

Question Rewriting for Conversational Question Answering

Svitlana Vakulenko

Task: Conversational QA

Q: Where is Xi'an?

A: Shaanxi, China

Q: What is its GDP?

A: 95 Billion USD

Q: What is the share in the province GDP?

A: 41.8%

Task: Conversational QA

Q: Where is Xi'an?

A: Shaanxi, China

Q: What is its GDP?

A: 95 Billion USD

Q: What is the share in the province GDP?

A: 41.8%

Anaphora

Ellipsis



@svakulenkov



Approach: Question Rewriting

Q: Where is Xi'an?

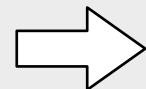
A: Shaanxi, China

Q: What is its GDP?

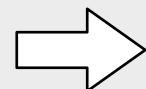
A: 95 Billion USD

Q: What is the share in the province GDP?

A: 41.8%



What is Xi'an's GDP?

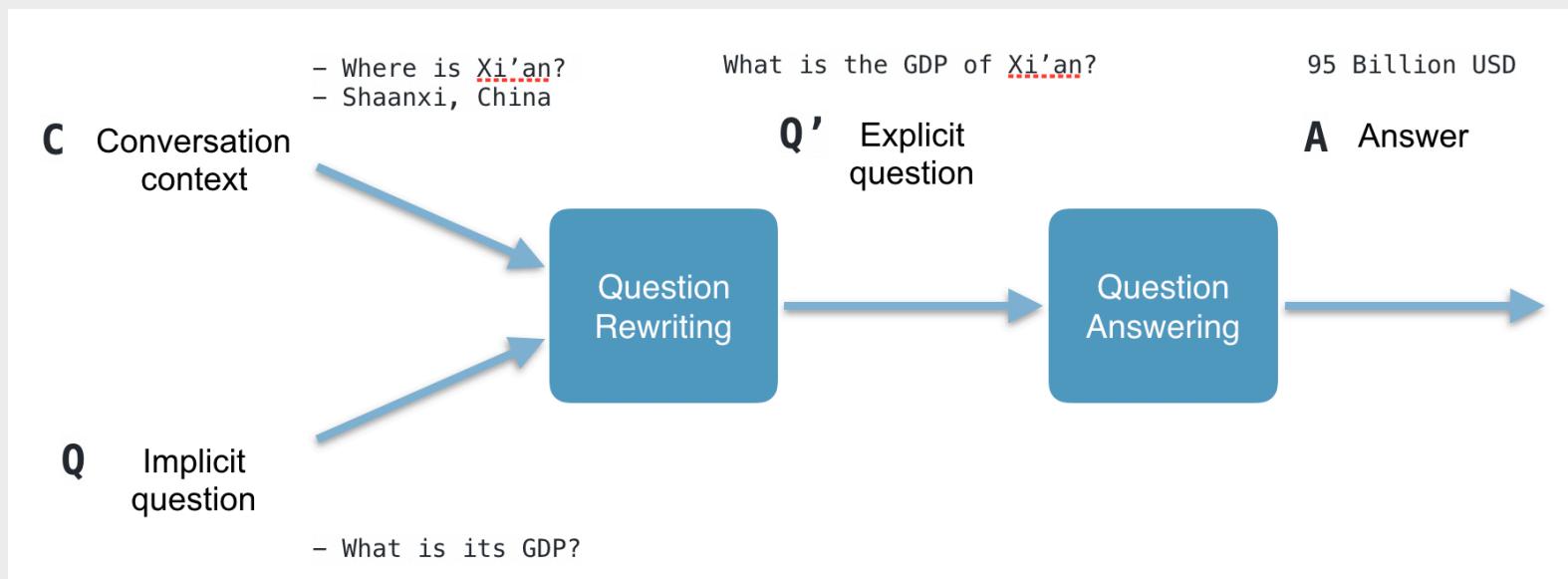


What is the share of Xi'an in the Shaanxi province GDP?



@svakulenkov

Architecture: Conversational QA



Approach: Unsupervised

■ [Mele et al., 2020]

Input: Where is **Xi'an**? <SEP> Shaanxi, China <SEP> What is **its** GDP?

Output: (its, Xian)

QR: What is **Xi'an** GDP?

- co-reference resolution + heuristics
- identify topics as nouns in dependency parses
- identify topic shifts via lexical cues



Approach: Supervised Sequence Generation

- Transformer++ [Vakulenko et al., 2021]

Input: Where is Xi'an? <SEP> Shaanxi, China <SEP> What is its GDP?

