A Copy-Augmented Generative Model for Open-Domain Question Answering

Shuang Liu[1], Dong Wang[2], Xiaoguang Li[1], Minghui Huang[2], Meizhen Ding[2]

[1] Huawei Noah's Ark Lab

[2] AI Application Research Center (AARC), Huawei Technologies Co., Ltd

ACL 2022

Open-Domain QA – Problem Setup

Open-Domain QA

Answering natural language factoid question from an open set of domains

Characteristics

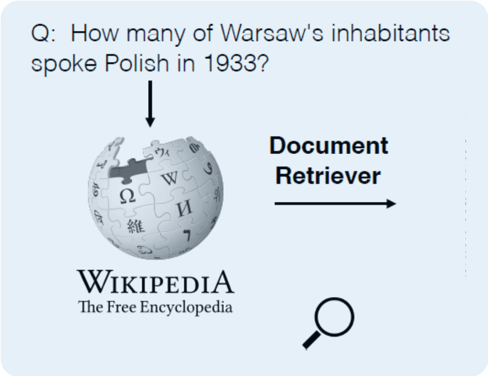
- Wikipedia as knowledge source
- Factoid question answering
- Short and concise answer
- Textual QA

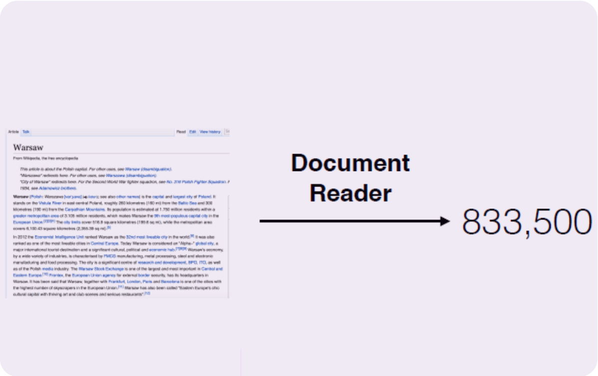
Datasets

- NaturalQuestions [Kwiatkowski et al., 2019]
- TriviaQA [Joshiet al., 2017]

Figure source: [Zhu et al., 2021]

Open-Domain QA – Two-stage Approach





Retriever

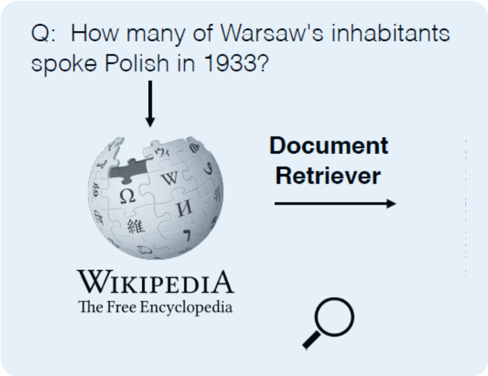
- **Sparse**
  - TF-IDF
  - BM25
- **Dense**
  - BERT

Reader

- **Extractive**
  - BERT
  - ELECTRA
- **Generative**
  - BART
  - T5
  - GPT3

Figure source: [Chen et al., 2017]

