Backtracing: Retrieving the Cause of the Query

Rose E. Wang* Pawan Wirawarn Omar Khattab Noah Goodman Dorottya Demszky

*rewang@cs.stanford.edu, Stanford University

Motivation

Backtracing: Given the corpus X and query q, retrieve the sentence that most likely caused the query.

Corpus X (e.g., lecture transcript) **Query** q (e.g., student question)

I have a question, if I project the projection again that's the same point that is P^2=P. But if I keep doing such it should tell P^3=P^4=P^n=P, and this property holds for Identity matrix. Is my logic correct?

What did I say that triggered this student's question?

[...] The projection is the same point. So that means that if I project twice, I get the same answer as I did in the first project. So those are the two properties that tell me I'm looking at a projection matrix. [...]

- While information retrieval (IR) systems may provide answers for user queries, they do not directly assist content creators (e.g., teachers) identify segments that caused a user to ask those questions.
- Identifying the cause of a query is challenging because of **1.** lack of explicit labeling, **2.** large size of corpus, and **3.** required domain expertise to understand both the query and corpus.
- We introduce the task of ***backtracing***, in which systems retrieve the text segment that most likely caused a user query.

Contributions

Task. We formalize *backtracing*: Retrieve the text segment that most likely caused the user query.

Benchmark. We develop a heterogeneous benchmark for backtracing: retrieving the cause of student confusion in the LECTURE setting, reader curiosity in the NEWS ARTICLE setting, and user emotion in the CONVERSATION setting.

Evaluations. We evaluate a suite of popular retrieval systems and show that there is room for improvement in current retrieval methods. This suggests that backtracing is not only challenging but also requires new retrieval approaches.

Backtracing Task

Given corpus of N sentences $X = \{x_1, \ldots, x_N\}$ and query q, backtracing selects

$$\hat{t} = \arg\max_{t \in 1 \ldots N} p(t | x_1, \ldots, x_N, q) \quad (1)$$

where x_t is the t^{th} sentence in corpus X and p is a probability distribution over the corpus indices, given the corpus and the query.

This task intuitively translates to: Given a lecture transcript and student question, retrieve the lecture sentence(s) that most likely caused the student to ask that question.

Backtracing Benchmark

	Lecture	News Article	Conversation
Corpus	**Lecturer:** What matrix do I multiply by to get the identity if I have A here? OK, that'll be simple but so basic. [...] This product turns out to be better than this one. Let me take a typical case here. [...] Maybe rather than saying left of A, left of U, let me write down again what I mean. EA is U, whereas A is LU. OK. Let me make the point now in words. The order that the matrices come for L is the right order. The two and the five don't sort of interfere to produce this ten. In the right order, the multipliers just sit in the matrix L. That's the point. That the, so that if I want to know L, I've no work to do. I just keep a record of what those multipliers were. And that gives me L. So I'll draw the, so what's the order? So let me state it. So this is the A equal LU. So if no row exchanges, the multipliers, [...]	**Journalist:** In a last-ditch effort to keep its sales force and customer base, Integrated Resources Inc. said it agreed in principle to transfer ownership of its broker-dealer subsidiary to two of its top executives. The financial-services firm, struggling since summer to avoid a bankruptcy-law filing after missing interest payments on about $1 billion of debt, will retain the right to regain the subsidiary. It said it will exercise that right only if it sells substantially all of its other core businesses. It also can sell the right to regain the subsidiary to another party. Also, the broker-dealer subsidiary, Integrated Resources Equity Corp., was renamed Royal Alliance Associates Inc. Because of Integrated's widely reported [...]	**Person A:** Hi, I made a reservation for mid-size vehicle. The name is Jimmy Fox. **Person B:** I'm sorry, we have no mid-size available at the moment. **Person A:** I don't understand, I made a reservation, do you have my reservation? **Person B:** Yes, we do, unfortunately we ran out of cars. **Person A:** But the reservation keeps the car here. That's why you have the reservation. **Person B:** I know why we have reservations.
Query	**Student:** Can someone explain why A=LU is *better* than EA=U?	**Reader:** Was is necessary to rename the subsidiary?	**Person A (anger):** I don't think you do. If you did, I'd have a car.

Figure 1. The text context that causes the query is the green-highlighted sentence. Popular retrieval systems retrieve incorrect contexts shown in red.

LECTURE. Retrieve the cause of student confusion [3].

NEWS ARTICLE. Retrieve the cause of reader curiosity [1].

CONVERSATION. Retrieve the cause of user emotion (e.g., anger) [2].

	# sentences	Lecture	News Article	Conversation
Corpus X	Total	11042	2125	8263
	Average	525.8	19.0	12.3
Query q	Total	210	1382	671
	Average	30.9	7.1	11.6

Results

- **The best-performing models achieve modest accuracies.** Measuring causal relevance is challenging and markedly different from existing retrieval tasks.
- **The methods do not generalize across domains.** For instance, while a cross-encoder method performs well on the NEWS ARTICLE domain with top-3 85% accuracy, it only manages top-3 15% accuracy on the CONVERSATION domain.

		Lecture @1	Lecture @3	News Article @1	News Article @3	Conversation @1	Conversation @3
	Random	0	2	6	21	11	31
	Edit	4	8	7	18	1	16
	Bi-Encoder (Q&A)	23	37	48	71	1	32
	Bi-Encoder (`all-MiniLM`)	26	40	49	75	1	37
	Cross-Encoder	22	39	66	**85**	1	15
	Re-ranker	**30**	44	66	**85**	1	21
	`gpt-3.5-turbo-16k`	15	N/A	**67**	N/A	**47**	N/A
Single-sentence $p(q\|x_t)$	GPT2	21	34	43	64	3	46
	GPTJ	23	42	**67**	**85**	5	**65**
	OPT 6B	**30**	**43**	66	82	2	56
Autoregressive $p(q\|x_{\leq t})$	GPT2	11	16	9	18	5	54
	GPTJ	14	24	55	76	8	60
	OPT 6B	16	26	52	73	18	**65**
ATE $p(q\|X) - p(q\|X/\{x_t\})$	GPT2	13	21	51	68	2	24
	GPTJ	8	18	**67**	79	3	18
	OPT 6B	2	6	64	76	3	22

Table 1. **Accuracy in percentage (%).** The best models in each column are bolded. For each dataset, we report the top-1 and 3 accuracies.

More results and analysis in the paper!

References

[1] Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. Inquisitive question generation for high level text comprehension. *arXiv preprint arXiv:2010.01657*, 2020.

[2] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332, 2021.

[3] Rose Wang, Pawan Wirawarn, Noah Goodman, and Dorottya Demszky. Sight: A large annotated dataset on student insights gathered from higher education transcripts. In *Proceedings of Innovative Use of NLP for Building Educational Applications*, 2023.