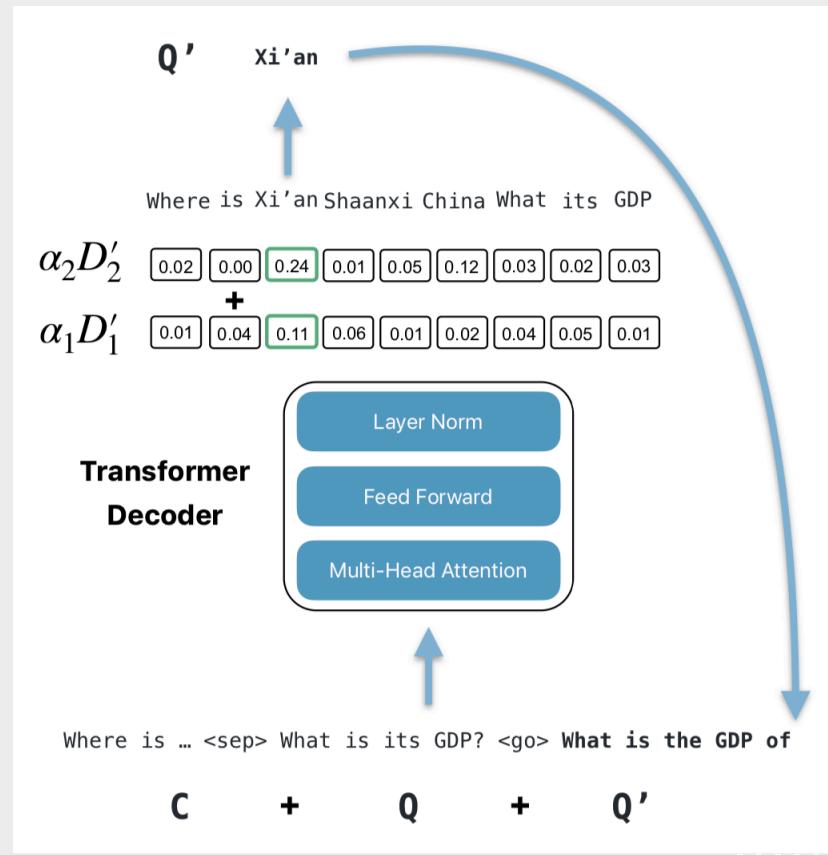
Output: What is Xi'an's GDP?

- CANARD 35K conversational questions + rewrites
- fine-tune GPT2

Approach: Supervised Sequence Generation

Transformer++

$$D' = \sum_{i=0}^m \alpha_i D'_i$$





Approach: Weakly Supervised Generation

- [Yu et al., 2020]

Input: Where is Xi'an? <SEP> Shaanxi, China <SEP> What is its GDP?

Output: What is Xi'an's GDP?

- MS MARCO 152K sessions -> rule-based/self-learn conversations
- fine-tune GPT2

Approach: Supervised Classification

- QuReTeC [Voskarides et al., 2020]

Input: Where is **Xi'an**? <SEP> Shaanxi, China <SEP> What is its GDP?

Output: 0 0 1 0 0

QR: What is its GDP? **Xi'an**

- CANARD 35K questions
- fine-tune BERT

Results: TREC CAsT 2019

[Vakulenko et al., 2021]

Table 4: Comparison with the state-of-the-art results reported on the TREC CAsT test set.

Approach	NDCG@3
Mele et al. [21]	0.397
Voskarides et al. [37]	0.476
Yu et al. [41]	0.492
Ours	0.529

