Retrieval-based LMs-I

Large Language Models: Introduction and Recent Advances

ELL881 · AIL821

Yatin Nandwani
Research Scientist, IBM Research

Outline

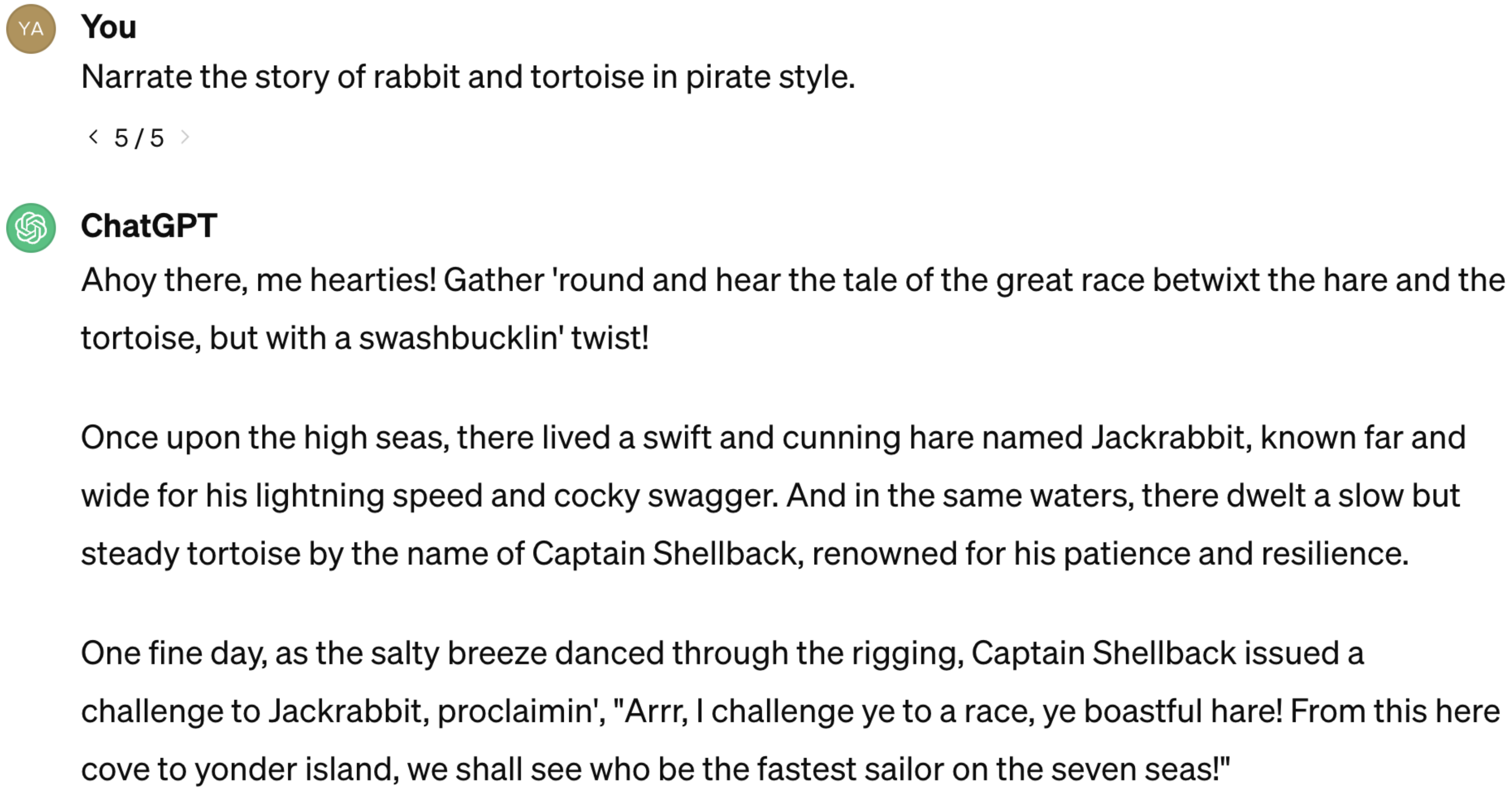
- Motivation
    - Drawbacks of Parametric LLMs – *hallucination, verification …*
    - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- kNN, RETRO, REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training …*
- Limitations – *lost in the middle, still hallucinating, retriever failures …*

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- kNN, RETRO, REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training …*
- Limitations – *lost in the middle, still hallucinating, retriever failures …*

Parametric LLMs

Parametric LLMs – amazing creative writers!



You

Narrate the story of rabbit and tortoise in pirate style.

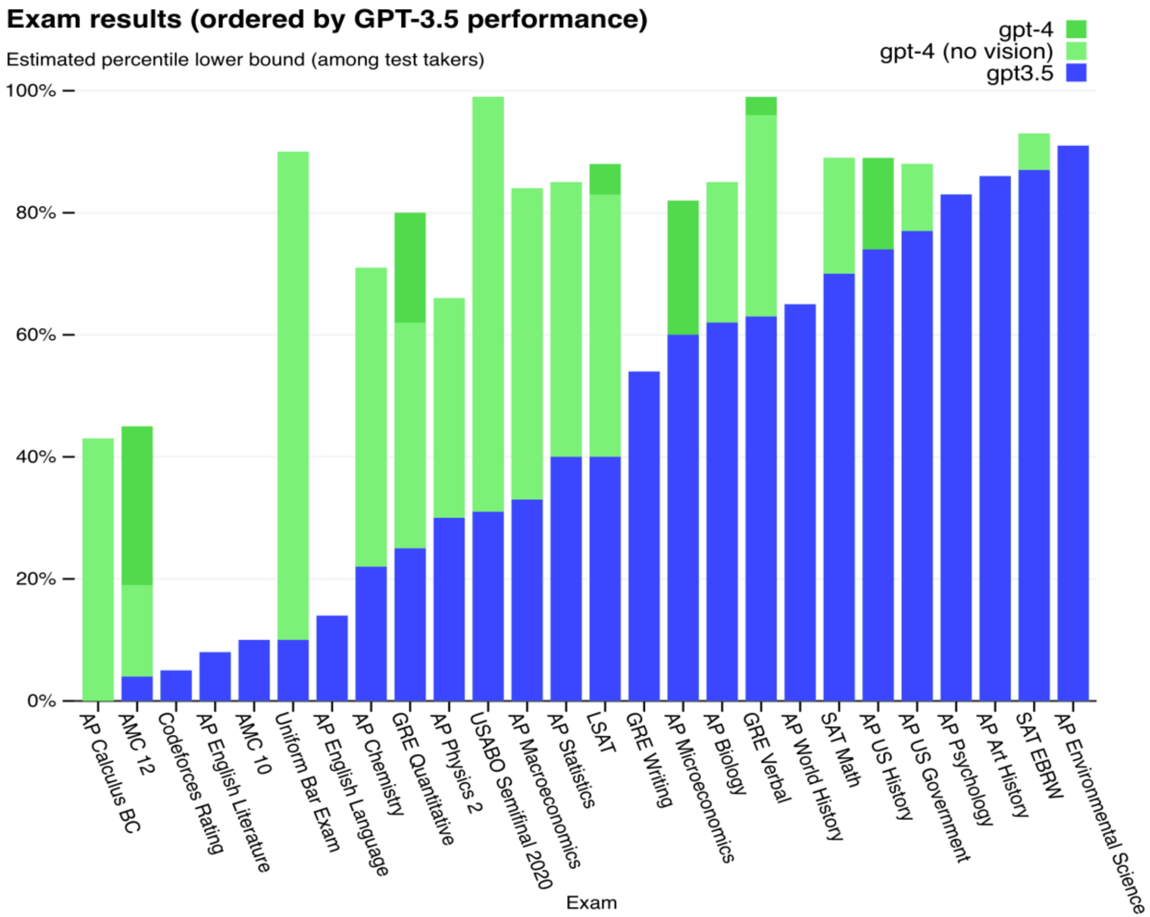
ChatGPT

Ahoy there, me hearties! Gather 'round and hear the tale of the great race betwixt the hare and the tortoise, but with a swashbucklin' twist!

Once upon the high seas, there lived a swift and cunning hare named Jackrabbit, known far and wide for his lightning speed and cocky swagger. And in the same waters, there dwelt a slow but steady tortoise by the name of Captain Shellback, renowned for his patience and resilience.

One fine day, as the salty breeze danced through the rigging, Captain Shellback issued a challenge to Jackrabbit, proclaimin', "Arrr, I challenge ye to a race, ye boastful hare! From this here cove to yonder island, we shall see who be the fastest sailor on the seven seas!"

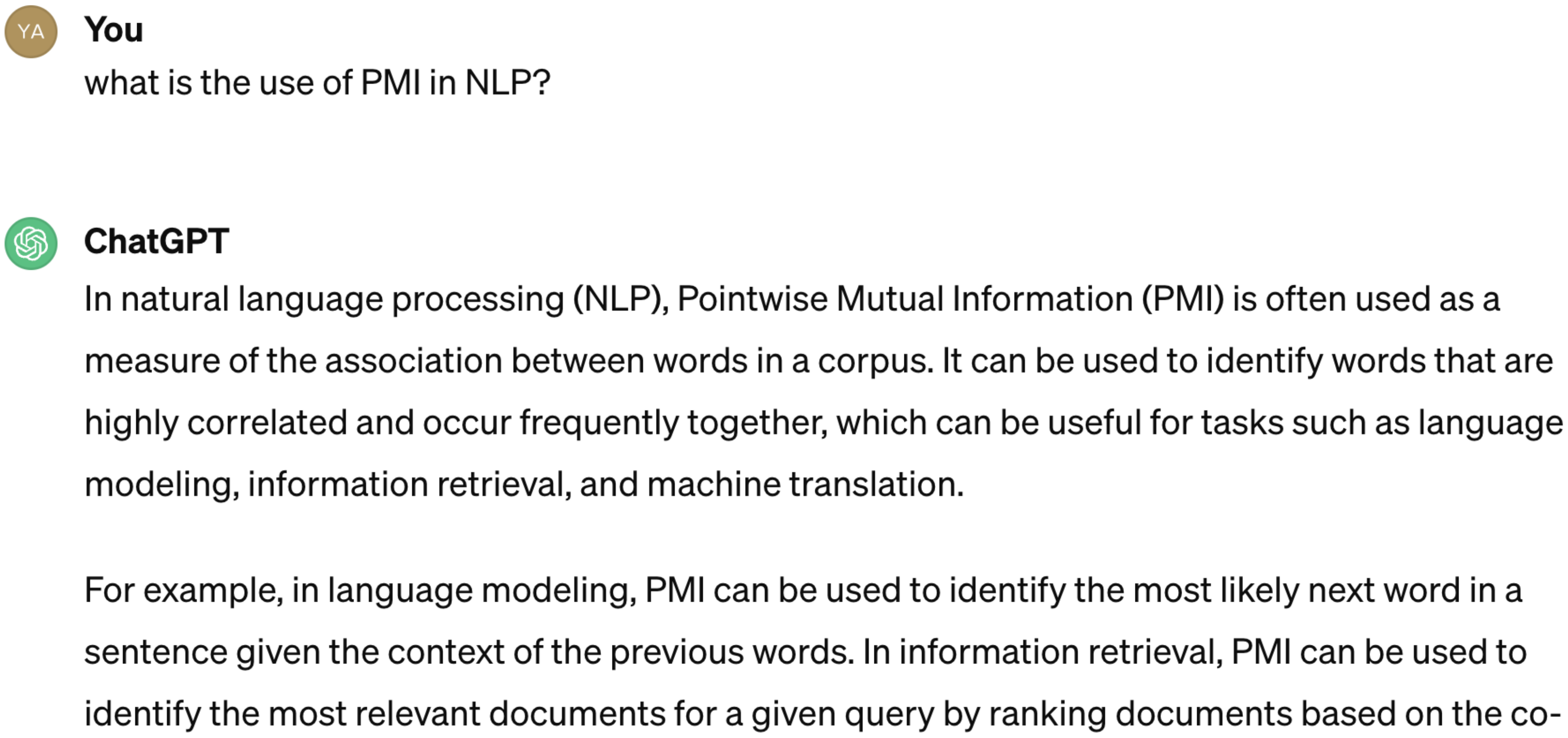
Parametric LLMs – Stellar performance in exams!



OpenAI – GPT-4 Technical Report
https://arxiv.org/abs/2303.08774

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Hallucinations in Parametric LLMs



You

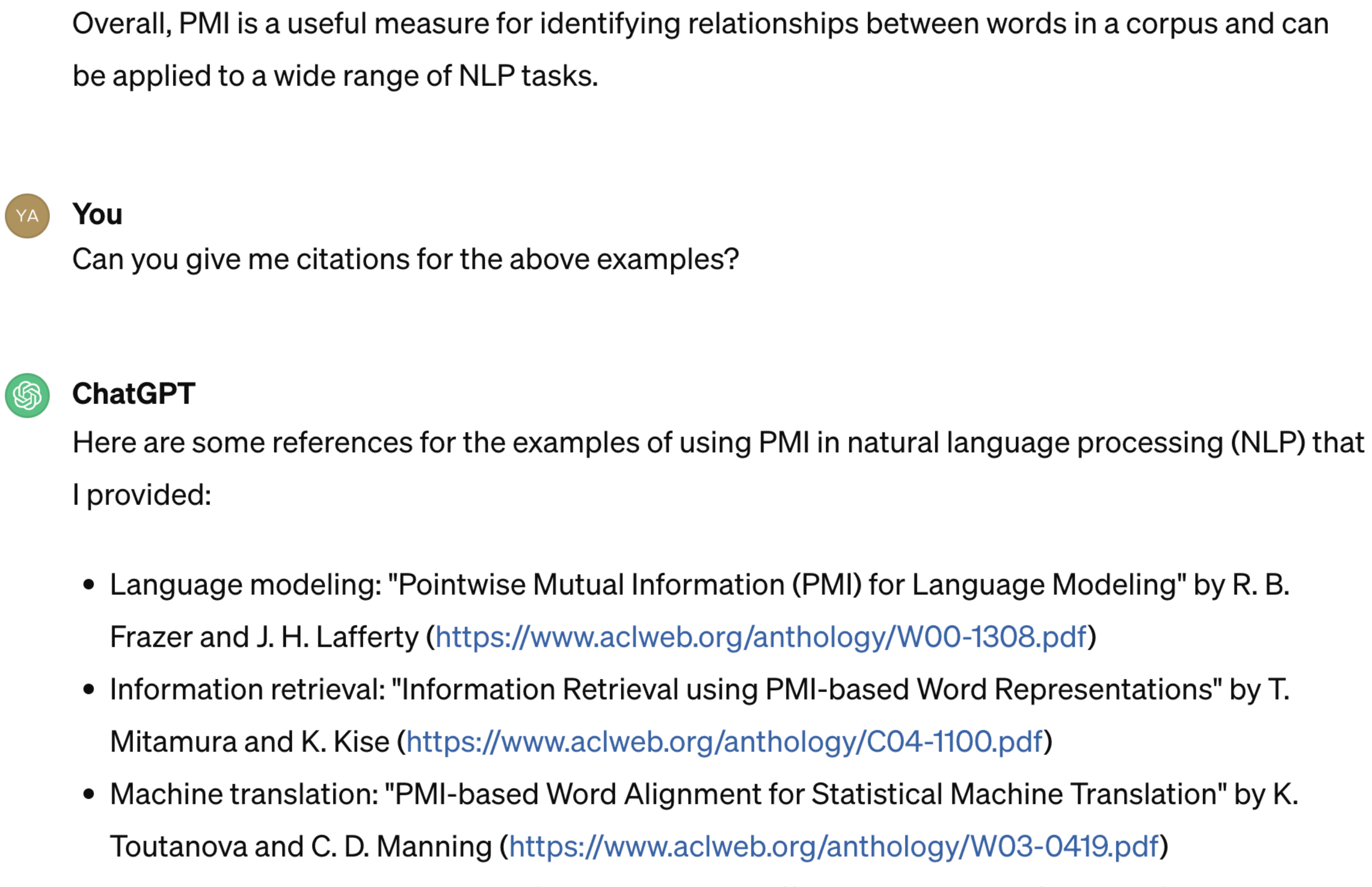
what is the use of PMI in NLP?

ChatGPT

In natural language processing (NLP), Pointwise Mutual Information (PMI) is often used as a measure of the association between words in a corpus. It can be used to identify words that are highly correlated and occur frequently together, which can be useful for tasks such as language modeling, information retrieval, and machine translation.

For example, in language modeling, PMI can be used to identify the most likely next word in a sentence given the context of the previous words. In information retrieval, PMI can be used to identify the most relevant documents for a given query by ranking documents based on the co-

Hallucinations in Parametric LLMs



Overall, PMI is a useful measure for identifying relationships between words in a corpus and can be applied to a wide range of NLP tasks.

You

Can you give me citations for the above examples?

ChatGPT

Here are some references for the examples of using PMI in natural language processing (NLP) that I provided:

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Hallucinations in Para

Overall, PMI is a useful measure for identifying relation

be applied to a wide range of NLP tasks.

You

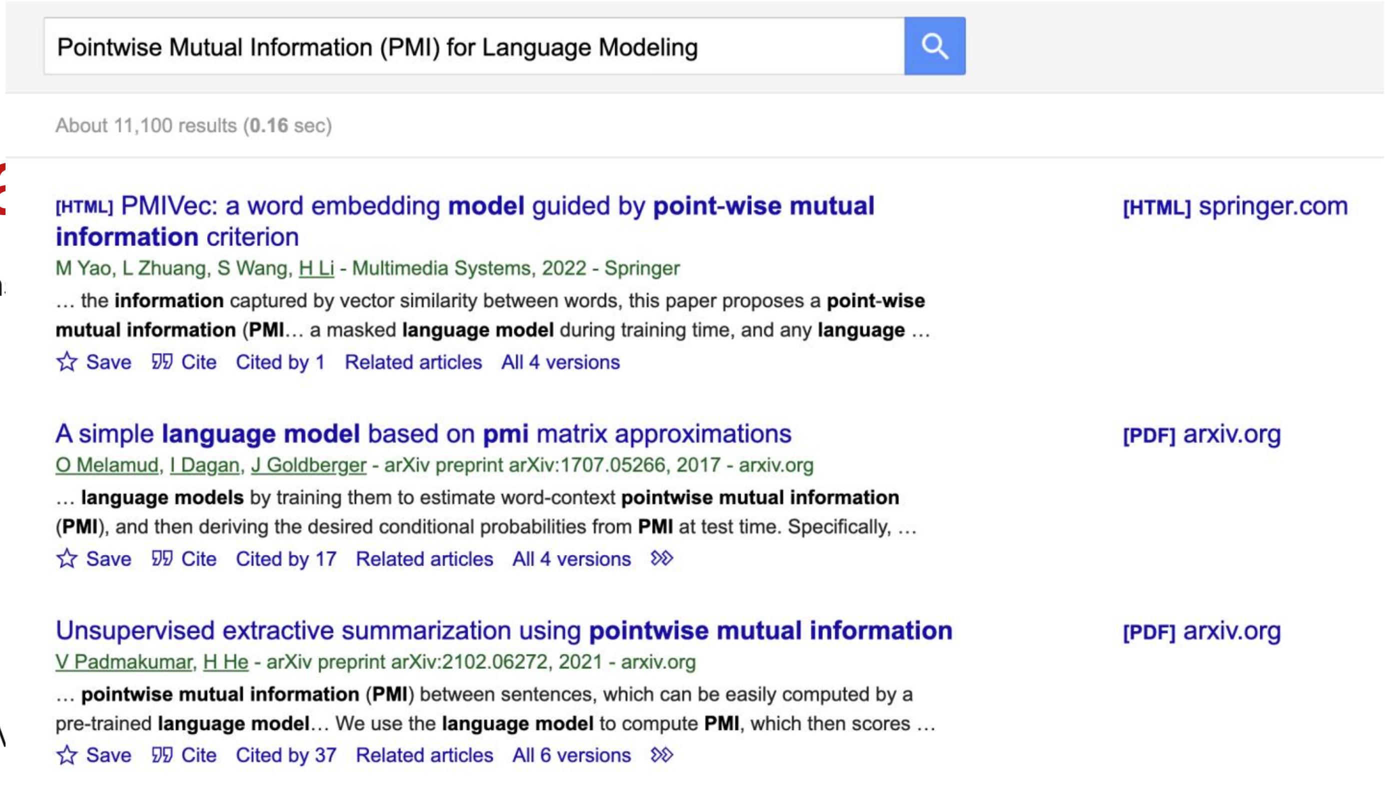
Can you give me citations for the above examples?

ChatGPT

Here are some references for the examples of using PM

I provided:

- Language modeling "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)



Pointwise Mutual Information (PMI) for Language Modeling

About 11,100 results (**0.16** sec)

[HTML] PMIVec: a word embedding **model** guided by **point-wise mutual information** criterion — [HTML] springer.com

M Yao, L Zhuang, S Wang, H Li - Multimedia Systems, 2022 - Springer

… the **information** captured by vector similarity between words, this paper proposes a **point-wise mutual information** (**PMI**… a masked **language model** during training time, and any **language** …

☆ Save 99 Cite Cited by 1 Related articles All 4 versions

A simple **language model** based on **pmi** matrix approximations — [PDF] arxiv.org

O Melamud, I Dagan, J Goldberger - arXiv preprint arXiv:1707.05266, 2017 - arxiv.org

… **language models** by training them to estimate word-context **pointwise mutual information** (**PMI**), and then deriving the desired conditional probabilities from **PMI** at test time. Specifically, …

☆ Save 99 Cite Cited by 17 Related articles All 4 versions »

Unsupervised extractive summarization using **pointwise mutual information** — [PDF] arxiv.org

V Padmakumar, H He - arXiv preprint arXiv:2102.06272, 2021 - arxiv.org

… **pointwise mutual information** (**PMI**) between sentences, which can be easily computed by a pre-trained **language model**… We use the **language model** to compute **PMI**, which then scores …

☆ Save 99 Cite Cited by 37 Related articles All 6 versions »

Hallucinations in

Overall, PMI is a useful measure for identifying
be applied to a wide range of NLP tasks.

