



INFO-I590 Fundamentals and Applications of LLMs

RAG

Retrieval-Augmented Generation

Jisun An

Standard Prompting

- Combine a prompt template together with an input

Please answer this question:

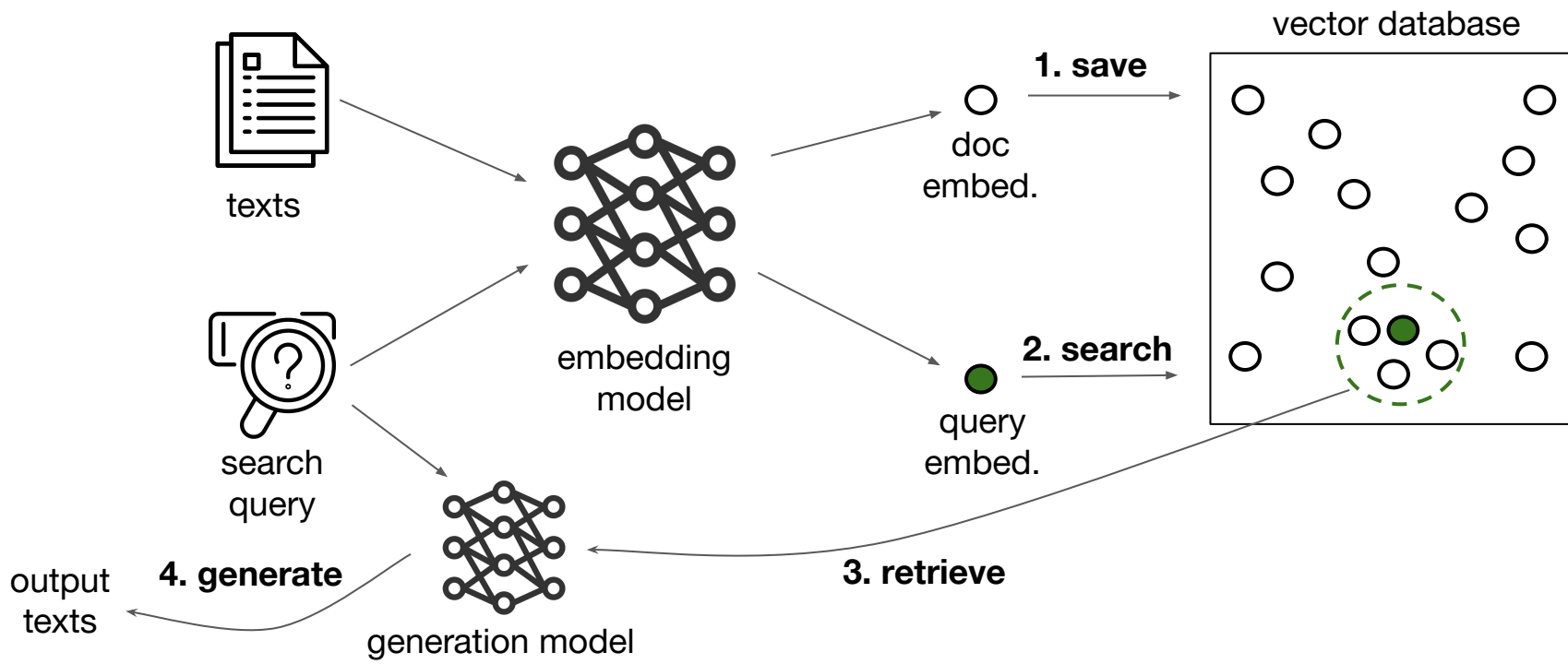
I think Vin Diesel has been a voice actor for several characters in TV series, do you know what their names are?

Problems

- Accuracy issues:
 - Knowledge cutoffs: parameters are usually only updated to a particular time
 - Private data: data stored in private text or data repositories not suitable for training
 - Learning failures: even for data that the model was trained on, it might not be sufficient to get the right answer
- *Verifiability issues*: It is hard to tell if the answer is correct (Hallucination)

Retrieval-Augmented Generation

- Efficiently **retrieve** relevant passages based on a query
- **Generate** an answer using the retrieved passages



Sparse Retrieval

Sparse Retrieval

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

$q = \text{what is nlp}$	$d_1 = \text{what is life ?}$ candy is life !	$d_2 = \text{nlp is an acronym for}$ $\text{natural language processing}$	$d_3 = \text{I like to do}$ $\text{good research on nlp}$
$\begin{pmatrix} \text{what} & 0.33 \\ \text{candy} & 0 \\ \text{nlp} & 0.33 \\ \text{is} & 0.33 \\ \text{language} & 0 \\ \dots & \dots \end{pmatrix}$	$\begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \dots \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0.125 \\ 0 \\ \dots \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0 \\ 0 \\ \dots \end{pmatrix}$
	$q \cdot d_1 = 0.165$	$q \cdot d_2 = 0.0825$	$q \cdot d_3 = 0.0413$

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

Term Weighting (See Manning et al. 2009)

- Some terms are more important than others; low-frequency words are often more important
- Term frequency - in-document frequency (TF-IDF)

