Precise Zero-Shot Dense Retrieval without Relevance Labels

Luyu Gao[1], **Xueguang Ma**[2], Jummy Lin2, Jamie Callan[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]David R. Cheriton School of Computer Science, University of Waterloo
ACL 2023
2024.08.05

발제자: 윤예준 (yeayen789@gmail.com)

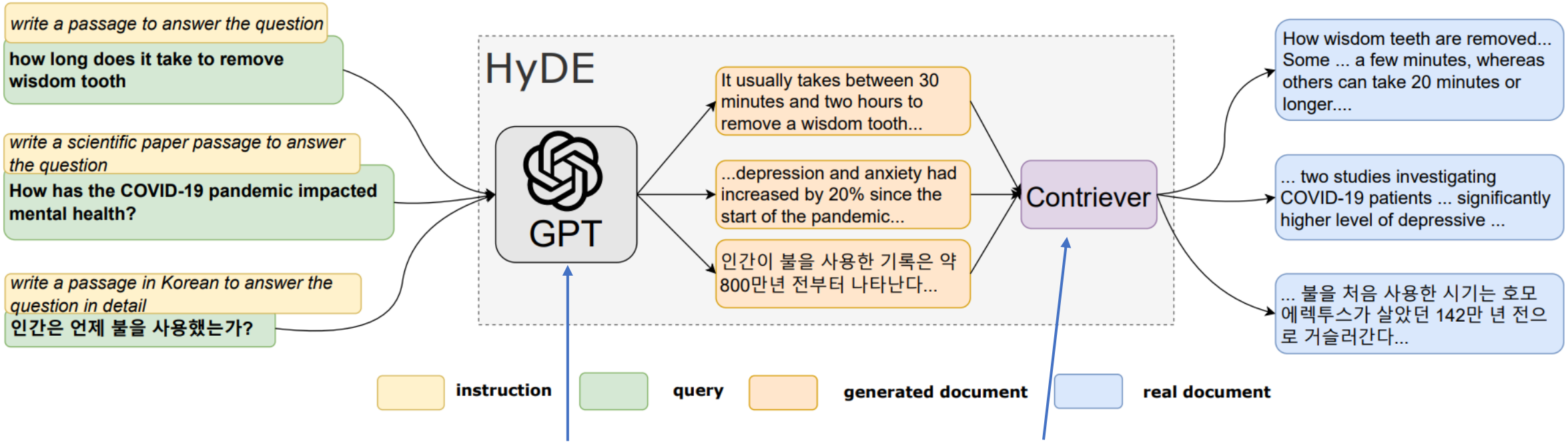
Background

- Semantic embedding similarities를 사용하는 **dense retrieval**은
  web search, question answering, fact verification 등 다양한 작업에서 좋은 성능을 보임
- 그러나 fully **zero-shot** dense retrieval은 여전히 어려움

**Relevance supervision 없이, 즉시 사용 가능하며, 다양한 작업에 효과적인
fully zero-shot dense retrieval system 구축을 목표**

HYDE – Hypothetical Document Embeddings

- Self-supervised representation learning method로부터 시작
  - LLMs는 강력한 NLU, NLG 능력을 가지고 있음
    -> 특히 instruct tuning된 LLM은 다양한 unseen instruction에 대한 zero-shot generalize 능력을 가지고 있음
  - 대조 목적 함수로 학습된 retriever는 inner products로 document간 similarity를 계산할 수 있음



HYDE – Hypothetical Document Embeddings

- Hypothetical document를 생성
- 생성된 document는 **not real**하며 **factual errors**를 포함할 수 있지만 관련 문서와 'like'할 것이라 기대

Please write a passage to answer the question.
Question: what is wifi vs bluetooth
Passage: Wi-Fi and Bluetooth are two different types of wireless technology. Wi-Fi is a local area network technology that uses radio waves to provide high-speed internet and network connections. Bluetooth is a short-range wireless technology that allows for the exchange of data between two devices over short distances.

