

Retrieval Augmented Language Generation

Songwei Dong, Kefan Song

Outline

- Generalization through Memorization: Nearest Neighbor Language Models
- Dense Passage Retrieval for Open-Domain Question Answering
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

GENERALIZATION THROUGH MEMORIZATION: NEAREST NEIGHBOR LANGUAGE MODELS

Urvashi Khandelwal^{†,*}, Omer Levy[‡], Dan Jurafsky[†], Luke Zettlemoyer[‡] & Mike Lewis[‡]

[†]Stanford University

[‡]Facebook AI Research

{urvashik, jurafsky}@stanford.edu

{omerlevy, lsz, mikelewis}@fb.com

Hypothesis: Knowing Text Similarity is Easier than Predicting the Next Token

Consider this next token prediction example:

Obama's birthplace is __

Hypothesis: Knowing Text Similarity is Easier than Predicting the Next Token

Consider this next token prediction example:

Obama's birthplace is ____

Obama was born in ____

Hypothesis: Knowing Text Similarity is Easier than Predicting the Next Token

Consider this next token prediction example:

“Obama’s birthplace is”



High Similarity

“Obama was born in”

Idea: Assist Next Token Prediction with a Datastore and Text Similarity

Consider this next token prediction example:

Obama's birthplace is __



High Similarity

Obama was born in __



Obama was born in Hawaii.

Motivation1: Rare Patterns Can Be Memorized Explicitly

- Transformers are expressive enough to memorize the training data (by training without dropout), but at a cost of generalization.

Motivation1: Rare Patterns Can Be Memorized Explicitly

- Transformers are expressive enough to memorize the training data (by training without dropout), but at a cost of generalization.
- Training transformers with dropout achieves generalization, but may not memorize rare patterns.

Motivation1: Rare Patterns Can Be Memorized Explicitly

- Transformers are expressive enough to memorize the training data (by training without dropout), but at a cost of generalization.
- Training transformers with dropout achieves generalization, but may not memorize rare patterns.
- Text similarity + datastore allow for rare patterns to be memorized and effectively recalled, without compromising the language model's generalization.

Motivation2: Domain Adaptation without Retraining

By updating the datastore on the target domain training set, we can achieve domain adaptation without retraining the Language Model.

Method: KNN-LM

1. Given a context c , retrieve k most similar contexts from the datastore.
2. The targets following these k contexts are used to improve the prediction of the next token for context c .





Training Contexts c_i	Targets v_i
Obama was senator for	Illinois
Barack is married to	Michelle
Obama was born in	Hawaii
...	...
Obama is a native of	Hawaii


Test Context x	Target
Obama's birthplace is	?

Method: KNN-LM

Similarity between contexts are computed by

1. Forward pass all contexts to an LM to obtain vectorized intermediate representations

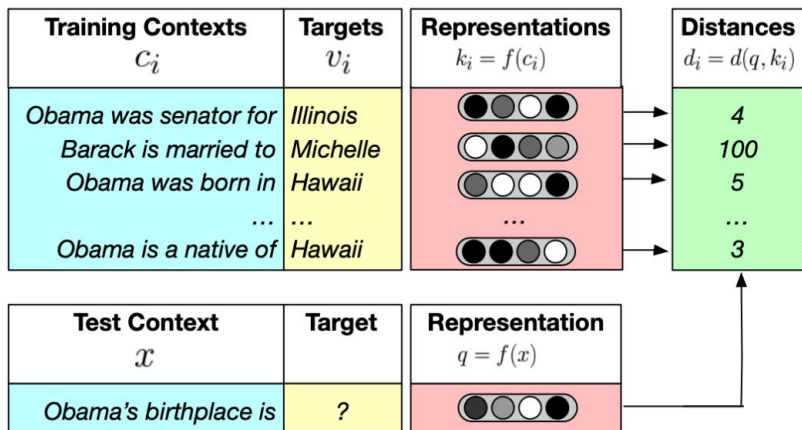
Training Contexts c_i	Targets v_i	Representations $k_i = f(c_i)$
Obama was senator for	Illinois	
Barack is married to	Michelle	
Obama was born in	Hawaii	
...
Obama is a native of	Hawaii	

Test Context x	Target	Representation $q = f(x)$
Obama's birthplace is	?	