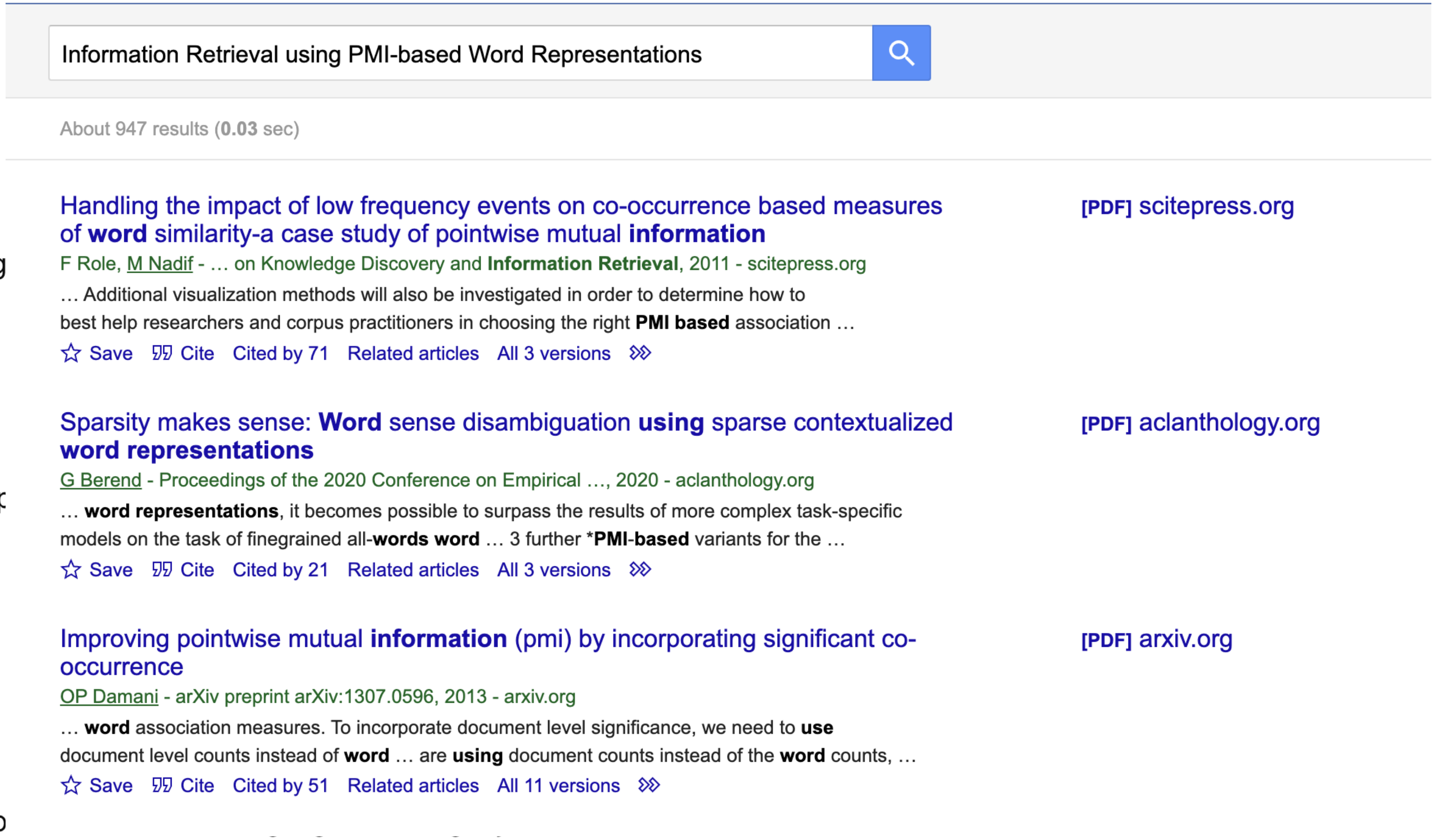
YA **You**

Can you give me citations for the above examp

ChatGPT

Here are some references for the examples of
I provided:

- Language modeling: "Pointwise Mutual Info
  Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)



Information Retrieval using PMI-based Word Representations

About 947 results (**0.03** sec)

Handling the impact of low frequency events on co-occurrence based measures of **word** similarity-a case study of pointwise mutual **information** — **[PDF]** scitepress.org

F Role, M Nadif - … on Knowledge Discovery and **Information Retrieval**, 2011 - scitepress.org

… Additional visualization methods will also be investigated in order to determine how to best help researchers and corpus practitioners in choosing the right **PMI based** association …

☆ Save 99 Cite Cited by 71 Related articles All 3 versions

Sparsity makes sense: **Word** sense disambiguation **using** sparse contextualized **word representations** — **[PDF]** aclanthology.org

G Berend - Proceedings of the 2020 Conference on Empirical …, 2020 - aclanthology.org

… **word representations**, it becomes possible to surpass the results of more complex task-specific models on the task of finegrained all-**words word** … 3 further ***PMI-based** variants for the …

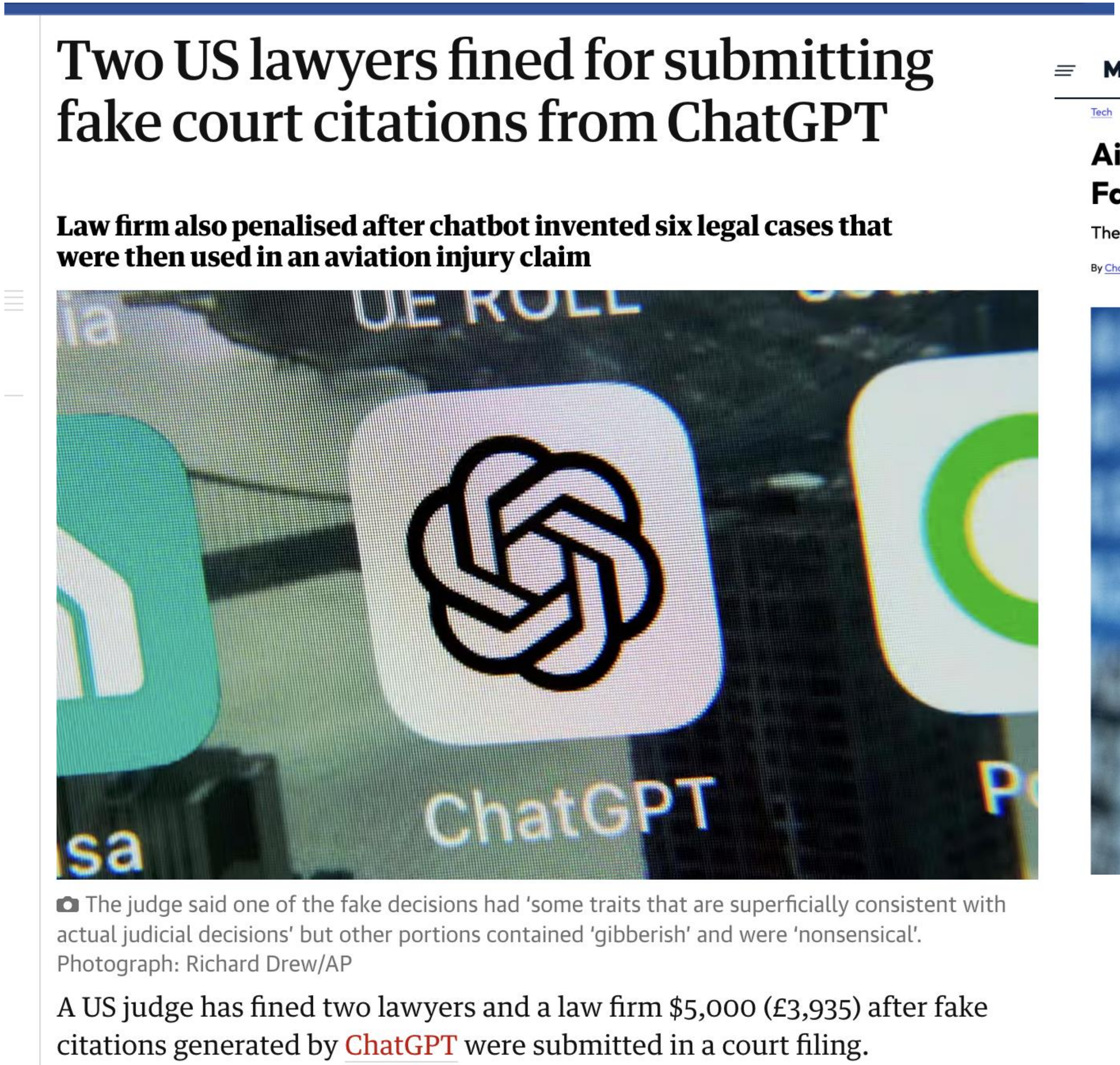
☆ Save 99 Cite Cited by 21 Related articles All 3 versions

Improving pointwise mutual **information** (pmi) by incorporating significant co-occurrence — **[PDF]** arxiv.org

OP Damani - arXiv preprint arXiv:1307.0596, 2013 - arxiv.org

… **word** association measures. To incorporate document level significance, we need to **use** document level counts instead of **word** … are **using** document counts instead of the **word** counts, …

☆ Save 99 Cite Cited by 51 Related articles All 11 versions

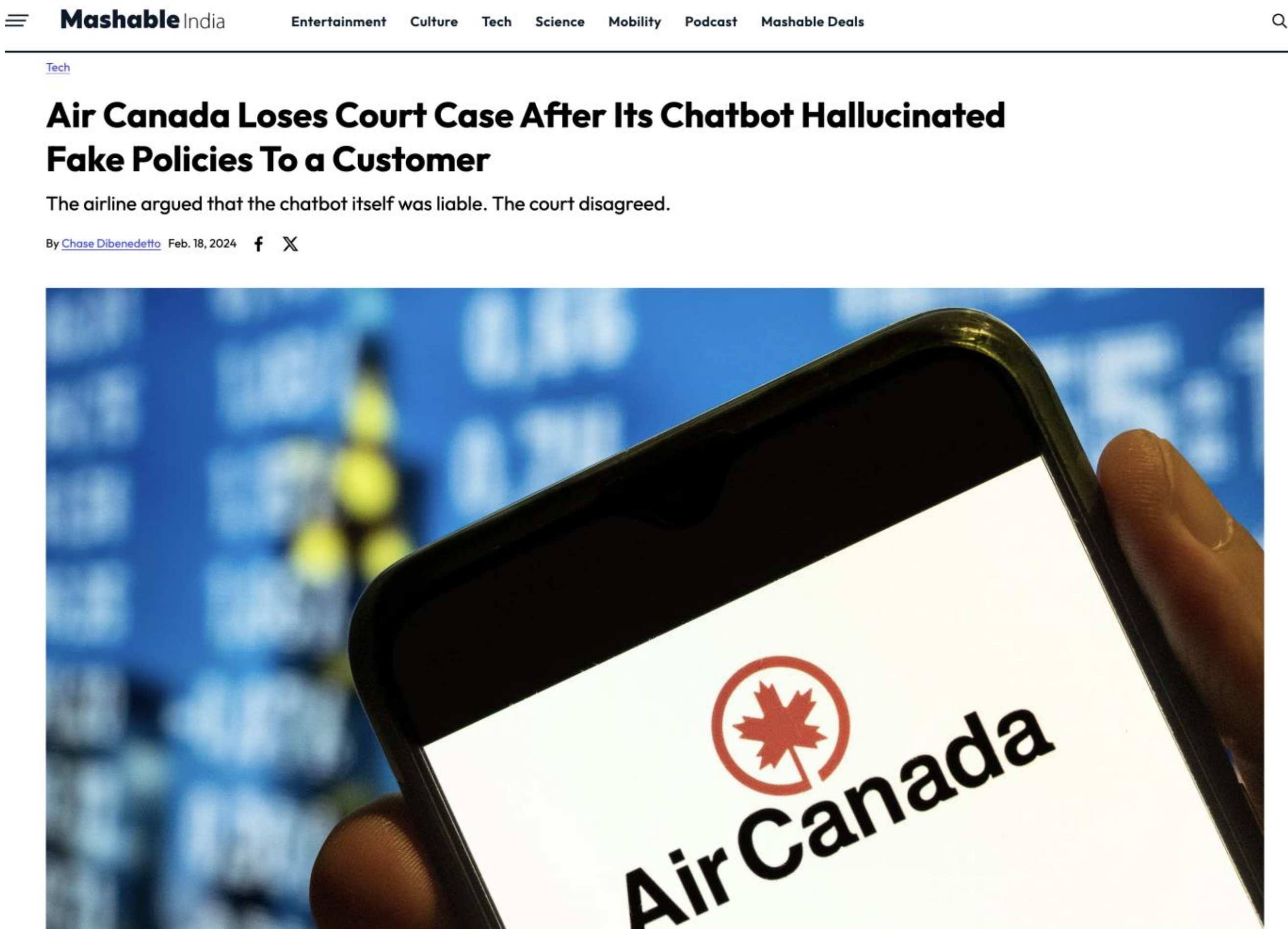


Two US lawyers fined for submitting fake court citations from ChatGPT

Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim

The judge said one of the fake decisions had 'some traits that are superficially consistent with actual judicial decisions' but other portions contained 'gibberish' and were 'nonsensical'. Photograph: Richard Drew/AP

A US judge has fined two lawyers and a law firm $5,000 (£3,935) after fake citations generated by ChatGPT were submitted in a court filing.



Tech

Air Canada Loses Court Case After Its Chatbot Hallucinated Fake Policies To a Customer

The airline argued that the chatbot itself was liable. The court disagreed.

By Chase Dibenedetto Feb. 18, 2024

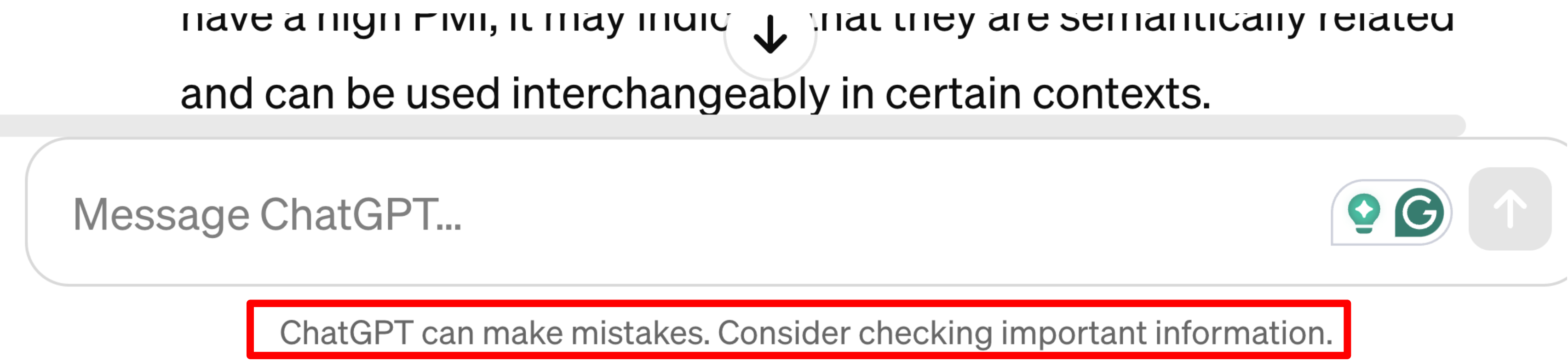
Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Core Limitations of Parametric LLMs

- Hallucinations

Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues



have a high PMI, it may indic hat they are semantically related and can be used interchangeably in certain contexts.

Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

Verifiability

Overall, PMI is a useful measure for identifying re

be applied to a wide range of NLP tasks.

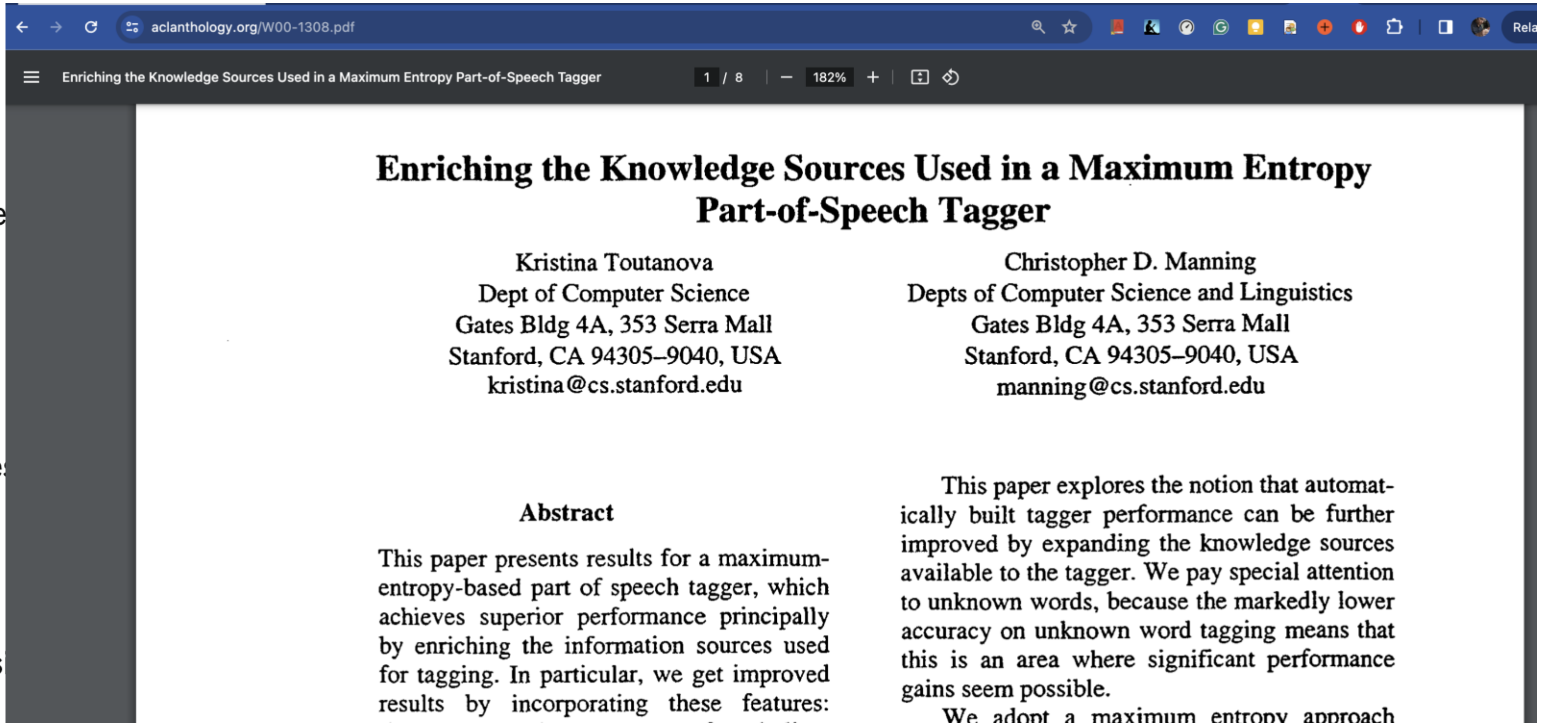
You

Can you give me citations for the above example

ChatGPT

Here are some references for the examples of us

I provided:



aclanthology.org/W00-1308.pdf
Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger
1 / 8
182%
Rela

Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger

Kristina Toutanova
Dept of Computer Science
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
kristina@cs.stanford.edu

Christopher D. Manning
Depts of Computer Science and Linguistics
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
manning@cs.stanford.edu

Abstract

This paper presents results for a maximum-entropy-based part of speech tagger, which achieves superior performance principally by enriching the information sources used for tagging. In particular, we get improved results by incorporating these features:

This paper explores the notion that automatically built tagger performance can be further improved by expanding the knowledge sources available to the tagger. We pay special attention to unknown words, because the markedly lower accuracy on unknown word tagging means that this is an area where significant performance gains seem possible.

We adopt a maximum entropy approach

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Verifiability

Overall, PMI is a useful measure for identifying re

be applied to a wide range of NLP tasks.