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

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

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

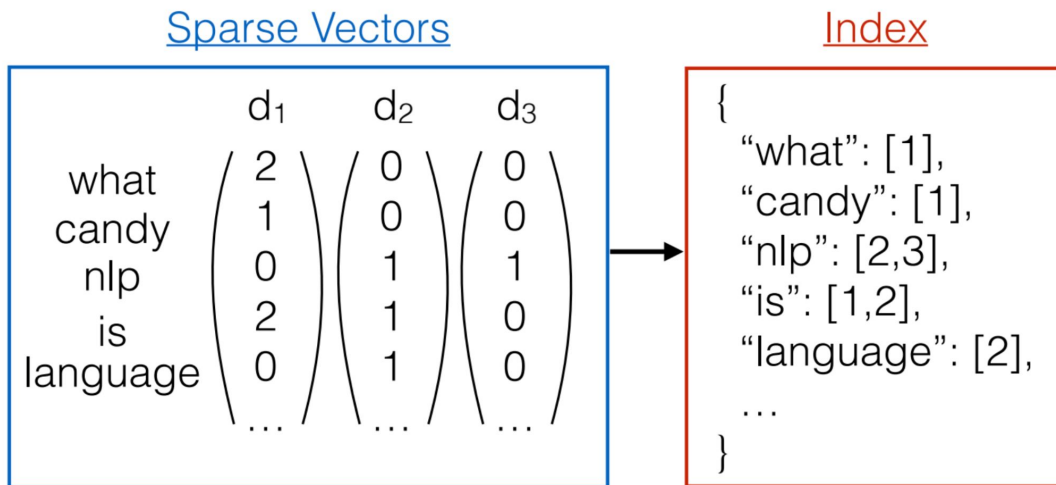
$$\text{BM-25}(t, d) = \text{IDF}(t) \cdot \frac{\text{freq}(t, d) \cdot (k_1 + 1)}{\text{freq}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}} \right)}$$