Method: KNN-LM

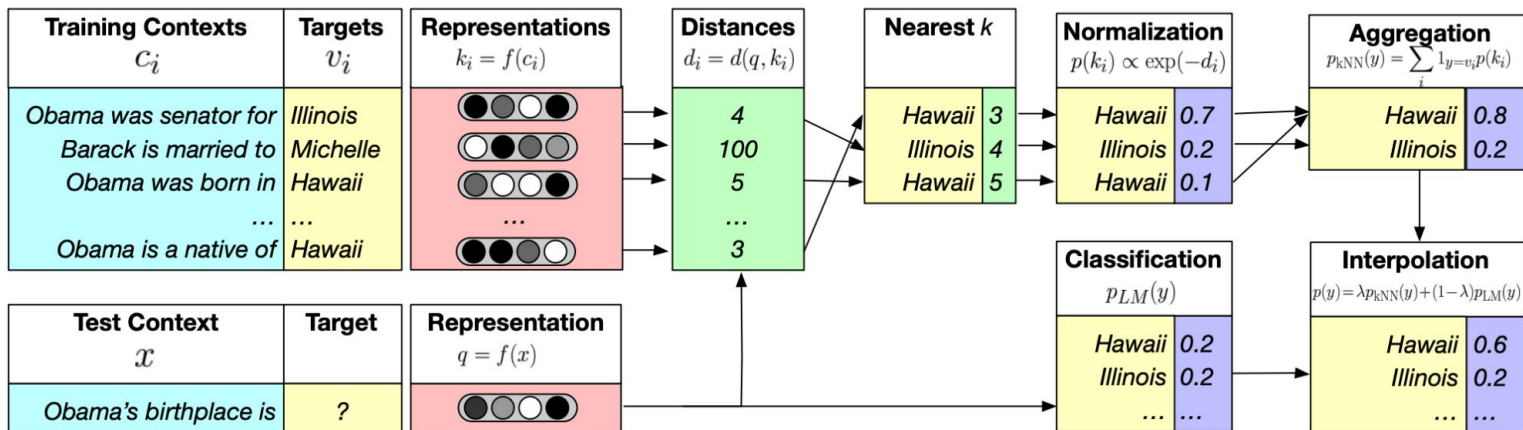
Similarity between contexts are computed by

1. Forward pass all contexts to an LM to obtain vectorized intermediate representations
2. Compute a squared L2 distance between test context and datastore contexts.



Method: KNN-LM

1. Normalize the nearest k targets' distances and aggregate the same targets to obtain the KNN distribution.
2. Perform a linear interpolation of the KNN distribution and the original LM's distribution.



Method: KNN-LM

Normalize the nearest k targets' distances and aggregate the same targets:

$$p_{\text{kNN}}(y|x) \propto \sum_{(k_i, v_i) \in \mathcal{N}} \mathbb{1}_{y=v_i} \exp(-d(k_i, f(x)))$$

Linear interpolation of the KNN distribution and the original LM's distribution:

$$p(y|x) = \lambda p_{\text{kNN}}(y|x) + (1 - \lambda) p_{\text{LM}}(y|x)$$

Results: Using Training Data as Datastore (1/3)

Performance on WIKITEXT-103

Model	Perplexity (\downarrow)		# Trainable Params
	Dev	Test	
Baevski & Auli (2019)	17.96	18.65	247M
+Transformer-XL (Dai et al., 2019)	-	18.30	257M
+Phrase Induction (Luo et al., 2019)	-	17.40	257M
Base LM (Baevski & Auli, 2019)	17.96	18.65	247M
+ k NN-LM	16.06	16.12	247M
+Continuous Cache (Grave et al., 2017c)	17.67	18.27	247M
+ k NN-LM + Continuous Cache	15.81	15.79	247M

Results: Using More Data without Retraining (2/3)

Performance on WIKI-3B

Training Data	Datastore	Perplexity (\downarrow)	
		Dev	Test
WIKI-3B	-	16.11	15.17
WIKI-100M	-	20.99	19.59
WIKI-100M	WIKI-3B	14.61	13.73

Results: Domain Adaptation without Retraining (3/3)

Performance on BOOKS

Model	Perplexity (\downarrow)		# Trainable Params
	Dev	Test	
Base LM (Baevski & Auli, 2019)	14.75	11.89	247M
+ k NN-LM	14.20	10.89	247M

Analysis: Effective for Memorizing Rare Patterns (1/3)

Rare Patterns:

