CausalQA: A Benchmark for Causal Question Answering

October 12–17, 2022

Alexander Bondarenko[⋆] Magdalena Wolska[†] Stefan Heindorf[‡] Lukas Blübaum[‡]
Axel-Cyrille Ngonga Ngomo[‡] Benno Stein[†] Pavel Braslavski[§,¶] Matthias Hagen[⋆] Martin Potthast[‖]

[⋆]Martin-Luther-Univeristät Halle-Wittenberg [†]Bauhaus-Universität Weimar
[‡]Paderborn University [§]Ural Federal University [¶]HSE University [‖]Leipzig University

www.webis.de

The 29th International Conference on Computational Linguistics (COLING '22); Gyeongju, Republic of Korea.

CausalQA: A Benchmark for Causal Question Answering

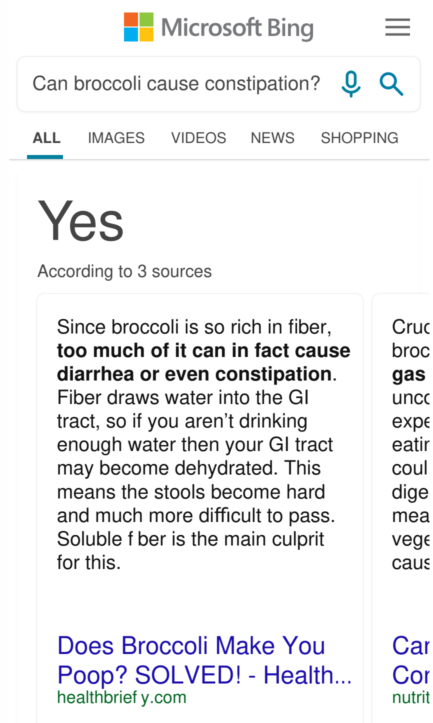
Motivation

- At least 5% of questions submitted to search engines ask about causality.
- Existing benchmark datasets for causal QA are comparably small.
- Causal QA is hampered by a lack of specialized, large-scale resources.

CausalQA: A Benchmark for Causal Question Answering

Motivation

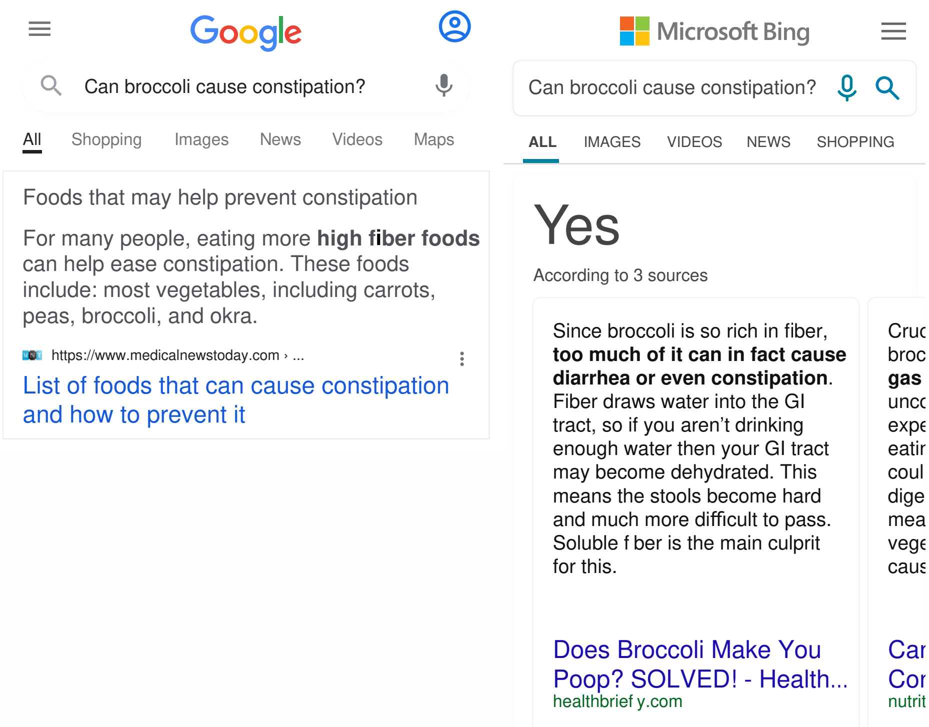
- At least 5% of questions submitted to search engines ask about causality.
- Existing benchmark datasets for causal QA are comparably small.
- Causal QA is hampered by a lack of specialized, large-scale resources.