Please write a scientific paper passage to answer the question.
Question: how long does coronavirus remain stable on surfaces?
Passage: Recent studies have revealed that the novel coronavirus, SARS-CoV-2, can remain stable on surfaces for up to nine days. In a study conducted at the National Institute of Allergy and Infectious Diseases (NIAID), researchers found that the virus was stable on plastic and stainless steel surfaces for up to three days, and on cardboard for up to 24 hours.

Please write a passage in Japanese to answer the question in detail.
Question: 海底ケーブルが初めて結ばれたのはどこ？
Passage: 海底ケーブルが初めて結ばれたのは1866年、イギリスとフランスをつなぐケーブルが結ばれたアイルランドの海岸沿いにあるヴァレンタイン島です。ヴァレンタイン島はアイルランドの西海岸に位置しており、ケーブルはヴァレンタイン島からフランスのブレスト湾まで延びました。このケーブルは、イギリスとフランス間の通信を可能にしました。

HYDE – Hypothetical Document Embeddings

- Unsupervised contrastive encoder $f = \text{enc}_d = \text{enc}_{\text{con}}$
- Document vector $\mathbf{v_d} = f(d) \quad \forall d \in D_1 \cup D_2 \cup ... \cup D_L$
- Hypothetical documents $g(q,\ \text{INST}) = \text{InstructLM}(q,\ \text{INST})$
- Query vector $\mathbb{E}[\mathbf{v}_{q_{ij}}] = \mathbb{E}[f(g(q_{ij}, \text{INST}_i))]$ ➡ $\hat{\mathbf{v}}_{q_{ij}} = \frac{1}{N} \sum_{\hat{d}_k \sim g(q_{ij}, \text{INST}_i)} f(\hat{d}_k)$ ➡ $\hat{\mathbf{v}}_{q_{ij}} = \frac{1}{N+1}[\sum_{k=1}^{N} f(\hat{d}_k) + f(q_{ij})]$

$$= \frac{1}{N} \sum_{k=1}^{N} f(\hat{d}_k)$$

- Document retrieval $\text{sim}(\mathbf{q}_{ij}, \mathbf{d}) = \langle \hat{\mathbf{v}}_{q_{ij}}, \mathbf{v}_d \rangle \quad \forall d \in D_i$

Experiment

- Implementation
  - Zero-shot setting
  - LLM: InstructGPT, GPT-3
  - Retrieval model: Contriever, mContriever

- Datasets
  - Web search: TREC DL19 and DL20 (MS MARCO 기반)
  - Low resource: Scifact(scientific paper abstracts), Arguana(argument retrieval), TREC-COVIDE(COVID-19 scientific papers), FiQA(financial articles), DBPedia(entity retrieval), TREC-NEWS(news articles), Climate-Fever(climate fact verification)
  - Non-English: Swahili, Korean, Japanese and Bengali from Mr.TyDi, TyDiQA

Results

- Web Search
  - BM25가 Contriever보다 성능이 좋지만 HyDE가 BM25보다 높은 성능을 보여줌
  - Supervised와 비교해도 경쟁력을 가지는 것을 알 수 있음

	DL19			DL20		
	mAP	nDCG@10	Recall@1k	mAP	nDCG@10	Recall@1k
Unsupervised						
BM25	30.1	50.6	75.0	28.6	48.0	78.6
Contriever	24.0	44.5	74.6	24.0	42.1	75.4
HyDE	**41.8**	**61.3**	**88.0**	**38.2**	**57.9**	**84.4**
Supervised						
DPR	36.5	62.2	76.9	41.8	**65.3**	81.4
ANCE	37.1	**64.5**	75.5	40.8	64.6	77.6
Contriever-ft	41.7	62.1	83.6	**43.6**	63.2	**85.8**