- Factual Knowledge,
- Names,
- near duplicates

Test Context	$(p_{\text{kNN}} = 0.995, p_{\text{LM}} = 0.025)$		Test Target
<i>For Australians and New Zealanders the Gallipoli campaign came to symbolise an important milestone in the emergence of both nations as independent actors on the world stage and the development of a sense of national identity. Today, the date of the initial landings, 25 April, is known as Anzac Day in Australia and New Zealand and every year thousands of people gather at memorials in both nations, as well as Turkey, to...</i>			honour
Training Set Context	Training Set Target	Context Probability	
<i>Despite this, for Australians and New Zealanders the Gallipoli campaign has come to symbolise an important milestone in the emergence of both nations as independent actors on the world stage and the development of a sense of national identity. Today, the date of the initial landings, 25 April, is a public holiday known as Anzac Day in Australia and New Zealand and every year thousands of people gather at memorials in both nations, and indeed in Turkey, to ...</i>	honour	0.995	
<i>On the anniversary date of his death, every year since 1997, thousands of people gather at his home in Memphis to...</i>	celebrate	0.0086	
<i>Twenty-five years after Marseille's death, fighter pilot veterans of World War II gathered to...</i>	honour	0.0000041	

Analysis: Effective for Memorizing Rare Patterns (1/3)

Rare Patterns:

- Factual Knowledge,
- Names,
- near duplicates

Test Context ($p_{\text{kNN}} = 0.959, p_{\text{LM}} = 0.503$)	Test Target	
<i>U2 do what they're best at, slipping into epic rock mode, playing music made for the arena". In two other local newspaper reviews, critics praised the song's inclusion in a sequence of greatest hits. For the PopMart Tour of 1997—...</i>	1998	
Training Set Context	Training Set Target	Context Probability
<i>Following their original intent, "Sunday Bloody Sunday" was not played during any of the forty-seven shows on the Lovetown Tour in 1989. The song reappeared for a brief period during the Zoo TV Tour, and late during the second half of PopMart Tour (1997—...</i>	1998	0.936
<i>They are 6 times Champions and they won the Challenge Cup in 1938, and have experienced two previous stretches in the Super League, 1997—...</i>	2002	0.0071
<i>About \$40 million (\$61.4 million in 2018 dollars) was spent on the property acquisition. After weather-related construction delays due to the El Nino season of the winter of 1997—...</i>	1998	0.0015
<i>This made it the highest-rated season of The X-Files to air as well as the highest rated Fox program for the 1997—...</i>	98	0.00000048

Analysis: Effective for Memorizing Rare Patterns (1/3)

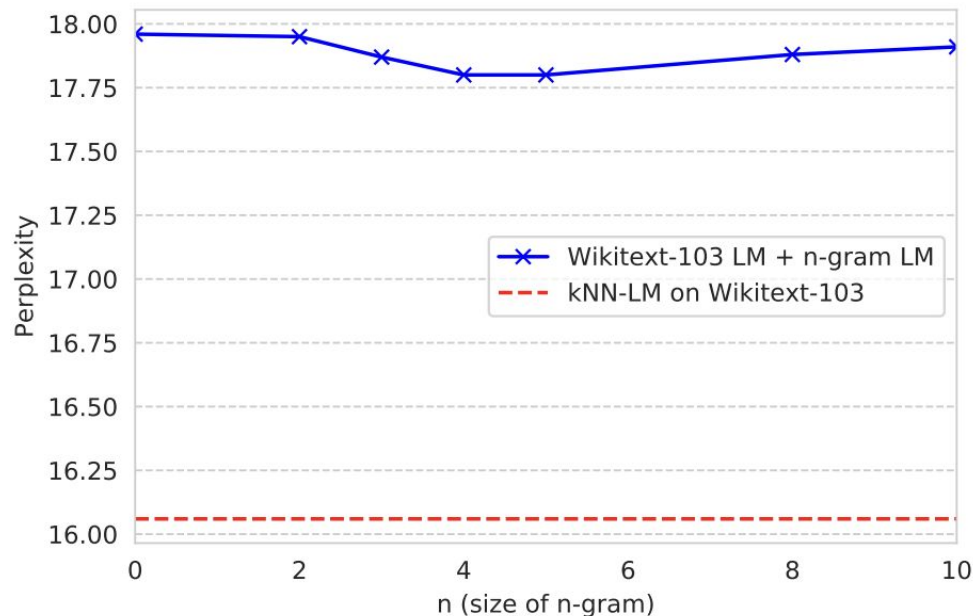
Rare Patterns:

