Deep Retriever: Information Retrieval for Multi-hop Question Answering

Vera Lin, Leo Mehr, Zijian Wang
{veralin, leomehr, zijwang}@stanford.edu
Stanford University | CS224N | Winter 2019

Problem

- **Open-domain QA**
- **Multi-hop QA**
- **Information retrieval w/** elasticsearch

We propose a deep learning-based IR pipeline that is designed for multi-hop question answering with Elasticsearch.

Task

In which city did Mark Zuckerberg go to college?

Mark Zuckerberg -> Harvard -> Cambridge, MA

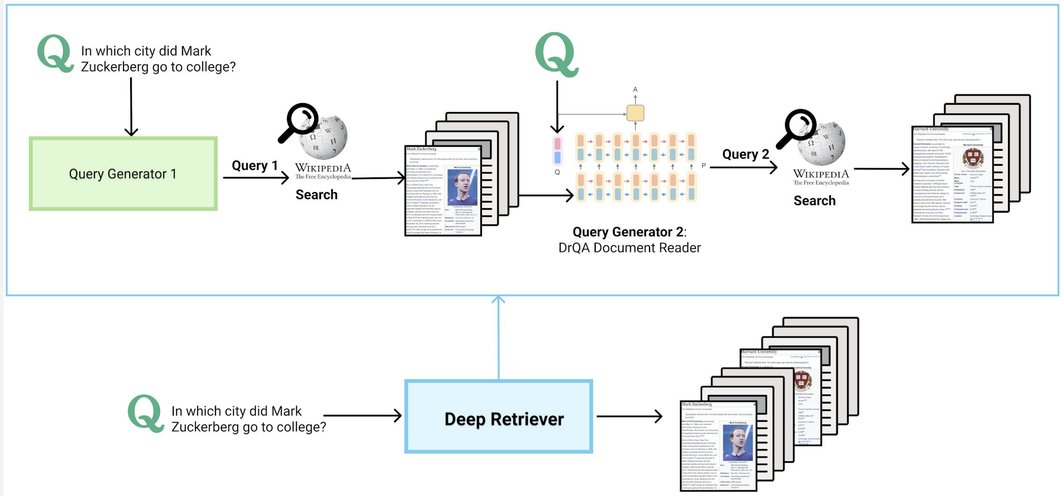
Data

- Adapted HotpotQA dataset

Split	Questions
`train-train`	72,356
`train-dev`	9,044
`train-test`	9,045
`dev`	3,702
`test`	3,703
`Total`	97,850

- 482,021 paragraphs

Approach



End-to-End Performance

Retrieval System	Split	Answer		Sup Fact		Joint	
		EM	F_1	EM	F_1	EM	F_1
HotpotQA	dev	25.14	34.63	4.97	37.04	2.59	16.88
	test	22.95	32.44	3.83	35.84	1.73	15.09
Baseline ES	dev	25.12	34.67	5.97	36.37	2.84	18.26
	test	23.04	32.07	5.64	35.26	2.78	16.75
Deep Retriever	dev	27.99	37.67	6.64	38.79	3.68	20.80
	test	**26.23**	**35.68**	**6.50**	**36.98**	**3.47**	**19.37**

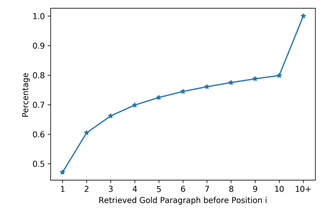
Table 4: End-to-end performance comparison in `full wiki` setting. (cf. Table 4 in [4])

- Baseline ES comparable to HotpotQA
- Deep Retriever boosts performance:
  - ~20% better than Baseline ES, up to ~100% better than HotpotQA

IR Performance

	Split	HotpotQA	Baseline ES	Deep Retriever
Hits@10	dev	56.06	49.55	52.94
	test	55.88	48.29	51.65