Open-Domain QA – Two-stage Approach



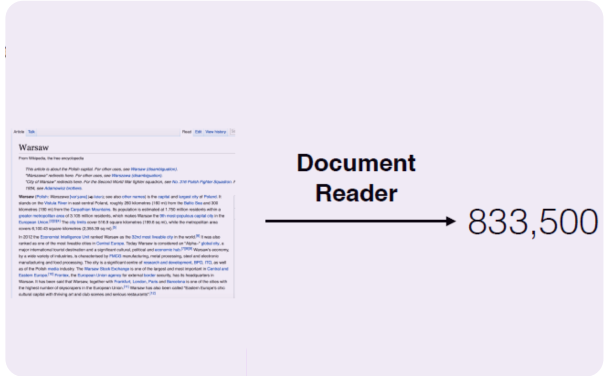
Retriever

Sparse

- TF-IDF
- BM25

Dense

- BERT



Reader

Extractive

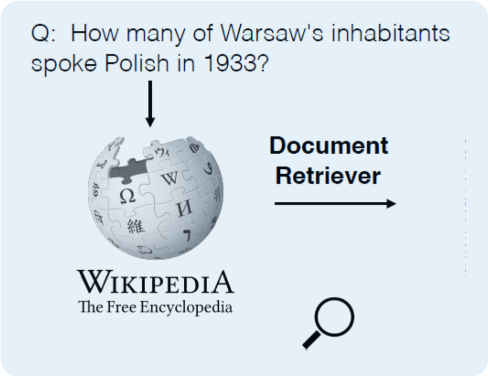
- BERT
- ELECTRA

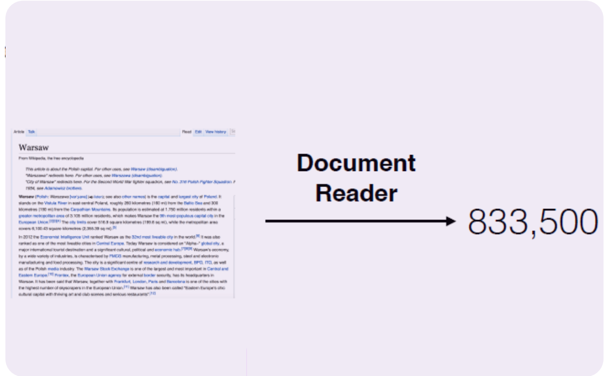
Generative

- BART
- T5
- GPT3

Figure source: [Chen et al., 2017]

Open-Domain QA – Two-stage Approach





Retriever

Sparse
- TF-IDF
- BM25

Dense
- BERT

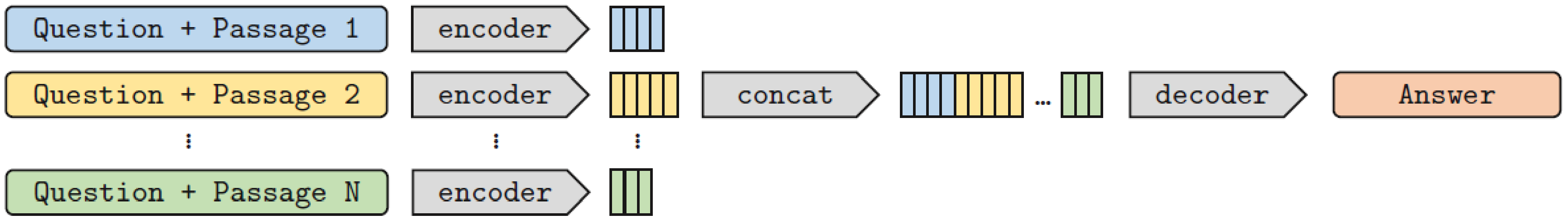
Reader

Extractive
- BERT
- ELECTRA

Generative
- BART
- T5
- GPT3

Figure source: [Chen et al., 2017]

Related Work



FiD [Izacard et al., 2020a]

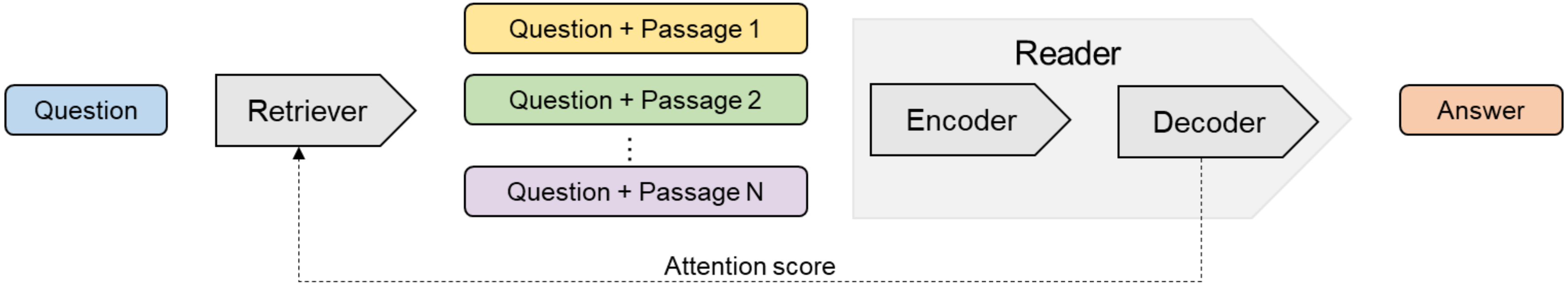
- Fusion-in-Decoder
- Retriever: DPR
- Reader: T5
- Generative model works well on aggregating evidence from multiple passages