- Factual Knowledge,
- Names,
- near duplicates

Test Context ($p_{\text{kNN}} = 0.624, p_{\text{LM}} = 0.167$)	Test Target	
<i>Lord Strathcona awarded Gauthier a scholarship in 1906 that allowed her to return to Europe and continue her vocal studies. She returned there and continued both to study and give performances. Her first operatic performance came in 1909 in Pavia, Italy as Micaela in Bizet's...</i>	Carmen	
Training Set Context	Training Set Target	Context Probability
<i>Despite poor relations with the orchestra, Mahler brought five new operas to the theatre, including Bizet's...</i>	Carmen	0.356
<i>The fourth movement of An die Jugend (1909), for instance, uses two of Niccolo Paganini's Caprices for solo violin (numbers 11 and 15), while the 1920 piece Piano Sonatina No. 6 (Fantasia da camera super Carmen) is based on themes from Georges Bizet's...</i>	opera	0.0937
<i>It also hosted the Ballet of her Majesty's Theatre in the mid-19th century, before returning to hosting the London premieres of such operas as Bizet's...</i>	Carmen	0.0686

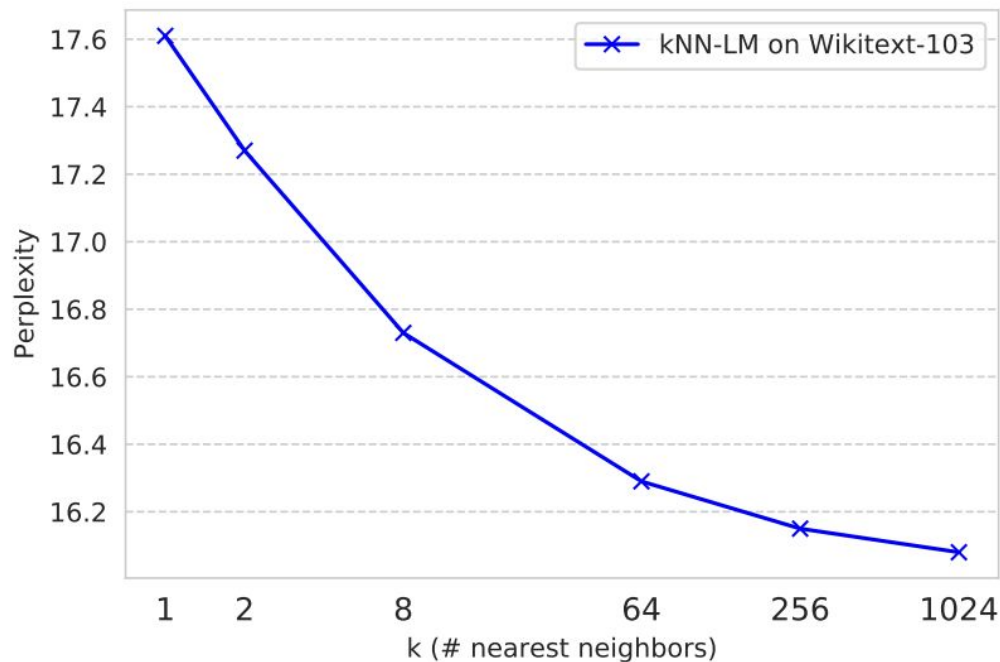
Analysis: How to Best Compute the Similarity? (2/3)

Simple N-Gram vs Neural Representation



Analysis: How to Choose the Number of Nearest Neighbors? (3/3)

More Neighbors Improves Performance



Summary:

- KNN-LM outperforms standard methods by directly querying training examples at test time
- The non-parametric nature allows domain adaptation without training
- Shows that representation learning for text similarity is easier than next token prediction.

Limitations:

- KNN-LM assumes the most similar context from the corpus will be followed by the desired target.
- It is only evaluated on the text prediction task measured by perplexity.
- The datastore construction also incurs additional costs.

Dense Passage Retrieval for Open-Domain Question Answering

**Vladimir Karpukhin^{*}, Barlas Oğuz^{*}, Sewon Min[†], Patrick Lewis,
Ledell Wu, Sergey Edunov, Danqi Chen[‡], Wen-tau Yih**

Facebook AI [†]University of Washington [‡]Princeton University

{vladk, barlaso, plewis, ledell, edunov, scotttyih}@fb.com
sewon@cs.washington.edu
danqic@cs.princeton.edu

Design a more efficient retrieval module for the RAG in open-domain QA tasks.

Why do we need retrieval?

- (Open-Domain QA) is an NLP task where a system answers questions without being restricted to a specific domain. It requires broad knowledge coverage.
- We can not just use all documents for each question generation since model can not process such a large amount of documents.