You

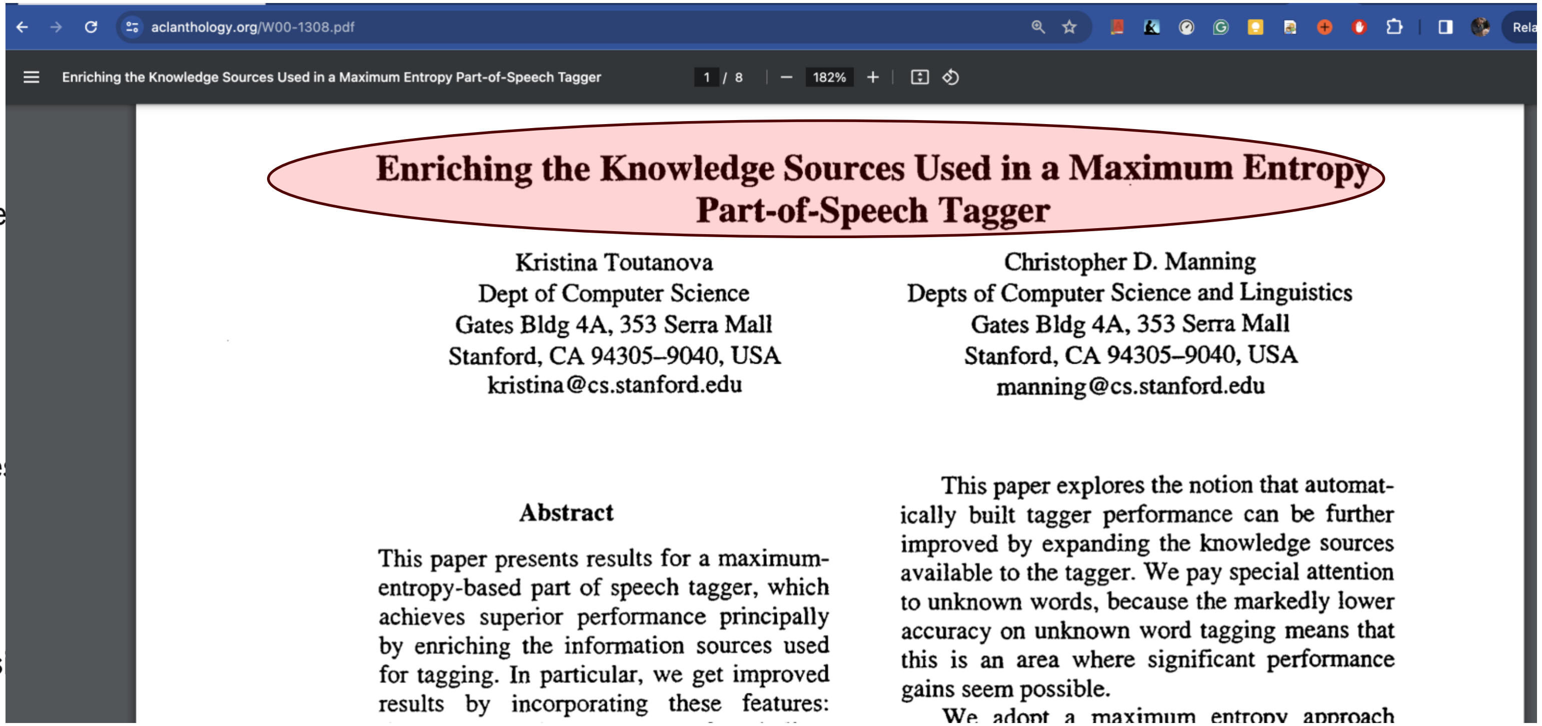
Can you give me citations for the above example

ChatGPT

Here are some references for the examples of us

I provided:

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)



aclanthology.org/W00-1308.pdf
Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger
1 / 8
182%
Rela

Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger

Kristina Toutanova
Dept of Computer Science
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
kristina@cs.stanford.edu

Christopher D. Manning
Depts of Computer Science and Linguistics
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
manning@cs.stanford.edu

Abstract

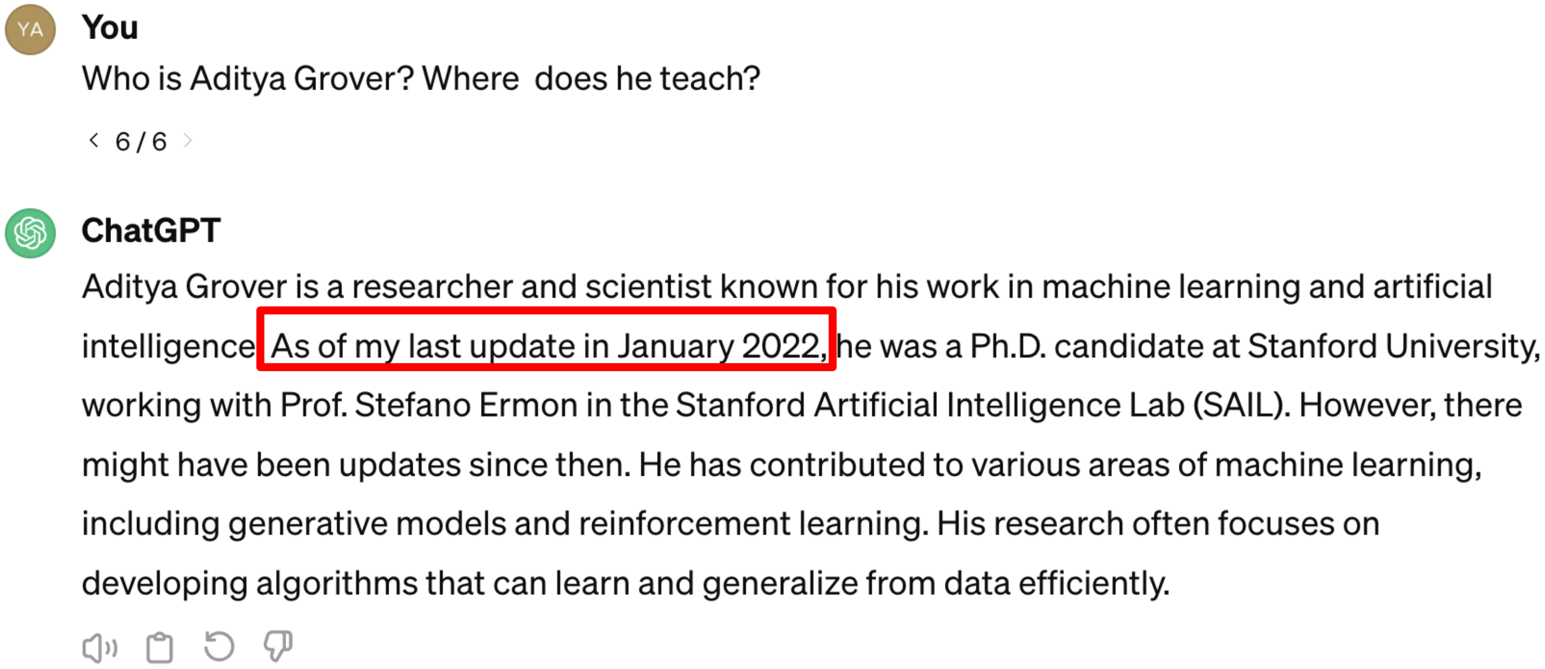
This paper presents results for a maximum-entropy-based part of speech tagger, which achieves superior performance principally by enriching the information sources used for tagging. In particular, we get improved results by incorporating these features:

This paper explores the notion that automatically built tagger performance can be further improved by expanding the knowledge sources available to the tagger. We pay special attention to unknown words, because the markedly lower accuracy on unknown word tagging means that this is an area where significant performance gains seem possible.

We adopt a maximum entropy approach

Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues
- Knowledge cut-offs



You

Who is Aditya Grover? Where does he teach?

ChatGPT

Aditya Grover is a researcher and scientist known for his work in machine learning and artificial intelligence. As of my last update in January 2022, he was a Ph.D. candidate at Stanford University, working with Prof. Stefano Ermon in the Stanford Artificial Intelligence Lab (SAIL). However, there might have been updates since then. He has contributed to various areas of machine learning, including generative models and reinforcement learning. His research often focuses on developing algorithms that can learn and generalize from data efficiently.

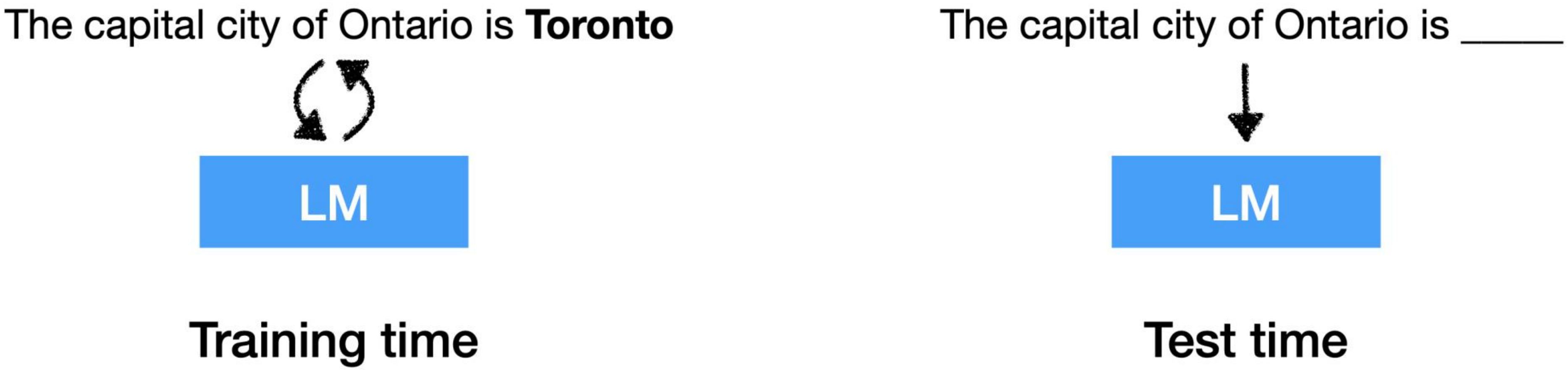
Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues
- Knowledge cut-offs
- Learning failures

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- kNN, RETRO, REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training …*
- Limitations – *lost in the middle, still hallucinating, retriever failures …*

Parametric LLMs – Training vs Test


The capital city of Ontario is **Toronto**

LM

Training time

The capital city of Ontario is _____

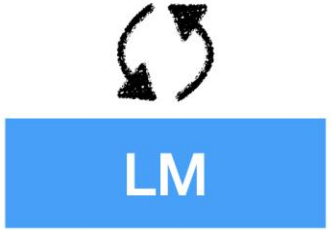
LM

Test time

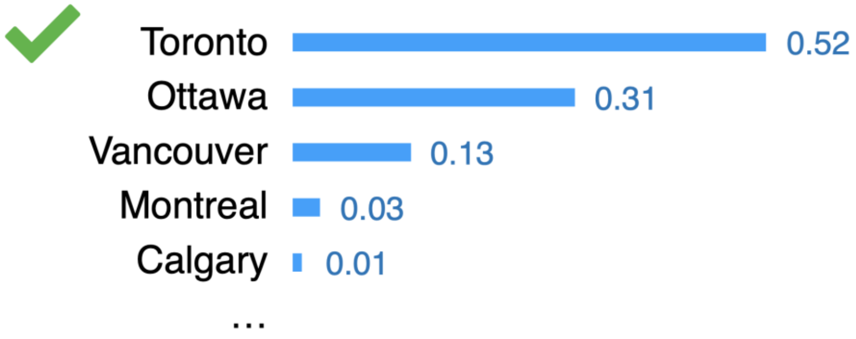
Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Parametric LLMs – Training vs Test

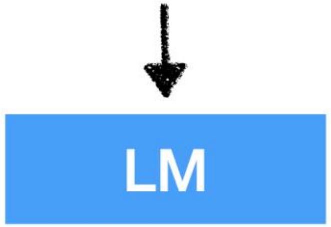
The capital city of Ontario is **Toronto**



Training time



The capital city of Ontario is _____



Test time

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Closed Book vs Open Book Exams

Parametric LLMs

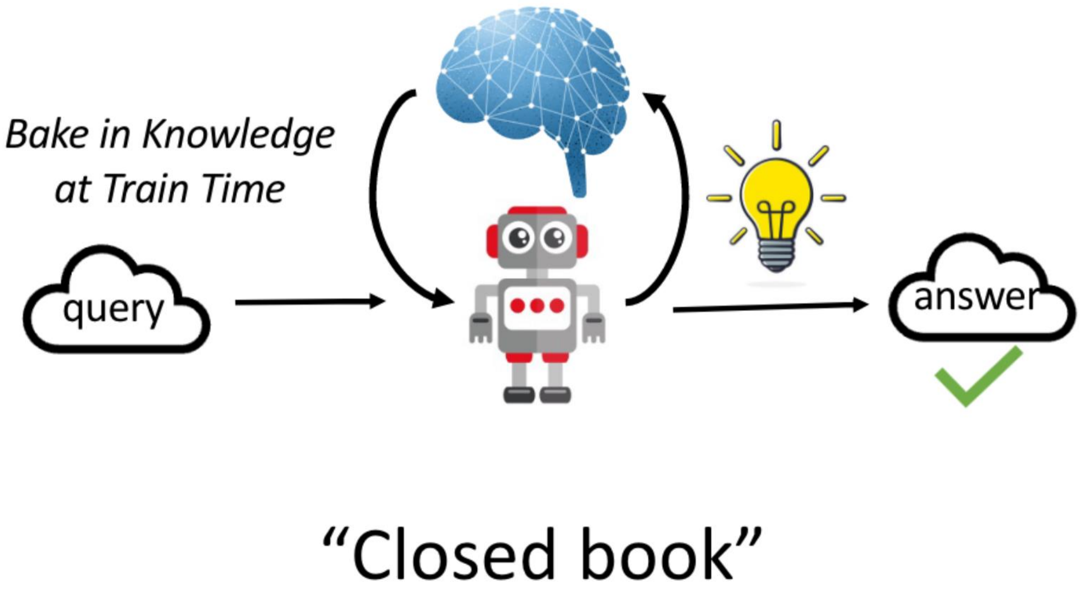


Image source: http://arxiv.org/abs/2403.10131

Closed Book vs Open Book Exams

Parametric LLMs

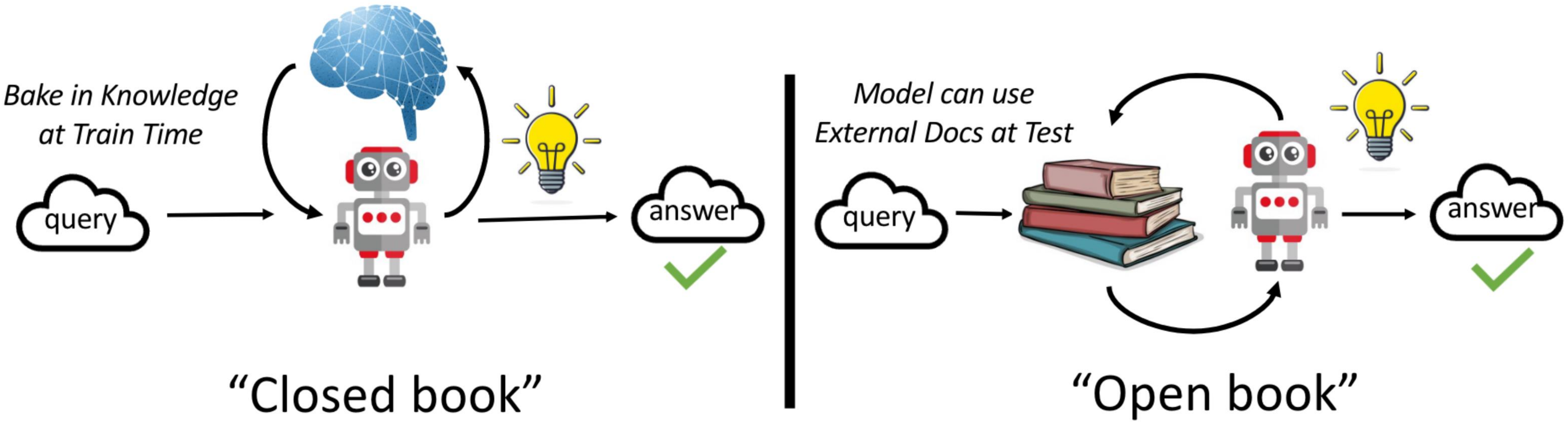

Bake in Knowledge at Train Time
query
answer
"Closed book"
Model can use External Docs at Test
query
answer
"Open book"

Image source: http://arxiv.org/abs/2403.10131

Closed Book vs Open Book Exams

Parametric LLMs

Retrieval-based LLMs

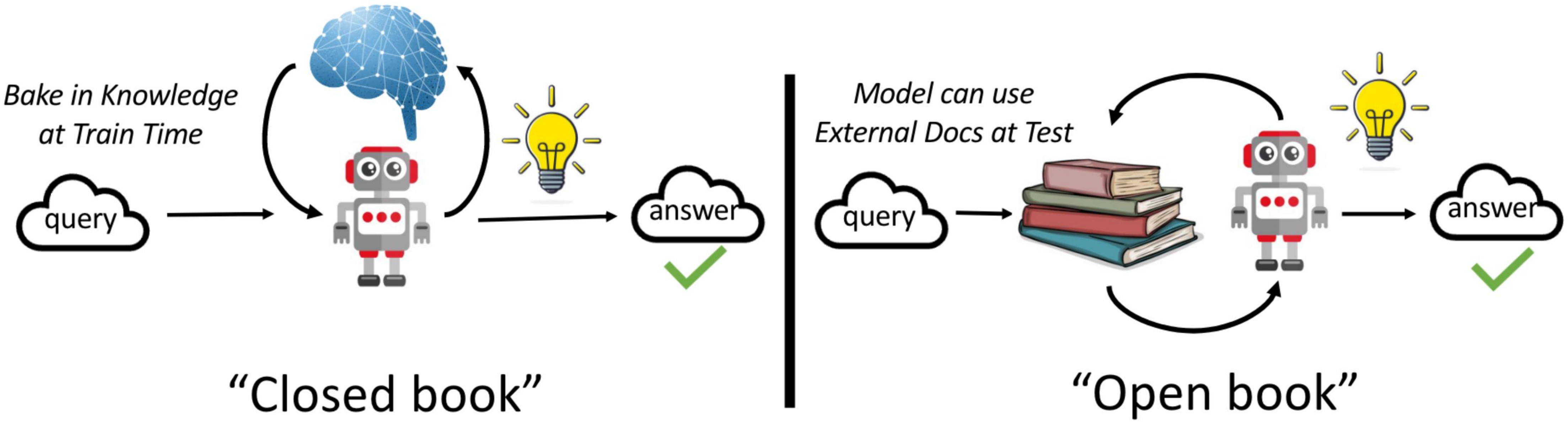
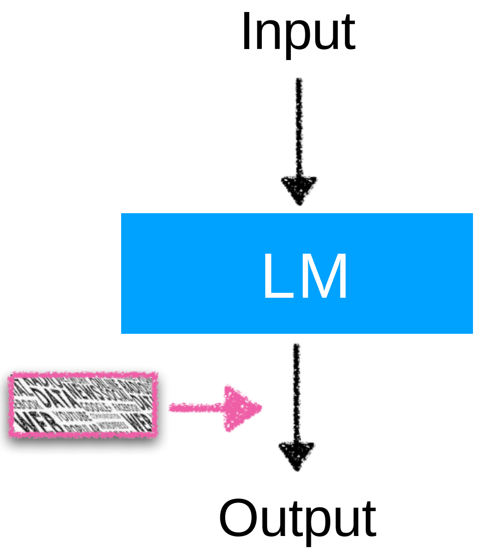


Image source: http://arxiv.org/abs/2403.10131

How to use the Book?

- Output interpolations - After solving the question yourself?

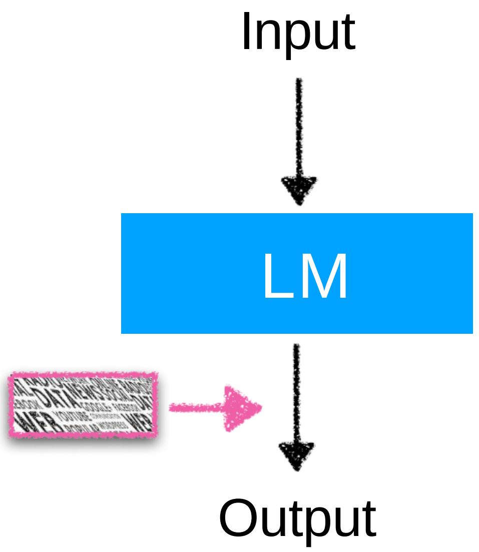


Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?

kNN LMs



Input

LM

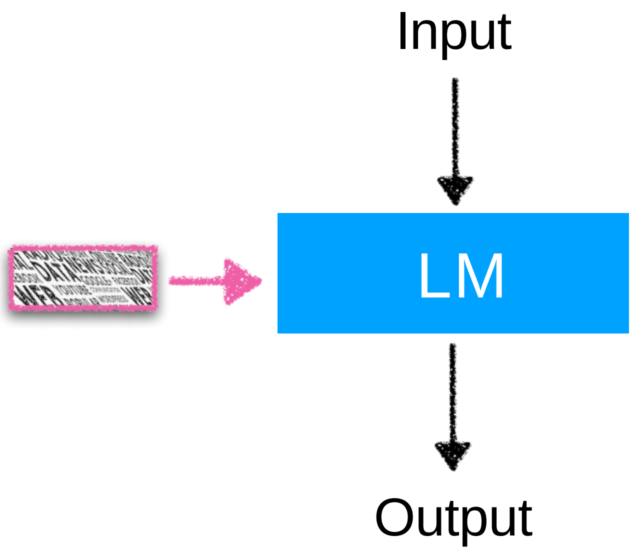
Output

Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?

- Intermediate fusion – modify the LM architecture to be aware of the book?



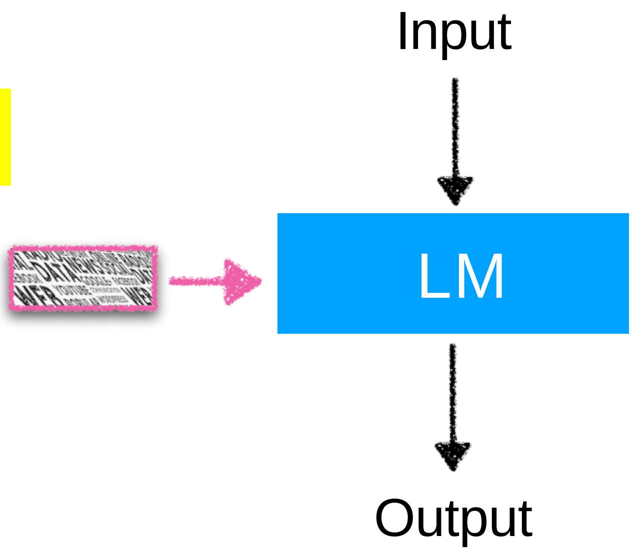
Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?

- Intermediate fusion – modify the LM architecture to be aware of the book?