Table 3: Information retrieval performance comparisons



Query 2 Generator Performance

Split	Answer-EM	Answer-F_1
train-dev	61.43	67.22
train-test	60.30	65.86

Table 1: Performance comparisons in train

Split	Answer-EM	Answer-F_1
dev	47.63	53.54
test	47.30	52.95

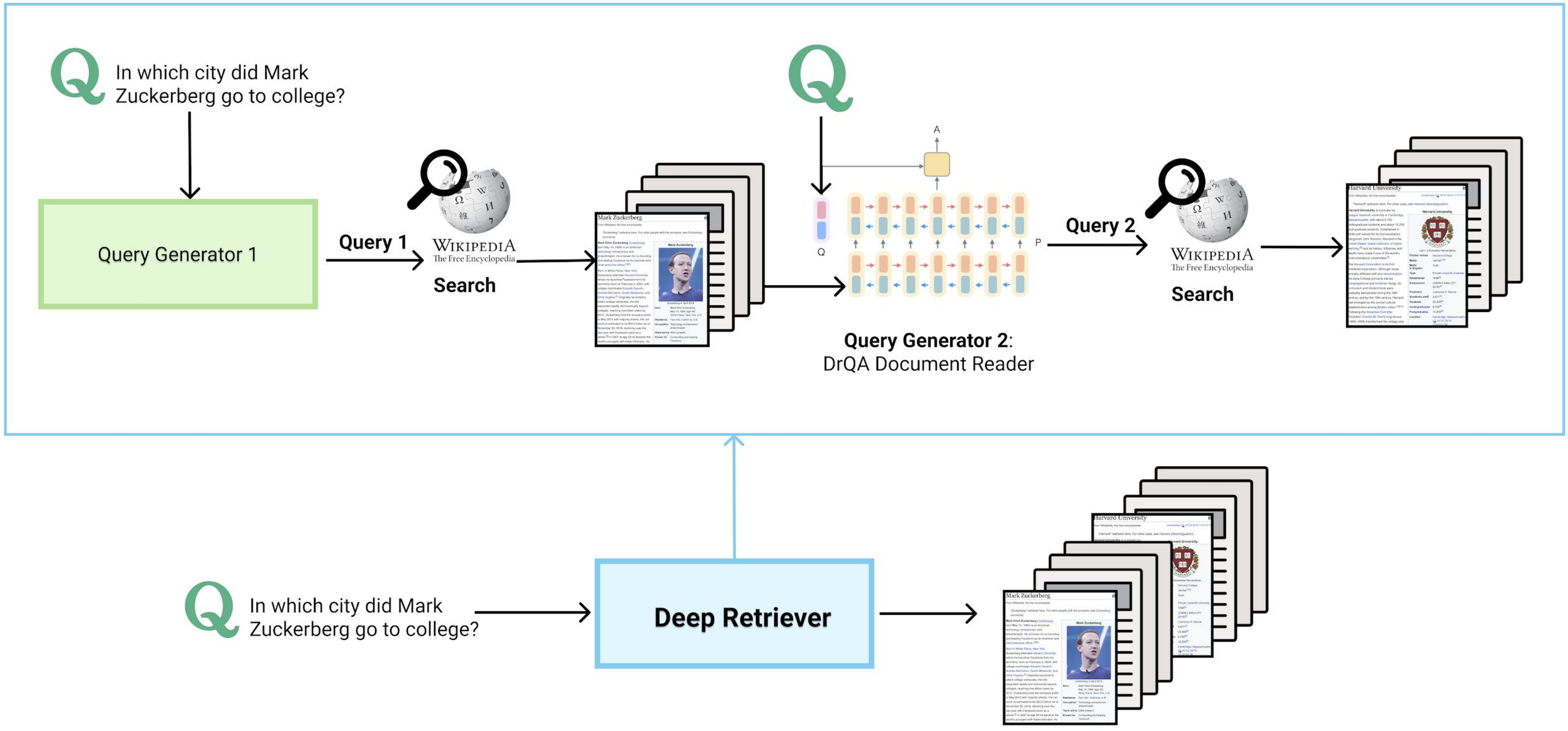
Table 2: Performance comparisons in dev

Conclusions

- Create a *new* dataset with heuristically-inferred labels for query generation tasks
- Build a deep-learning based pipeline to perform *multi-hop* retrievals
- Increase EM by *24.8%* and F1 by *15.6%* over the existing baselines

References

[1]Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1870–1879, 2017.
[2]Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 784–789, 2018.
[3]Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016
[4]Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov,and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018



	Split	HotpotQA	Baseline ES	Deep Retriever
Hits@10	dev	56.06	49.55	52.94
	test	55.88	48.29	51.65