Term Frequency Saturation

Document length normalization

Inverted Index

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

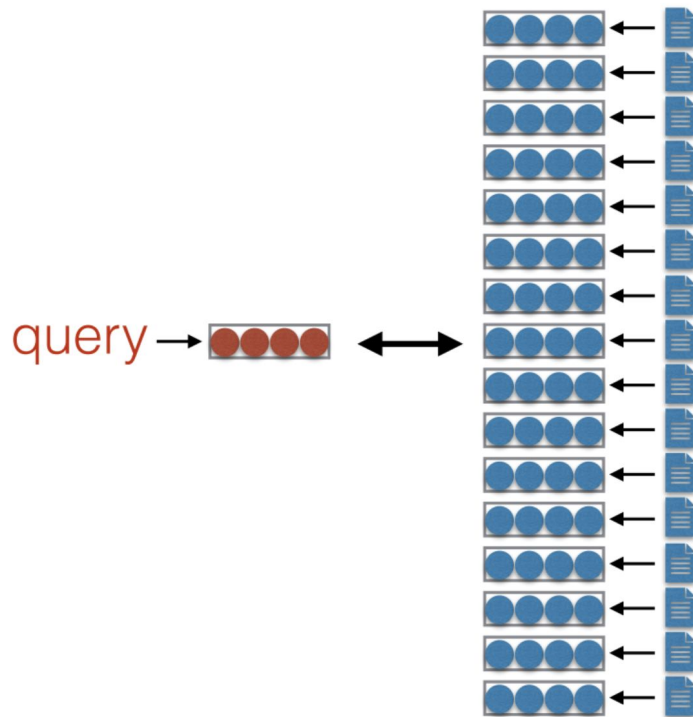


- Example software: Apache Lucene

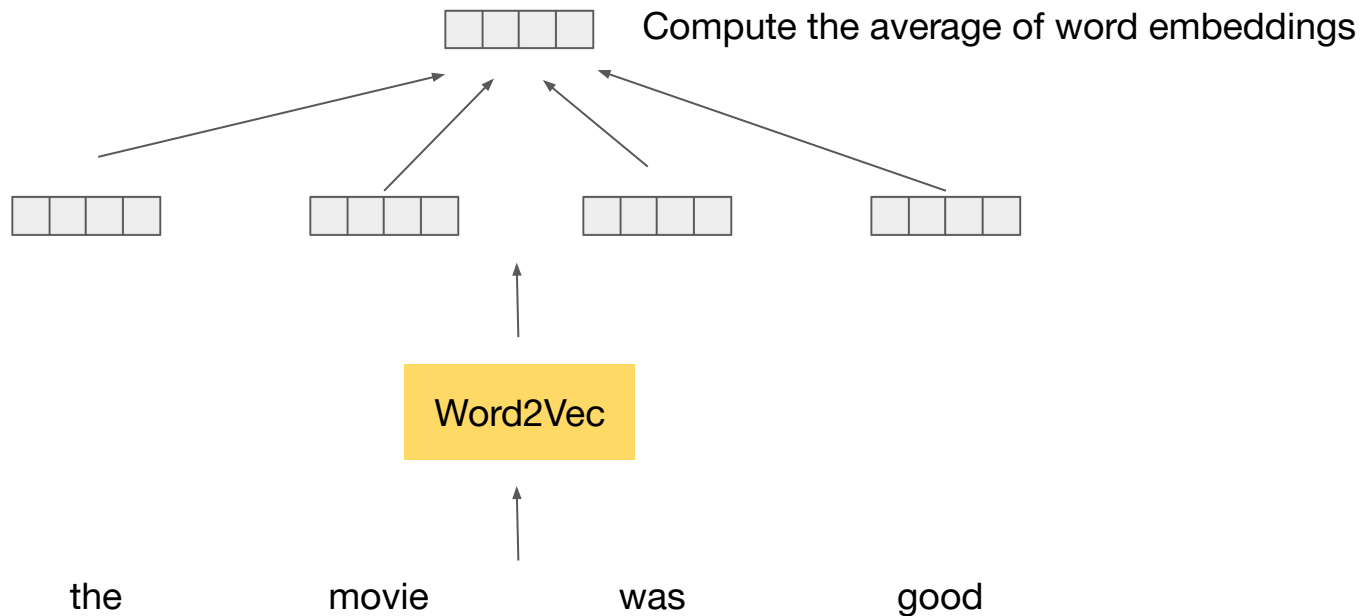
Dense Retrieval

Dense Retrieval

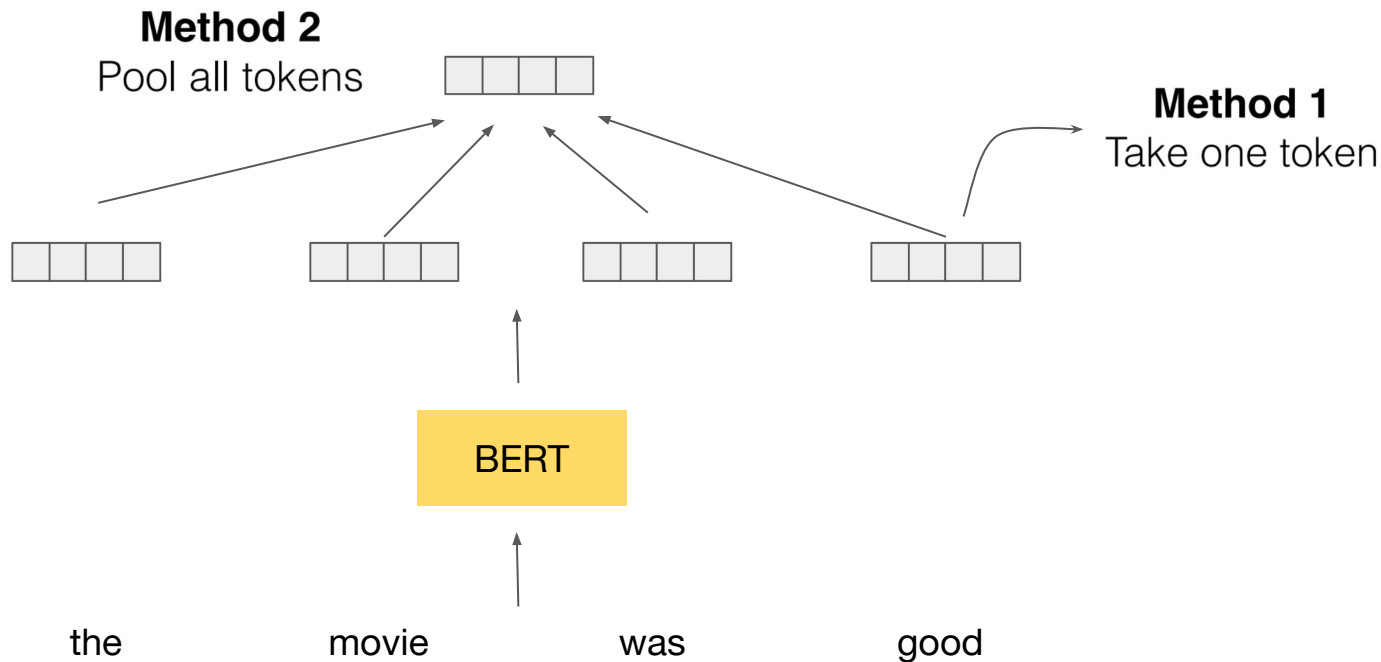
- Encode document/query into dense embeddings and find the nearest neighbors
- Can use:
 - Out-of-the-box embeddings
 - Learned embeddings



Creating Query/Document Embeddings (1)



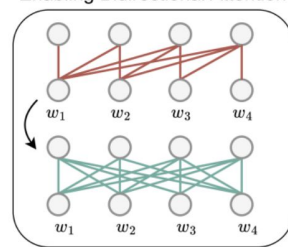
Creating Query/Document Embeddings (2)



Bidirectional vs. Unidirectional Attention

- **Bidirectional Attention:** Use a masked language model like BERT, RoBERTa, ModernBERT etc. as base
- **LLM2Vec** (BehnamGhader 2024): Transform any decoder-only LLM into a strong text encoder. Remove mask and use/train
- **Echo Embeddings** (Springer et al 2024): Repeat the string multiple times in a unidirectional model

Enabling Bidirectional Attention



Autoregressive embeddings do not encode context from later tokens



Repetition enables embeddings to encode context from later tokens



Caveats of Dense Retrieval + Solutions

Caveats of Dense Retrieval

- Returns results even if no exact answer exists
 - Solution: Set a relevance threshold to filter out low-confidence results
- Answers may span multiple sentences
 - Solution: Optimize **text chunking** strategies
- Struggles with finding exact phrases
 - Solution: Use **hybrid search** (semantic + keyword search)
- Models trained on one domain (e.g., Wikipedia) may not perform well in another (e.g., legal texts)
 - Solution: **Train on domain-specific data**

Chunking long texts

Each sentence is a chunk

Interstellar (film)

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Interstellar is a 2014 epic science fiction drama film directed by Christopher Nolan, who co-wrote it with his brother Jonathan. It stars Matthew McConaughey, Anne Hathaway, Jessica Chastain, Bill Irwin, Ellen Burstyn, Michael Caine, and Matt Damon. Set in a dystopian future where Earth is suffering from catastrophic drought and famine, the film follows a group of astronauts who travel through a wormhole near Saturn in search of a new home for humankind.