RETRO



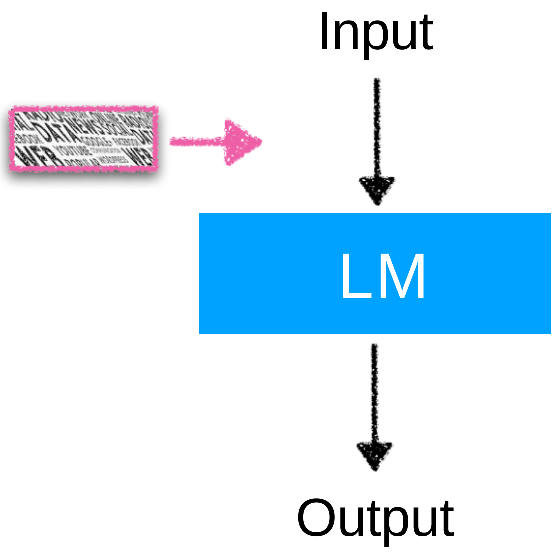
Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?

- Intermediate fusion – modify the LM architecture to be aware of the book?

- Input augmentation (RAG) - Before you start solving?

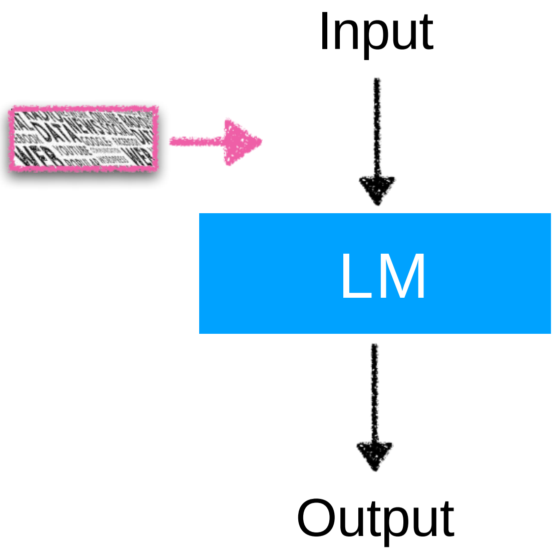


Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?
- Intermediate fusion – modify the LM architecture to be aware of the book?
- Input augmentation (RAG) - Before you start solving?

RAG,
REALM



Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Output interpolations - After solving the question yourself?

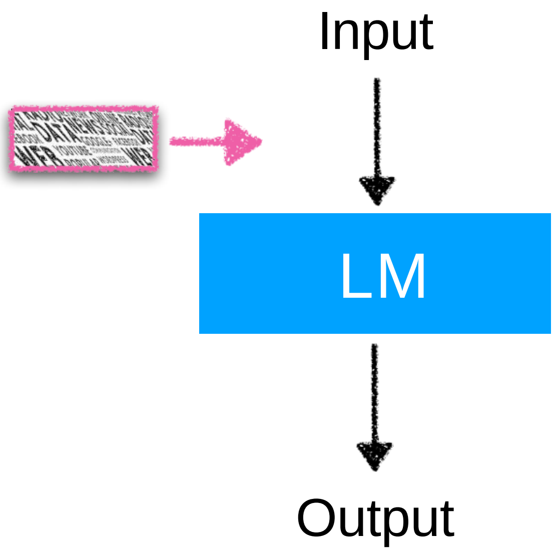
kNN LMs

- Intermediate fusion – modify the LM architecture to be aware of the book?

RETRO

- Input augmentation (RAG) - Before you start solving?

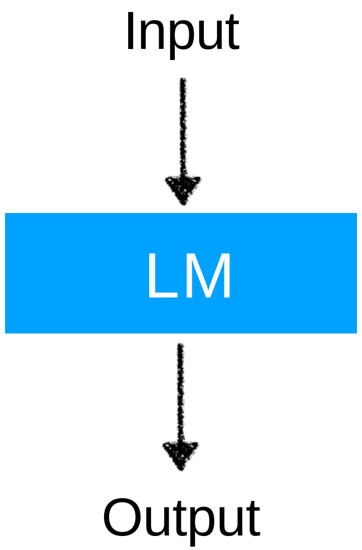
RAG,
REALM



Content credit: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

How to use the Book?

- Need to search to use the book



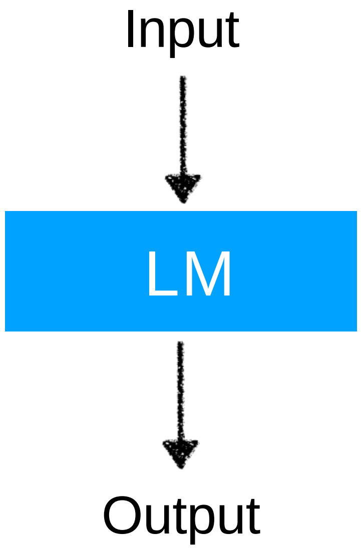
Input

LM

Output

How to use the Book?

- Need to search to use the book – Retrieval


Input

LM

Output

How do you search in a book?

- Inverted index at the end?

Retrieval Methods

- Sparse retrieval
- Document-level dense retrieval
- Token-level dense retrieval
- Cross-encoder reranking
- Differentiable search index (DSI)
- Table of Contents based search
- Black-box retrieval (just ask Google/Bing)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

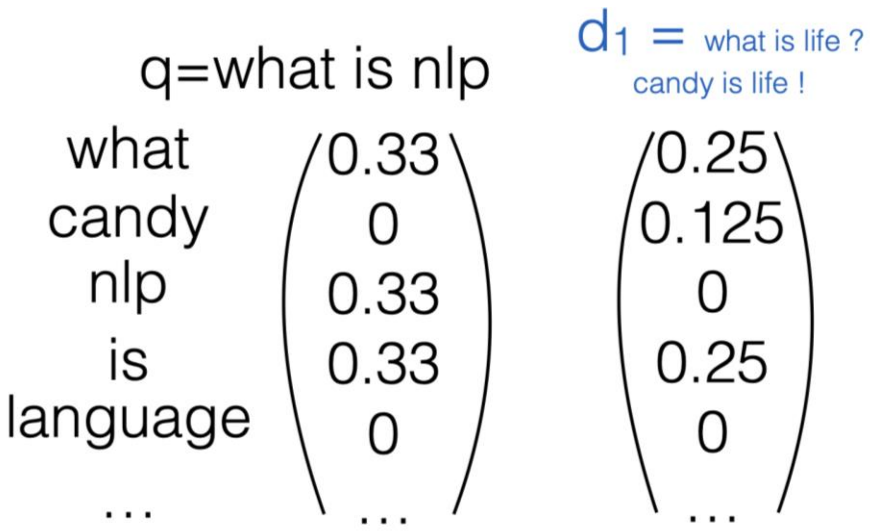
$$q=\text{what is nlp}$$

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix}$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)



q=what is nlp

d_1 = what is life ? candy is life !

$$\begin{array}{l} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \dots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \dots \end{pmatrix} \quad \begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \dots \end{pmatrix}$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

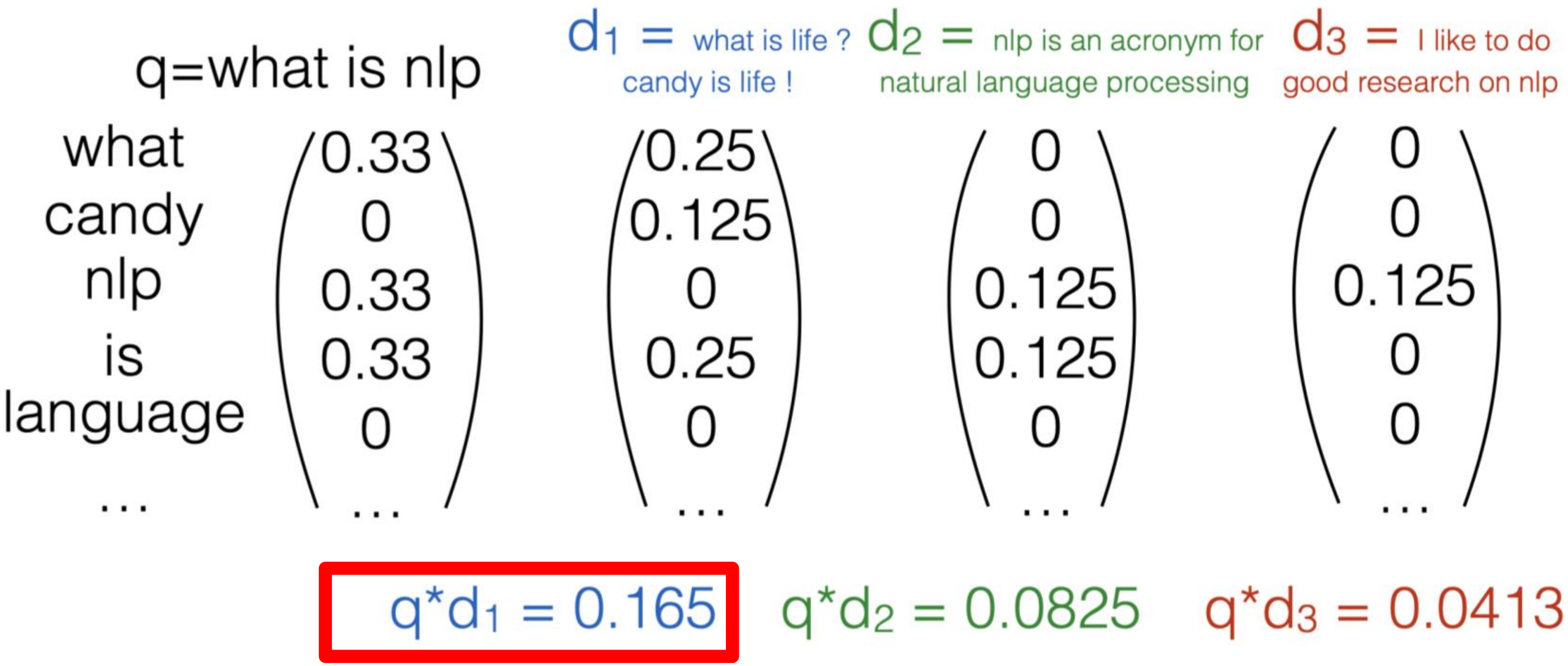
d_3 = I like to do good research on nlp

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0.125 \\ 0 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0 \\ 0 \\ \ldots \end{pmatrix}$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)



q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \ldots \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0.125 \\ 0 \\ \ldots \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0 \\ 0 \\ \ldots \end{pmatrix}$$

$\boxed{q^*d_1 = 0.165}$ $q^*d_2 = 0.0825$ $q^*d_3 = 0.0413$

- Find the document with the highest inner-product or cosine similarity in the document collection

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

$$\mathrm{TF}(t,d) = \frac{\mathrm{freq}(t,d)}{\sum_{t'} \mathrm{freq}(t',d)}$$

$$\mathrm{IDF}(t) = \log\left(\frac{|D|}{\sum_{d' \in D} \delta(\mathrm{freq}(t,d') > 0)}\right)$$

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

$$\mathrm{TF}(t,d) = \frac{\mathrm{freq}(t,d)}{\sum_{t'} \mathrm{freq}(t',d)} \qquad \mathrm{IDF}(t) = \log\left(\frac{|D|}{\sum_{d' \in D} \delta(\mathrm{freq}(t,d') > 0)}\right)$$

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

- BM25: TF term similar to smoothed count-based LMs

$$\text{BM-25}(t,d) = \mathrm{IDF}(t) \cdot$$

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

$$\mathrm{TF}(t,d) = \frac{\mathrm{freq}(t,d)}{\sum_{t'} \mathrm{freq}(t',d)} \qquad \mathrm{IDF}(t) = \log\left(\frac{|D|}{\sum_{d' \in D} \delta(\mathrm{freq}(t,d') > 0)}\right)$$

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

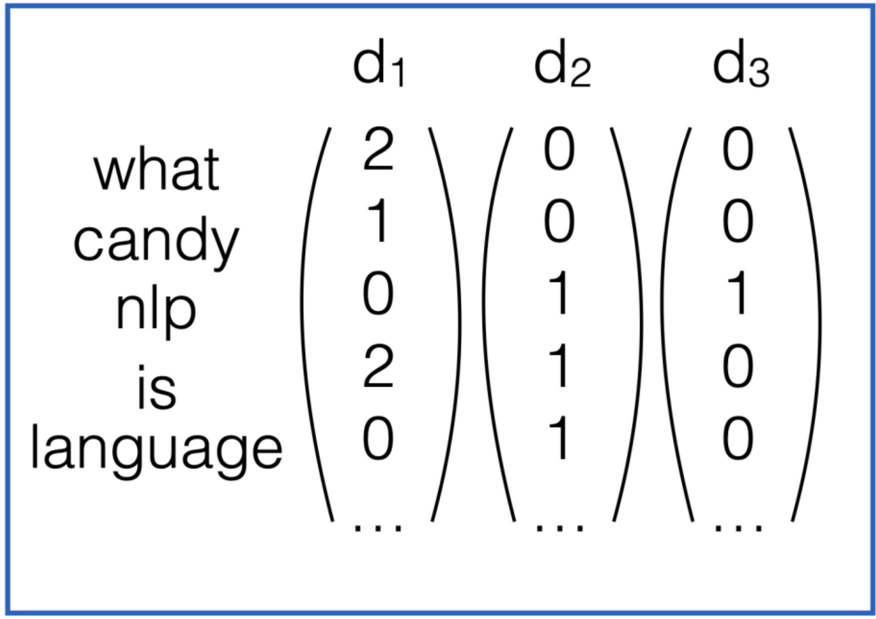
- BM25: TF term similar to smoothed count-based LMs

$$\text{BM-25}(t,d) = \mathrm{IDF}(t) \cdot \frac{\mathrm{freq}(t,d) \cdot (k_1 + 1)}{\mathrm{freq}(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)}$$

Inverted Index

- A data structure that allows for efficient sparse lookup of vectors

Sparse Vectors



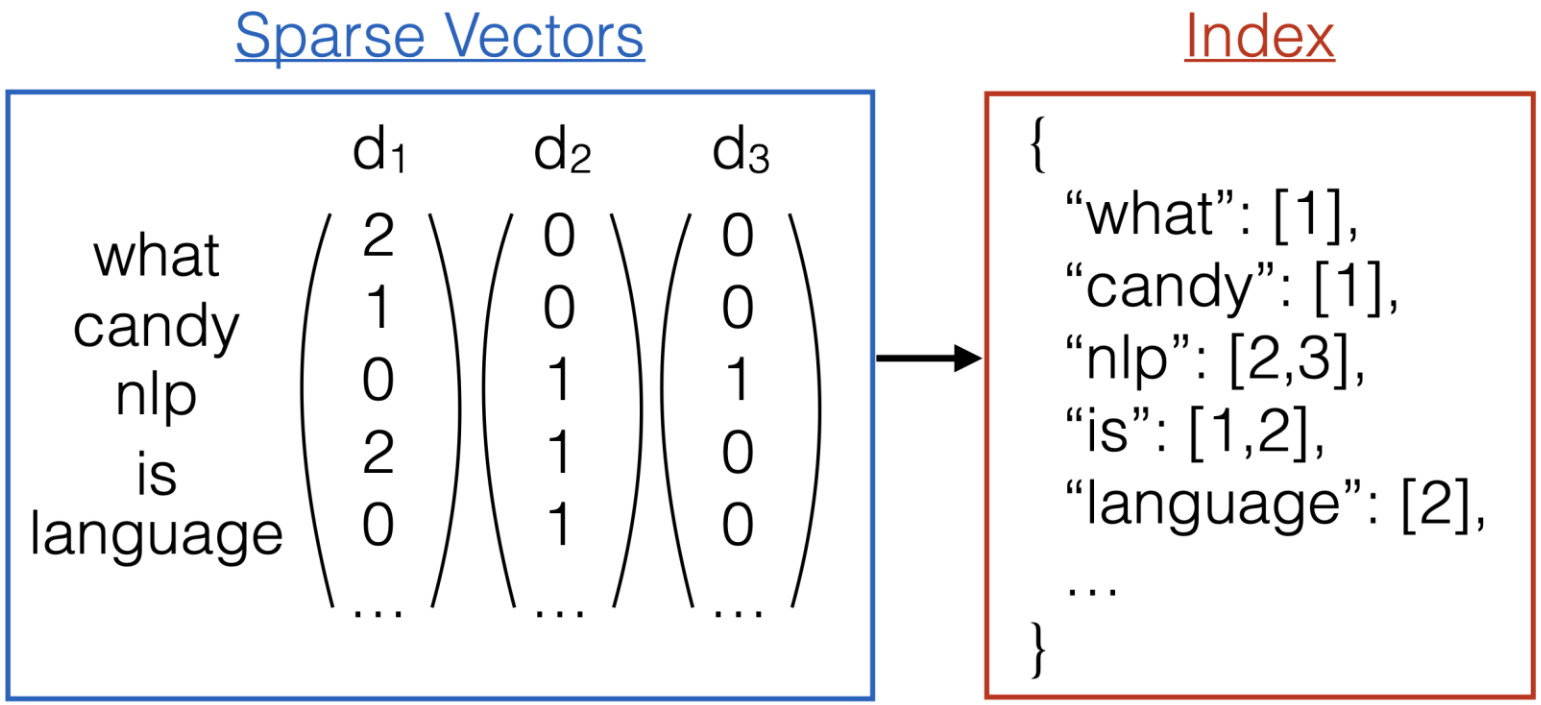
$$\begin{array}{lccc} & d_1 & d_2 & d_3 \\ \text{what} & 2 & 0 & 0 \\ \text{candy} & 1 & 0 & 0 \\ \text{nlp} & 0 & 1 & 1 \\ \text{is} & 2 & 1 & 0 \\ \text{language} & 0 & 1 & 0 \\ & \cdots & \cdots & \cdots \end{array}$$

- Example software: Apache Lucene

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Inverted Index

- A data structure that allows for efficient sparse lookup of vectors



Sparse Vectors

$$\begin{array}{lccc} & d_1 & d_2 & d_3 \\ \text{what} & 2 & 0 & 0 \\ \text{candy} & 1 & 0 & 0 \\ \text{nlp} & 0 & 1 & 1 \\ \text{is} & 2 & 1 & 0 \\ \text{language} & 0 & 1 & 0 \\ & \dots & \dots & \dots \end{array}$$

Index

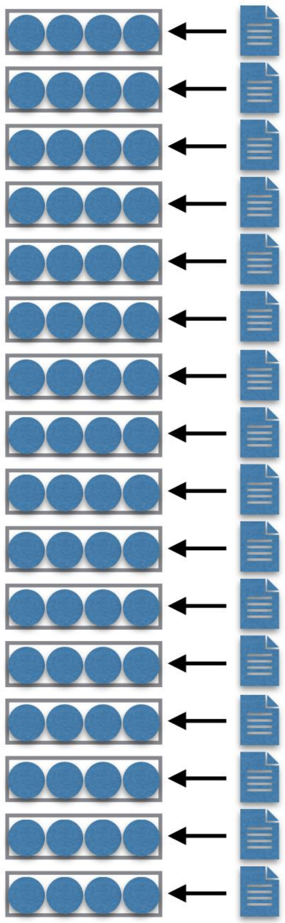
```
{
  “what”: [1],
  “candy”: [1],
  “nlp”: [2,3],
  “is”: [1,2],
  “language”: [2],
  …
}
```

- Example software: Apache Lucene

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

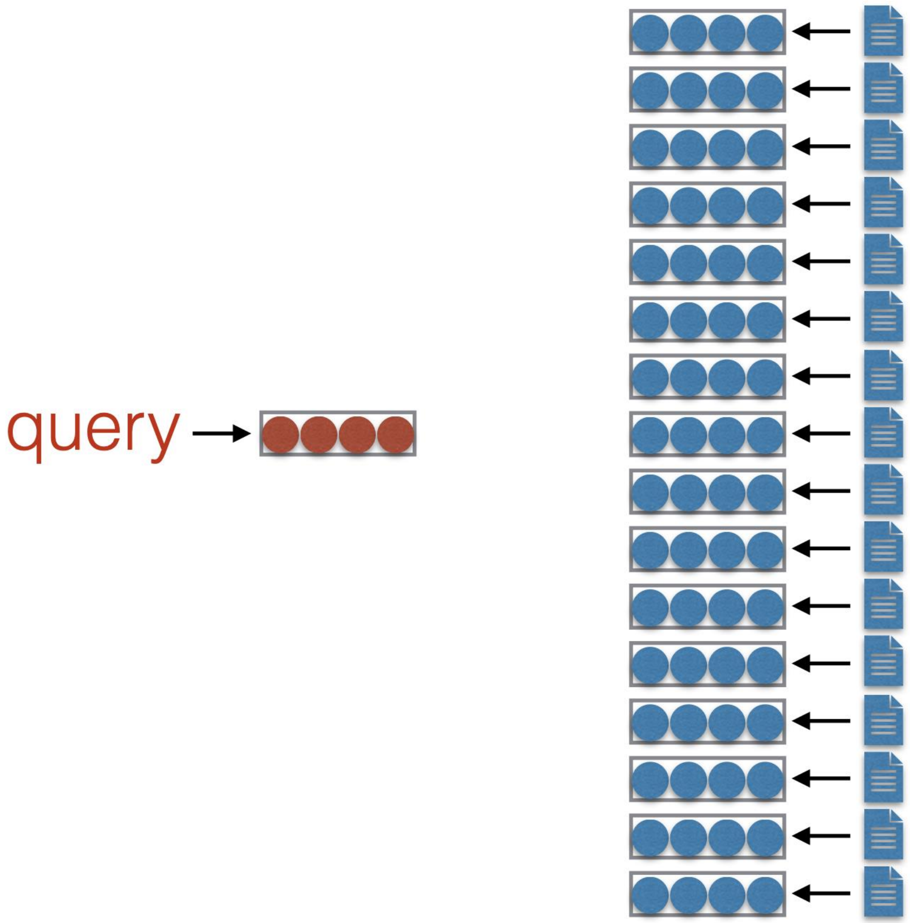
- Encode all documents using a LM and index them (one time task). Can use:
  - ✓ Out-of-the-box embeddings. E.g. BERT
  - ✓ Learned embeddings (covered later)



Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓ Out-of-the-box embeddings. E.g. BERT
  - ✓ Learned embeddings (covered later)
- At test time:
  - Encode Query

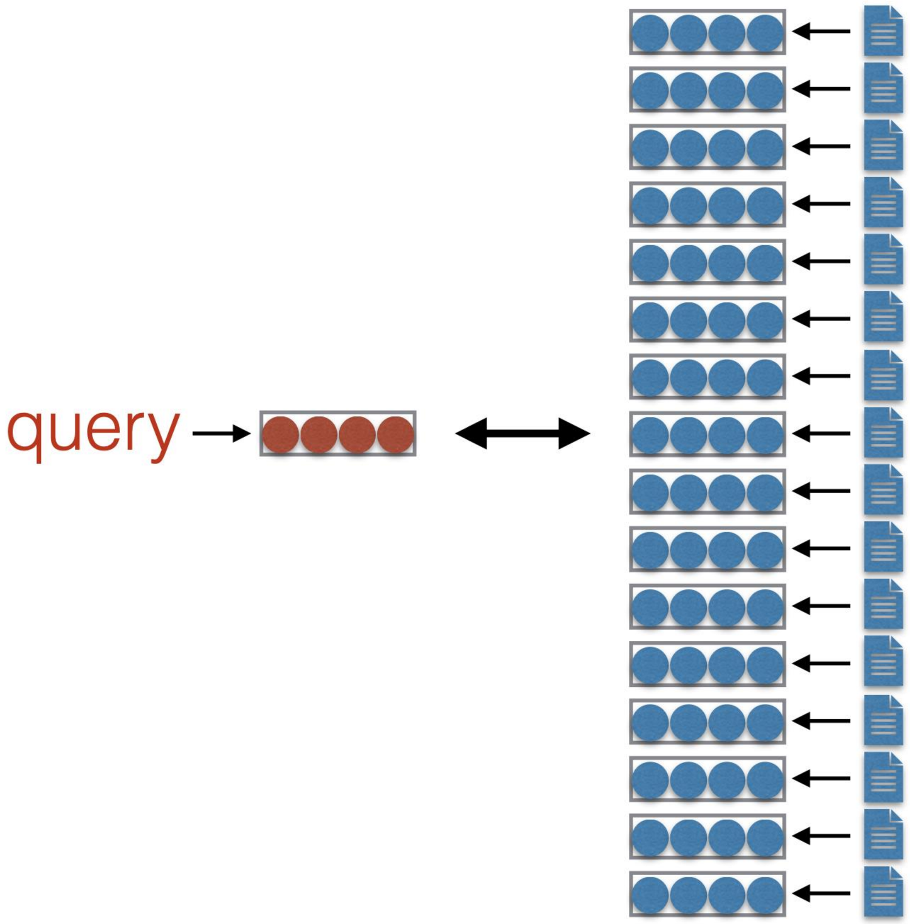


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓ Out-of-the-box embeddings. E.g. BERT
  - ✓ Learned embeddings (covered later)

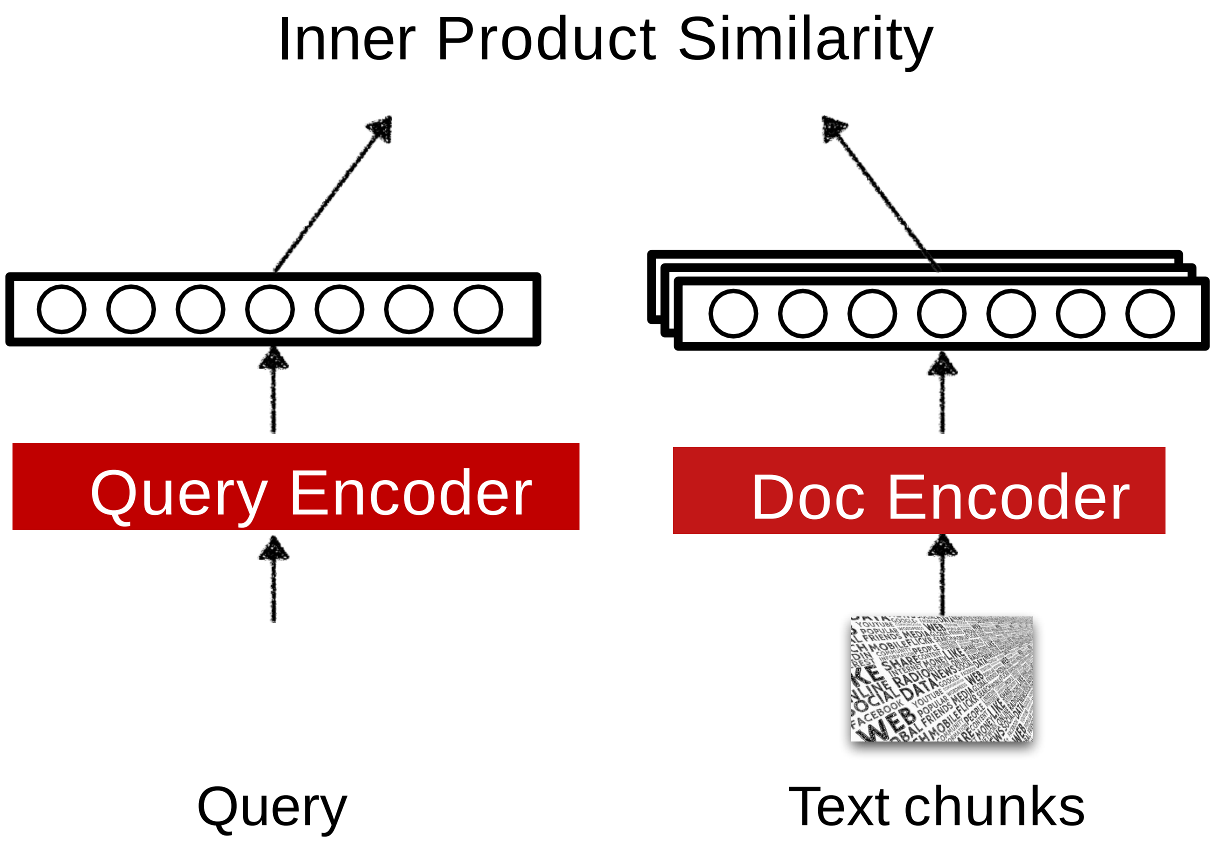
- At test time:
  - Encode Query
  - Find similar documents



Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



Inner Product Similarity

Query Encoder

Doc Encoder

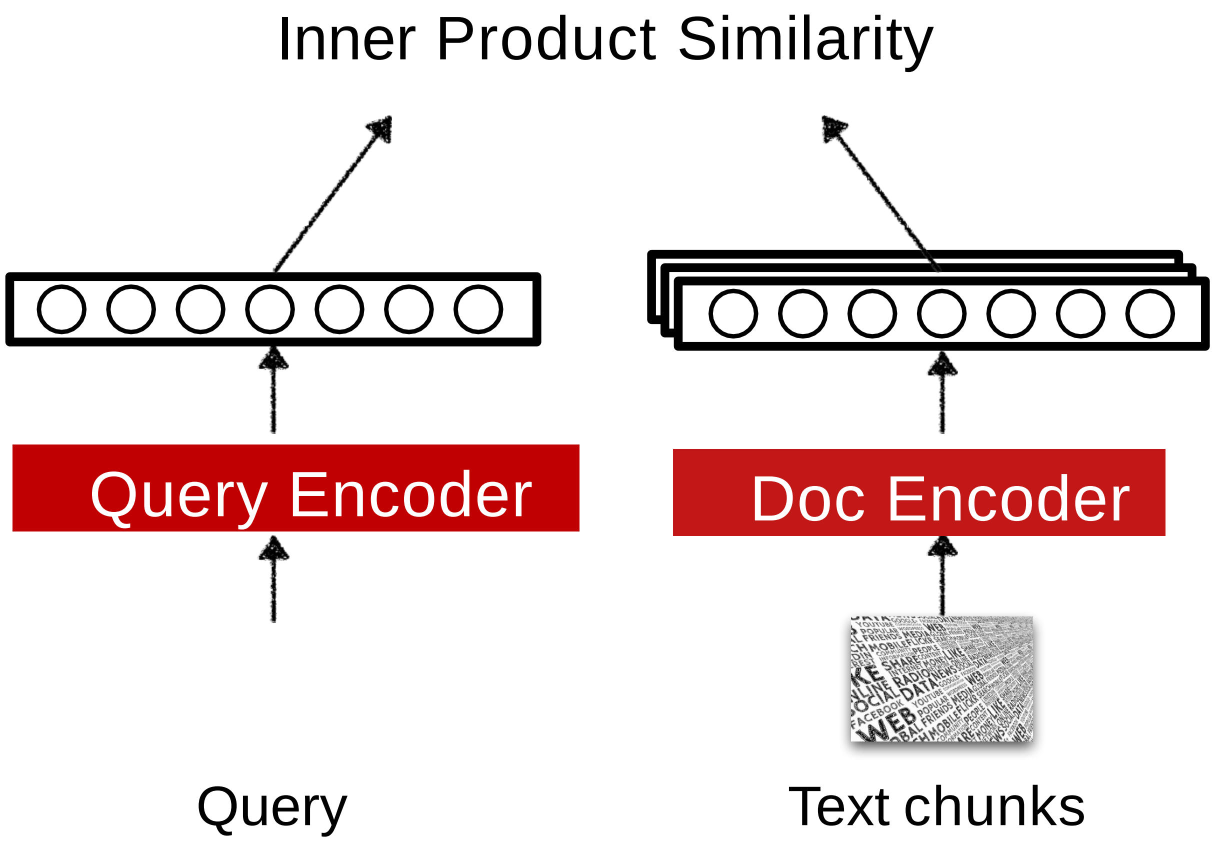
Query

Text chunks

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



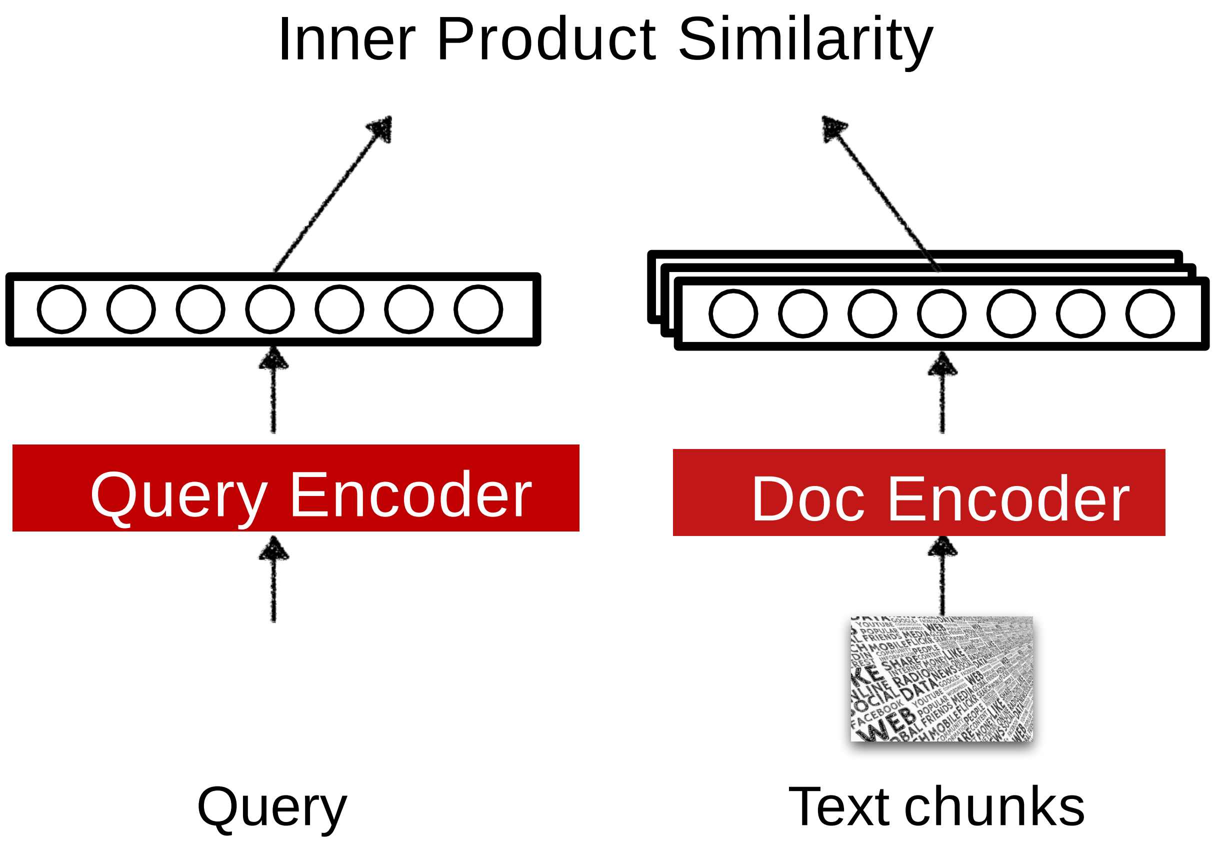
$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

$$= -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



$$L(q, \boxed{p^+}, p_1^-, p_2^-, \ldots, p_n^-)$$

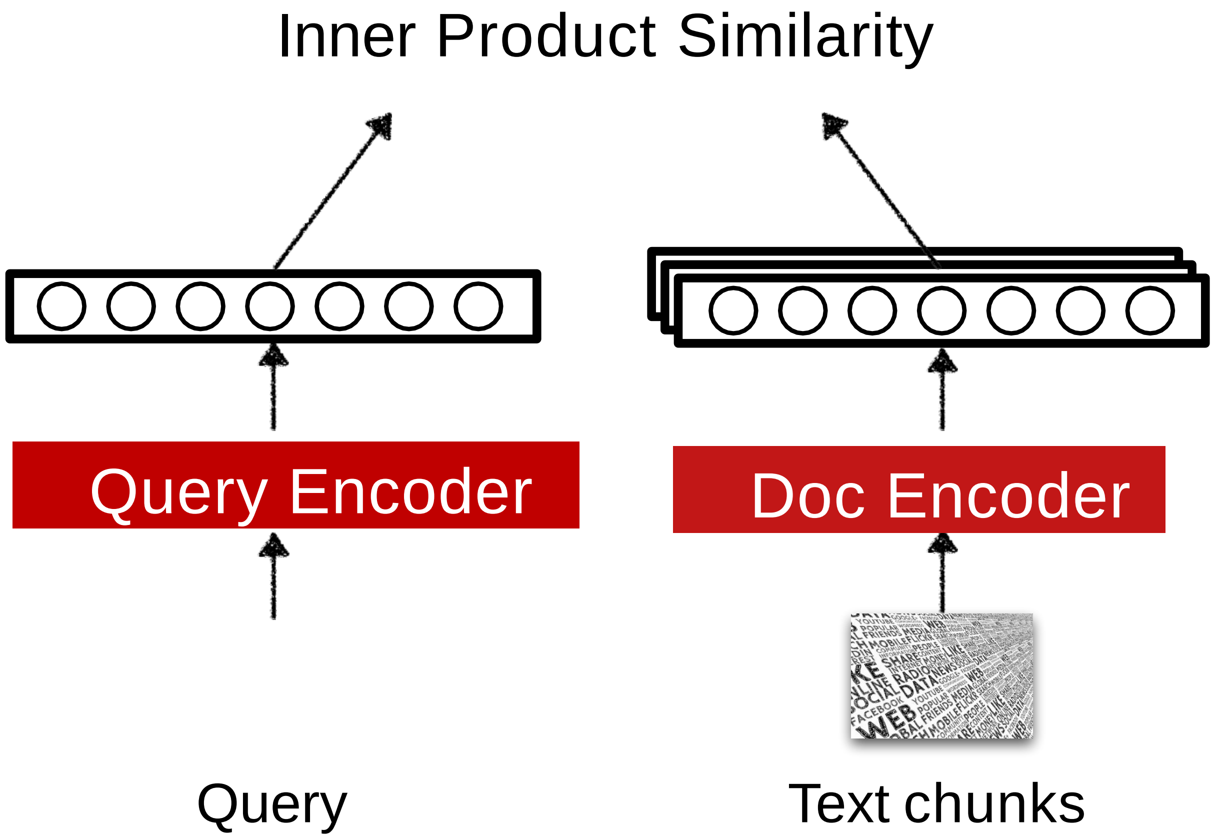
Positive passage

$$= -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



Inner Product Similarity

Query Encoder

Doc Encoder

Query

Text chunks

Negative passages

Too expensive to consider all negatives!

$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-)$$

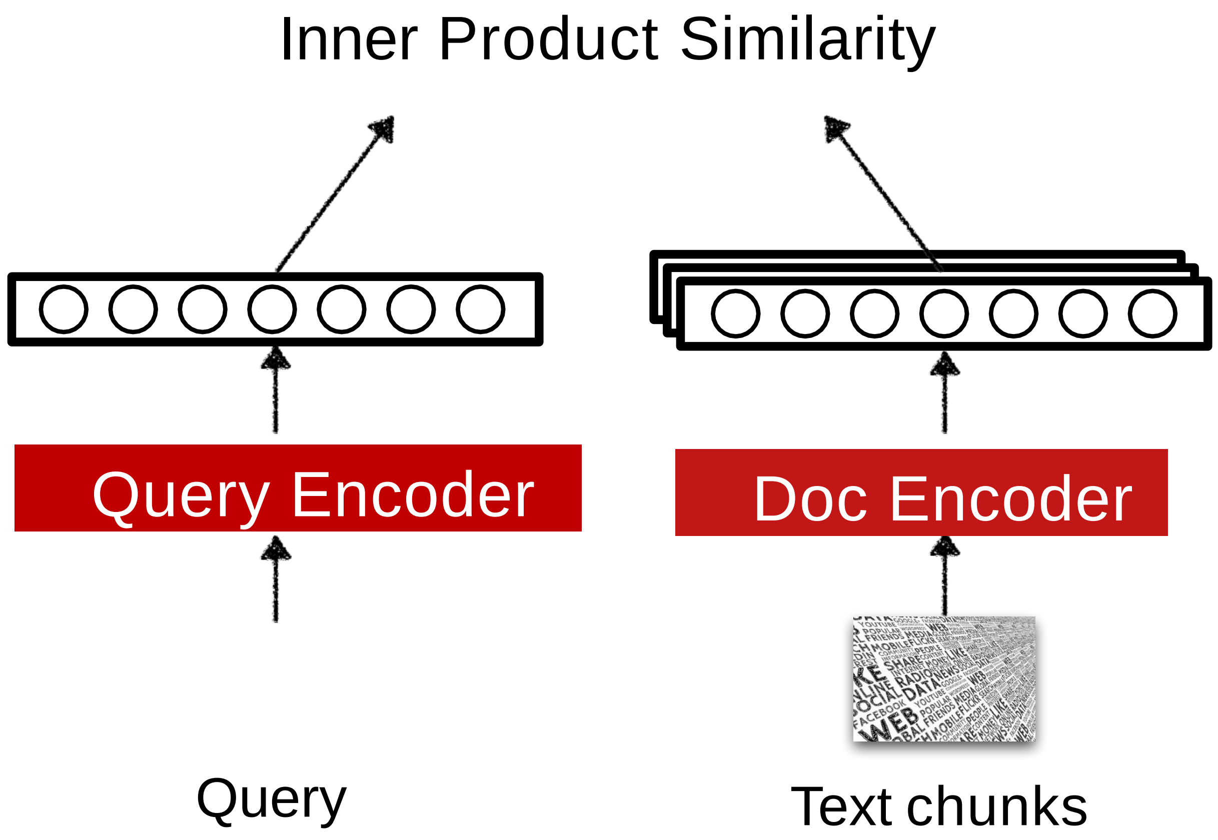
Positive passage

$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\text{sim}(q, p_j^-))}$$

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



Inner Product Similarity

Query Encoder

Doc Encoder

Query

Text chunks

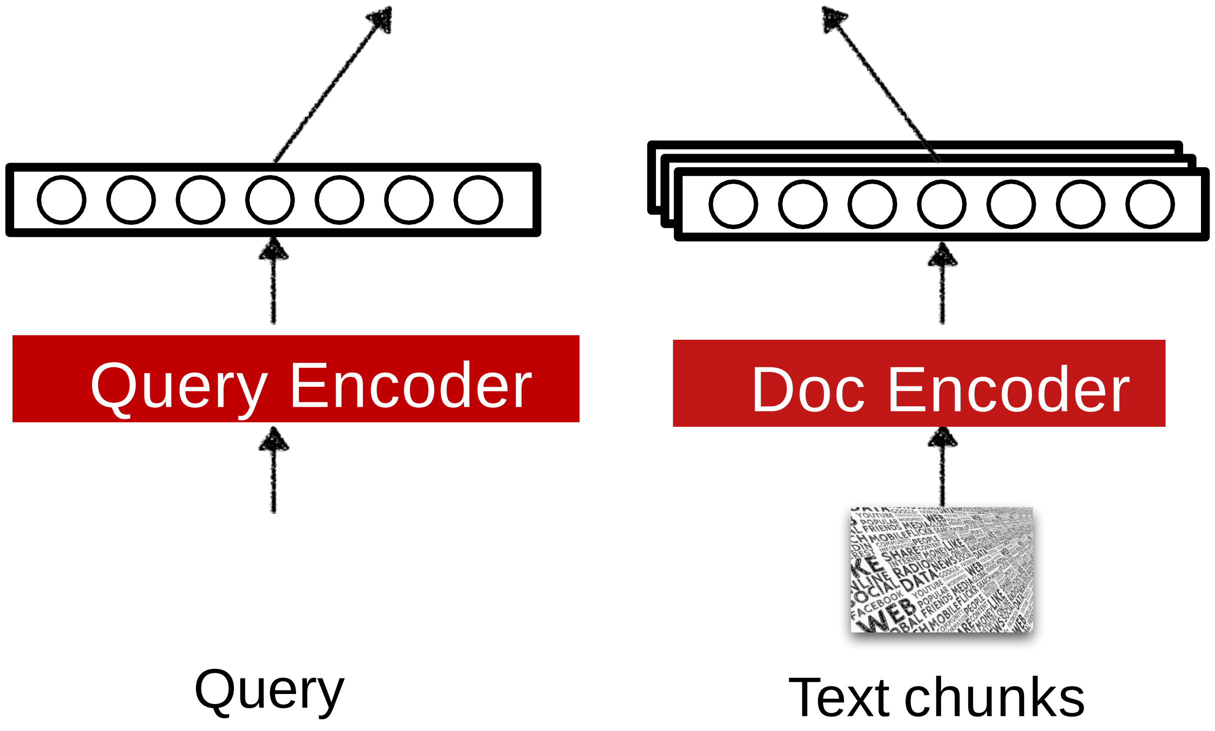
$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Contrastive learning

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.