Model	NQ	TriviaQA		SQuAD Open	
	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	29.8	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-	53.0	60.9
Path Retriever (Asai et al., 2020)	31.7	-	-	**56.5**	**63.8**
Graph Retriever (Min et al., 2019b)	34.7	55.8	-	-	-
Hard EM (Min et al., 2019a)	28.8	50.9	-	-	-
ORQA (Lee et al., 2019)	31.3	45.1	-	20.2	-
REALM (Guu et al., 2020)	40.4	-	-	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-	36.7	-
SpanSeqGen (Min et al., 2020)	42.5	-	-	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0	-	-
T5 (Roberts et al., 2020)	36.6	-	60.5	-	-
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2	-	-
Fusion-in-Decoder (base)	48.2	65.0	77.1	53.4	60.6
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**	**56.7**	63.2

Table 1: Comparison to state-of-the-art. On TriviaQA, we report results on the open domain test set (left), and on the hidden test set (right), competitions.codalab.org/competitions/17208#results).

[Izacard et al., 2020a]

Related Work



FiD-KD [Izacard et al., 2020b]

- Fusion-in-Decoder
- Retriever: DPR
- Reader: T5
- Leverage attention scores of reader model as synthetic labels for retriever system

Model	NQ dev.	NQ test	TriviaQA dev.	TriviaQA test
DPR (Karpukhin et al., 2020)	-	41.5	-	57.9
RAG (Lewis et al., 2020b)	-	44.5	-	56.1
ColBERT-QA (Khattab et al., 2020)	-	48.2	-	63.2
Fusion-in-Decoder (T5 base) (Izacard & Grave, 2020)	-	48.2	-	65.0
Fusion-in-Decoder (T5 large) (Izacard & Grave, 2020)	-	51.4	-	67.6
Ours (starting from BERT, T5 base)	39.3	40.0	62.5	62.7
Ours (starting from BM25, T5 base)	47.9	48.9	67.7	67.7
Ours (starting from DPR, T5 base)	48.0	49.6	68.6	68.8
Ours (starting from DPR, T5 large)	**51.9**	**53.7**	**71.9**	**72.1**

Table 2: Comparison to state-of-the-art models on NaturalQuestions and TriviaQA.

[Izacard et al., 2020b]

Motivation

Question: where was a hologram for the king filmed?

Title: A Hologram for the King (film)
Production was set to begin in first quarter of 2014. Principal photography commenced on March 6, 2014 in Morocco. Filming also took place in Hurghada in Egypt, as well as in Berlin and Düsseldorf in Germany. Shooting wrapped in June 2014.