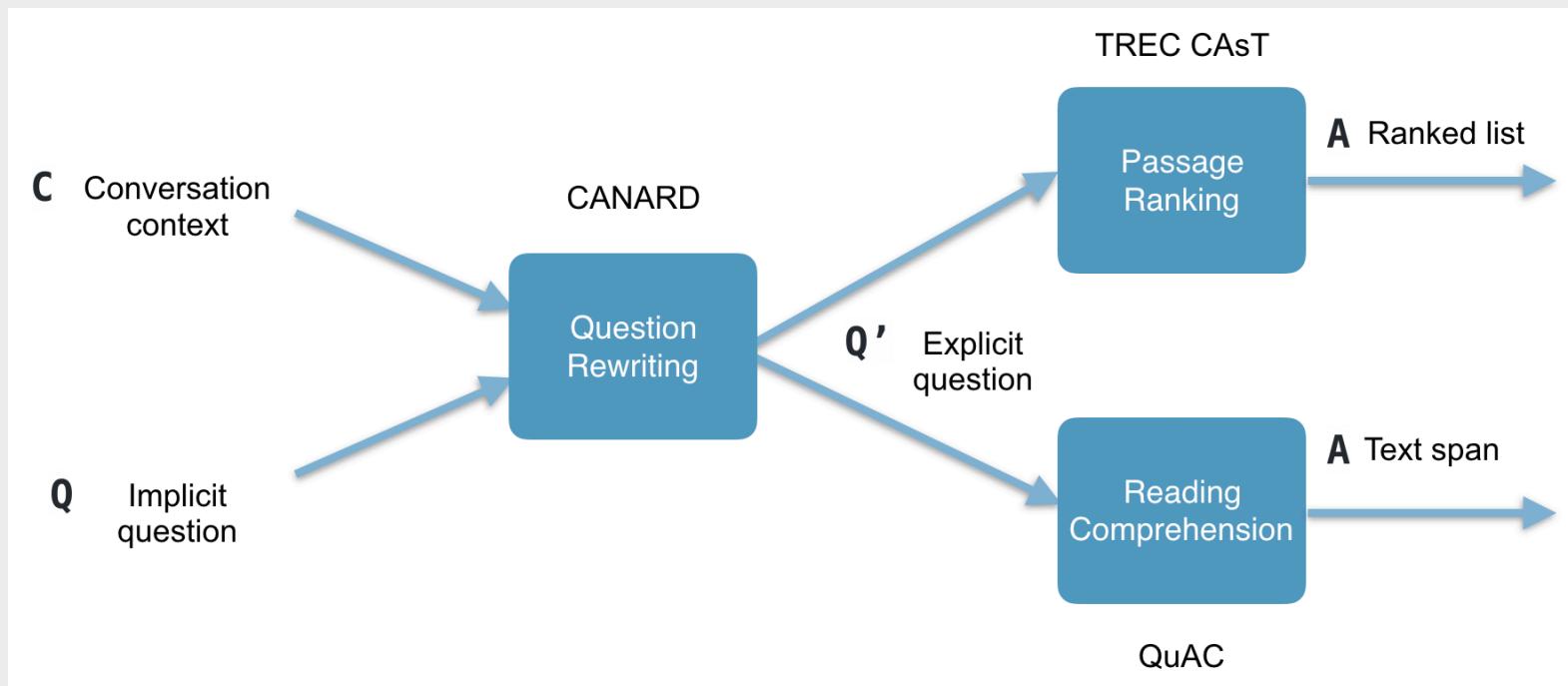
Results: TREC CAsT 2019

[Vakulenko et al., 2021]

Table 3: Retrieval QA results on the TREC CAsT test set.

QA Input	QA Model	MAP	MRR	NDCG@3
Original	Anserini	0.172	0.403	0.265
Original + 1-DT*	+BERT	0.230	0.535	0.378
Original + 2-DT*		0.245	0.576	0.404
Original + 3-DT*		0.238	0.575	0.401
Co-reference		0.201	0.473	0.316
PointerGenerator		0.183	0.451	0.298
CopyTransformer		0.284	0.628	0.440
Transformer++		0.341	0.716	0.529
Human		0.405	0.879	0.589

Architecture: Conversational QA



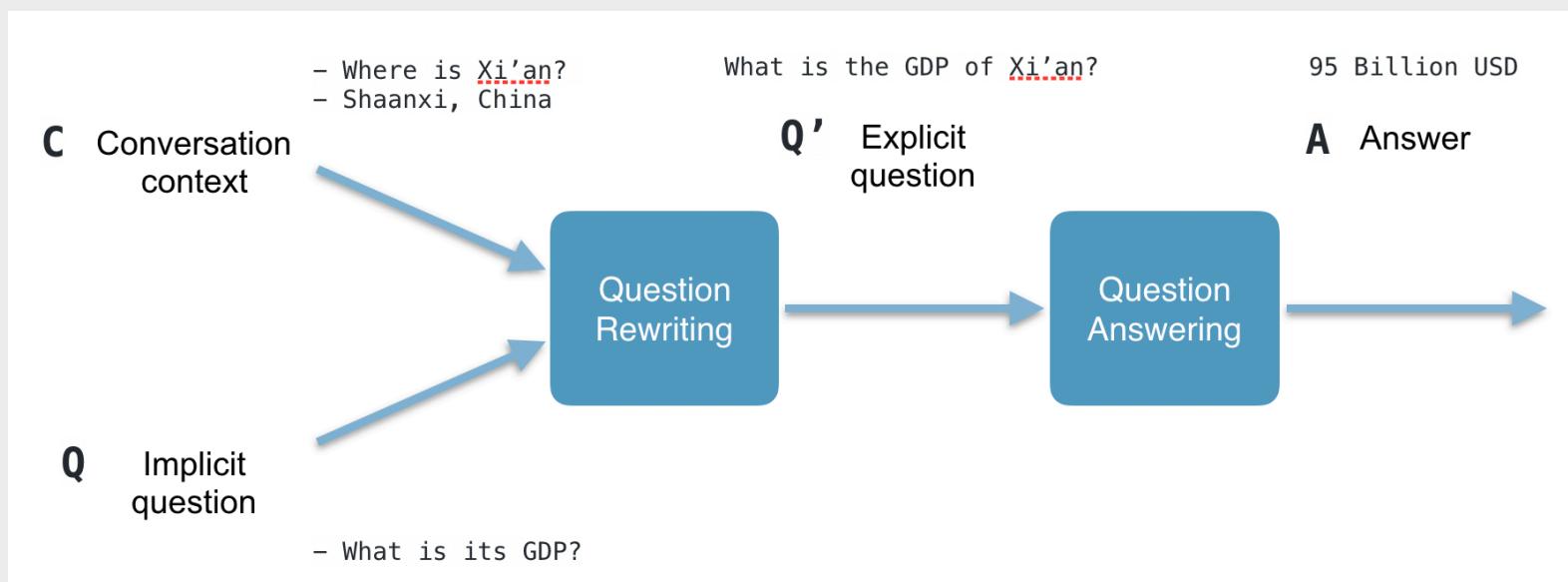
Results: TREC CAsT 2019 & CANARD

[Vakulenko et al., 2021]

Table 5: Extractive QA results on the CANARD test set. F1 and EM is calculated for both answerable and unanswerable questions, while NA Acc only for unanswerable questions.