The screenplay had its origins in a script Jonathan developed in 2007, and was originally set to be directed by Steven Spielberg. Theoretical physicist Kip Thorne was an executive producer and scientific consultant on the film, and wrote the tie-in book *The Science of Interstellar*. Cinematographer Hoyte van Hoytema shot it on 35 mm movie film in the Panavision anamorphic format and IMAX 70 mm. Filming began in late 2013 and took place in Alberta, Klausur, and Los Angeles. *Interstellar* uses extensive practical and miniature effects, and the company DNEG created additional digital effects.

Interstellar premiered in Los Angeles on October 26, 2014. In the United States, it was first released on film stock, expanding to venues using digital projectors. The film received positive reviews from critics and grossed over \$681 million worldwide (\$709 million after subsequent re-releases), making it the tenth-highest grossing film of 2014. Thorne's computer-generated depiction of a black hole in the film has also received commendation from astronomers and physicists. *Interstellar* was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades.

Theatrical release poster

Directed by Christopher Nolan
Written by Jonathan Nolan
Produced by Emma Thomas
Directed by Christopher Nolan
Produced by Emma Thomas
Screenplay by Jonathan Nolan
Story by Jonathan Nolan
Music by Junkie XL
Edited by Lisa A. Johnson
Production companies Warner Bros. Entertainment Inc.
Distributed by Warner Bros. Entertainment Inc.



Each paragraph is a chunk

Interstellar (film)

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Interstellar is a 2014 epic science fiction drama film directed by Christopher Nolan, who co-wrote it with his brother Jonathan. It stars Matthew McConaughey, Anne Hathaway, Jessica Chastain, Bill Irwin, Ellen Burstyn, Michael Caine, and Matt Damon. Set in a dystopian future where Earth is suffering from catastrophic drought and famine, the film follows a group of astronauts who travel through a wormhole near Saturn in search of a new home for humankind.

The screenplay had its origins in a script Jonathan developed in 2007, and was originally set to be directed by Steven Spielberg. Theoretical physicist Kip Thorne was an executive producer and scientific consultant on the film, and wrote the tie-in book *The Science of Interstellar*. Cinematographer Hoyte van Hoytema shot it on 35 mm movie film in the Panavision anamorphic format and IMAX 70 mm. Filming began in late 2013 and took place in Alberta, Klausur, and Los Angeles. *Interstellar* uses extensive practical and miniature effects, and the company DNEG created additional digital effects.

Interstellar premiered in Los Angeles on October 26, 2014. In the United States, it was first released on film stock, expanding to venues using digital projectors. The film received positive reviews from critics and grossed over \$681 million worldwide (\$709 million after subsequent re-releases), making it the tenth-highest grossing film of 2014. Thorne's computer-generated depiction of a black hole in the film has also received commendation from astronomers and physicists. *Interstellar* was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades.

Theatrical release poster

Directed by Christopher Nolan
Written by Jonathan Nolan
Produced by Emma Thomas
Directed by Christopher Nolan
Produced by Emma Thomas
Screenplay by Jonathan Nolan
Story by Jonathan Nolan
Music by Junkie XL
Edited by Lisa A. Johnson
Production companies Warner Bros. Entertainment Inc.
Distributed by Warner Bros. Entertainment Inc.



Overlapping window of sentences

Interstellar (film)

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Interstellar is a 2014 epic science fiction drama film directed by Christopher Nolan, who co-wrote it with his brother Jonathan. It stars Matthew McConaughey, Anne Hathaway, Jessica Chastain, Bill Irwin, Ellen Burstyn, Michael Caine, and Matt Damon. Set in a dystopian future where Earth is suffering from catastrophic drought and famine, the film follows a group of astronauts who travel through a wormhole near Saturn in search of a new home for humankind.

The screenplay had its origins in a script Jonathan developed in 2007, and was originally set to be directed by Steven Spielberg. Theoretical physicist Kip Thorne was an executive producer and scientific consultant on the film, and wrote the tie-in book *The Science of Interstellar*. Cinematographer Hoyte van Hoytema shot it on 35 mm movie film in the Panavision anamorphic format and IMAX 70 mm. Filming began in late 2013 and took place in Alberta, Klausur, and Los Angeles. *Interstellar* uses extensive practical and miniature effects, and the company DNEG created additional digital effects.

Interstellar premiered in Los Angeles on October 26, 2014. In the United States, it was first released on film stock, expanding to venues using digital projectors. The film received positive reviews from critics and grossed over \$681 million worldwide (\$709 million after subsequent re-releases), making it the tenth-highest grossing film of 2014. Thorne's computer-generated depiction of a black hole in the film has also received commendation from astronomers and physicists. *Interstellar* was nominated for five awards at the 87th Academy Awards, winning Best Visual Effects, and received numerous other accolades.

Theatrical release poster

Directed by Christopher Nolan
Written by Jonathan Nolan
Produced by Emma Thomas
Directed by Christopher Nolan
Produced by Emma Thomas
Screenplay by Jonathan Nolan
Story by Jonathan Nolan
Music by Junkie XL
Edited by Lisa A. Johnson
Production companies Warner Bros. Entertainment Inc.
Distributed by Warner Bros. Entertainment Inc.



Hybrid Search

- Keyword search (sparse retrieval, e.g., TF-IDF, BF25)
 - Matches exact words; precise but misses synonyms.
- Semantic search (dense retrieval)
 - Understands meaning; flexible but may retrieve loosely related results.
- Hybrid Search
 - Combines both for better accuracy and coverage.

