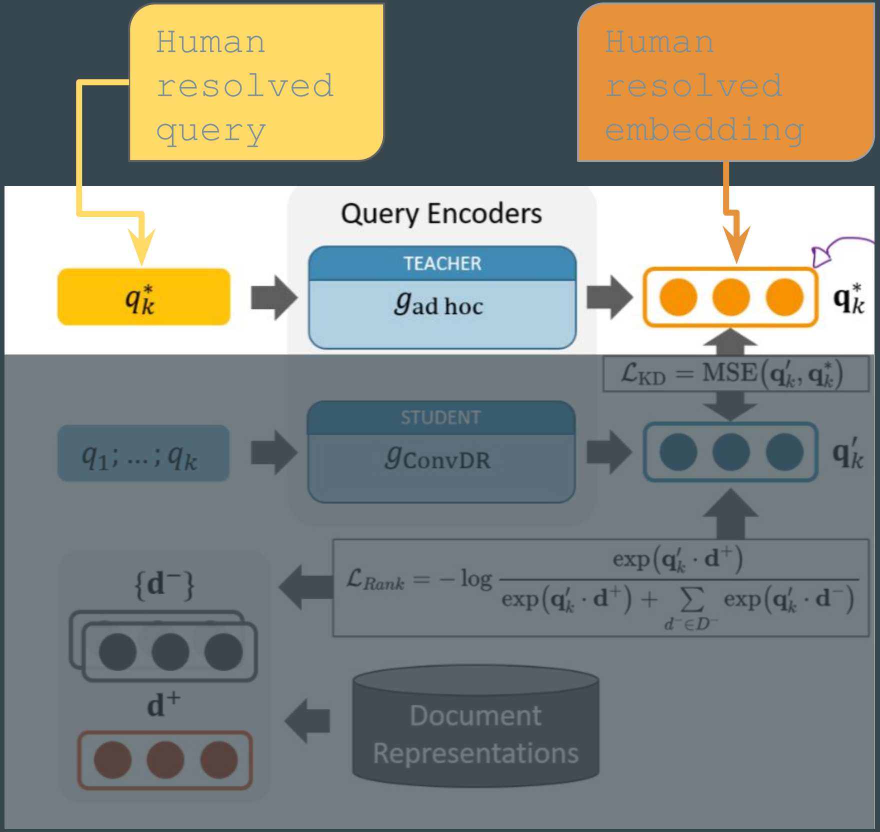


Training (FS Conv DPR [11])

F.S. Conv.DPR [11] : Training

➔ **Teacher:** produces "ideal" embedding (based on resolution)

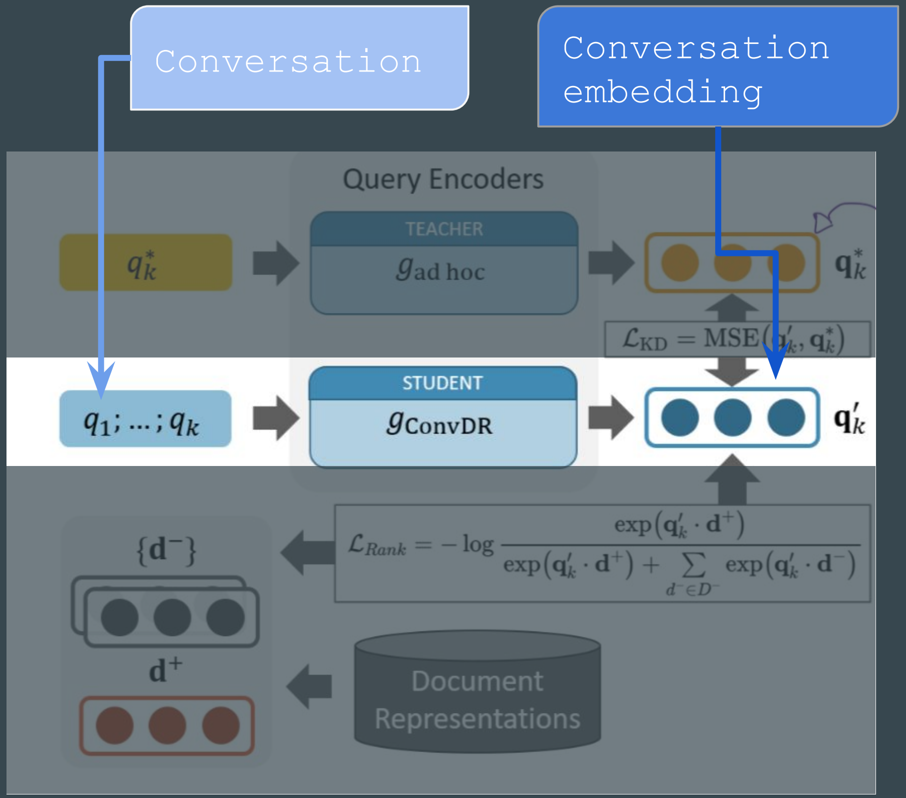
➔ Teacher: pre-trained **ANCE ranker**



Training (FS Conv DPR [11])