Inner Product Similarity

Query Encoder

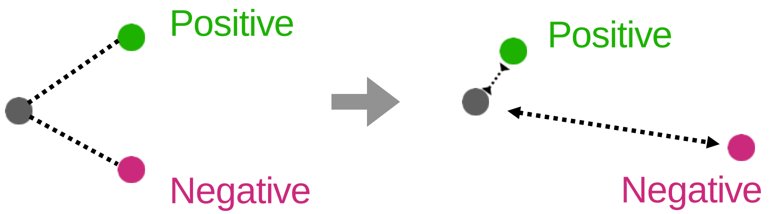
Doc Encoder

Query

Text chunks

$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Contrastive learning



Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

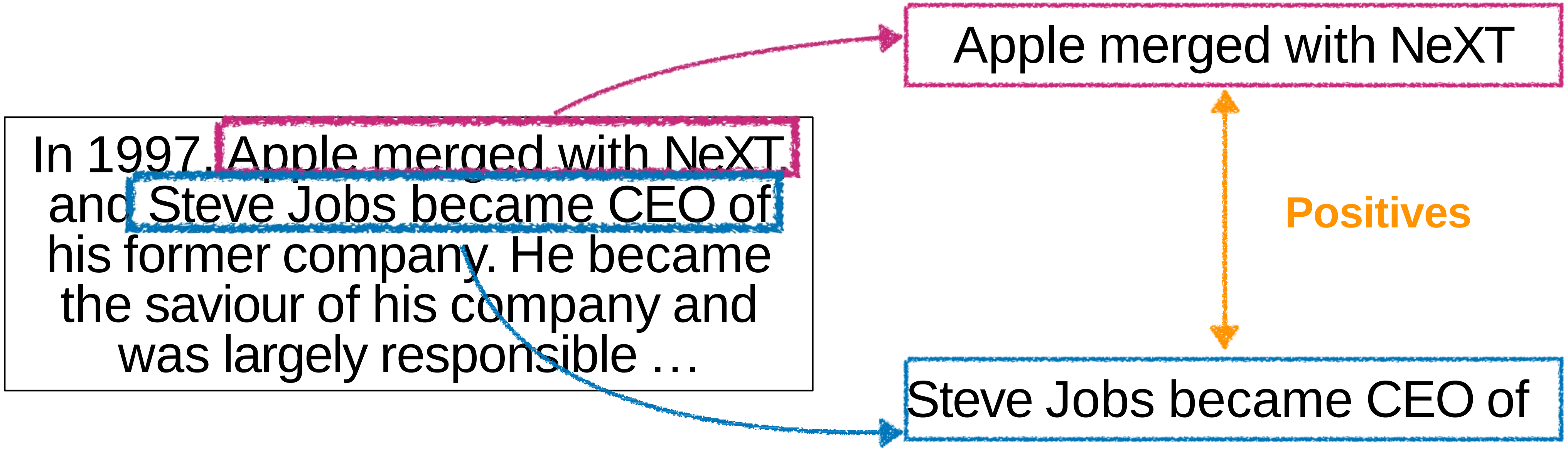
- Select positive and negative documents, train using a contrastive loss

- **DPR** (Karpukhin et al. 2020): learn encoders based on a BM25 hard negatives and in-batch negatives.

- **Contriever** (Izacard et al. 2022): contrastive learning using two random spans as positive pairs - **Unsupervised** dense retrieval model.

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Independent Cropping in **Contriever** (Izacard et al. 2022):



In 1997, Apple merged with NeXT, and Steve Jobs became CEO of his former company. He became the saviour of his company and was largely responsible …

Apple merged with NeXT

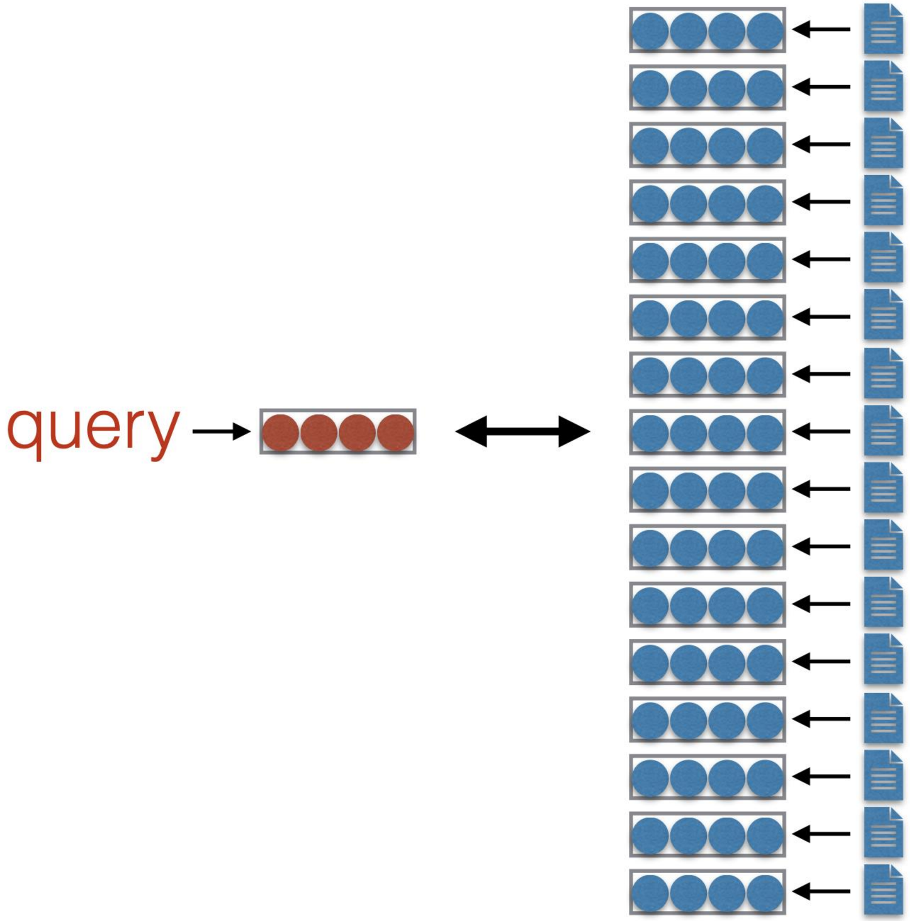
Positives

Steve Jobs became CEO of

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓ Out-of-the-box embeddings. E.g. BERT
  - ✓ Learned embeddings (covered later)

- At test time:
  - Encode Query
  - Find similar documents

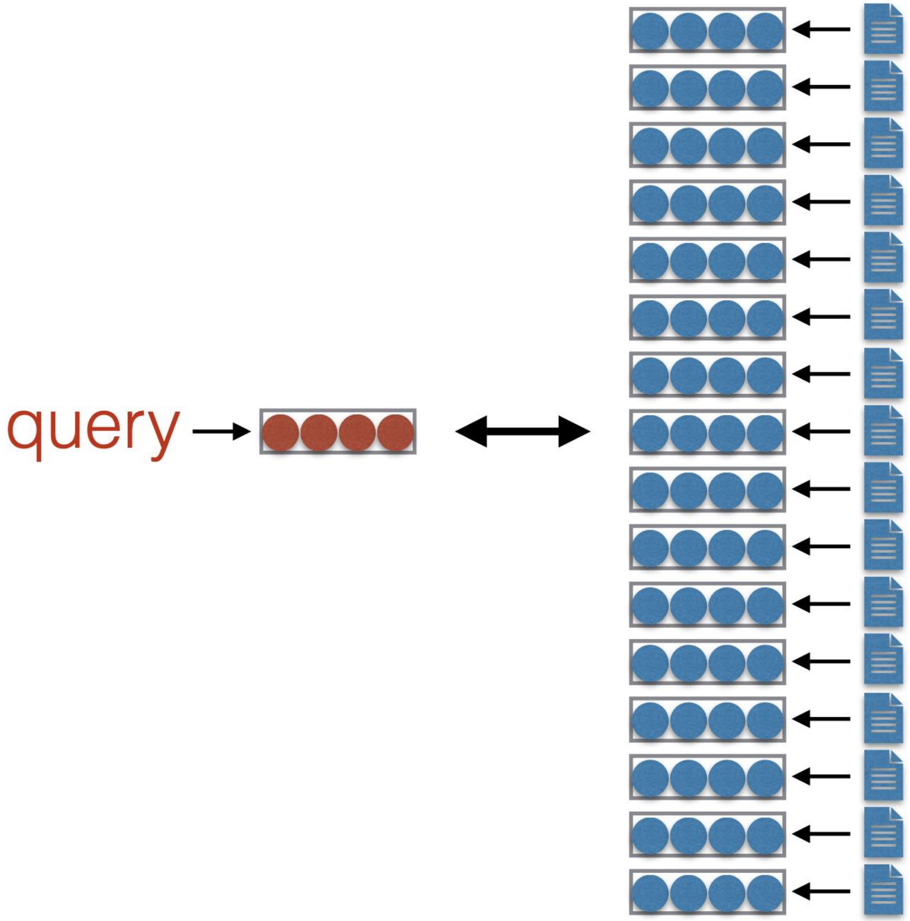


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓ Out-of-the-box embeddings. E.g. BERT
  - ✓ Learned embeddings (covered later)

- At test time:
  - ▪ Encode Query
  - ▪ Use Nearest Neighbor Search to find similar documents



Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

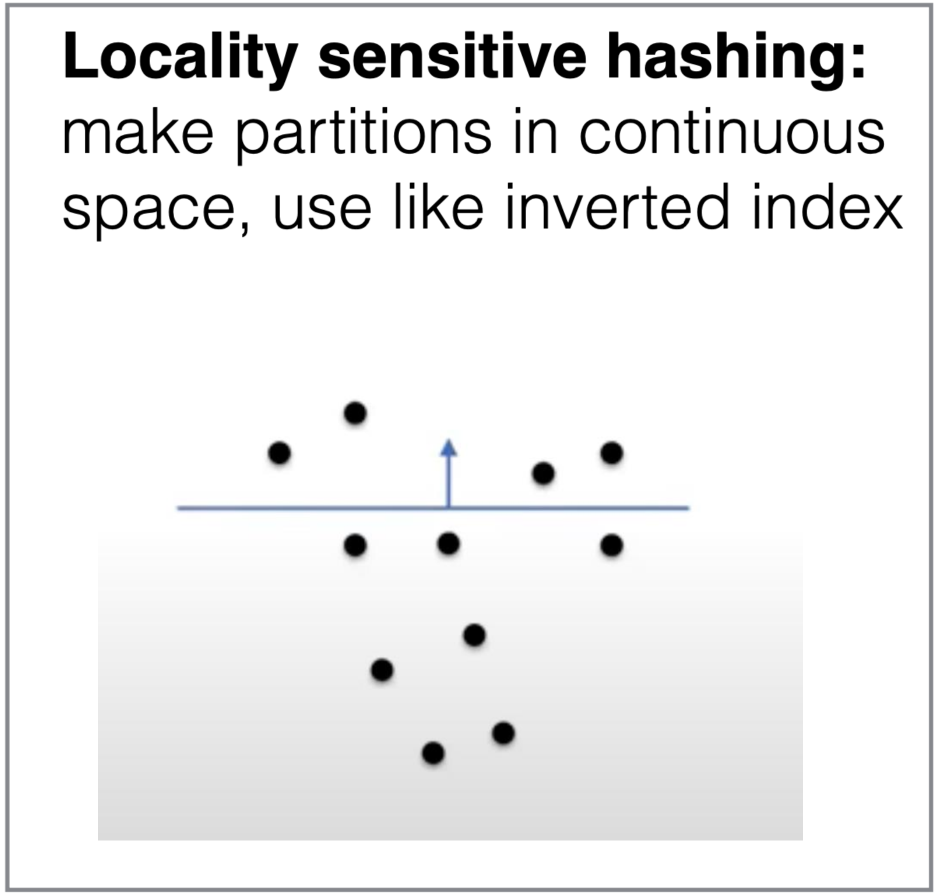
Approximate Nearest Neighbor Search
Maximum Inner Product Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

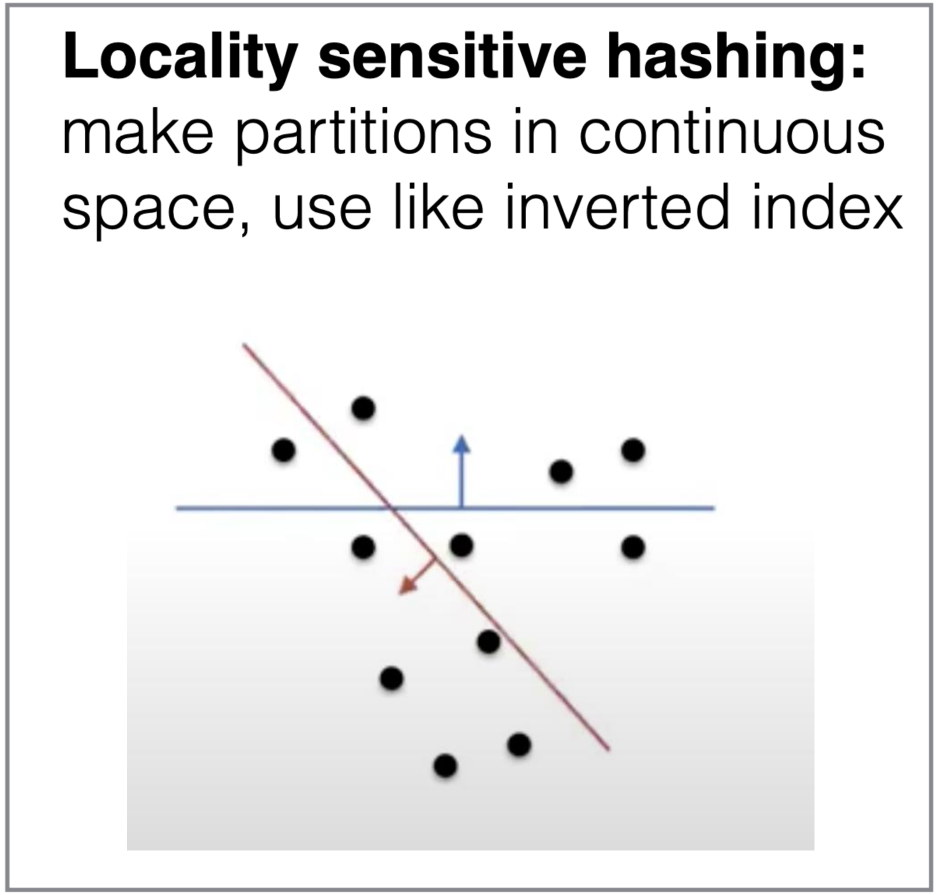
Approximate Nearest Neighbor Search (MIPS)



Locality sensitive hashing: make partitions in continuous space, use like inverted index

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

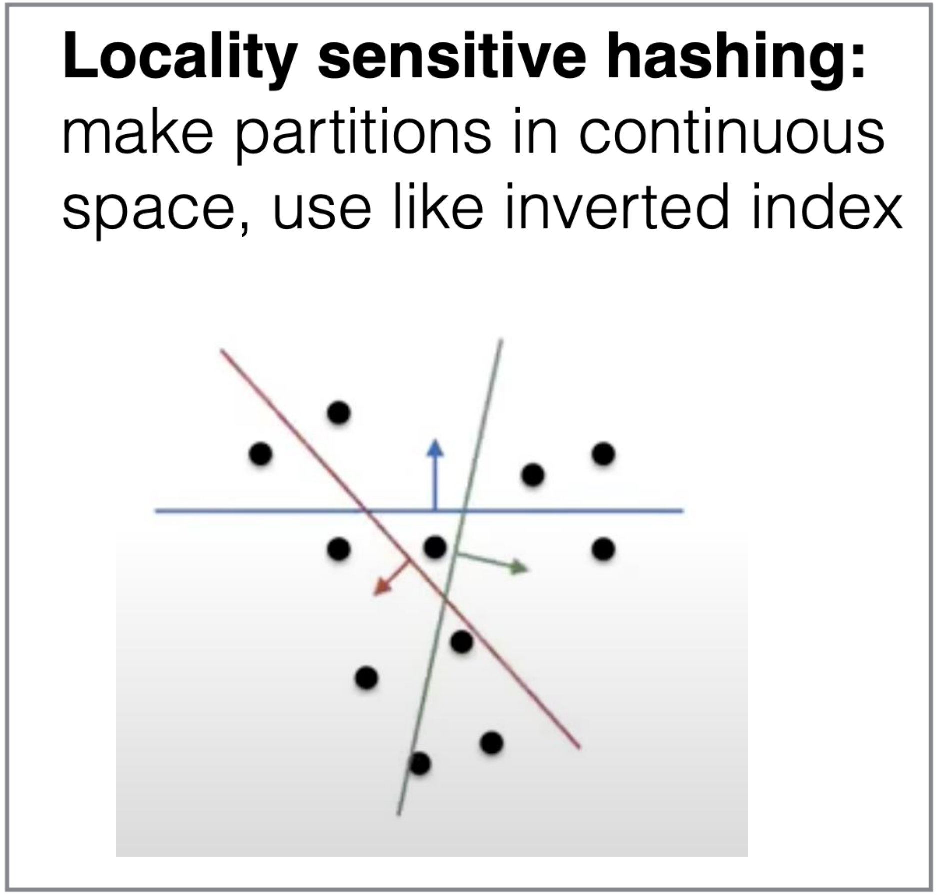
Approximate Nearest Neighbor Search (MIPS)



Locality sensitive hashing: make partitions in continuous space, use like inverted index

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

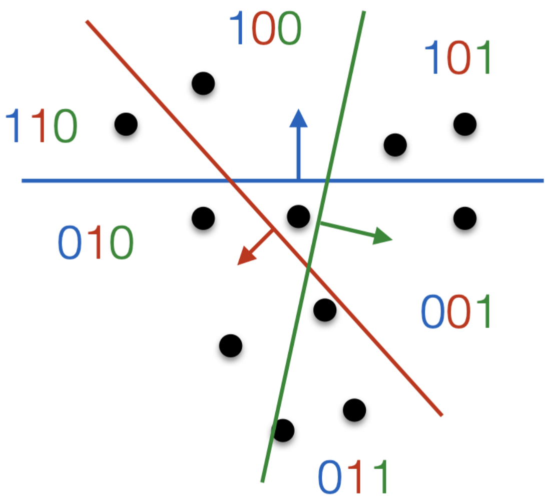


Locality sensitive hashing: make partitions in continuous space, use like inverted index

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

Locality sensitive hashing: make partitions in continuous space, use like inverted index

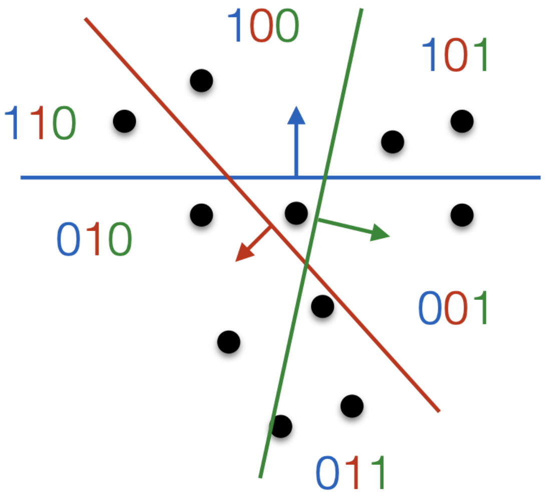


Graph-based search: create "hubs" and search from there

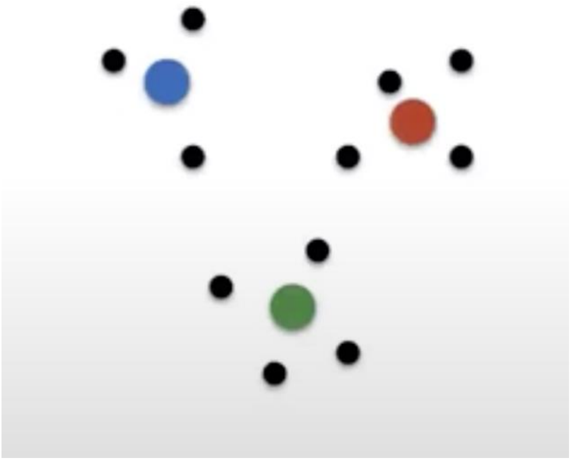
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

Locality sensitive hashing: make partitions in continuous space, use like inverted index