Answer: Hurghada in Egypt, Berlin and Düsseldorf in Germany

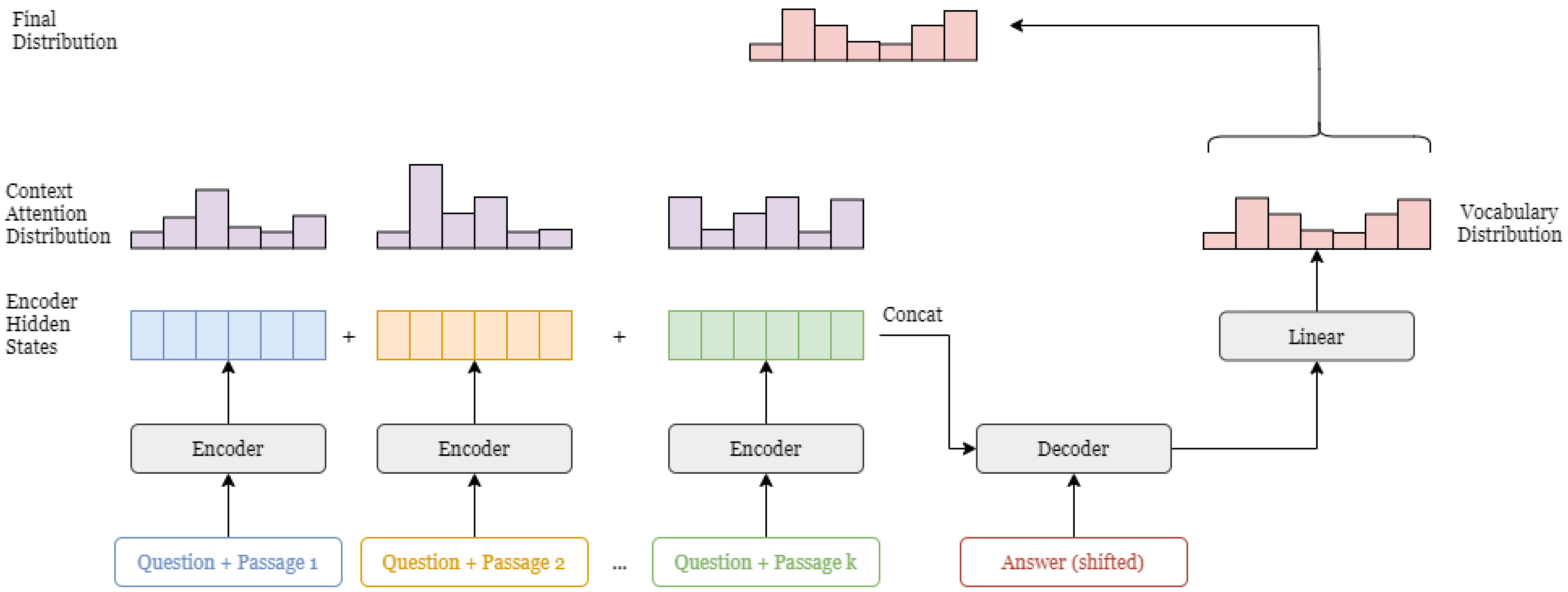
FiD generated: Dubai in Germany

Motivation

- Generative reader
  - Pros
    - Has the ability to generate answer that does not appear in retrieved passages
    - Integrates multi-passages information
  - Cons
    - Hallucination problem (generated text might be factually incorrect or not faithful to the input)
    - Out-of-vocabulary (OOV)
- Extractive reader
    - Consistent with inputs
- Can we combine the extractive and generative readers?

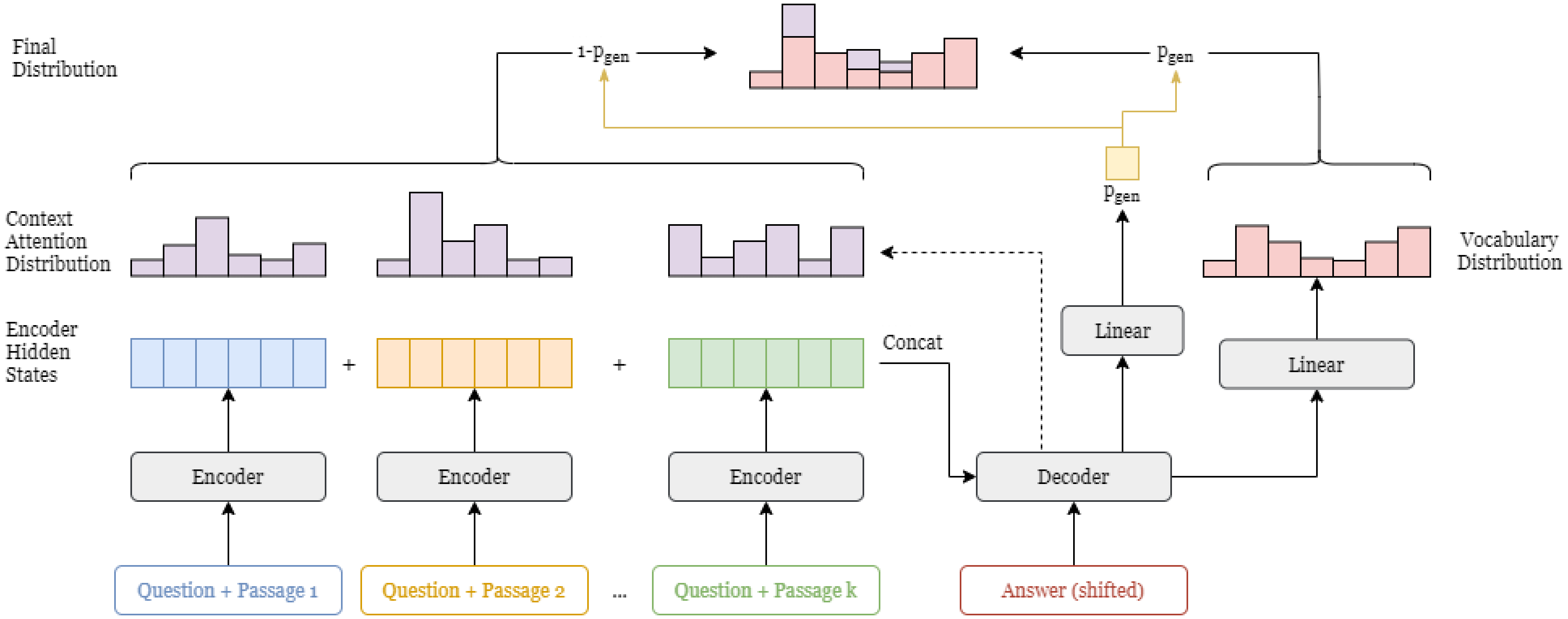
Model -- FiD

- Model architecture of FiD [Izacard et al., 2020]