F.S. Conv.DPR [11]: Training

→ **Student:** produces embedding from conversation

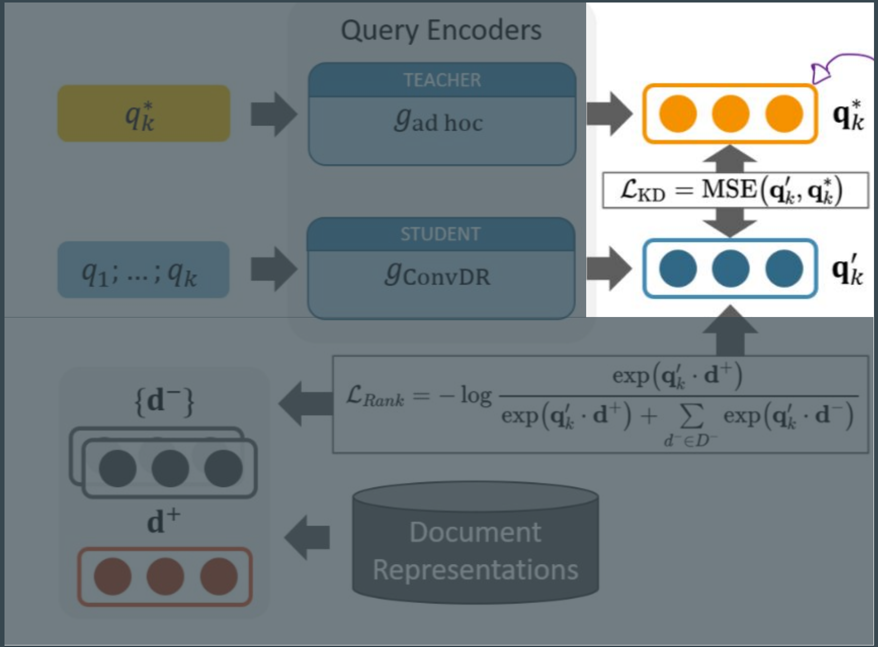


Training (FS Conv DPR [11])

F.S. Conv.DPR [11] : Training

➔ **Knowledge Distillation loss** L_{KD} **:**

Teaching the student to **produce "good embeddings" from conversation**

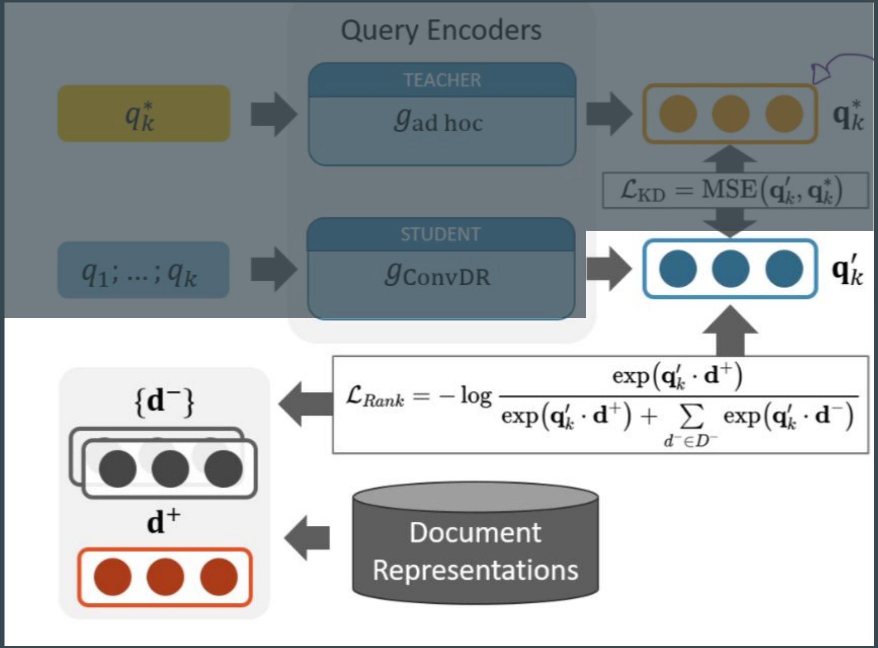


Training (FS Conv DPR [11])

F.S. Conv.DPR [11] : Training

➔ **Ranking loss** L_{Rank} **:**

Further fine-tuning of ranker in this task (Negative log likelihood)



Training (FS Conv DPR [11])

Which loss is more important?