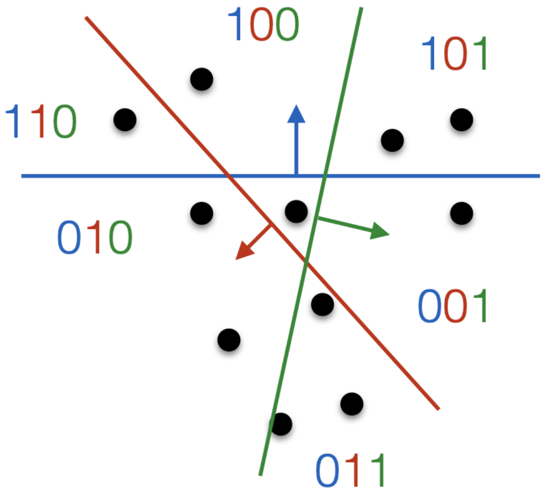
Graph-based search: create "hubs" and search from there



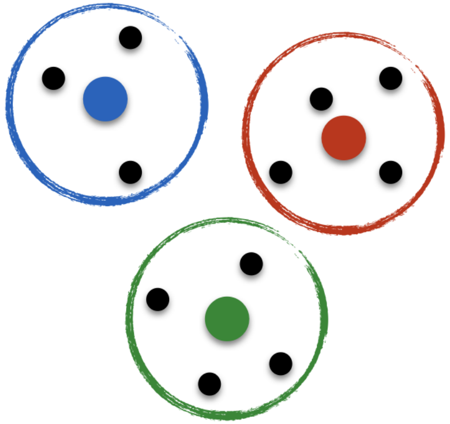
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

Locality sensitive hashing: make partitions in continuous space, use like inverted index

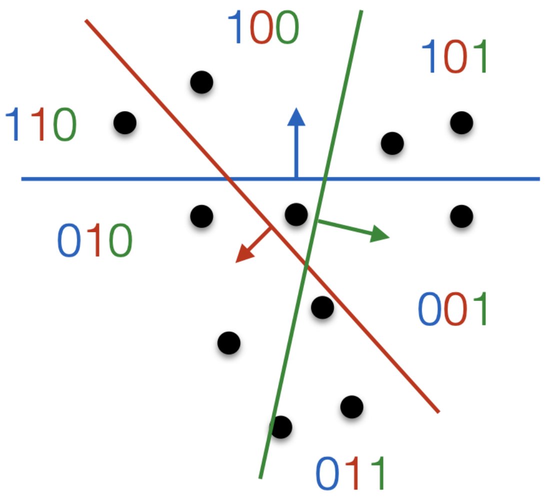


Graph-based search: create "hubs" and search from there

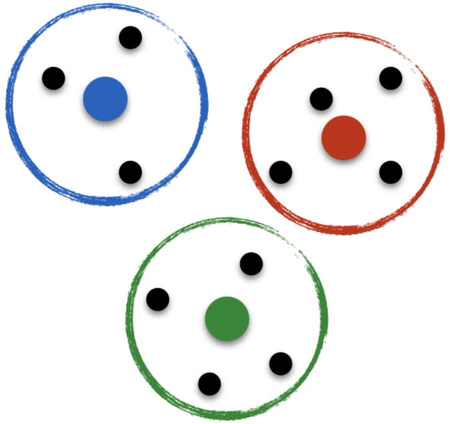


Approximate Nearest Neighbor Search (MIPS)

Locality sensitive hashing: make partitions in continuous space, use like inverted index

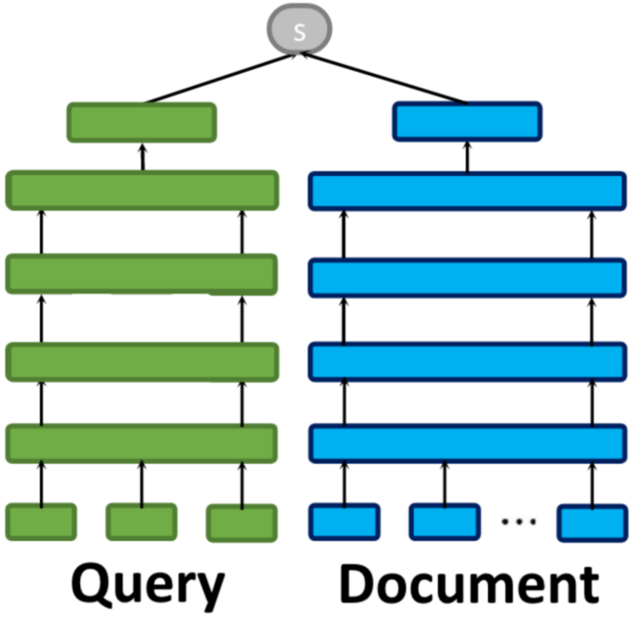


Graph-based search: create "hubs" and search from there



- Software: ANNOY (Spotify), FAISS

Bi-Encoder Scoring

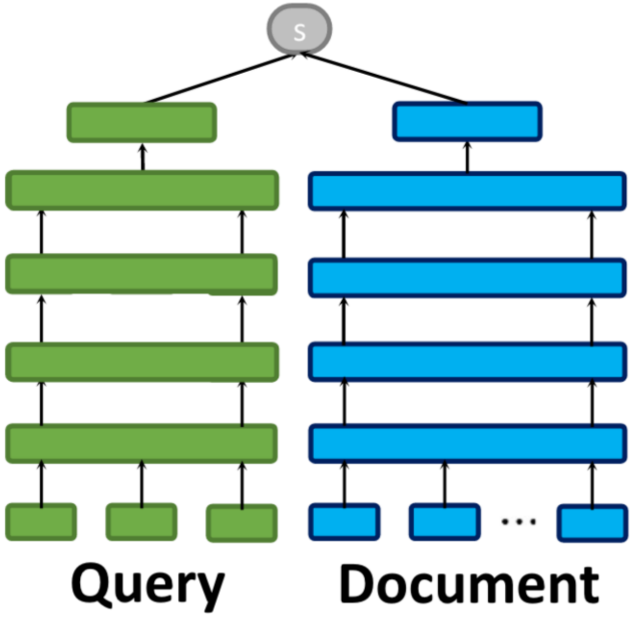


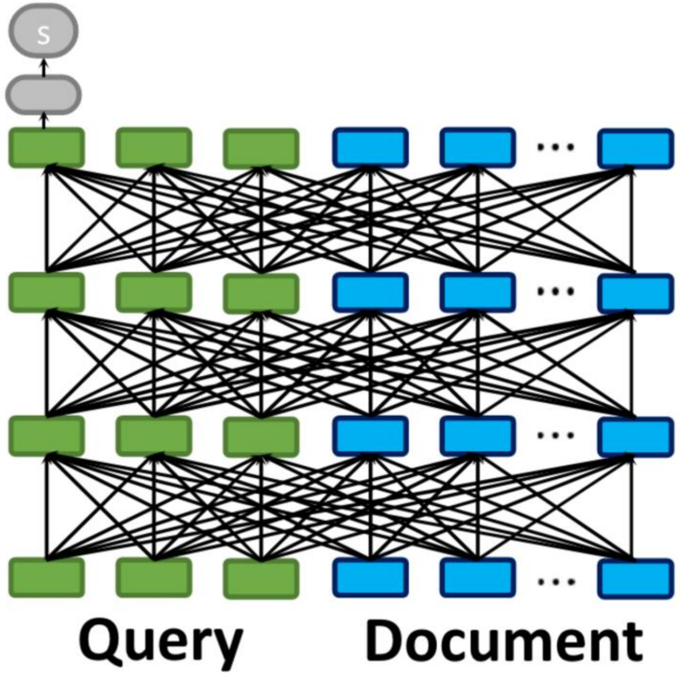
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Cross-Encoder Reranking

- Jointly encode both queries and documents using neural model (Nogueira et al. 2019)



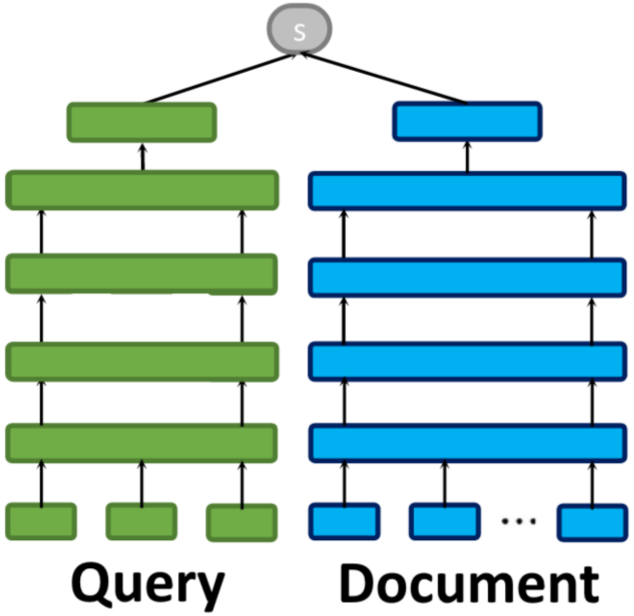


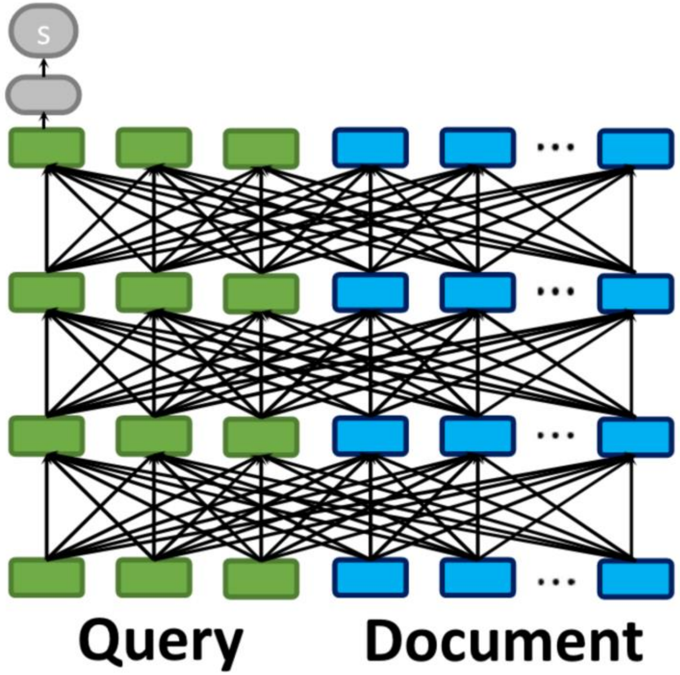
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Cross-Encoder Reranking

- Jointly encode both queries and documents using neural model (Nogueira et al. 2019)

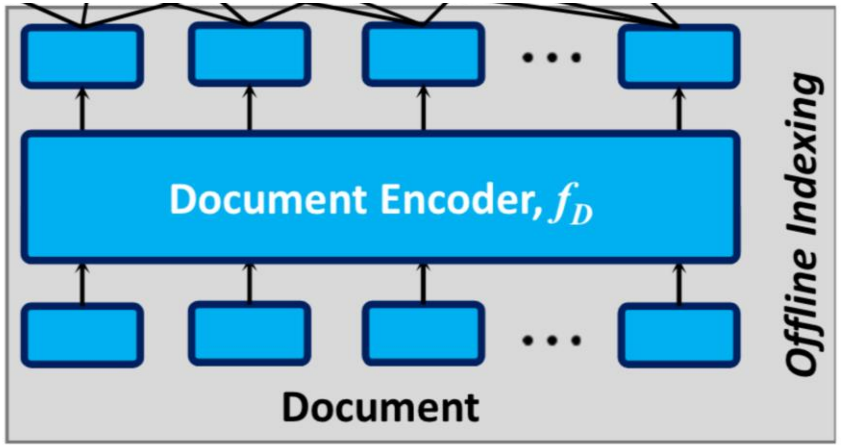




- Precludes approximate nearest neighbour lookup, so can only be used on small number of candidates

Figure from Khattab et al. (2020)

Token-level Dense Retrieval



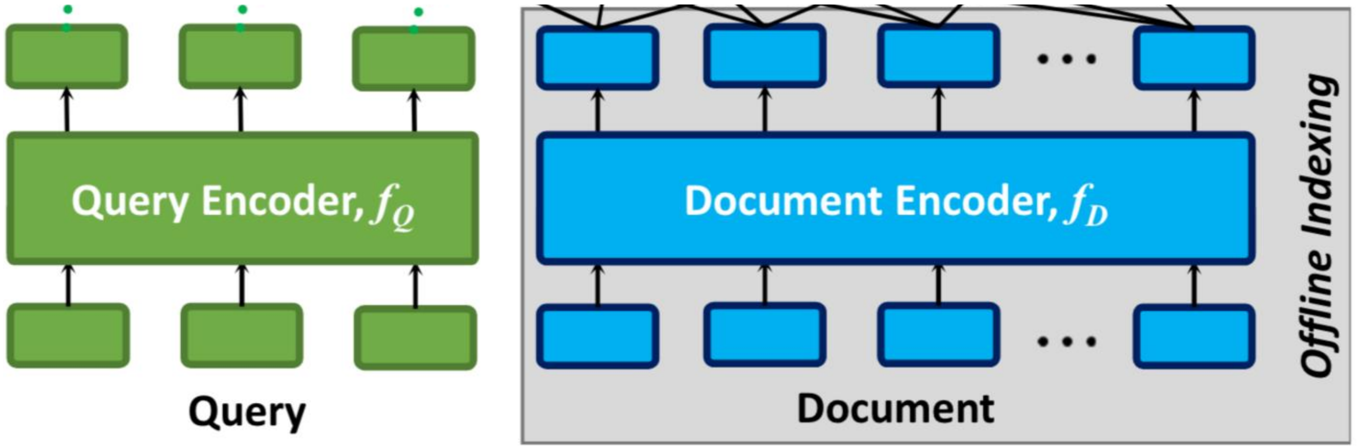
ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Token-level Dense Retrieval



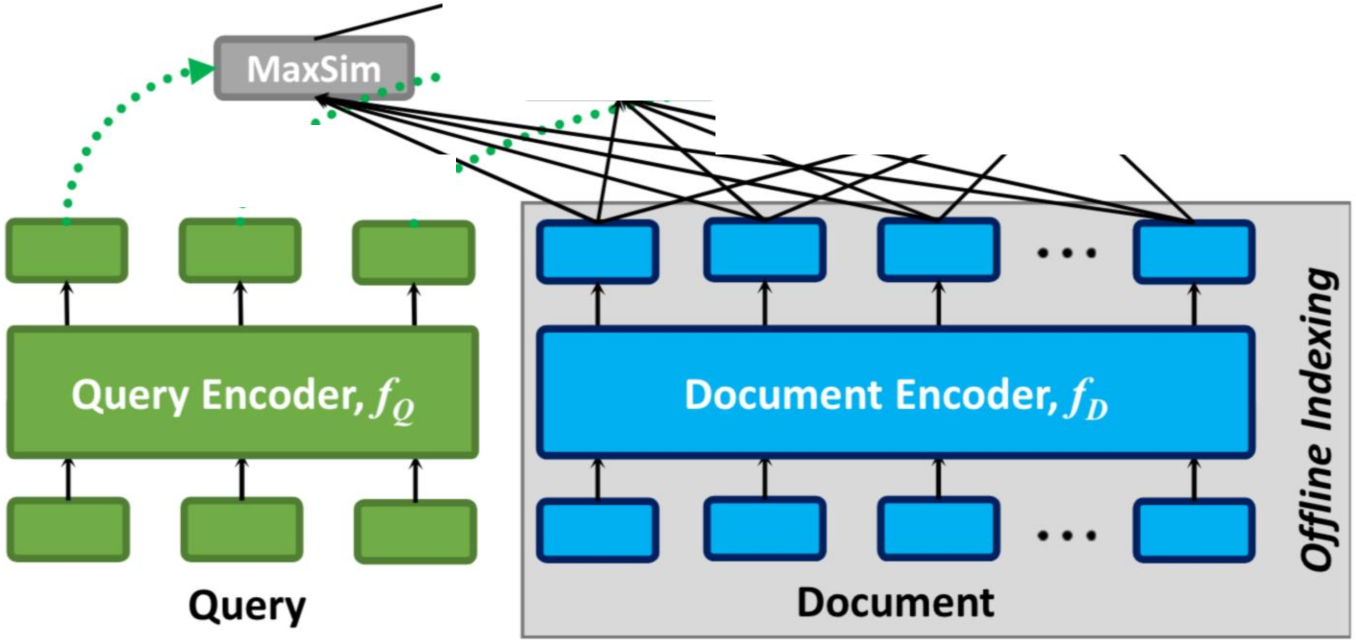
ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Token-level Dense Retrieval



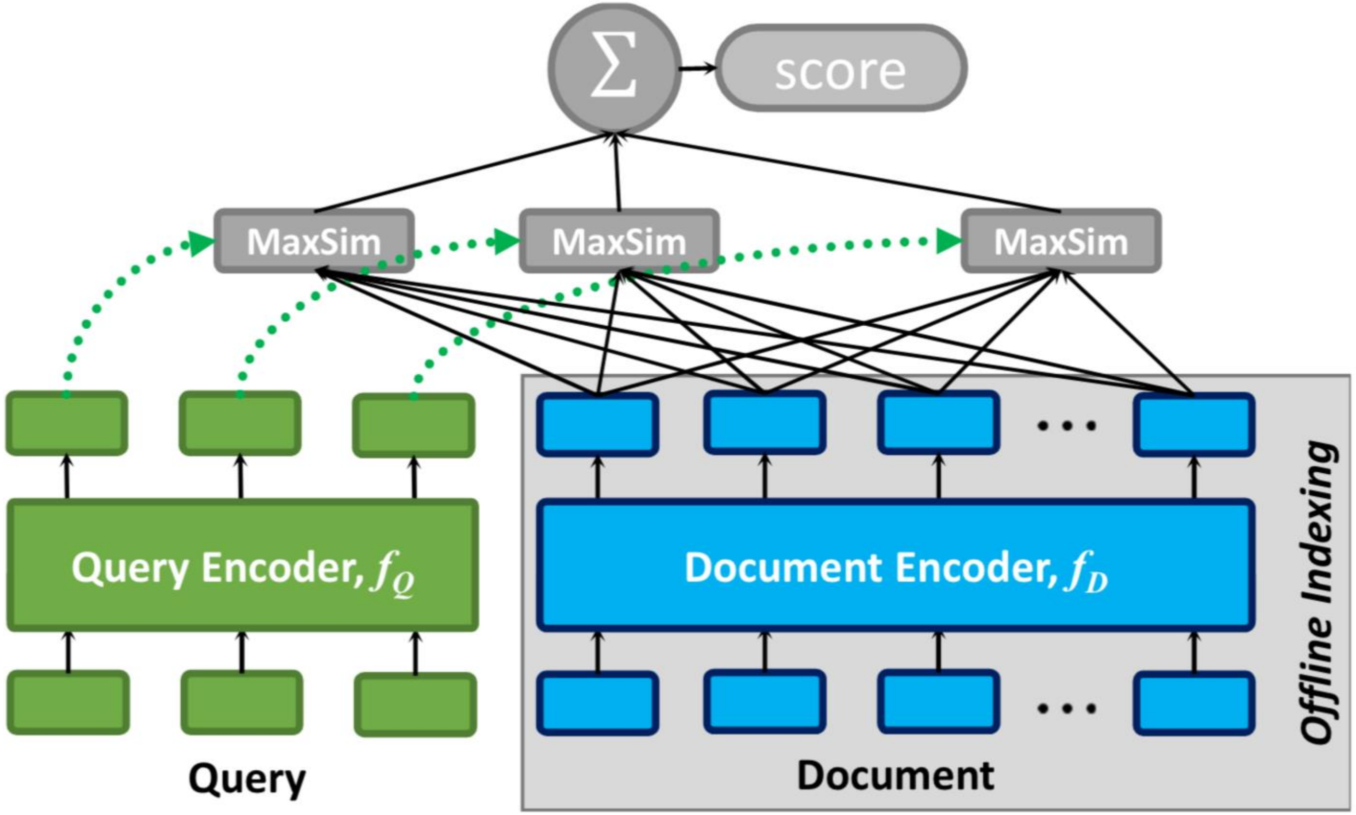
ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Token-level Dense Retrieval



ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

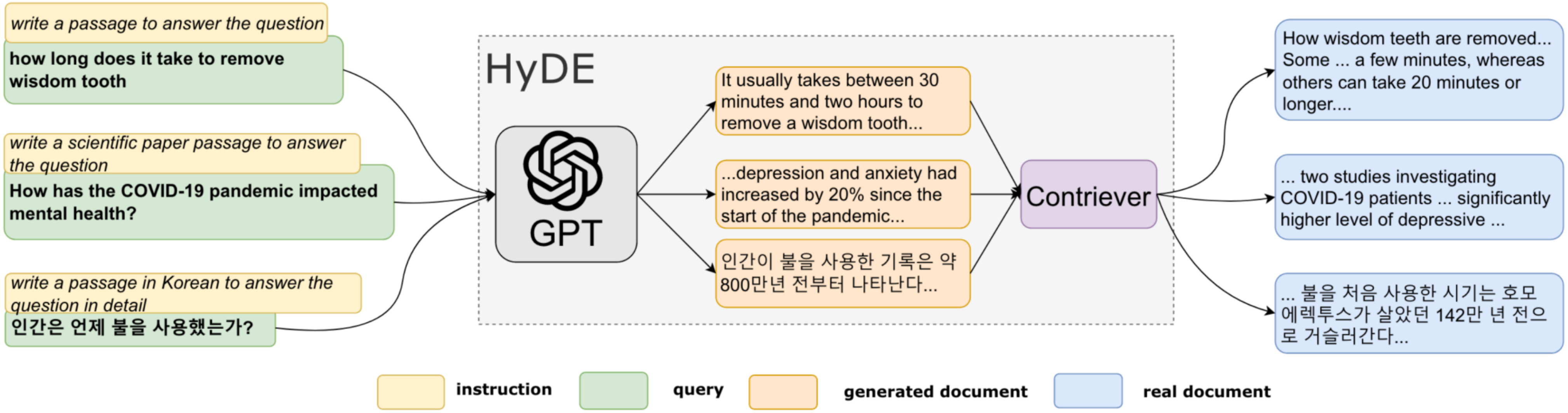
Significantly more effective (but more costly) than single-vector retrieval

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Hypothetical Document Embeddings (Gao et al. 2023)

- Generate a “hypothetical document” for the query using an LLM, and try to look it up
- Can be easier than trying to match under-specified query



write a passage to answer the question

how long does it take to remove wisdom tooth

write a scientific paper passage to answer the question

How has the COVID-19 pandemic impacted mental health?

write a passage in Korean to answer the question in detail

인간은 언제 불을 사용했는가?