Previous works: Best Matching 25 (BM25)

The BM25 relevance score between a document D and a query Q is calculated as:

$$\text{score}(D, Q) = \sum_{t \in Q} \text{IDF}(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgD}})}$$

Where:

- $f(t, D)$ = Term frequency (TF) of term t in document D .
- $|D|$ = Length of document D (i.e., total words in the document).
- avgD = Average document length across the collection.
- k_1 = A tuning parameter that controls TF saturation (commonly set between **1.2 and 2**).
- b = A tuning parameter that controls document length normalization (commonly set to **0.75**).

IDF: a metric that measures the importance of a term.

$$IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5} + 1$$

- t is the term (query word).
- N is the total number of documents in the corpus.
- $n(t)$ is the number of documents that contain the term t .

Consider a corpus with **10,000** documents:

- The term "**computer**" appears in **5,000** documents:

$$IDF("computer") = \log \frac{10000 - 5000 + 0.5}{5000 + 0.5} + 1 = \log \frac{5000.5}{5000.5} + 1 = 1$$

- The term "**quantum**" appears in only **10** documents:

$$IDF("quantum") = \log \frac{10000 - 10 + 0.5}{10 + 0.5} + 1 = \log \frac{9990.5}{10.5} + 1$$

The second value is significantly larger than the first, emphasizing the rarity of "quantum."

Motivation: Problems of BM25

A high BM25 score does not always mean that the query and the document are closely related, since BM25 only considers term matching instead of semantic similarity.

Question	Passage received by BM25
What is the body of water between England and Ireland?	Title:British Cycling ... England is not recognised as a region by the UCI, and there is no English cycling team outside the Commonwealth Games. For those occasions, British Cycling selects and supports the England team. Cycling is represented on the Isle of Man by the Isle of Man Cycling Association. Cycling in Northern Ireland is organised under Cycling Ulster, part of the all-Ireland governing body Cycling Ireland . Until 2006, a rival governing body existed, ...

Is there a retrieval method based on semantic similarity?

Dense Passage Retrieval (DPR) for Open-Domain Question Answering

STEPS:

1. Use two independent Encoders to encode query and documents into token sequence and get their [CLS] token.
2. Compute the similarity by using the [CLS] token and retrieve top-k documents

BERT style model

- Encoder: text \longrightarrow tokens
- [CLS] token only exists in BERT style models

After passing through multiple Transformer layers, each token in the sequence is assigned a **final hidden representation**.

$$H = [h_{\text{CLS}}, h_1, h_2, \dots, h_n, h_{\text{SEP}}]$$

where:

- h_{CLS} represents the special [CLS] token's final hidden state.
- h_1, h_2, \dots, h_n are contextual embeddings for individual tokens.

The [CLS] Token Represent the Entire Text

How [CLS] token is processed in BERT models?

During self-attention computations, [CLS] token interacts with all other tokens in the sequence at every Transformer layer

Text1. the most famous basketball player

Text2. Michael Jordan

Compute Similarity

- Use the dot product of [CLS] tokens to measure similarity, which is reasonable.
- The larger the dot product between two vectors, the more aligned their directions are, which implies that they have more similar meanings.
- For each query, the top-k passages with the highest dot product scores are selected as the retrieval results.

$$\text{sim}(q, p) = E_Q(q)^\top E_P(p). \quad (1)$$

Train

$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^- \rangle\}_{i=1}^m$ is the training dataset that consists of m instances.

One relevant (positive) passage and n irrelevant (negative) passages

$$\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-) \\ &= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}. \end{aligned} \tag{2}$$

Experimental Results

Training	Retriever	Top-20					Top-100				
		NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	85.0	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	71.5	83.8	84.5	80.5	92.7	81.3
Multi	DPR	79.4	78.8	75.0	89.1	51.6	86.0	84.7	82.9	93.9	67.6
	BM25 + DPR	78.0	79.9	74.7	88.5	66.2	83.9	84.4	82.3	94.1	78.6

Table 2: Top-20 & Top-100 retrieval accuracy on test sets, measured as the percentage of top 20/100 retrieved passages that contain the answer. *Single* and *Multi* denote that our Dense Passage Retriever (DPR) was trained using individual or combined training datasets (all the datasets excluding SQuAD). See text for more details.