- L_{Rank}
- L_{KD}

Ranking vs. Knowledge Distillation losses

Method	CAsT 2019		OR-QuAC	
	NDCG@3	W/T/L	MRR@5	W/T/L
First Stage Retrieval Only				
ConvDR (Zero-Shot)	0.202	13/30/57	0.568	24/61/16
ConvDR (KD)	**0.466**	38/39/24	0.519	19/63/18
ConvDR (Rank)	0.084	3/19/78	0.588	29/52/19
ConvDR (Multi-Task)	0.157	12/19/69	**0.616**	30/56/14

Ablation study ([11])

- ➔ **CAsT 19:** Fine-tuning w/o knowledge distillation fails
  - ◆ Few-shot setting
- ➔ **OR-QuAC:** Ranking loss is useful
  - ◆ why?
    - # training data?
    - dataset bias: all relevant passages from Wikipedia article

F.S. Conv.DPR [11] : Initial retrieval results

Ranker	MRR	NDCG @ 3
QuReTeC + BM25	-	0.17
FS Conv QR + BM25	-	0.15
FS Conv DPR	**0.50**	**0.34**
Human queries + BM25	0.45	0.30
Human queries + ANCE	**0.59**	**0.42**

Cast-20: First-stage Ranking Results

- ➔ largely outperforms other query rewriting methods when BM25 is the ranker
- ➔ w.r.t. **oracles** (human resolutions):
  - ◆ +15% wrt. BM25
  - ◆ -20% wrt. **ANCE** oracle

F.S. Conv.DPR [11] : Re-ranking results

Ranker	MRR	NDCG @ 3
QuReTeC + BERT		
h2oloo (T5-based)	**0.62**	**0.49**
FS Conv QR + BERT	-	0.34
FS Conv DPR + BERT	**0.54**	**0.39**
Human queries + BM25 + BERT	0.63	0.47
Human queries + ANCE + BERT	0.66	0.48

Cast-20: Re-ranking Results

$$\mathbf{H}_0(q_k, d) = \mathrm{BERT}([\mathrm{CLS}] \circ q_1 \circ \ldots [\mathrm{SEP}] \circ q_k \circ [\mathrm{SEP}] \circ d \circ [\mathrm{SEP}]),$$

Re-ranking method

- After **re-ranking**, improvements fade away
  - **ANCE 1st stage ranking** vs … **reranking**
- Waterloo's T5-based pipeline outperforms all approaches, even human!
- What's a fair evaluation?

Conclusions

- Discussed various methods for Conv Passage Retrieval
  - lexical vs. neural
- Evaluation varies on various ranking components
  - what's a **fair comparison**?
- Open problems:
  - Handling **previous answers / long context**
  - **Robustness** to pipeline changes
  - Training in **few-shot** setting?
  - **other types of conversations**?

References

[1] Reddy, S., Chen, D., & Manning, C. D. (2019). Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7, 249-266.

[2] ter Hoeve, M., Sim, R., Nouri, E., Fourney, A., de Rijke, M., & White, R. W. (2020, March). Conversations with documents: An exploration of document-centered assistance. In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (pp. 43-52).

[3] Aliannejadi, Mohammad, et al. "Asking clarifying questions in open-domain information-seeking conversations." Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. 2019.

[4] Krasakis, Antonios Minas, et al. "Analysing the effect of clarifying questions on document ranking in conversational search." Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. 2020.

References

[5] Hashemi, Helia, Hamed Zamani, and W. Bruce Croft. "Guided transformer: Leveraging multiple external sources for representation learning in conversational search." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[6] Zhang, Yongfeng, et al. "Towards conversational search and recommendation: System ask, user respond." Proceedings of the 27th acm international conference on information and knowledge management. 2018.

[7] Dalton, Jeffrey, Chenyan Xiong, and Jamie Callan. "TREC CAsT 2019: The conversational assistance track overview." arXiv preprint arXiv:2003.13624 (2020).

[8] Voskarides, Nikos, et al. "Query resolution for conversational search with limited supervision." Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020.

References

[9] Yu, Shi, et al. "Few-shot generative conversational query rewriting." Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020.

[10] Vakulenko, Svitlana, et al. "A comparison of question rewriting methods for conversational passage retrieval." arXiv preprint arXiv:2101.07382 (2021).

[11] Yu, Shi, et al. "Few-Shot Conversational Dense Retrieval." arXiv preprint arXiv:2105.04166 (2021).

[12] Lin, Sheng-Chieh, Jheng-Hong Yang, and Jimmy Lin. "TREC 2020 Notebook: CAsT Track."