Reciprocal Rank Fusion (RRF)

- Reciprocal Rank Fusion (RRF) is an ensemble ranking method used to combine multiple ranked lists of search results.
- Given multiple ranked lists (e.g., from different retrieval models), RRF assigns each document a score based on its rank in each list using the following formula:

$$\text{RRF Score}(d) = \sum_{i=1}^N \frac{1}{k + \text{rank}_i(d)}$$

where:

- N = number of ranking lists
- $\text{rank}_i(d)$ = rank of document d in the i -th ranked list
- k = a small constant (typically 60) to prevent dominance by top-ranked results

RRF Example

Reciprocal Rank Fusion		
BM25	Vector	Results 1/RANK
A	B	$A = 1/1 + 1/3 = 1.3$
B	C	$B = 1/2 + 1/1 = 1.5$
C	A	$C = 1/3 + 1/2 = 0.83$

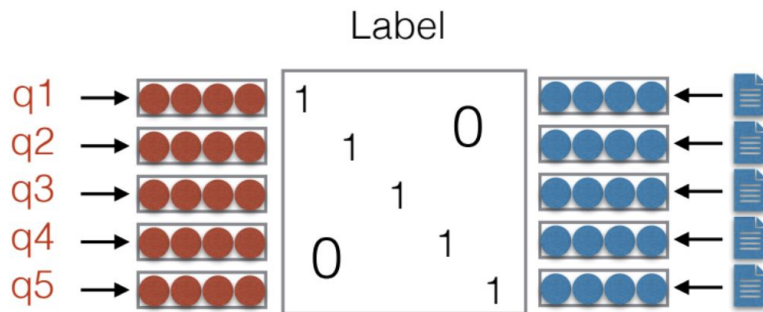
Learning Retrieval-oriented Embeddings

- Basic idea: move the positive documents closer, negative documents farther away
- Select positive and negative documents, train using a contrastive loss (e.g. triplet loss)

$$L = \max(d(A, P) - d(A, N) + m, 0)$$

How to get negative examples? - In-batch negatives

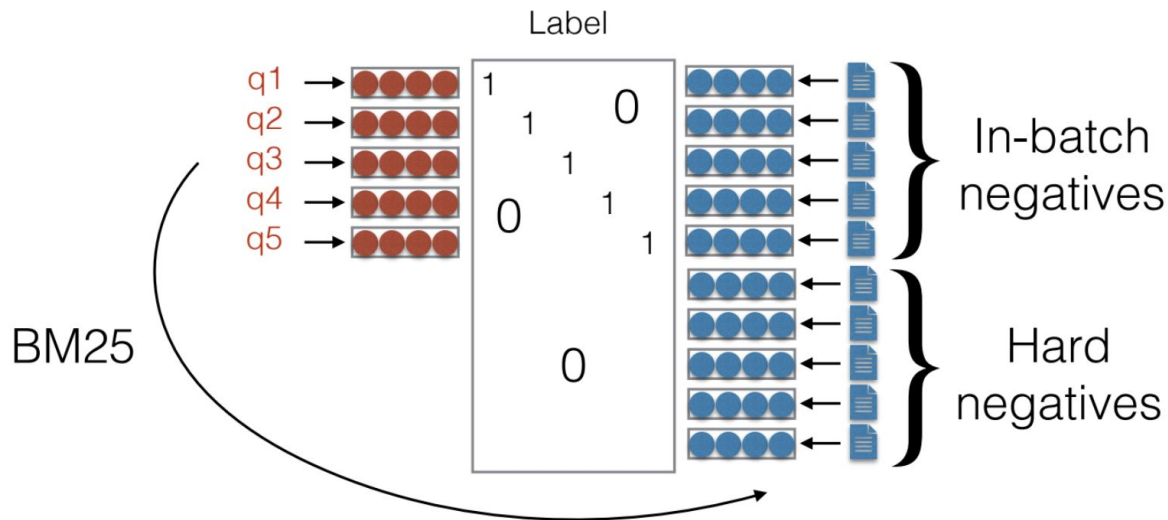
- Create a batch of queries and associated documents
- Treat all other documents in the batch as negative examples



- Problem: not enough hard examples

Negative examples - Hard negative mining

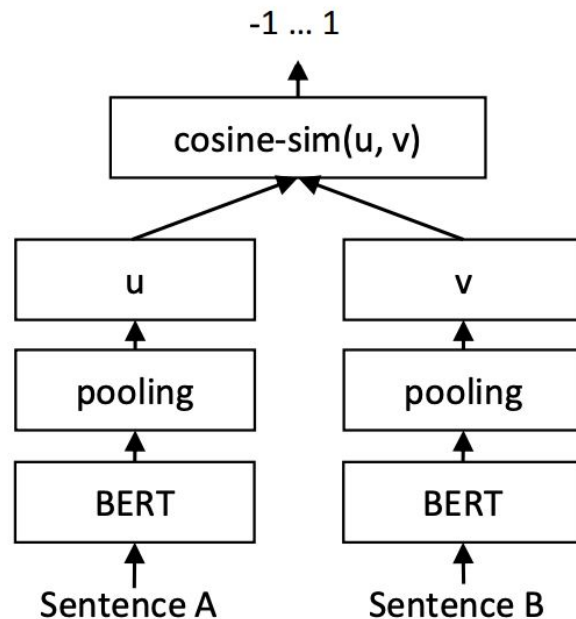
- Use a weaker retriever (e.g. BM25) to more examples and treat them as negatives