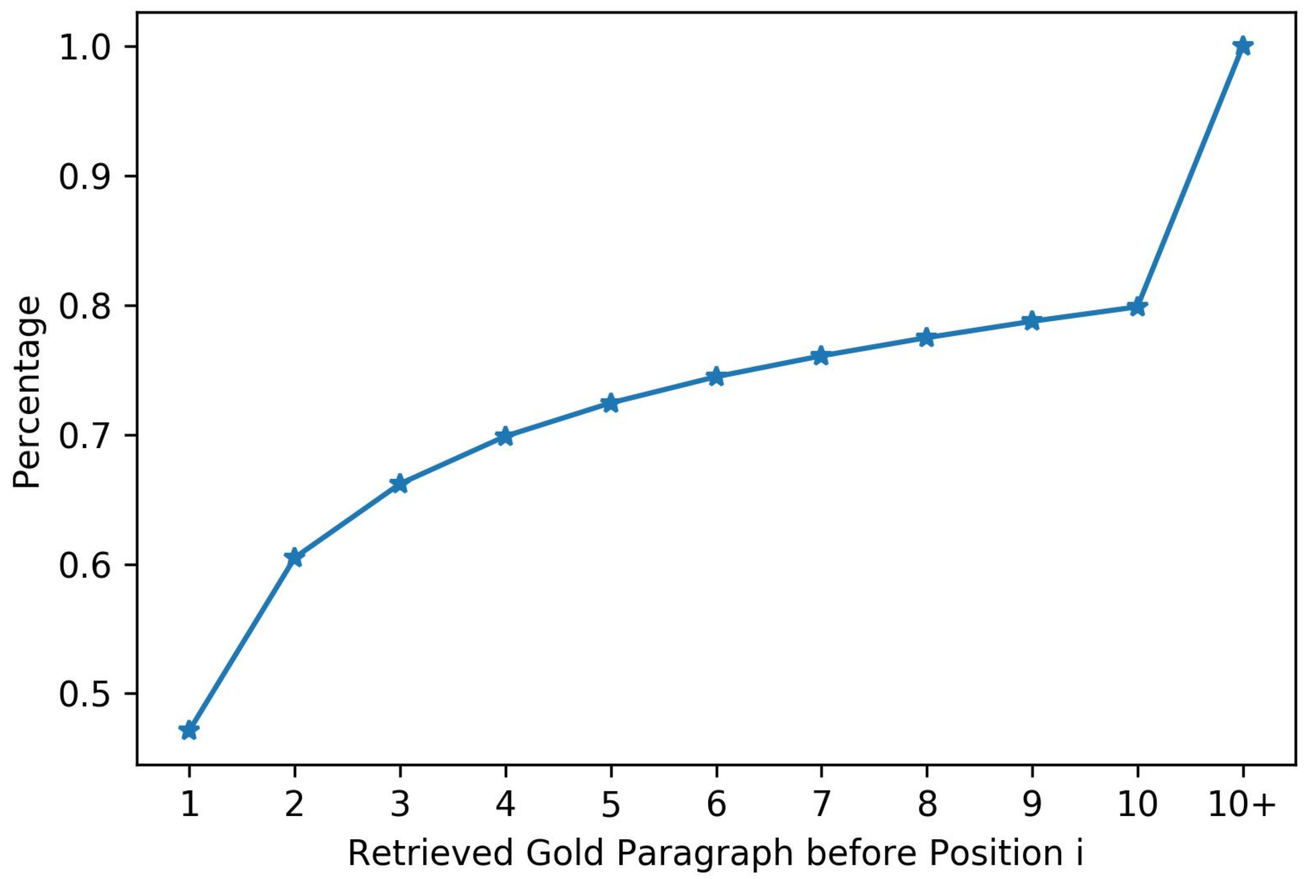
Table 3: Information retrieval performance comparisons

Split	Answer-EM	Answer-F_1
train-dev	61.43	67.22
train-test	60.30	65.86

Table 1: Performance comparisons in train

Split	Answer-EM	Answer-F_1
dev	47.63	53.54
test	47.30	52.95

Table 2: Performance comparisons in dev



elasticsearch

Split	Size
`train-train`	72356
`train-dev`	9044
`train-test`	9045
`dev`	3702
`test`	3703