Results

- Low-Resource Retrieval
  - HyDE의 뛰어난 성능을 보여줌

	Scifact	Arguana	Trec-Covid	FiQA	DBPedia	TREC-NEWS	Climate-Fever
				nDCG@10			
Unsupervised							
BM25	67.9	39.7	**59.5**	23.6	31.8	39.5	16.5
Contriever	64.9	37.9	27.3	24.5	29.2	34.8	15.5
HyDE	**69.1**	**46.6**	59.3	**27.3**	**36.8**	**44.0**	**22.3**
Supervised							
DPR	31.8	17.5	33.2	29.5	26.3	16.1	14.8
ANCE	50.7	41.5	**65.4**	30.0	28.1	38.2	19.8
Contriever-ft	67.7	44.6	59.6	**32.9**	**41.3**	42.8	**23.7**
				Recall@100			
Unsupervised							
BM25	92.5	93.2	**49.8**	54.0	46.8	44.7	42.5
Contriever	92.6	90.1	17.2	56.2	45.3	42.3	44.1
HyDE	**96.4**	**97.9**	41.4	**62.1**	**47.2**	**50.9**	**53.0**
Supervised							
DPR	72.7	75.1	21.2	34.2	34.9	21.5	39.0
ANCE	81.6	93.7	45.7	58.1	31.9	39.8	44.5
Contriever-ft	94.7	97.7	40.7	**65.6**	**54.1**	49.2	**57.4**

Results

- Multilingual Retrieval
  - 다양한 언어에서도 잘 동작하는 것을 보여줌
  - 그러나 LLM의 언어 역량에 따라 성능이 떨어질 수 있음 (가설)

	sw	ko	ja	bn
Unsupervised				
BM25	38.9	28.5	21.2	**41.8**
mContriever	38.3	22.3	19.5	35.3
HyDE	**41.7**	**30.6**	**30.7**	41.3
Supervised				
mDPR	7.3	21.9	18.1	25.8
mBERT	37.4	28.1	27.1	35.1
XLM-R	35.1	32.2	24.8	41.7
mContriever-ft	**51.2**	**34.2**	**32.4**	**42.3**

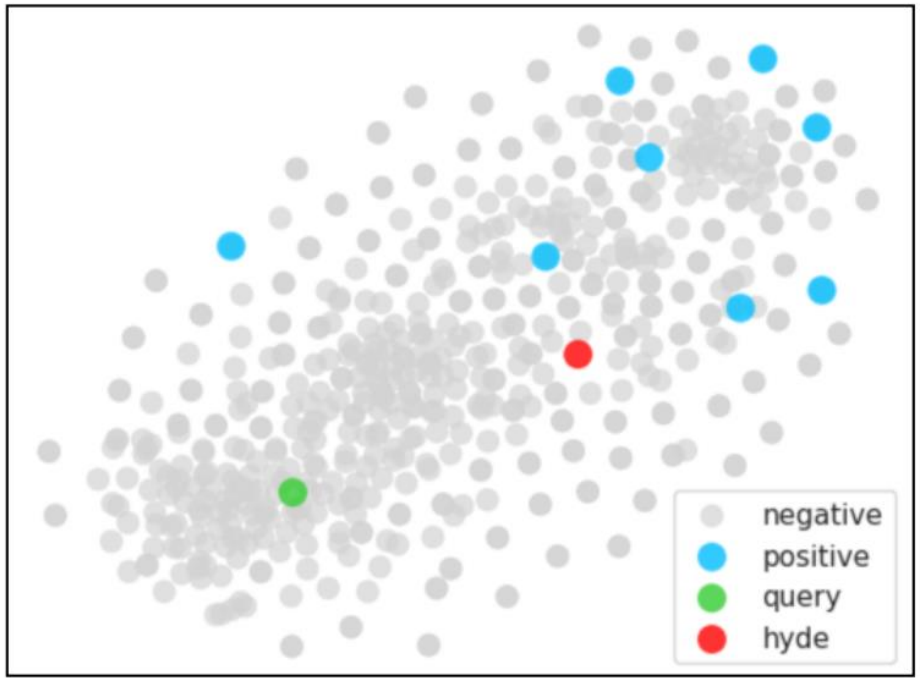
Analysis

Model	DL19		DL20	
	mAP	nDCG@10	mAP	nDCG@10
Contriever	24.0	44.5	24.0	42.1
HyDE				
w/ Flan-T5	32.1	48.9	34.7	52.9
w/ Cohere	34.1	53.8	36.3	53.8
w/ InstructGPT	**41.8**	**61.3**	**38.2**	**57.9**

