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

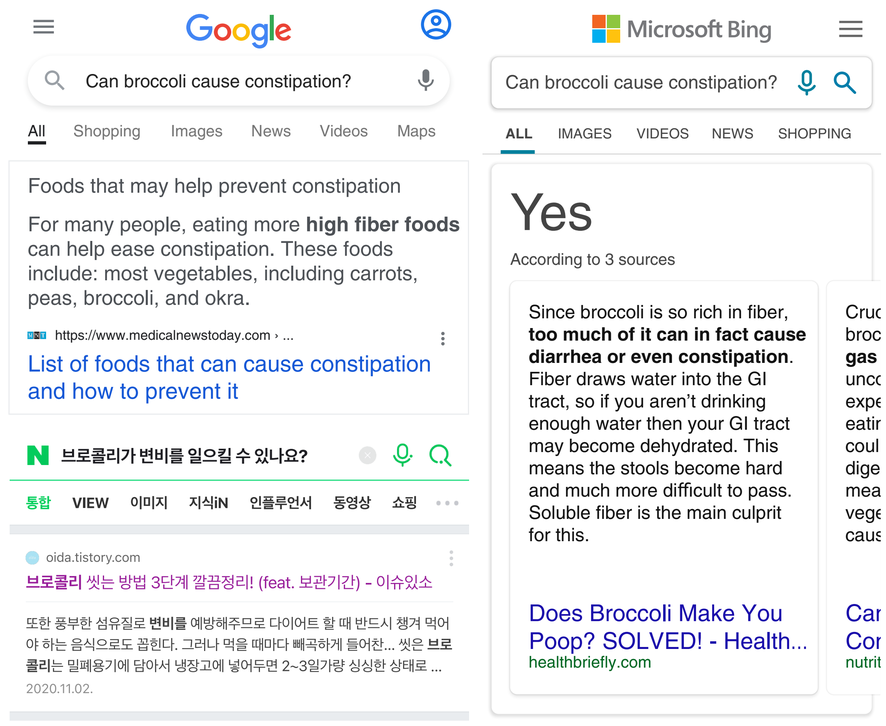
CausalQA: A Benchmark for Causal Question Answering

Motivation

- At least 5% of questions submitted to search engines ask about cause–effect relationships in some way.
- Existing benchmark datasets for causal QA are comparably small.
- Causal QA is hampered by a lack of specialized, large-scale resource.

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 new taxonomy.
- Baseline question answering experiments on the dataset.



Google's and Naver's top results both disagree with that of Bing.

Webis Causal Question Answering Dataset

Characteristics of the QA datasets used to create Webis-CausalQA-22

Dataset	Type		Size			Length (Words)	
	Question source	Answer	Questions	Causal questions		Caus. qu.	Answ.
PAQ	Generated with BART	Term(s)	64,875,601	769,606	(1.2%)	9.6	2.7
GooAQ	Google's autocomplete	Term, Passage	5,030,530	146,286	(2.9%)	7.3	44.3
MS MARCO QnA	Bing query log	Passage	1,010,916	25,569	(2.5%)	6.4	17.5
Natural Questions	Google query log	Passage	315,203	1,208	(0.4%)	9.8	10.8
ELI5	Reddit questions	Passage	272,634	131,033	(48.0%)	32.5	99.0
SearchQA	Human-written	Term(s)	216,136	780	(0.4%)	16.8	1.8
SQuAD v.2.0	Human-written	Term(s)	142,192	3,209	(2.3%)	10.5	6.2
NewsQA	Human-written	Term(s)	119,633	652	(0.5%)	7.2	6.1
HotpotQA	Human-written	Term(s), Passages	112,662	390	(0.4%)	21.8	3.8
TriviaQA	Human-written	Term(s)	109,767	703	(0.6%)	19.4	3.1
Webis-CausalQA-22	Mixed	Mixed	72,205,274	1,079,436	(1.5%)	12.0	22.5

Lexical rules matching causal questions

(R1) `[why]`
Why does mosquito bite itch?

(R2) `[cause(s)?]`
What causes broken blood vessels?

(R3) `[how come|how did]`
How did the constellation Bootes get its name?

(R4) `[effect(s)?|affect(s)?]`
What was the effect of the silk road on religions?

(R5) `[lead(s)? to]`
What does increasing water vapor lead to?