- Problem: hard “negatives” might actually be positive

(reminder) Sentence-BERT (SBERT) (Reimers and Gurevych 2019)

- A modification of BERT designed for sentence embeddings
- Optimized for semantic similarity tasks via siamese or triplet networks, enabling fast & accurate sentence comparison
- Widely used in semantic search, clustering, and QA



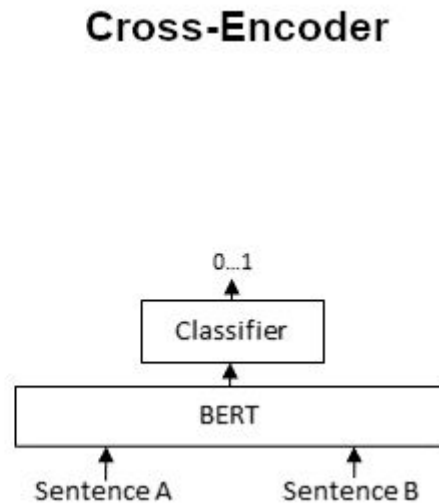
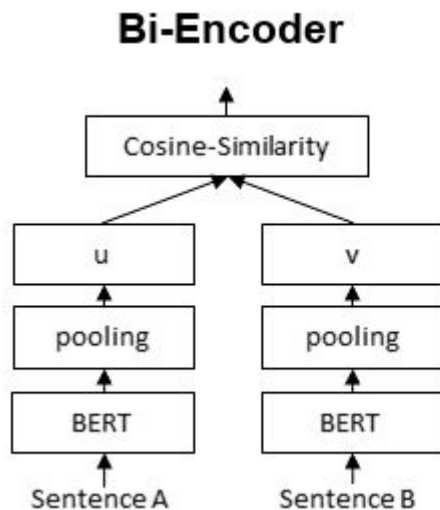
Dataset structures to train your SentenceTransformers model

Dataset structure	Example datasets(repo id in Hugging Face Hub)	Loss functions(imported from sentence_transformers)
Pair of sentences and a label indicating how similar they are	snli	ContrastiveLoss; SoftmaxLoss; CosineSimilarityLoss
Pair of positive (similar) sentences without a label	embedding-data/flickr30k_captions Quintets; embedding-data/coco_captions Quintets	MultipleNegativesRankingLoss; MegaBatchMarginLoss
Single sentence with an integer label	trec; yahoo_answers_topics	BatchHardTripletLoss; BatchAllTripletLoss; BatchHardSoftMarginTripletLoss; BatchSemiHardTripletLoss
Triplet (anchor, positive, negative) sentences	embedding-data/QQP_triplets	TripletLoss

Reranking

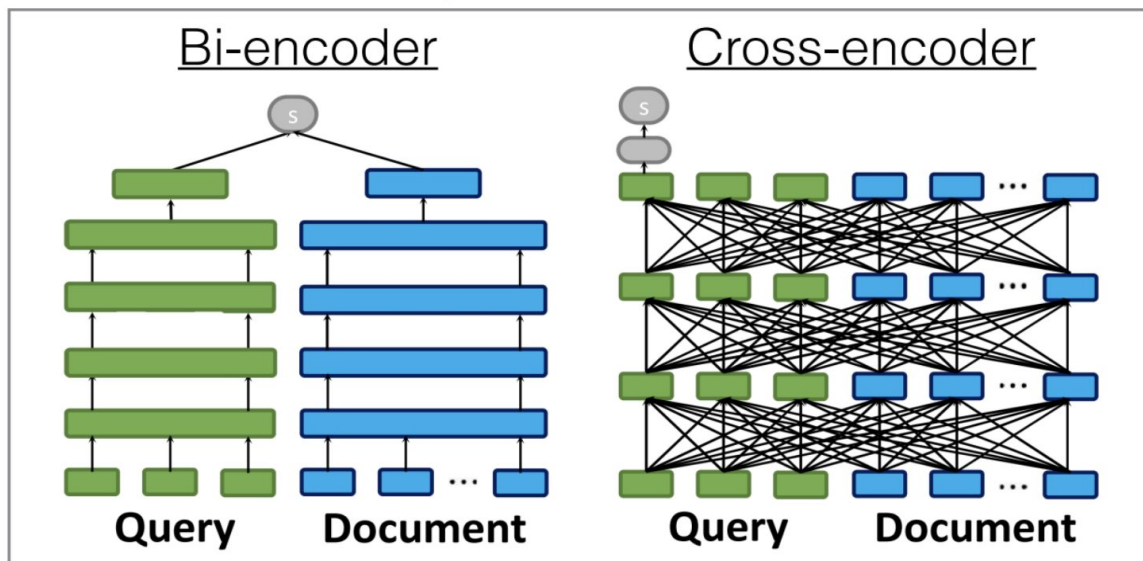
Bi-Encoder vs Cross-Encoder

- **Bi-Encoder:** Encodes sentences independently into embeddings, which are then compared using cosine similarity.
- **Cross-Encoder:** Jointly encode both queries and documents using neural model (Nogueira et al. 2019), directly outputting a similarity score between 0 and 1.