QA Input	Training Set	EM	F1	NA Acc
Original	MultiQA → CANARD-H	41.32	54.97	65.84
Original + 1-DT		43.15	57.03	68.64
Original + 2-DT		42.20	57.33	69.42
Original + 3-DT		43.29	57.87	71.50
Co-reference		42.70	57.59	66.20
PointerGenerator		41.93	57.37	63.16
CopyTransformer		42.67	57.62	68.02
Transformer++		43.39	58.16	68.29
Human		45.40	60.48	70.55

Architecture: Conversational QA





Approach: Error Analysis

NDCG@3

- **Original:** What types does olive oil contain? 0
- **QR:** What types of **fats** does olive oil contain? 0.9
- **Human:** What types of **unsaturated fats** does olive oil contain? 1

Approach: Error Analysis

Original	QR	Human
✗	✗	✗
✓	✗	✗
✗	✓	✗
✓	✓	✗
✗	✗	✓
✓	✗	✓
✗	✓	✓
✓	✓	✓

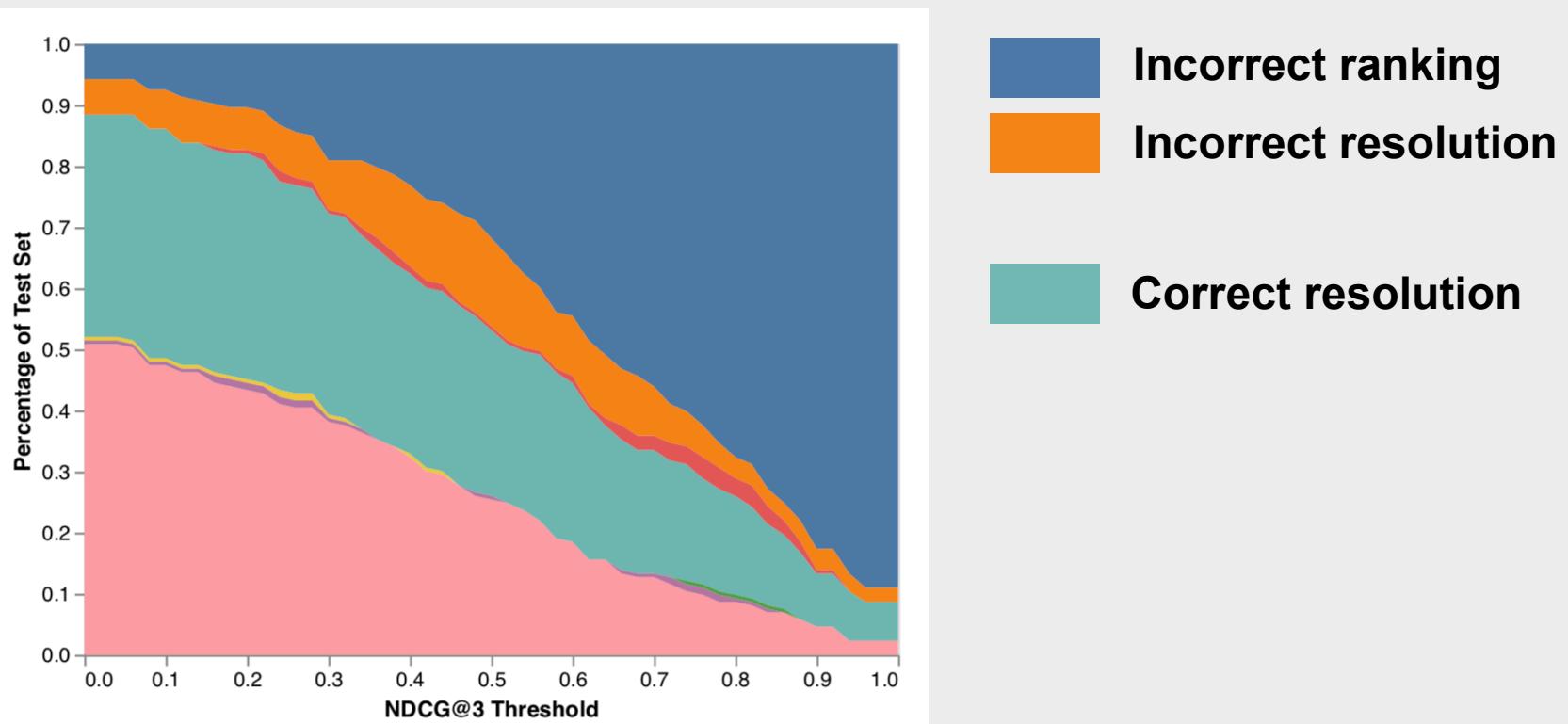
NDCG@3

■ **Original:** What types does olive oil contain? 0

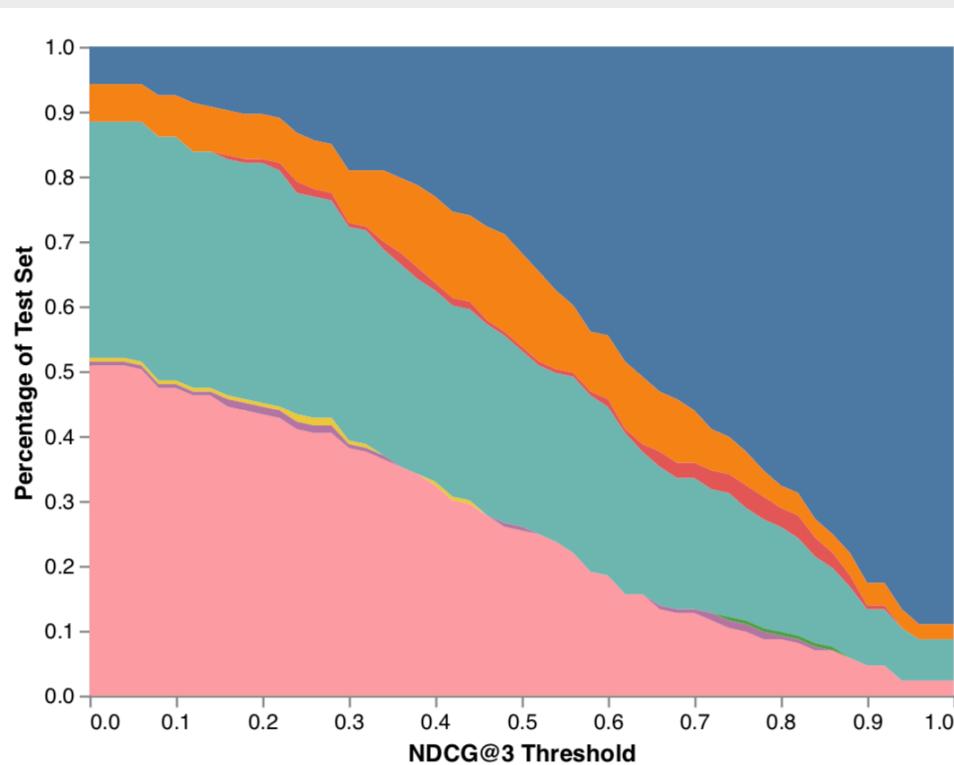
■ **QR:** What types of **fats** does olive oil contain? 0.9

■ **Human:** What types of unsaturated fats does 1
olive oil contain?

Approach: Error Analysis

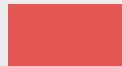
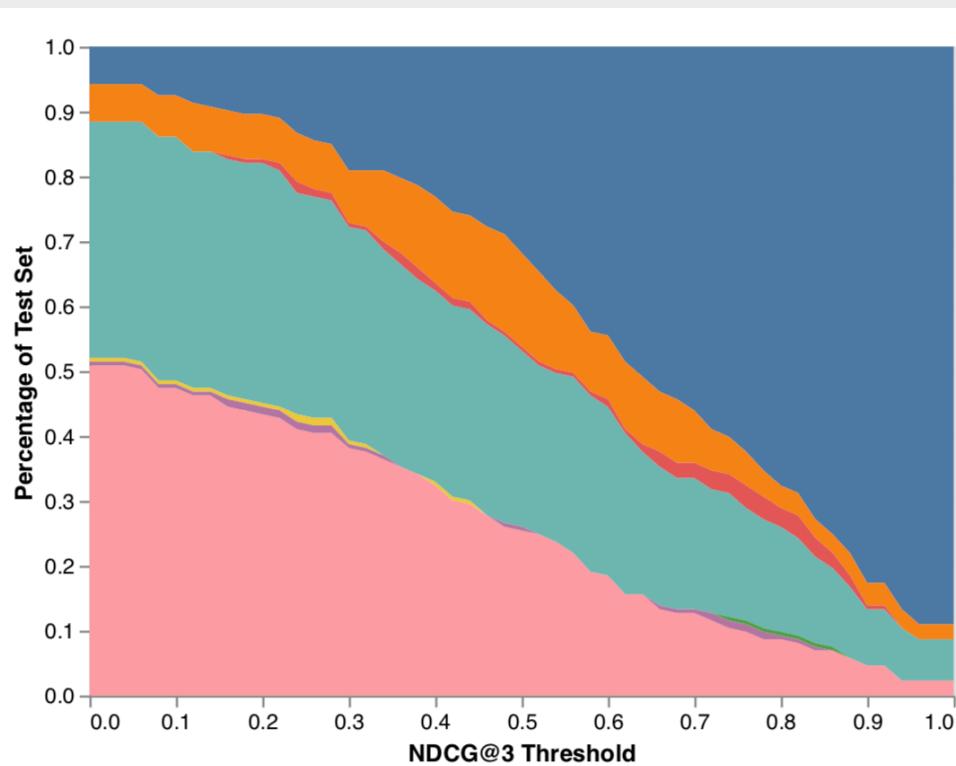