Combined retriever: BM25 + DPR

$$\text{BM25}(q,p) + \lambda \cdot \text{sim}(q,p)$$

Experimental Results

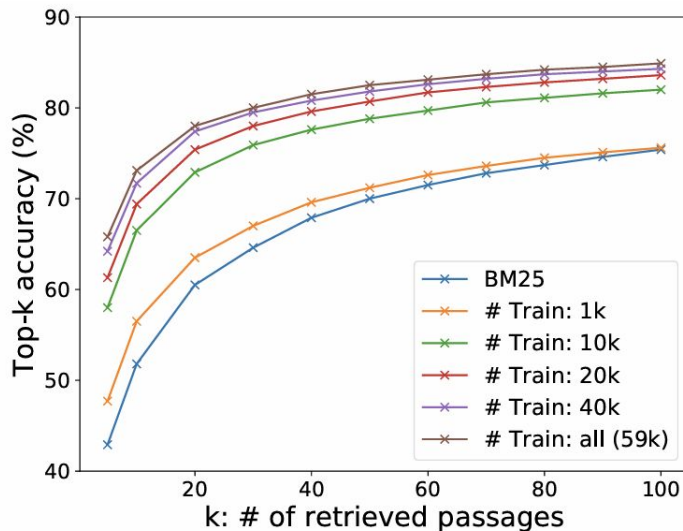


Figure 1: Retriever top- k accuracy with different numbers of training examples used in our dense passage retriever vs BM25. The results are measured on the development set of Natural Questions. Our DPR trained using 1,000 examples already outperforms BM25.

On the **NQ dataset**, DPR outperforms BM25 with only **1,000 training examples**.

As more training examples are used, **DPR's performance continues to improve**.

Limitations:

- In some cases BM25 has stronger performance than DPR

DPR relies on semantic matching, and if the model doesn't learn the semantics of some particular terms, it may miss relevant documentation.

- DPR is much slower than BM25, especially when database is large
- DPR needs to train a model and it also needs high quality query-passage pairs

Summary:

- **DPR enables better semantic retrieval**

Uses a **dual-encoder architecture**, encoding **queries and passages separately**.

Leverages **BERT-based Transformers**, allowing for **semantic understanding beyond keyword matching**.

- **DPR vs. BM25: Trade-offs**

DPR: Uses **dense vectors** for semantic retrieval but it requires a higher computation cost.

BM25: Uses **sparse retrieval (keyword matching)**, making it **faster and more scalable** for large databases.

Conclusion: DPR is better for open-domain QA & semantic search, while **BM25** excels in large-scale retrieval.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com

Motivation1: Knowledge Updates and Manipulation are Difficult for Pre-trained Language Models

PLMs that rely on the parametric knowledge solely have the following issues:

1. PLMs cannot easily expand or revise their memory
2. PLMs do not provide a grounded explanation for their generation
3. PLMs may produce “hallucinations”

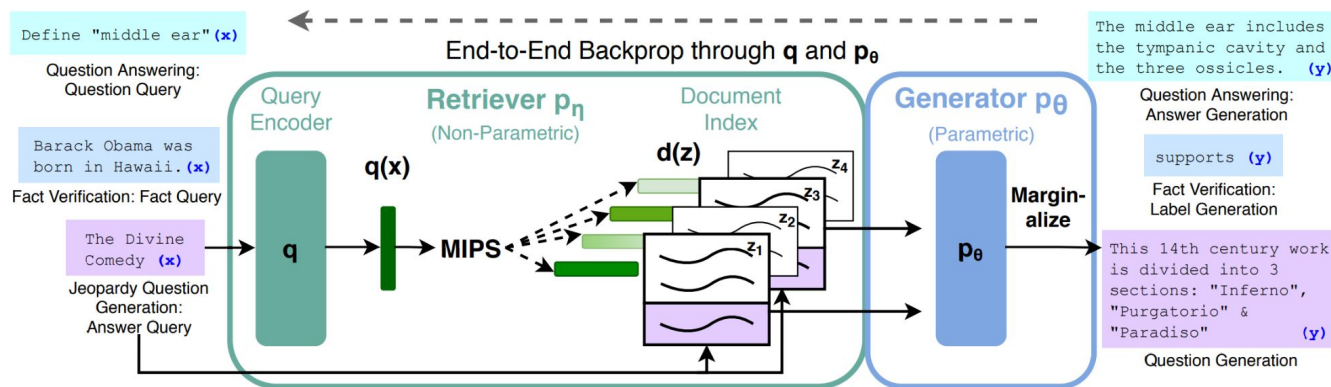
Method: Retrieval Augmented Generation (RAG)

Parametric Memory Generation Model:

Pre-trained seq2seq transformer, i.e. BART

Non-parametric Memory:

Dense vector index of Wikipedia + a neural retriever based on DPR



Retriever: Dense Passage Retrieval (DPR)