Model -- Our Approach

- Take advantages of attention scores to help to extract answers from passages
- Adopt Pointer-generator network [See et al., 2017]



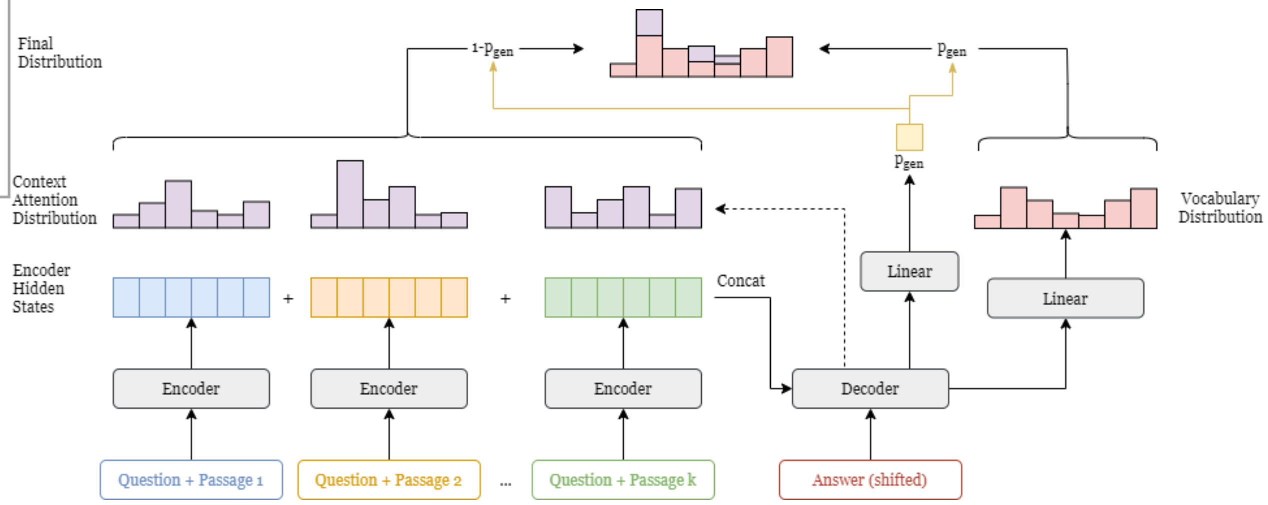
Model -- Our Approach

Probability of y_t $$P(y_t) = p_{gen} P_{vocab}(y_t) + (1 - p_{gen}) P_{ctx}(y_t)$$

p_{gen} $$p_{gen} = \sigma(w_e^T e_t + w_s^T s_t^L + b)$$

$$P_{vocab}(y_t) = softmax(W_E s_t^L)$$

$$P_{ctx}(y_t) = \sum_{j:\, x_{1:k,j} = y_t} \alpha_{t,j}^L$$



Experiments – Main Results

- Exact Match (EM) accuracy on NQ and Trivia datasets
- Achieve SOTA result on NQ and comparable result on Trivia using only ¼ of data as in FiD-KD
- Pointer-generator helps to generate answer accurately from limited number of passages

Model	**Reader Size**	**Top-k**	**NQ**	**TriviaQA**
DPR (BERT-base) (Karpukhin et al., 2020)	110M	24	41.5	57.9
RAG-Seq (BART-large) (Lewis et al., 2020b)	406M	50	44.5	56.8
FiD (T5-base) (Izacard and Grave, 2021b)	220M	100	48.2	65.0
FiD-KD (T5-base) (Izacard and Grave, 2021a)	220M	100	49.6	**68.8**
FiD-KD (Our implementation)	220M	25	48.5	67.5
FiD-PGN	220M	25	**51.4**	68.4

Table 3: Exact match (EM) scores on NQ and TriviaQA test sets. Top-k indicates the number of retrieved passages used during reader training. The performance of SOTA model is in **bold** and the second best model is in underline.