Approach: Error Analysis



Incorrect no resolution
Correct no resolution

Approach: Error Analysis



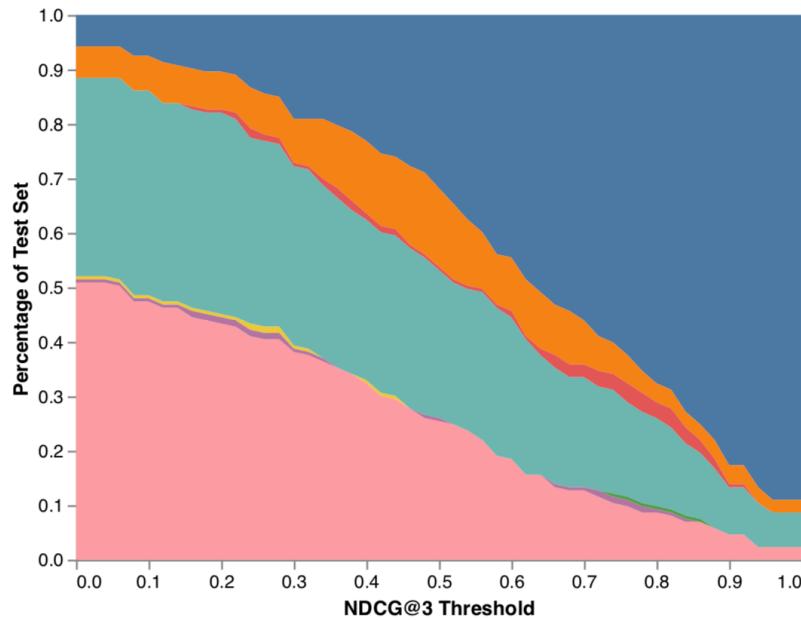
Superhuman resolution



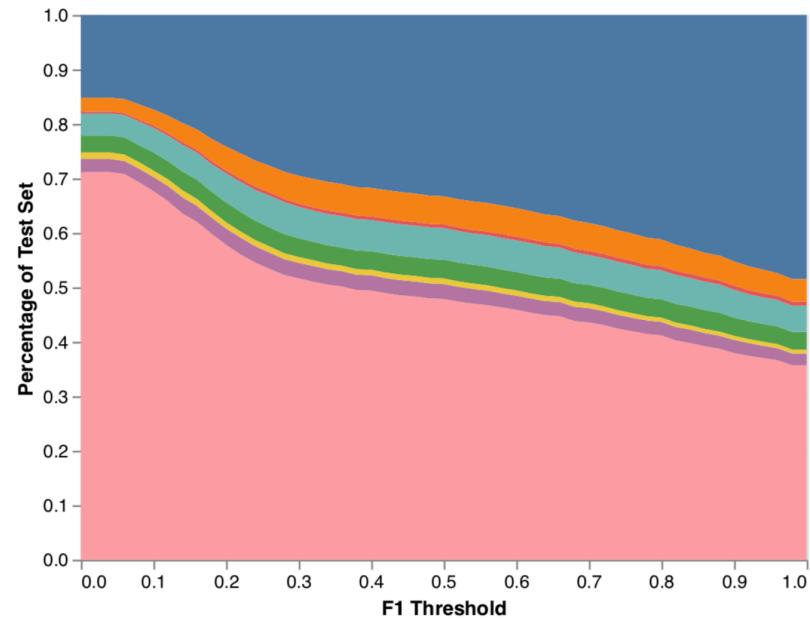
@svakulenkov

Results: Error Analysis

Passage ranking

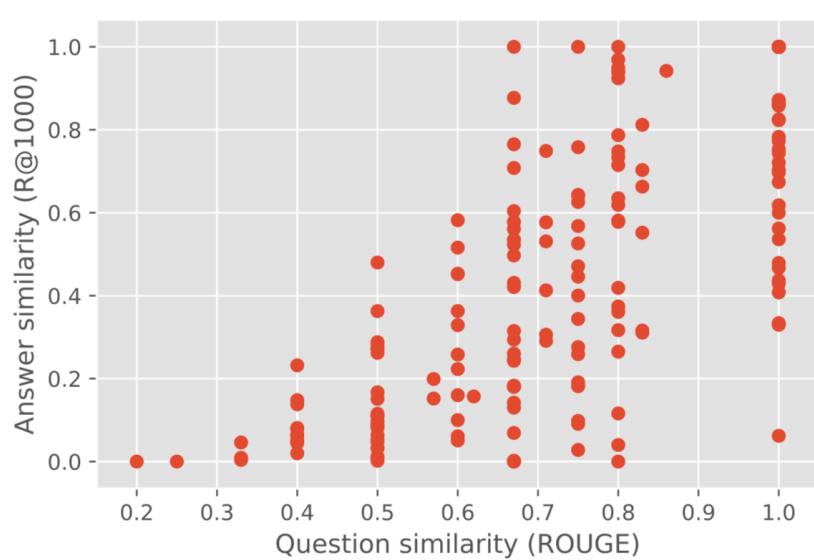


Reading comprehension

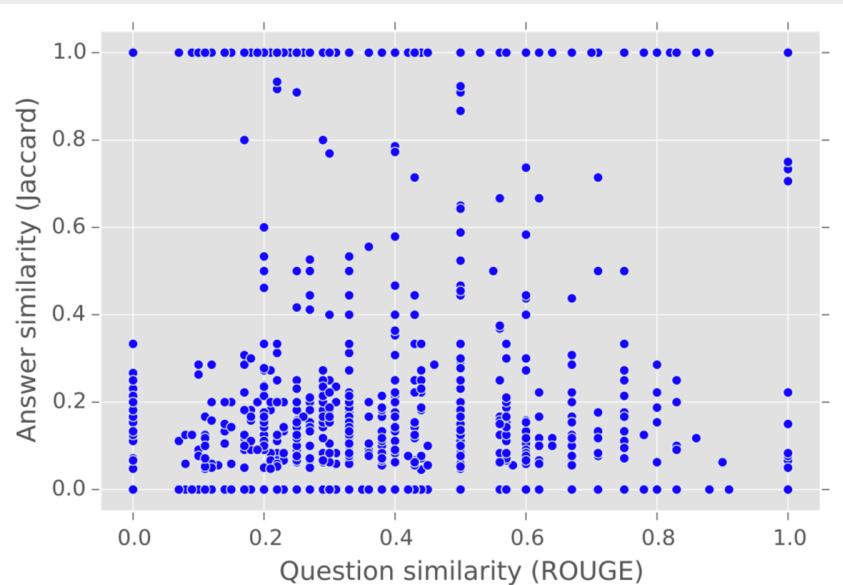


Results: Error Analysis

Passage ranking



Reading comprehension





Results: Error Analysis

QR	Human	ROUGE-1 R	R@1000
How is intermittent fasting related to keto diet	How is intermittent fasting related to keto?	1	0.48
<i>What's the biggest great whites ever caught</i>	What about for great whites?	0.6	1

QR	Human	ROUGE-1 R	Jaccard
<i>was kick out the jams well received?</i>	was Kick Out the Jams well received?	1	0.03
<i>did he do anything else?</i>	did Guy Lombardo have any other career highlights besides racing and the restaurant?	0.08	1



Dataset: TREC CAsT 2020

The TREC Conversational Assistance Track (CAsT)

Conversational search benchmark at TREC

[View On GitHub](#)