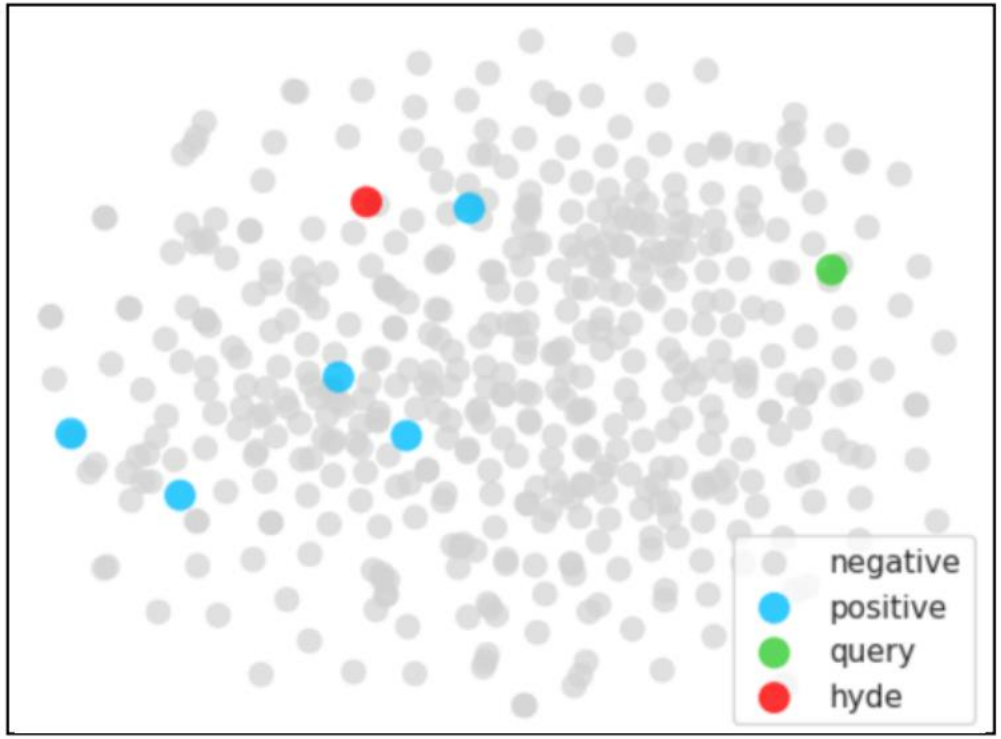
	Scifact	FiQA	DBPedia
Contriever	64.9	24.5	29.2
HyDE			
w/ InstructGPT	**69.1**	27.3	36.8
w/ GPT-3	65.9	**27.9**	**40.5**

Model	DL19		DL20	
	mAP	nDCG@10	mAP	nDCG@10
Contriever-ft	41.7	62.1	43.6	63.2
+ HyDE	48.6	67.4	46.9	63.5
GTR-XL	46.7	69.6	46.9	70.7
+ HyDE	**50.6**	**71.9**	**51.5**	**70.8**

Analysis



(a) Query example from **TREC-COVID**: ***What is the mechanism of inflammatory response and pathogenesis of COVID-19 cases?***



(b) Query example from **DBPedia**: ***Which mountains are higher than the Nanga Parbat?***

결론

- Relevance label 없이 완전한 unsupervised 방식으로 효과적인 dense retrieval 방법인 HyDE 제안
- HyDE는 no other relevance-free model에서는 제공하기 어려운 미세 조정 모델과 비슷한 성능을 제공
- LLM을 통해 가상 문서를 생성해야하기 때문에 high throughput or low latency에서는 맞지 않을 수 있음

느낀점

- 기존 dense retrieval의 문제점을 해결하기 위해 기존에 존재하는 방법들을 잘 적용했다고 생각 됨
- 질적 분석이 추가되었으면 함
  - LLM에 의해서 어떤 정보가 생성되었기에 query의 embedding이 좋아졌는지 모르겠음
  - dense retrieval를 통해 hallucination을 없앨 거라 기대한다고 했으나 분석이 존재하지 않음

Open Questions

- 어떻게 가상 문서를 생성해야 더 좋은 query를 만들 수 있는가?

감사합니다.