HyDE

GPT

It usually takes between 30 minutes and two hours to remove a wisdom tooth...

...depression and anxiety had increased by 20% since the start of the pandemic...

인간이 불을 사용한 기록은 약 800만년 전부터 나타난다...

Contriever

How wisdom teeth are removed... Some ... a few minutes, whereas others can take 20 minutes or longer....

... two studies investigating COVID-19 patients ... significantly higher level of depressive ...

... 불을 처음 사용한 시기는 호모 에렉투스가 살았던 142만 년 전으로 거슬러간다...

🟨 **instruction** 🟩 **query** 🟧 **generated document** 🟦 **real document**

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Retrieval Methods

- Sparse retrieval
- Document-level dense retrieval
- Token-level dense retrieval
- Cross-encoder reranking
- Differentiable search index (DSI)
- Table of Contents based search
- Black-box retrieval (just ask Google/Bing)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Retrieval Methods

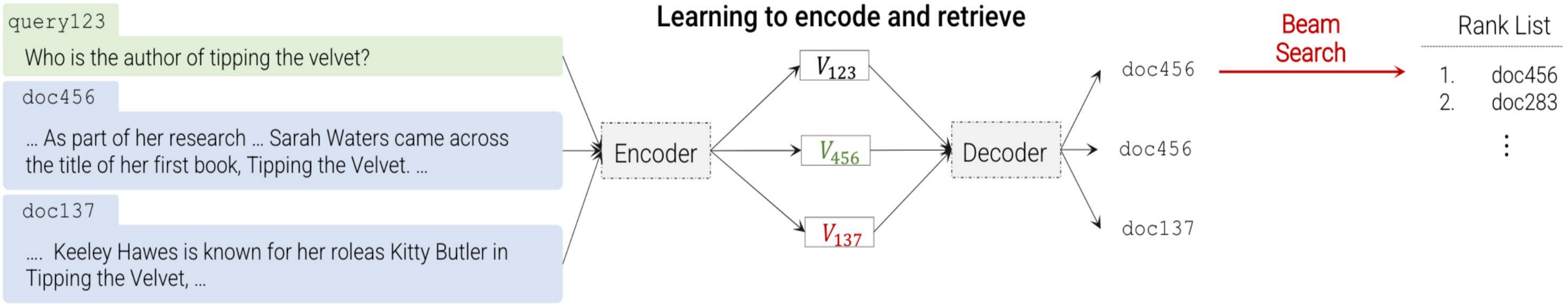
- Sparse retrieval
- Document-level dense retrieval
- Token-level dense retrieval
- Cross-encoder reranking
- Differentiable search index (DSI)
- Table of Contents based search
- Black-box retrieval (just ask Google/Bing)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Differentiable Search Index

- LLMs are powerful enough to memorize the entire corpus.
- Can we use them directly as retriever?

Differentiable Search Index



`query123`

Who is the author of tipping the velvet?

`doc456`

… As part of her research … Sarah Waters came across the title of her first book, Tipping the Velvet. …

`doc137`

…. Keeley Hawes is known for her roleas Kitty Butler in Tipping the Velvet, …

Learning to encode and retrieve

Rank List

1. doc456
2. doc283

⋮

Differentiable Search Index

- Fully parameterize multi-stage retrieve then rank pipeline with a single neural model
- Train a seq2seq LLM for:
  - **Indexing task:** Map each document to its doc id
  - **Retrieval task:** Map each query to the corresponding doc id

- How to represent a document?
- How to represent doc id?

Document Representation

- **Direct Indexing:** first L tokens of the document
- **Set indexing:** represent as set of words after removing stopwords.
- **Inverted Index:** Random contiguous chunks

Document Representation

Direct Indexing:	first L tokens of the document
Set indexing:	represent as set of words after removing stopwords.
Inverted Index:	Random contiguous chunks

Representation of doc ids

- **Unstructured Atomic Identifiers**
    - Use a new token to represent id of a document
    - Take softmax over the doc_id tokens
- **Naively Structured String Identifiers**
    - Decode the string representation of the doc_id
- **Semantically Structured Identifiers**
    - Create a hierarchical Tree structure over document embeddings.

Representation of doc ids

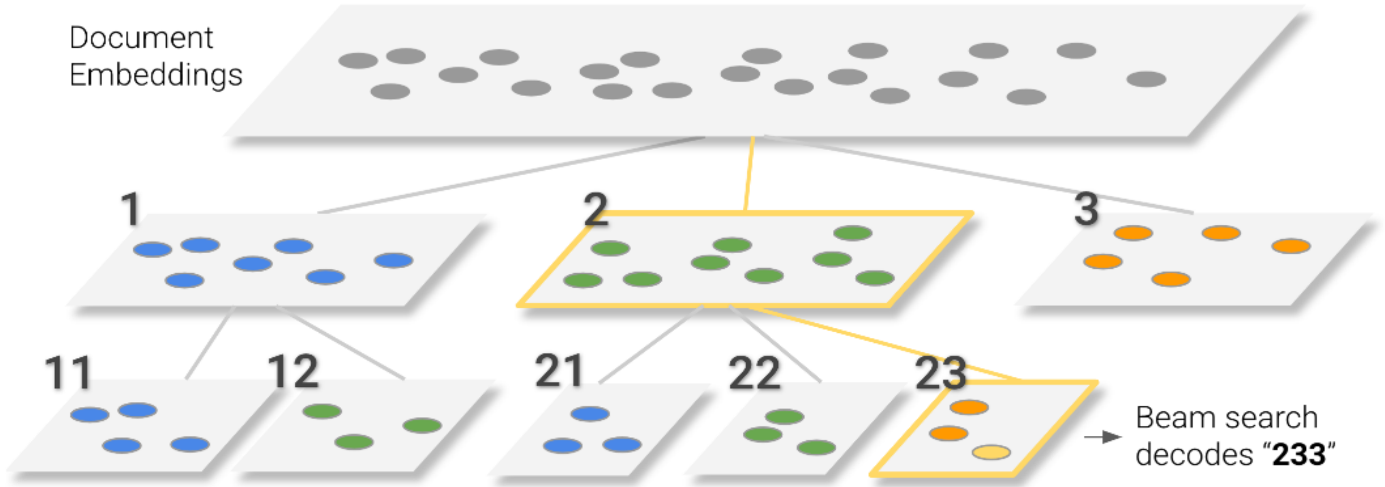
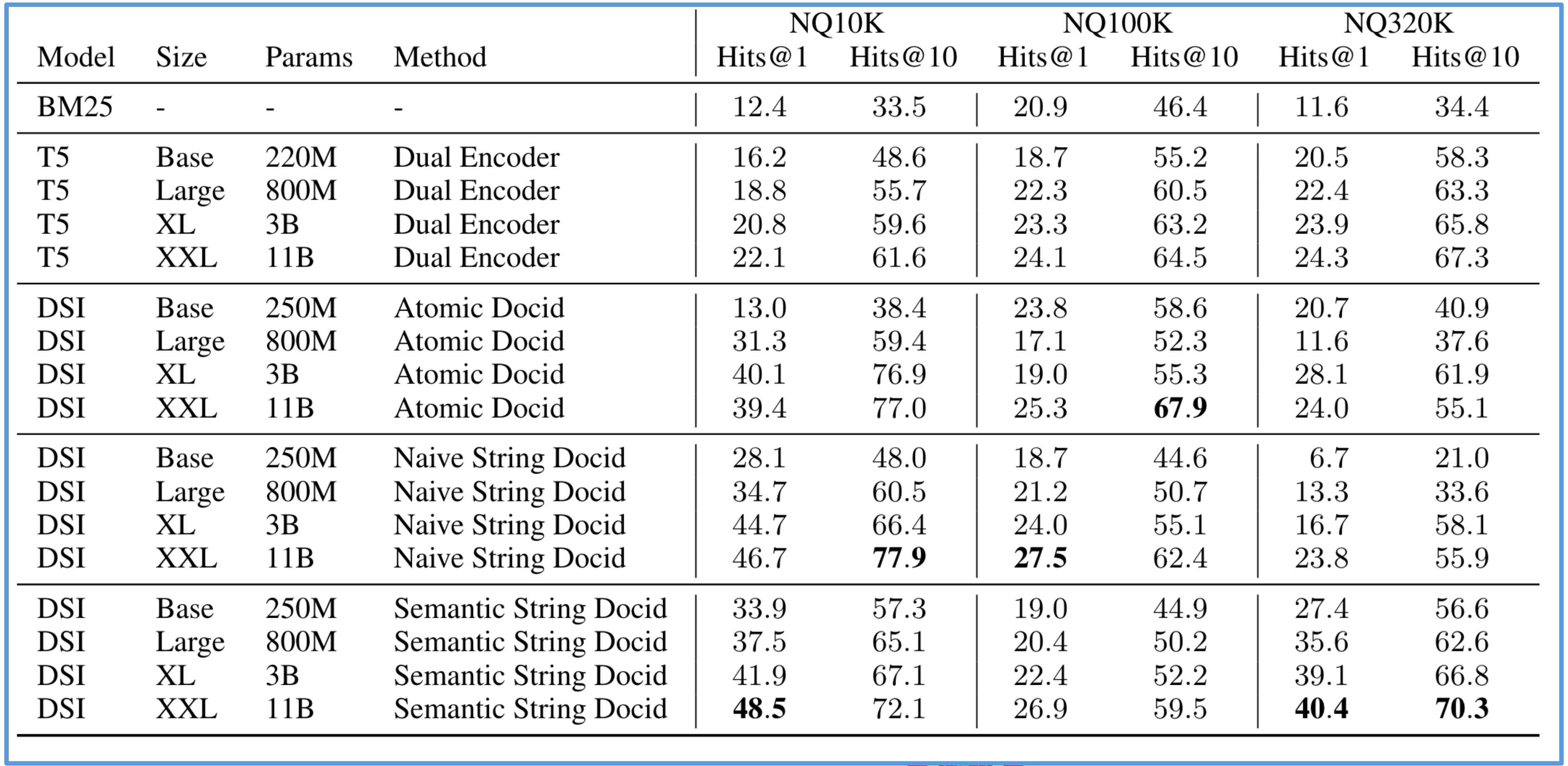


Figure 2: Visual example of a hierarchical clustering process used to assign semantically structured identifiers. During inference, beam search navigates this trie to decode the correct docid.

Results

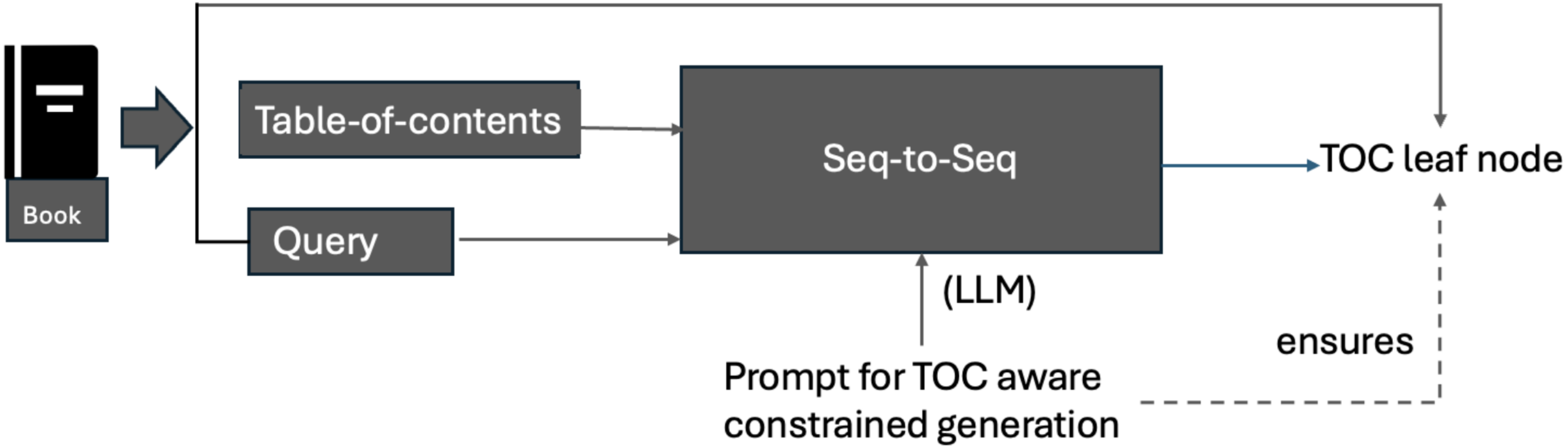


Model	Size	Params	Method	NQ10K Hits@1	NQ10K Hits@10	NQ100K Hits@1	NQ100K Hits@10	NQ320K Hits@1	NQ320K Hits@10
BM25	-	-	-	12.4	33.5	20.9	46.4	11.6	34.4
T5	Base	220M	Dual Encoder	16.2	48.6	18.7	55.2	20.5	58.3
T5	Large	800M	Dual Encoder	18.8	55.7	22.3	60.5	22.4	63.3
T5	XL	3B	Dual Encoder	20.8	59.6	23.3	63.2	23.9	65.8
T5	XXL	11B	Dual Encoder	22.1	61.6	24.1	64.5	24.3	67.3
DSI	Base	250M	Atomic Docid	13.0	38.4	23.8	58.6	20.7	40.9
DSI	Large	800M	Atomic Docid	31.3	59.4	17.1	52.3	11.6	37.6
DSI	XL	3B	Atomic Docid	40.1	76.9	19.0	55.3	28.1	61.9
DSI	XXL	11B	Atomic Docid	39.4	77.0	25.3	**67.9**	24.0	55.1
DSI	Base	250M	Naive String Docid	28.1	48.0	18.7	44.6	6.7	21.0
DSI	Large	800M	Naive String Docid	34.7	60.5	21.2	50.7	13.3	33.6
DSI	XL	3B	Naive String Docid	44.7	66.4	24.0	55.1	16.7	58.1
DSI	XXL	11B	Naive String Docid	46.7	**77.9**	**27.5**	62.4	23.8	55.9
DSI	Base	250M	Semantic String Docid	33.9	57.3	19.0	44.9	27.4	56.6
DSI	Large	800M	Semantic String Docid	37.5	65.1	20.4	50.2	35.6	62.6
DSI	XL	3B	Semantic String Docid	41.9	67.1	22.4	52.2	39.1	66.8
DSI	XXL	11B	Semantic String Docid	**48.5**	72.1	26.9	59.5	**40.4**	**70.3**

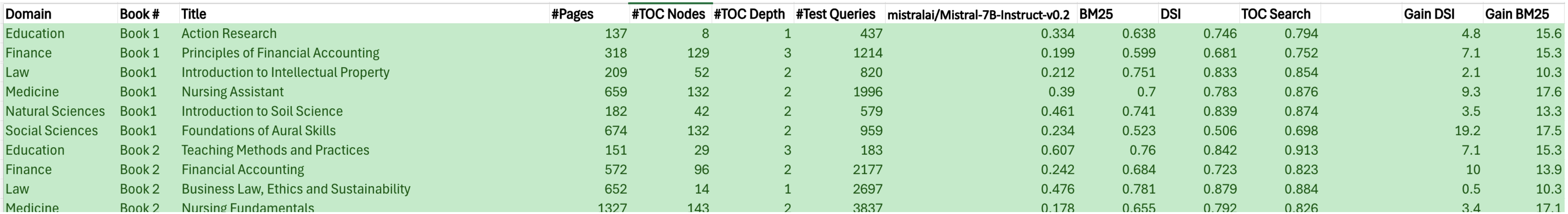
How do you search in a book?

- Inverted index at the end..
- Table of Contents?
- Prompt an LLM with ToC and a question, and ask it to generate the chapter title / subsection containing answer
- Can an LLM answer it zero-shot?
- Can we finetune an LLM to answer such queries?

ToC-aware Search (*under submission*)



Preliminary results (*under submission*)

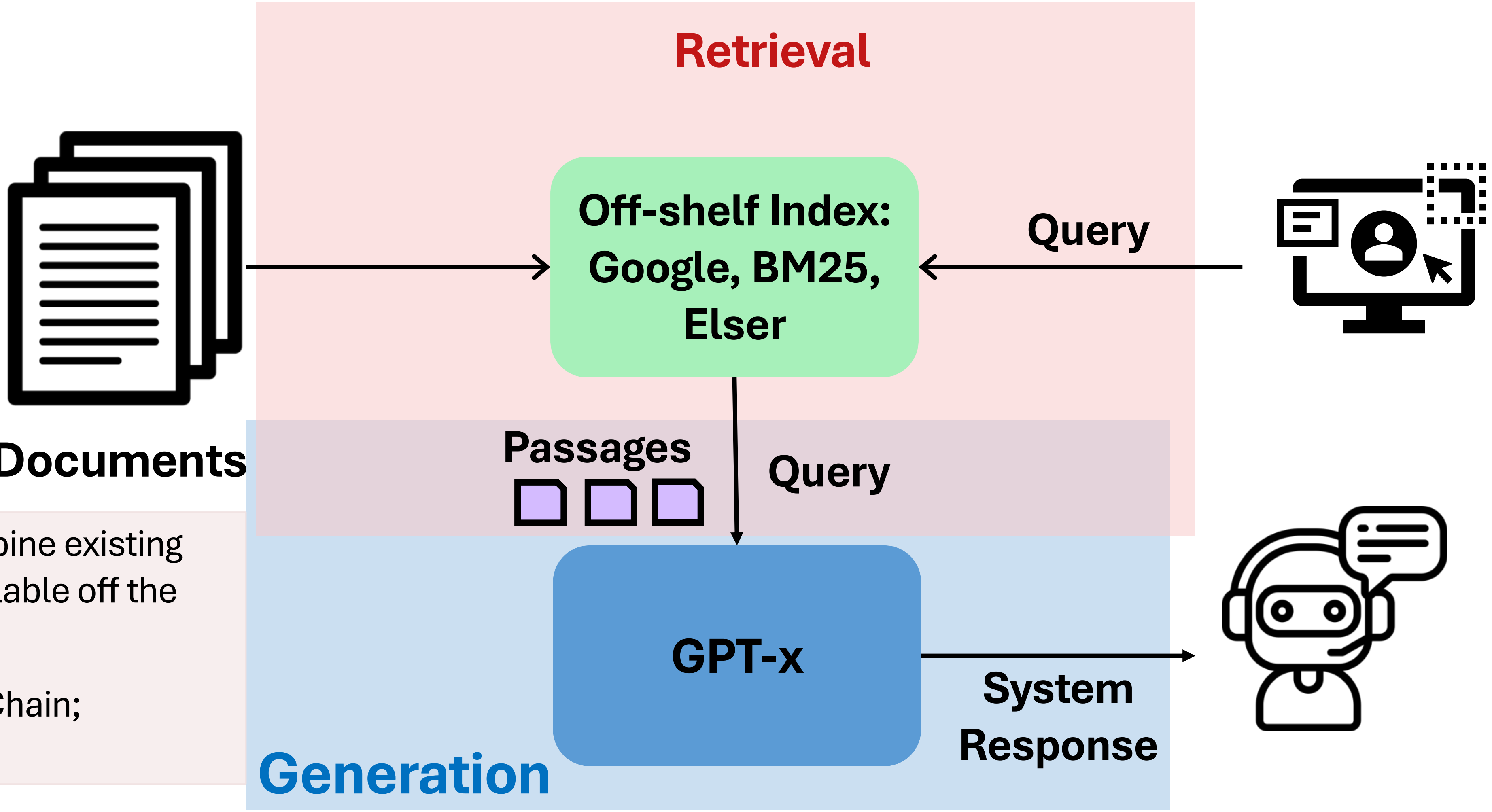


Domain	Book #	Title	#Pages	#TOC Nodes	#TOC Depth	#Test Queries	mistralai/Mistral-7B-Instruct-v0.2	BM25	DSI	TOC Search		Gain DSI	Gain BM25
Education	Book 1	Action Research	137	8	1	437	0.334	0.638	0.746	0.794		4.8	15.6
Finance	Book 1	Principles of Financial Accounting	318	129	3	1214	0.199	0.599	0.681	0.752		7.1	15.3
Law	Book1	Introduction to Intellectual Property	209	52	2	820	0.212	0.751	0.833	0.854		2.1	10.3
Medicine	Book1	Nursing Assistant	659	132	2	1996	0.39	0.7	0.783	0.876		9.3	17.6
Natural Sciences	Book1	Introduction to Soil Science	182	42	2	579	0.461	0.741	0.839	0.874		3.5	13.3
Social Sciences	Book1	Foundations of Aural Skills	674	132	2	959	0.234	0.523	0.506	0.698		19.2	17.5
Education	Book 2	Teaching Methods and Practices	151	29	3	183	0.607	0.76	0.842	0.913		7.1	15.3
Finance	Book 2	Financial Accounting	572	96	2	2177	0.242	0.684	0.723	0.823		10	13.9
Law	Book 2	Business Law, Ethics and Sustainability	652	14	1	2697	0.476	0.781	0.879	0.884		0.5	10.3
Medicine	Book 2	Nursing Fundamentals	1327	143	2	3837	0.178	0.655	0.792	0.826		3.4	17.1

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training …*
- Limitations – *lost in the middle, still hallucinating, retriever failures …*

Retrieval Based LLMs - Architecture



- ✓ Simply combine existing models available off the shelf!
- ✓ Tools: LangChain; LlamaIndex