CausalQA: A Benchmark for Causal Question Answering

Motivation

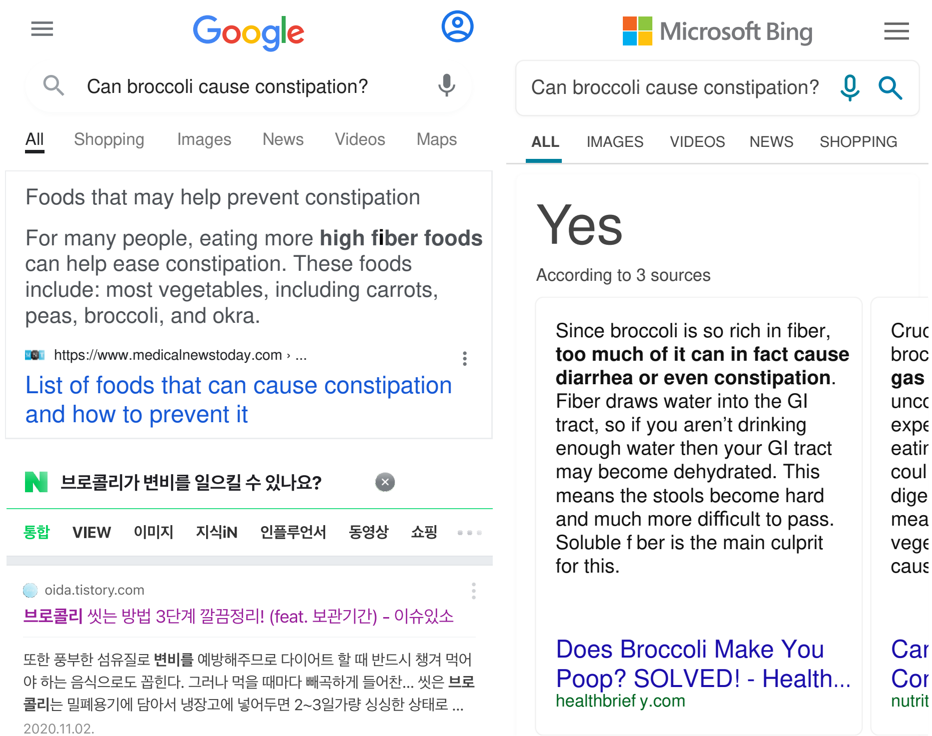
- At least 5% of questions submitted to search engines ask about causality.
- Existing benchmark datasets for causal QA are comparably small.
- Causal QA is hampered by a lack of specialized, large-scale resources.



CausalQA: A Benchmark for Causal Question Answering

Motivation

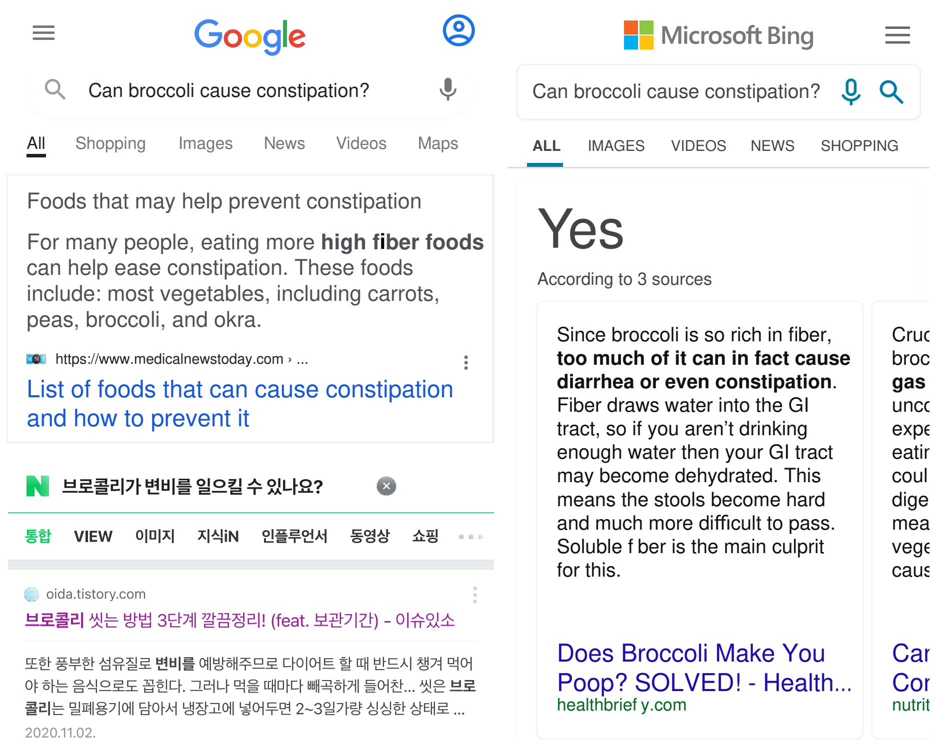
- At least 5% of questions submitted to search engines ask about causality.
- Existing benchmark datasets for causal QA are comparably small.
- Causal QA is hampered by a lack of specialized, large-scale resources.



