بسم الله الرحمن الرحيم

دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١

Information Retrieval – Summer 2021

9. BERT for Ranking

Tamer Elsayed

Qatar University

Today’s Roadmap

- monoBERT
- From Passages to Documents

- Multi-stage Rerankers
- Document Expansion

 for *Ranking*?

BERT for Ranking?

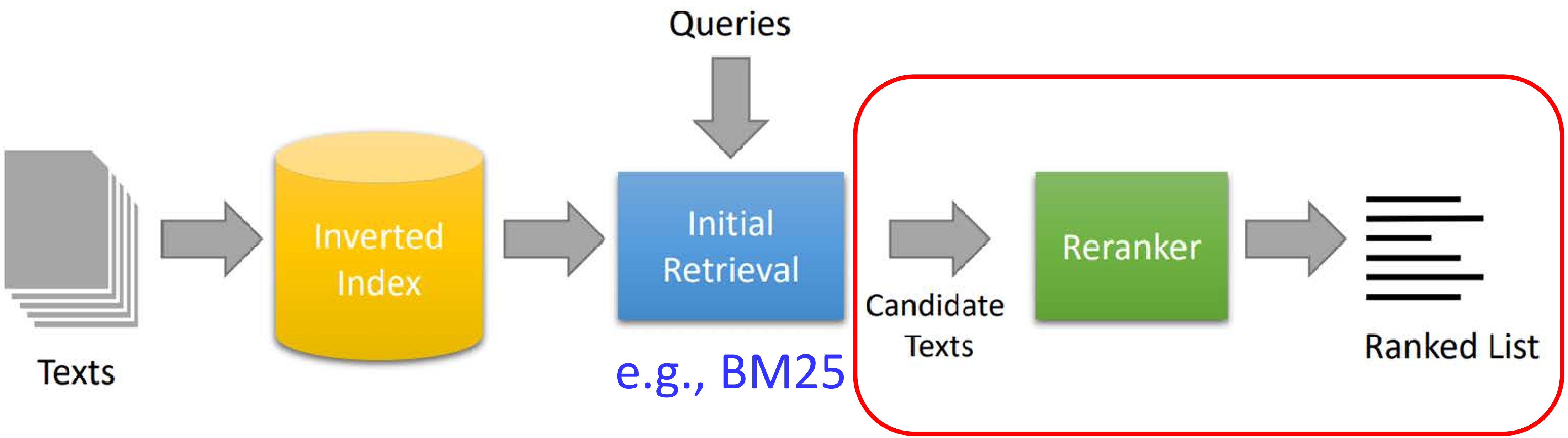
1. Ranking → classification problem
2. Sort the texts to be ranked based on the probability that each "item" belongs to the relevance class.