The probability of retrieving document z for input x is related to the similarity of the two neural representations.

$$p_{\eta}(z|x) \propto \exp(\mathbf{d}(z)^{\top} \mathbf{q}(x)) \quad \mathbf{d}(z) = \text{BERT}_d(z), \quad \mathbf{q}(x) = \text{BERT}_q(x)$$

Training:

Fine-tune the **query encoder** and the **BART generator** with backprop to minimize:

$$\sum_j -\log p(y_j|x_j)$$

RAG-Sequence Model

Using the same retrieved document to generate the complete sequence

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_{\eta}(z|x) p_{\theta}(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_{\eta}(z|x) \prod_i^N p_{\theta}(y_i|x, z, y_{1:i-1})$$

RAG-Token model

Many documents could be used for generating each token

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_{\eta}(z|x) p_{\theta}(y_i|x, z, y_{1:i-1})$$

Decoding RAG-Sequence Model:

Thorough Decoding

1. For each document z , obtain Y_z by Beam Search
2. Collect all sequences $Y = \bigcup_z Y_z$
3. Run forward pass to obtain the posterior probability $p_\theta(y \mid x, z')$ for every sequence.
4. Marginalize over z to estimate the probability of a sequence y given input x

$$p(y \mid x) = \sum_z p_\eta(z \mid x) \cdot p_\theta(y \mid x, z)$$

Decoding RAG-Sequence Model:

Fast Decoding

1. For each document z , obtain Y_z by Beam Search
2. Collect all sequences $Y = \bigcup_z Y_z$
3. Skip forward pass by assigning $p_\theta(y \mid x, z') = 0$ for sequences not generated by z .
4. Marginalize over z to estimate the probability of a sequence y

$$p(y \mid x) = \sum_z p_\eta(z \mid x) \cdot p_\theta(y \mid x, z)$$

Decoding RAG-Token Model:

Beam Search with the per-token likelihood

$$p'_\theta(y_i|x, y_{1:i-1}) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z_i|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

Results: Open-Domain QA (1/3)

Dataset: TriviaQA (TQA)

Task: Reading Comprehension

Question: “Who won Super Bowl XX?”

Gold passage: “Super Bowl XX was an American football game between the National Football Conference (NFC) champion Chicago Bears and the American Football Conference (AFC) champion New England Patriots ...”

Answer: “Chicago Bears”

Results: Open-Domain QA (1/3)

Dataset: TriviaQA (TQA)

Task: Open-Domain QA

Question: “Who won Super Bowl XX?”

Answer: “Chicago Bears”

Results: Open-Domain QA (1/3)

- Natural Questions (NQ)
- TriviaQA (TQA)
- WebQuestions (WQ)
- CuratedTrec (CT)

	Model	NQ	TQA	WQ	CT
Closed	T5-11B [52]	34.5	- / 50.1	37.4	-
Book	T5-11B+SSM[52]	36.6	- / 60.5	44.7	-
Open	REALM [20]	40.4	- / -	40.7	46.8
Book	DPR [26]	41.5	57.9 / -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	45.5	50.0
	RAG-Seq.	44.5	56.8/ 68.0	45.2	52.2

Results: Beyond Extractive QA (2/3)

- Abstractive QA: MSMARCO
- Jeopardy Question Generation: SearchQA
- Fact Verification: FEVER

Model	Jeopardy		MSMARCO		FVR3	FVR2
	B-1	QB-1	R-L	B-1	Label	Acc.
SotA	-	-	49.8*	49.9*	76.8	92.2*
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	17.3	22.2	40.1	41.5	72.5	<u>89.5</u>
RAG-Seq.	14.7	21.4	<u>40.8</u>	<u>44.2</u>		

Results: Beyond Extractive QA (2/3)

- RAG approaches the model with gold passages.
- RAG outperforms models without gold passages.

Model	Jeopardy		MSMARCO		FVR3	FVR2
	B-1	QB-1	R-L	B-1	Label	Acc.
SotA	-	-	49.8*	49.9*	76.8	92.2*
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	17.3	22.2	40.1	41.5	72.5	<u>89.5</u>
RAG-Seq.	14.7	21.4	<u>40.8</u>	<u>44.2</u>		

Results: Human Assessment of Jeopardy (3/3)

RAG generates answers with better factuality and specificity.