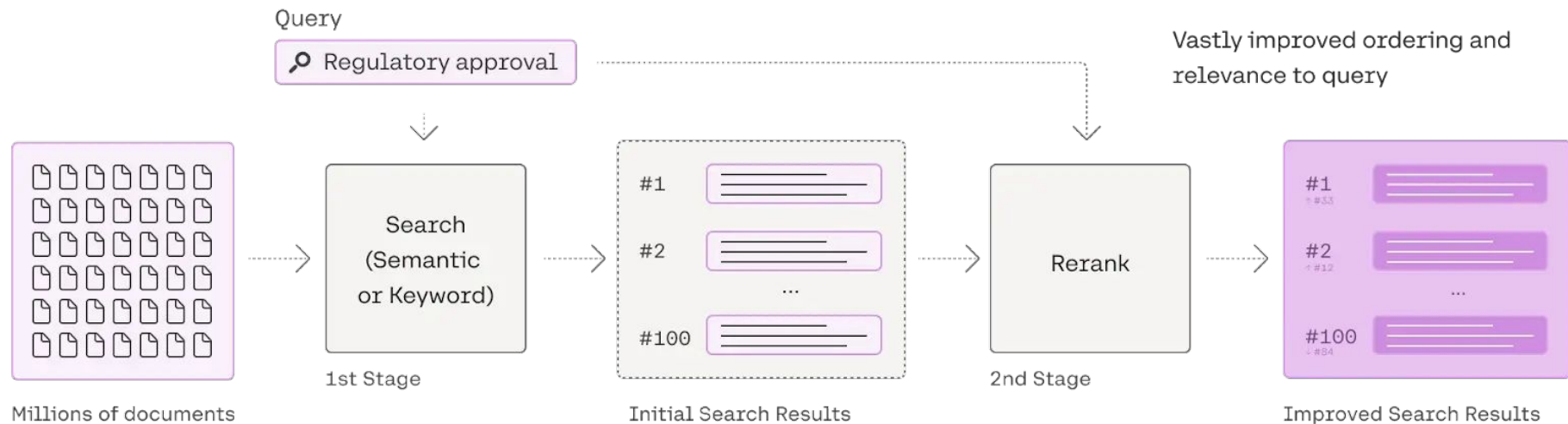


Bi-Encoder vs Cross-Encoder

Given 100 queries and 1,000 documents, how much computation is needed to identify similar sentences?



Reranking

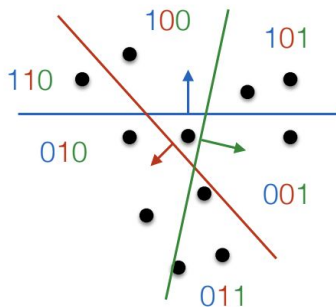


Vector Database

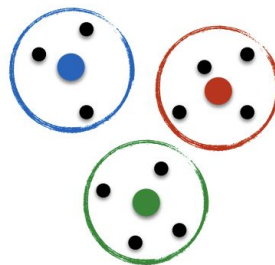
Approximate Nearest Neighbor Search

- Methods to retrieve embeddings in sub-linear time

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



Graph-based search: create
“hubs” and search from there



- Software: FAISS, ChromaDB

Evaluating Retrieval

Ranking Metrics

- What is good results:
 - Relevant information is included in the top of the ranking list
- Generally start with gold-standard relevance judgements and try to match them
- e.g.

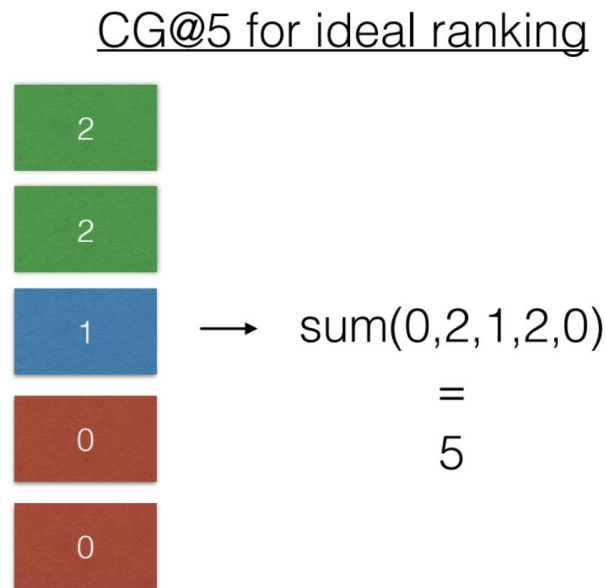
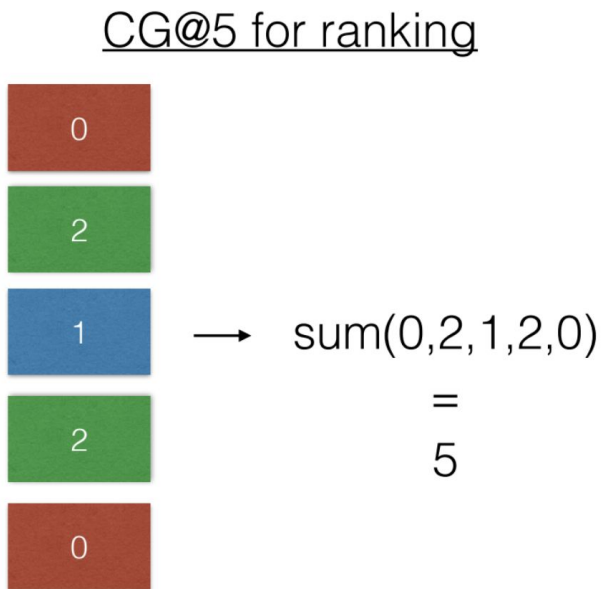
2: Highly Relevant

1: Somewhat Relevant

0: Not Relevant

Cumulative Gain (Hegde 22)