The screenshot shows the main page of the TREC CAsT 2020 website. The header is blue with the title "The TREC Conversational Assistance Track (CAsT)" and subtitle "Conversational search benchmark at TREC". Below the header is a yellow navigation bar with a "View On GitHub" button and a GitHub icon. The main content area has a light beige background. On the left, there's a sidebar with links for "TREC Conversational Assistance Track (CAsT)", "Year 2 (TREC 2020)", "Data", "Topics", "Baselines", "Collection", "Guidelines", "News", "Contact", "Important Dates", "Organizers", "Year 1 (TREC 2019)", "2019 Data", "Topics", and "Project updates and acknowledgments". At the bottom of the sidebar, it says "Hosted on GitHub Pages — theme by [Colab](#)". The main content area features a section titled "Year 2 (TREC 2020)" with subsections for "Data", "Topics", "Baselines", and "Collection", each containing a bulleted list of features.

Year 2 (TREC 2020)

Data

- 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.

Topics

- NEW - [BM25 + BERT baseline](#) - We provide a BM25 + BERT reranked baseline run for the raw utterances, automatically rewritten utterances, and the manually rewritten utterances.
- NEW - [Interactive web UI](#) - A simple web UI with the BM25 + BERT model used to create the baseline runs. No rewriting is performed.