$$P(\text{Relevant} = 1 | d_i, q)$$

Relevance Classification

Learning to Rank

A Simple Search Engine



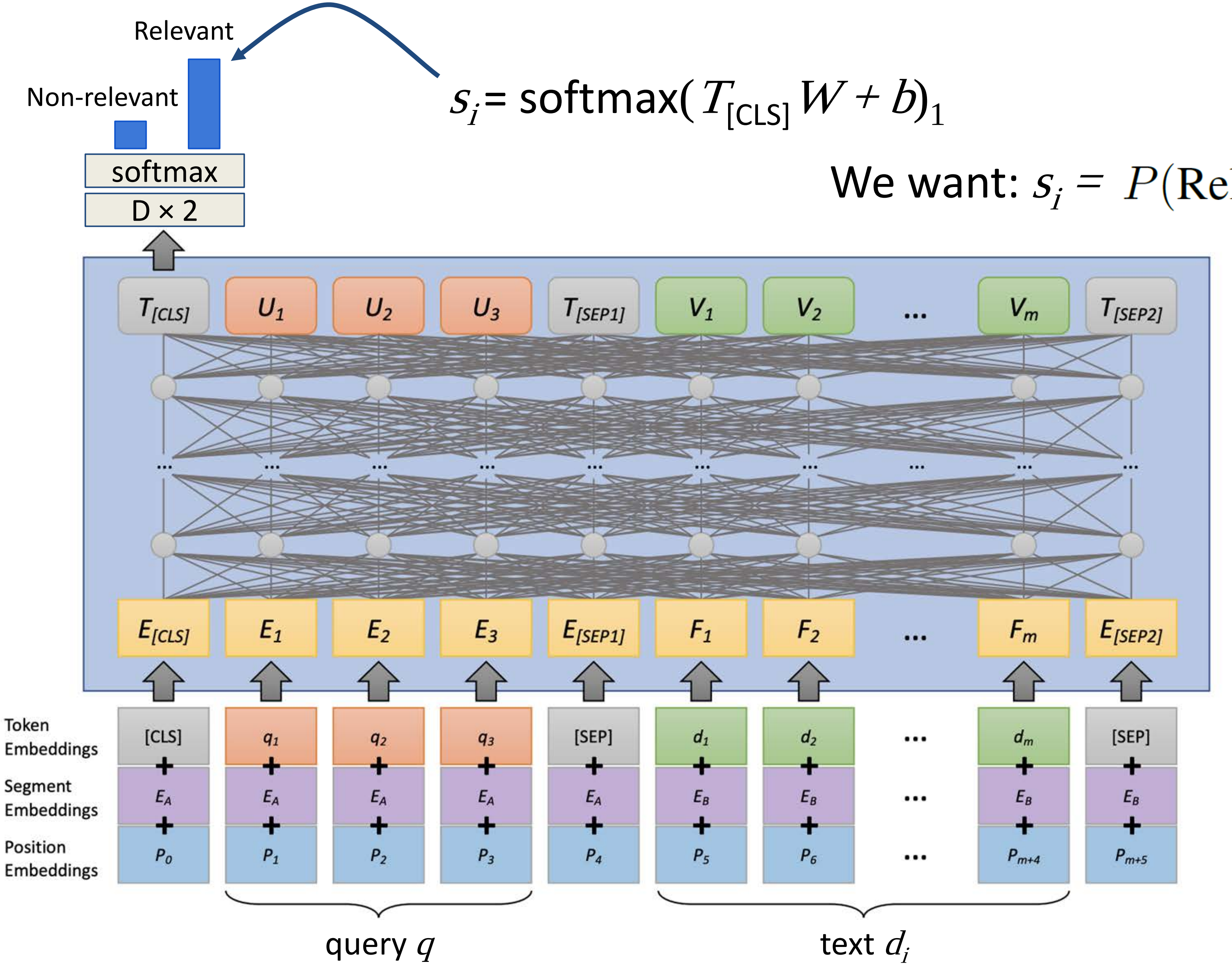

SEARCH
INFORMATION
RETRIEVAL
GOOGLE
CORPUS
QUERY
RANK
USEFUL
CRAWL
DATA
DOCUMENTS
DIFFICULT
DOCUMENT
DATABASE
WEB
INDEX
LARGE
IMPORTANT
COMPLEX
EFFICIENT
SCIENCE
PAGE
RETRIEVE
METADATA
SEARCHING
CRAWLER
DISTRIBUTED
GATHER
CONFUSING
MATHEMATICAL
AMAZING
COOL
HOW
FAST
FUN
SLOW
LIST
TEXT