(R6) `[what (will|might)? happen(s)?]∧[if|when]`
What happens if we drink very hot water?

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

Taxonomies of Causal Questions

Semantic Dimensions*

- **Questions about Antecedent**
  - Cause — Why does mosquito bite itch?
  - Goal — Why did Jean Valjean take care of Cosette?
  - Purpose — Why do gaming chairs have a race car design?
  - Enablement — How can FIFA be so blatantly corrupt?
- **Questions about Consequent**
  - Result — What does increasing water vapor lead to?
- **Questions about the Chain**
  - Verification — Would hydrophobic coating affect swimming?

Pragmatic Dimensions*

- **Solution seeking**
  - Problem solving — Why can't I log in into Facebook?
  - Problem prevention — What to do to prevent cancer?
  - Coping with problems — Why doesn't a director fire a stupid employee?
- **Knowledge Seeking**
  - Physical world — Why do chemical reactions depend on pH?
- **Opinion seeking**
  - Social issues — Why do men cheat on their wives?

*Excerpts from taxonomies. For complete versions, see the paper.

Causal Questions in Web Search

- Data: 1.5 billion question-like Yandex log entries .
- Ca. 82 million (about 5%) causal questions found with rules.
- "Why"-questions are most frequent (causal) questions.
- Most of the questions about causes or effects target causes of medical conditions or effects on health.
- Interesting: 90% of the "what happens if"-questions are about dream interpretation.

UnifiedQA on the Webis-CausalQA-22 corpus

Dataset	Original train/dev split										Random 90/10 split					
	N	Base model					Fine-tuned model					N	Fine-tuned model				
		ROUGE-L			Traditional		ROUGE-L			Traditional			ROUGE-L			Traditional	
		P	R	F_1	EM	F_1	P	R	F_1	EM	F_1		P	R	F_1	EM	F_1
PAQ	76,961	0.79	0.85	0.80	0.69	0.80	0.95	0.95	0.94	0.91	0.94	76,961	0.95	0.95	0.94	0.91	0.94
GooAQ	33	0.29	0.04	0.06	0.00	0.07	0.14	0.11	0.12	0.00	0.15	14,629	0.17	0.15	0.15	0.00	0.19
MS MARCO QnA	2,558	0.44	0.19	0.23	0.05	0.24	0.49	0.40	0.39	0.10	0.41	2,557	0.45	0.42	0.39	0.13	0.40
Natural Questions	71	0.14	0.05	0.06	0.01	0.07	0.34	0.37	0.33	0.18	0.34	121	0.37	0.34	0.32	0.16	0.33
ELI5	13,104	0.26	0.04	0.06	0.00	0.08	0.16	0.09	0.10	0.00	0.12	13,104	0.16	0.09	0.10	0.00	0.12
SearchQA	117	0.20	0.22	0.20	0.15	0.20	0.63	0.64	0.62	0.53	0.62	78	0.55	0.54	0.54	0.47	0.54
SQuAD v.2.0	252	0.79	0.81	0.78	0.63	0.78	0.84	0.84	0.83	0.66	0.83	321	0.96	0.96	0.95	0.93	0.95
NewsQA	29	0.57	0.55	0.53	0.31	0.53	0.65	0.56	0.58	0.45	0.58	66	0.76	0.76	0.73	0.58	0.73
HotpotQA	35	0.49	0.39	0.40	0.14	0.40	0.60	0.55	0.53	0.26	0.54	39	0.73	0.73	0.73	0.67	0.72
TriviaQA	66	0.37	0.35	0.34	0.26	0.35	0.43	0.41	0.40	0.27	0.40	71	0.44	0.43	0.42	0.28	0.42
Macro-averaged	93,226	0.43	0.35	0.35	0.23	0.35	0.52	0.49	0.48	0.34	0.49	107,947	0.55	0.54	0.53	0.41	0.53
Micro-averaged	93,226	0.70	0.72	0.68	0.58	0.68	0.82	0.81	0.81	0.75	0.81	107,947	0.73	0.72	0.72	0.65	0.73

Conclusions

- Dataset with 1.1M QA pairs to advance research in causal QA.
- Taxonomy of causal questions and rules to identify them.
- Baseline QA systems on the constructed dataset.

Future work:
- Combine text matching QA systems with causal inference.

github.com/webis-de/COLING-22