Baselines

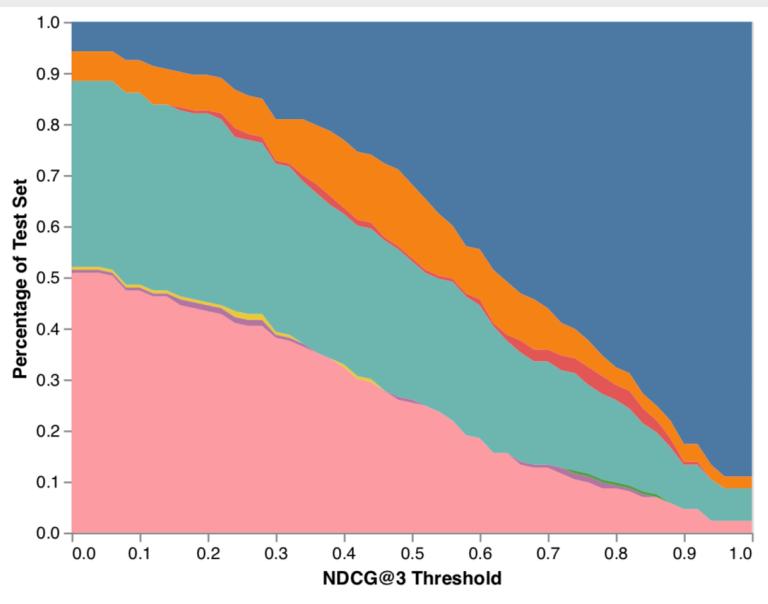
- The corpus is a combination of two standard TREC collections: MARCO Ranking passages and Wikipedia (TREC CAR).
- The [MS MARCO Passage Ranking collection](#) - This file only includes the passage id and passage text. For convenience, we also provide a passage id → URL mapping file in TSV format [pid to URL file](#).
- The [TREC CAR paragraph collection v2.0](#).

Collection

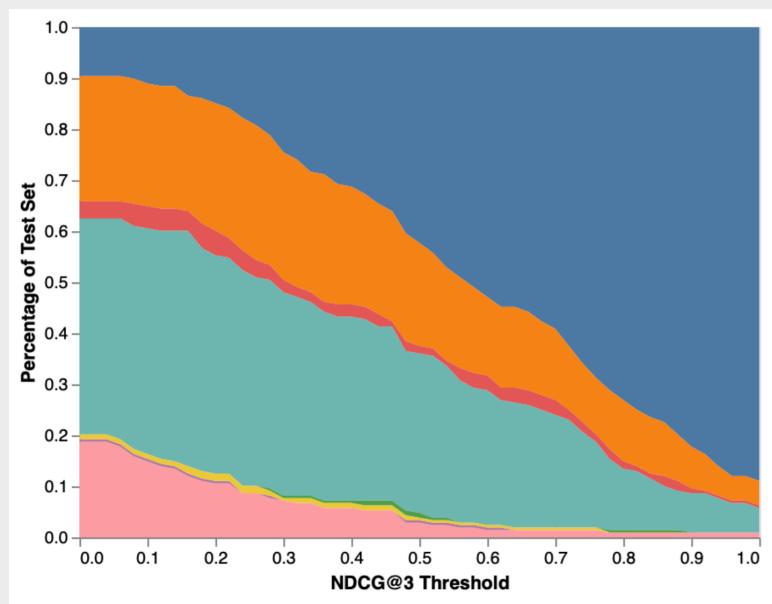
<https://www.treccast.ai>

CAsT Evolution

Transformer++ on CAsT 2019



QuReTeC on CAsT 2020



<https://github.com/svakulenkov/QRQA>



Results: QuReTeC on CAsT 2020

Question	QR	Answer
Who are some of the well-known Information Retrieval researchers?	-	Bruce Croft formed the Center for Intelligent Information Retrieval (CIIR) in 1991, since when he and his students have worked with more than 90 industry and government partners on research and technology projects and have produced more than 900 papers. Bruce Croft has made major contributions to most areas of information retrieval, including pioneering work in clustering, passage retrieval, sentence retrieval, and distributed search. One of the most important areas of work for Croft relates to ranking functions and retrieval models, where he has led the development of one of the major approaches to modeling search: language modelling. In later years, Croft also led the way in the development of feature-based ranking functions. Croft and his research group have also developed a series of search engines: InQuery, the Lemur toolkit, Indri, and Galago. These search engines are open source and offer unique capabilities that are not replicated in other research retrieval platforms source – consequently they are downloaded by hundreds of researchers world wide. As a consequence of his work, Croft is one of the most cited researchers in information retrieval.
What did he work on?	What did he work on? croft bruce	

References

- Svitlana **Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: Question Rewriting for Conversational Question Answering. **WSDM** 2021. (To appear)
- Svitlana **Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering. **SCAI@EMNLP** 2020: 7-16.
- Nikos **Voskarides**, Dan Li, Pengjie Ren, Evangelos Kanoulas, Maarten de Rijke: Query Resolution for Conversational Search with Limited Supervision. **SIGIR** 2020: 921-930.
- Shi **Yu**, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett et al. Few-Shot Generative Conversational Query Rewriting. **SIGIR** 2020: 1933-1936.
- Ida **Mele**, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego et al. Topic Propagation in Conversational Search. **SIGIR** 2020: 2057-2060.



Our Team



MSc Shayne Longpre
Apple Inc.



MSc Zhucheng (Michael) Tu
Apple Inc.



MSc Raviteja Anantha
Apple Inc.



MSc (soon Dr) Nikos Voskarides
University of Amsterdam

Conclusions: Question Rewriting

■ Modular



■ Reusable

■ Debuggable

■ Cheap