CausalQA: A Benchmark for Causal Question Answering

Contributions

- Webis-CausalQA-22 dataset with 1.1M causal questions and answers.
- A set of rules to identify causal questions at near-perfect precision.
- Analisys of causal questions and a new taxonomy.
- Baseline question answering experiments on the dataset.



CausalQA: A Benchmark for Causal Question Answering

Dataset Webis-CausalQA-22

- 10 existing QA datasets: PAQ, GooAQ, MS MARCO, SQuAD, etc.
- Datasets are well-known, large, and contain lexically diverse questions.

CausalQA: A Benchmark for Causal Question Answering

Dataset Webis-CausalQA-22

- 10 existing QA datasets: PAQ, GooAQ, MS MARCO, SQuAD, etc.
- Datasets are well-known, large, and contain lexically diverse questions.
- A question is *causal* if answering it requires:
  1. identifying causal chains,
  2. inference on those chains,
  3. verbalizing the causal relations involved when answering it.

CausalQA: A Benchmark for Causal Question Answering

Dataset Webis-CausalQA-22

- 10 existing QA datasets: PAQ, GooAQ, MS MARCO, SQuAD, etc.
- Datasets are well-known, large, and contain lexically diverse questions.
- A question is *causal* if answering it requires:
  1. identifying causal chains,
  2. inference on those chains,
  3. verbalizing the causal relations involved when answering it.
- 7 regex rules to identify causal questions:

R1 `[why]` e.g.: Why does mosquito bite itch?

CausalQA: A Benchmark for Causal Question Answering

Dataset Webis-CausalQA-22

- 10 existing QA datasets: PAQ, GooAQ, MS MARCO, SQuAD, etc.
- Datasets are well-known, large, and contain lexically diverse questions.
- A question is *causal* if answering it requires:
  1. identifying causal chains,
  2. inference on those chains,
  3. verbalizing the causal relations involved when answering it.
- 7 regex rules to identify causal questions:

R1 `[why]` e.g.: Why does mosquito bite itch?

R2 `[cause(s)?]` e.g.: What causes broken blood vessels?

⋮

R7 `[what (to do|should be done)]∧[if|to|when]`
e.g.: What to do if my Xbox won't connect to the Wi-Fi?

CausalQA: A Benchmark for Causal Question Answering

Dataset Webis-CausalQA-22

- 10 existing QA datasets: PAQ, GooAQ, MS MARCO, SQuAD, etc.
- Datasets are well-known, large, and contain lexically diverse questions.
- A question is *causal* if answering it requires:
  1. identifying causal chains,
  2. inference on those chains,
  3. verbalizing the causal relations involved when answering it.
- 7 regex rules to identify causal questions:

  R1 `[why]` e.g.: Why does mosquito bite itch?

  R2 `[cause(s)?]` e.g.: What causes broken blood vessels?

  ⋮

  R7 `[what (to do|should be done)]∧[if|to|when]`
  e.g.: What to do if my Xbox won't connect to the Wi-Fi?
- Webis-CausalQA-22 contains ca. 1.1 million causal QA pairs.

CausalQA: A Benchmark for Causal Question Answering

Benchmark

- UnifiedQA model, pre-trained and fine-tuned [Khashabi et al.; EMNLP ’20].

CausalQA: A Benchmark for Causal Question Answering

Benchmark

- UnifiedQA model, pre-trained and fine-tuned [Khashabi et al.; EMNLP '20].

Dataset	Random 90/10 split					
	N	Fine-tuned model				
		ROUGE-L			Traditional	
		P	R	F_1	EM	F_1
PAQ	76,961	0.95	0.95	0.94	0.91	0.94
GooAQ	14,629	0.17	0.15	0.15	0.00	0.19
MS MARCO QnA	2,557	0.45	0.42	0.39	0.13	0.40
Natural Questions	121	0.37	0.34	0.32	0.16	0.33
ELI5	13,104	0.16	0.09	0.10	0.00	0.12
SearchQA	78	0.55	0.54	0.54	0.47	0.54
SQuAD v. 2.0	321	0.96	0.96	0.95	0.93	0.95
NewsQA	66	0.76	0.76	0.73	0.58	0.73
HotpotQA	39	0.73	0.73	0.73	0.67	0.72
TriviaQA	71	0.44	0.43	0.42	0.28	0.42
Macro-averaged	107,947	0.55	0.54	0.53	0.41	0.53
Micro-averaged	107,947	0.73	0.72	0.72	0.65	0.73

CausalQA: A Benchmark for Causal Question Answering
Benchmark

- UnifiedQA model, pre-trained and fine-tuned [Khashabi et al.; EMNLP ’20].