Experiments – Main Results

- Exact Match (EM) accuracy on NQ and Trivia datasets
- Achieve SOTA result on NQ and comparable result on Trivia using only ¼ of data as in FiD-KD
- Pointer-generator helps to generate answer accurately from limited number of passages

Model	**Reader Size**	**Top-k**	**NQ**	**TriviaQA**
DPR (BERT-base) (Karpukhin et al., 2020)	110M	24	41.5	57.9
RAG-Seq (BART-large) (Lewis et al., 2020b)	406M	50	44.5	56.8
FiD (T5-base) (Izacard and Grave, 2021b)	220M	100	48.2	65.0
FiD-KD (T5-base) (Izacard and Grave, 2021a)	220M	100	<u>49.6</u>	**68.8**
FiD-KD (Our implementation)	220M	25	48.5	67.5
FiD-PGN	220M	25	**51.4**	<u>68.4</u>

Table 3: Exact match (EM) scores on NQ and TriviaQA test sets. Top-k indicates the number of retrieved passages used during reader training. The performance of SOTA model is in **bold** and the second best model is in <u>underline</u>.

Experiments -- Generation Probability

- p_{gen} in TriviaQA is always higher than in NQ
- TriviaQA model tend to produce tokens from vocabulary instead of extracting from passages

- Stated in [Rogers et al. (2021)]
  - TriviaQA - probing questions
  - NQ - information-seeking questions
- Our model performs better on information-seeking questions

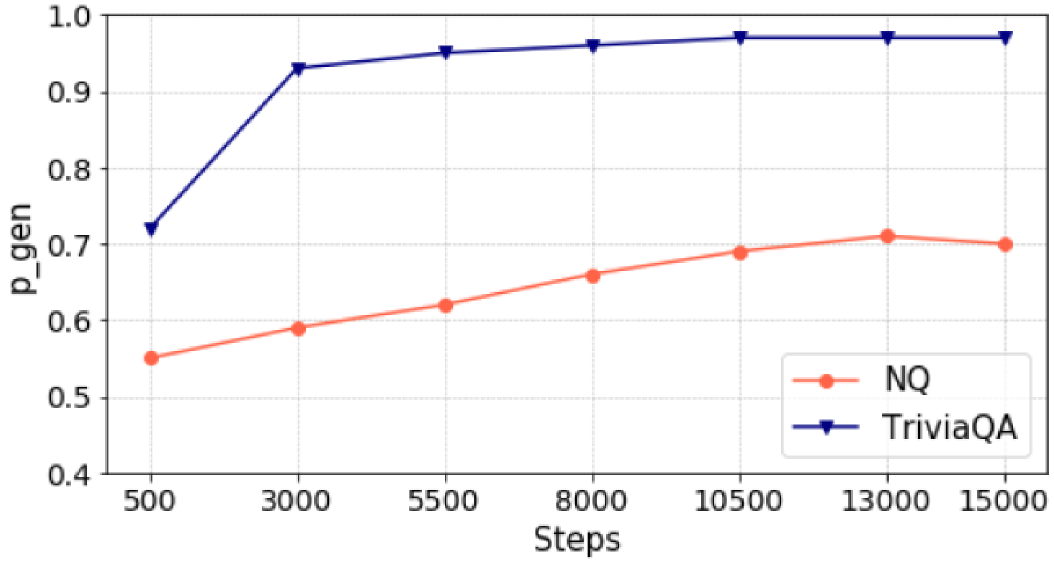


Figure 2: Generation probability p_{gen} over training steps on NQ and TriviaQA.

Experiments -- Test-train Overlap

- Our approach improves most over FiD reader on "No Overlap" category
- Better generalization ability to question answering

Dataset	Overlap Type	FiD	FiD-PGN	Δ
NQ	Total	48.5	**51.4**	2.9
	Question Overlap	73.5	**75.9**	2.4
	Answer Overlap Only	41.0	**45.1**	4.1
	No Overlap	28.8	**38.4**	9.6
TriviaQA	Total	67.5	**68.4**	0.9
	Question Overlap	88.4	**89.6**	1.2
	Answer Overlap Only	66.9	**68.4**	1.5
	No Overlap	41.5	**43.4**	1.9

Table 4: Test-train overlap evaluation on NQ and TriviaQA test sets. Exact match (EM) scores are reported.

Thanks!