- Sum of relevance score @ N values retrieved



Discounted Cumulative Gain (Hegde 22)

- Add a discount $1/\log_2(i+1)$ for lower ranked values

DCG@5 for ranking

0	* 1	
2	* 0.63	
1	* 0.5	→ 2.62
2	* 0.43	
0	* 0.38	

iDCG@5 for ideal ranking

2	* 1	
2	* 0.63	
1	* 0.5	→ 3.76
0	* 0.43	
0	* 0.38	

Normalized Discounted Cumulative Gain

- Makes sure that as you pick up more good docs you get a better score
- $nDCG = DCG/iDCG$

If 3rd item was
not relevant ('0') →

DCG@3=1.26
iDCG@3=3.76
nDCG@3=0.34

0	* 1	2	* 1
2	* 0.63	2	* 0.63
1	* 0.5	1	* 0.5
2	* 0.43	0	* 0.43
0	* 0.38	0	* 0.38

<u>DCG@2</u>	<u>iDCG@2</u>
1.26	3.26

nDCG@2
0.386

<u>DCG@3</u>	<u>iDCG@3</u>
1.76	3.76

nDCG@3
0.468

Other Metrics

- Mean Average Precision: The average precision at which each relevant document is retrieved

no	0/1
yes	1/2
yes	2/3 → 0.638
yes	3/4
no	3/5

- Recall@N: The proportion of relevant documents retrieved within the top-N results, relative to the total number of relevant documents available.

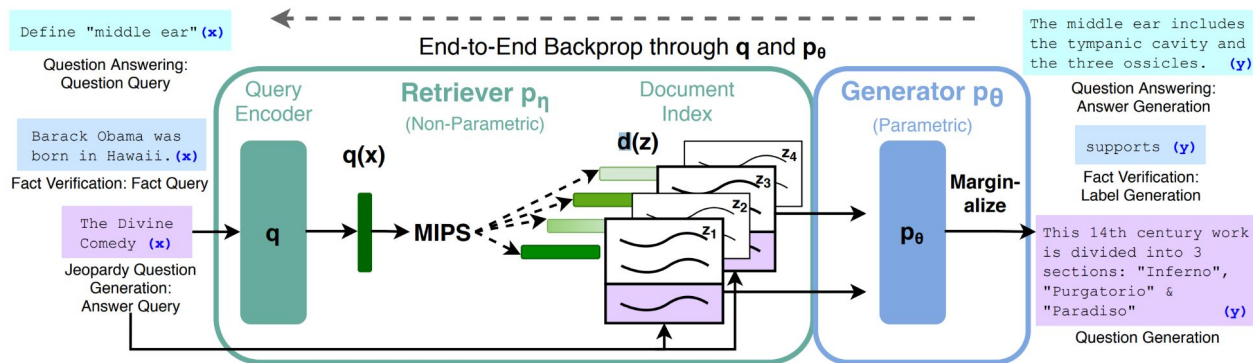
no	If there are 10 relevant documents for a query,
yes	
yes	$R@1 = 0$
yes	$R@2 = 1/10 = 0.1$
yes	$R@3 = 2/10 = 0.2$
no	$R@4 = 3/10 = 0.3$
	$R@5 = 3/19 = 0.3$

Retriever-Generator Models

Retriever + Generator End-to-end Training (“RAG”)

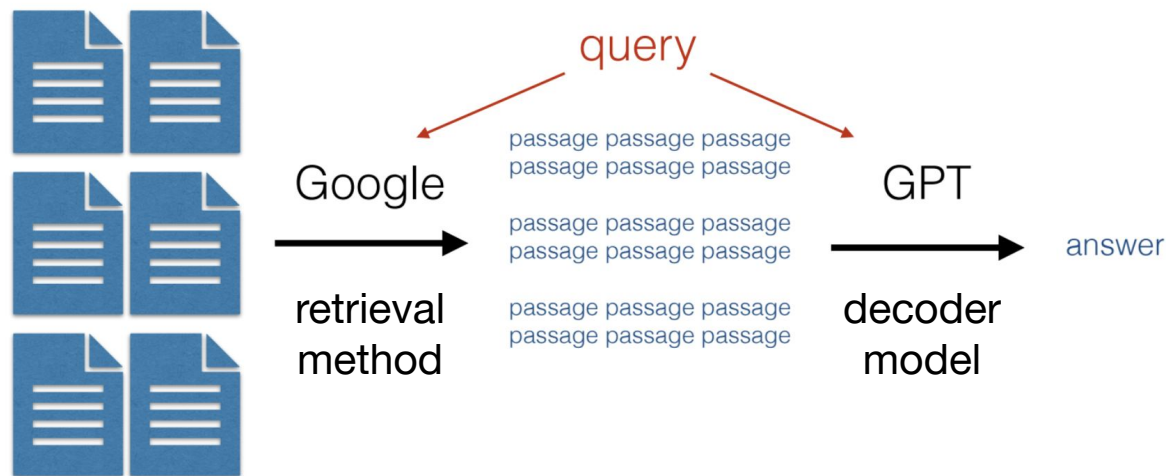
(Lewis et al. 2020)

- Train the retriever and generator to improve accuracy
- Generator: Maximize the generation likelihood given retrieved document(s)
- Retriever: Optimize mixture weights over documents to maximize overall likelihood (training applies only to the query encoder)



Simple: Just Chain Retrieval+Generator

- Use an out-of-the-box retriever and out-of-the-box generator



- Passages are concatenated to the context

Any Questions?