Dataset	Random 90/10 split					
	N	Fine-tuned model				
		ROUGE-L			Traditional	
		P	R	F_1	EM	F_1
PAQ	76,961	**0.95**	**0.95**	**0.94**	**0.91**	**0.94**
GooAQ	14,629	0.17	0.15	0.15	0.00	0.19
MS MARCO QnA	2,557	0.45	0.42	0.39	0.13	0.40
Natural Questions	121	0.37	0.34	0.32	0.16	0.33
ELI5	13,104	0.16	0.09	0.10	0.00	0.12
SearchQA	78	0.55	0.54	0.54	0.47	0.54
SQuAD v. 2.0	321	**0.96**	**0.96**	**0.95**	**0.93**	**0.95**
NewsQA	66	0.76	0.76	0.73	0.58	0.73
HotpotQA	39	0.73	0.73	0.73	0.67	0.72
TriviaQA	71	0.44	0.43	0.42	0.28	0.42
Macro-averaged	107,947	**0.55**	**0.54**	**0.53**	**0.41**	**0.53**
Micro-averaged	107,947	**0.73**	**0.72**	**0.72**	**0.65**	**0.73**

CausalQA: A Benchmark for Causal Question Answering

Benchmark

- UnifiedQA model, pre-trained and fine-tuned [Khashabi et al.; EMNLP ’20].
- Highest F_1 on SQuAD is 0.93, while that of humans is 0.89 [Rajpurkar et al.; ACL ’18].

Dataset	**Random 90/10 split**					
	N	Fine-tuned model				
		ROUGE-L			Traditional	
		P	R	F_1	EM	F_1
PAQ	76,961	**0.95**	**0.95**	**0.94**	**0.91**	**0.94**
GooAQ	14,629	0.17	0.15	0.15	0.00	0.19
MS MARCO QnA	2,557	0.45	0.42	0.39	0.13	0.40
Natural Questions	121	0.37	0.34	0.32	0.16	0.33
ELI5	13,104	0.16	0.09	0.10	0.00	0.12
SearchQA	78	0.55	0.54	0.54	0.47	0.54
SQuAD v. 2.0	321	**0.96**	**0.96**	**0.95**	**0.93**	**0.95**
NewsQA	66	0.76	0.76	0.73	0.58	0.73
HotpotQA	39	0.73	0.73	0.73	0.67	0.72
TriviaQA	71	0.44	0.43	0.42	0.28	0.42
Macro-averaged	107,947	**0.55**	**0.54**	**0.53**	**0.41**	**0.53**
Micro-averaged	107,947	**0.73**	**0.72**	**0.72**	**0.65**	**0.73**

CausalQA: A Benchmark for Causal Question Answering

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries.

CausalQA: A Benchmark for Causal Question Answering

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries.
- Ca. 82 million (about 5%) causal questions found with the rules.

CausalQA: A Benchmark for Causal Question Answering

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries.
- Ca. 82 million (about 5%) causal questions found with the rules.
- “Why”-questions are most frequent (causal) questions
  e.g.: Why can’t I log in into VKontakte? (cat. *problem solving*).

CausalQA: A Benchmark for Causal Question Answering

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries.
- Ca. 82 million (about 5%) causal questions found with the rules.
- "Why"-questions are most frequent (causal) questions
  e.g.: Why can't I log in into VKontakte? (cat. *problem solving*).
- Most of the questions about causes or effects target causes of medical conditions or effects on health (cat. *problem prevention*).

CausalQA: A Benchmark for Causal Question Answering

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries.
- Ca. 82 million (about 5%) causal questions found with the rules.
- “Why”-questions are most frequent (causal) questions
  e.g.: Why can’t I log in into VKontakte? (cat. *problem solving*).
- Most of the questions about causes or effects target causes of medical conditions or effects on health (cat. *problem prevention*).
- 90% of the “what happens if”-questions are about dream interpretation
  e.g.: What will happen, if one dreams of pregnancy?

Conclusions

- Dataset with 1.1M QA pairs to advance research in causal QA.
- Rules to identify causal questions and search engine log analysis.
- Taxonomy of causal questions.
- Baseline QA systems on the constructed dataset.

CausalQA: A Benchmark for Causal Question Answering

Conclusions

- Dataset with 1.1M QA pairs to advance research in causal QA.
- Rules to identify causal questions and search engine log analysis.
- Taxonomy of causal questions.
- Baseline QA systems on the constructed dataset.

Future Work

- Combine text matching QA systems with causal inference.

  Code and data: `github.com/webis-de/COLING-22`

- Dataset with 1.1M QA pairs to advance research in causal QA.
- Rules to identify causal questions and search engine log analysis.
- Taxonomy of causal questions.
- Baseline QA systems on the constructed dataset.

Future Work

- Combine text matching QA systems with causal inference.

Code and data: `github.com/webis-de/COLING-22`

thank you!