	Factuality	Specificity
BART better	7.1%	16.8%
RAG better	42.7%	37.4%

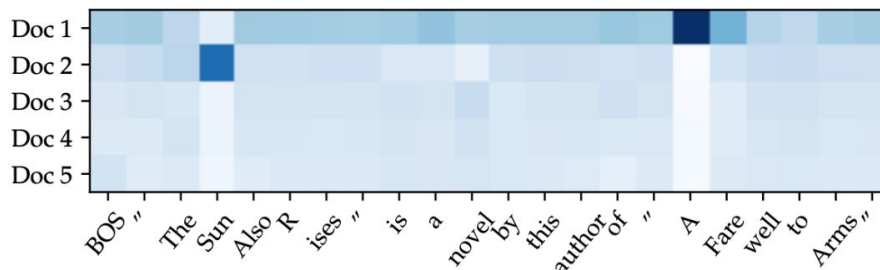
Analysis: Retriever Decides the Answer (1/3)

Different Factual Answers rely on different documents:

The non-parametric component helps to guide the generation

Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel "**A Farewell to Arms**" (1929) ...

Document 2: ... artists of the 1920s "Lost Generation" expatriate community. His debut novel, "**The Sun Also Rises**", was published in 1926.



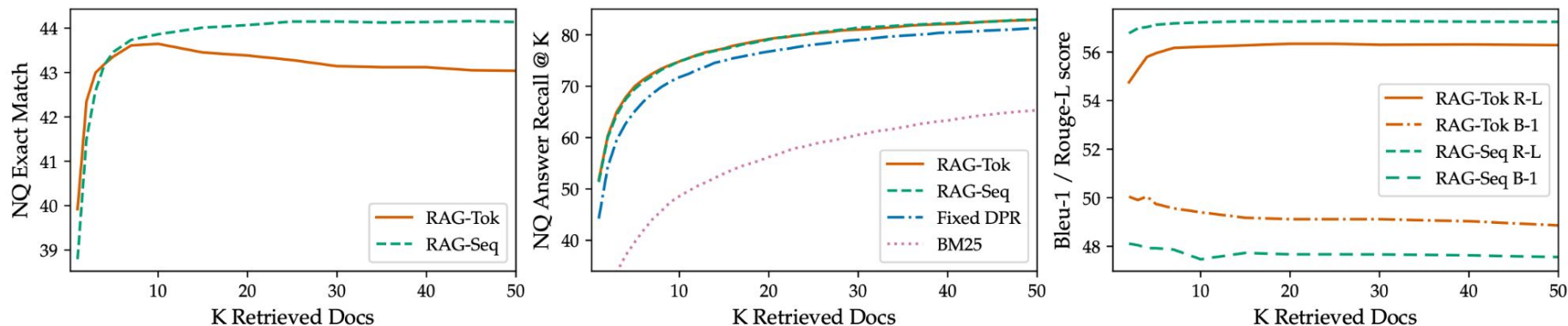
Analysis: Fine-tuning Retriever is Essential (2/3)

RAG with a frozen retriever has worse performance

Model	NQ	TQA Exact Match	WQ	CT	Jeopardy-QGen B-1	QB-1	MSMarco R-L	B-1	FVR-3 Label Accuracy	FVR-2
RAG-Token-BM25	29.7	41.5	32.1	33.1	17.5	22.3	55.5	48.4	75.1	91.6
RAG-Sequence-BM25	31.8	44.1	36.6	33.8	11.1	19.5	56.5	46.9		
RAG-Token-Frozen	37.8	50.1	37.1	51.1	16.7	21.7	55.9	49.4	72.9	89.4
RAG-Sequence-Frozen	41.2	52.1	41.8	52.6	11.8	19.6	56.7	47.3		
RAG-Token	43.5	54.8	46.5	51.9	17.9	22.6	56.2	49.4	74.5	90.6
RAG-Sequence	44.0	55.8	44.9	53.4	15.3	21.5	57.2	47.5		

Analysis: Minimal Effect of Retrieving more documents (3/3)

Performance peaks for RAG-Token at 10 retrieved documents.



Summary:

- We introduced RAG a hybrid generation model with access to parametric and non-parametric memory
- RAG can be fine-tuned end-to-end and achieves state of the art performance on open-domain QA
- Human prefers RAG over BART for its factual and specific answers.

Limitations:

- Current experiments show that achieving the best performance requires a uniquely fine-tuned retriever for each task.
- Both RAG-Sequence and RAG-token do not update the retrieved documents according to the new generated text.
- The retriever will retrieve documents “greedily” during test time, relying on a single query embedding, whereas human may first explore documents, then reflect and refine their query to retrieve the more relevant documents for the question.