MONOBERT

monoBERT: BERT reranker

$$s_i = \text{softmax}(T_{[\text{CLS}]} W + b)_1$$

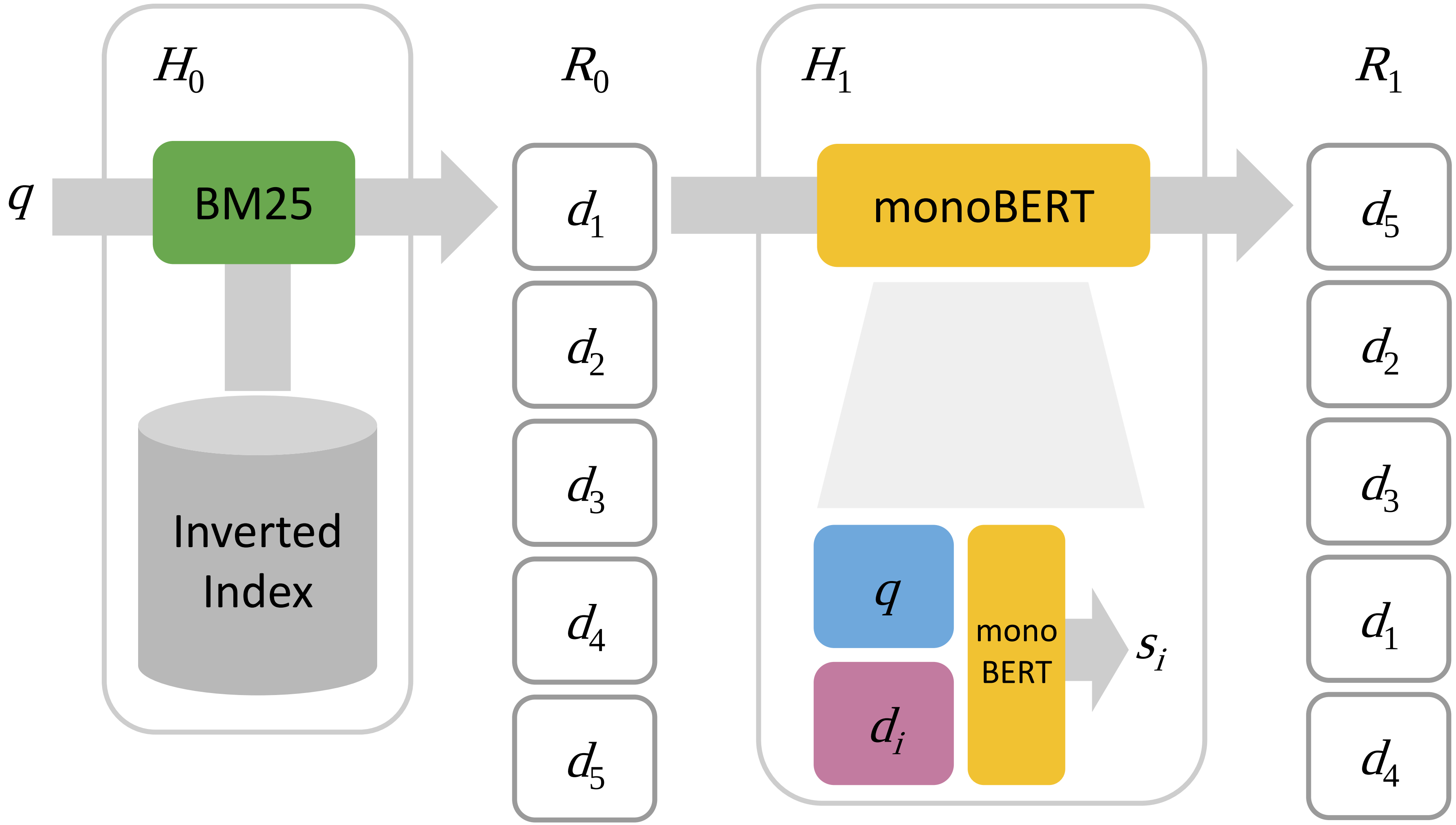
We want: $s_i = P(\text{Relevant} = 1 | d_i, q)$



Training monoBERT

Loss: $L = -\sum_{j \in J_{\text{pos}}} \log(s_j) - \sum_{j \in J_{\text{neg}}} \log(1 - s_j)$

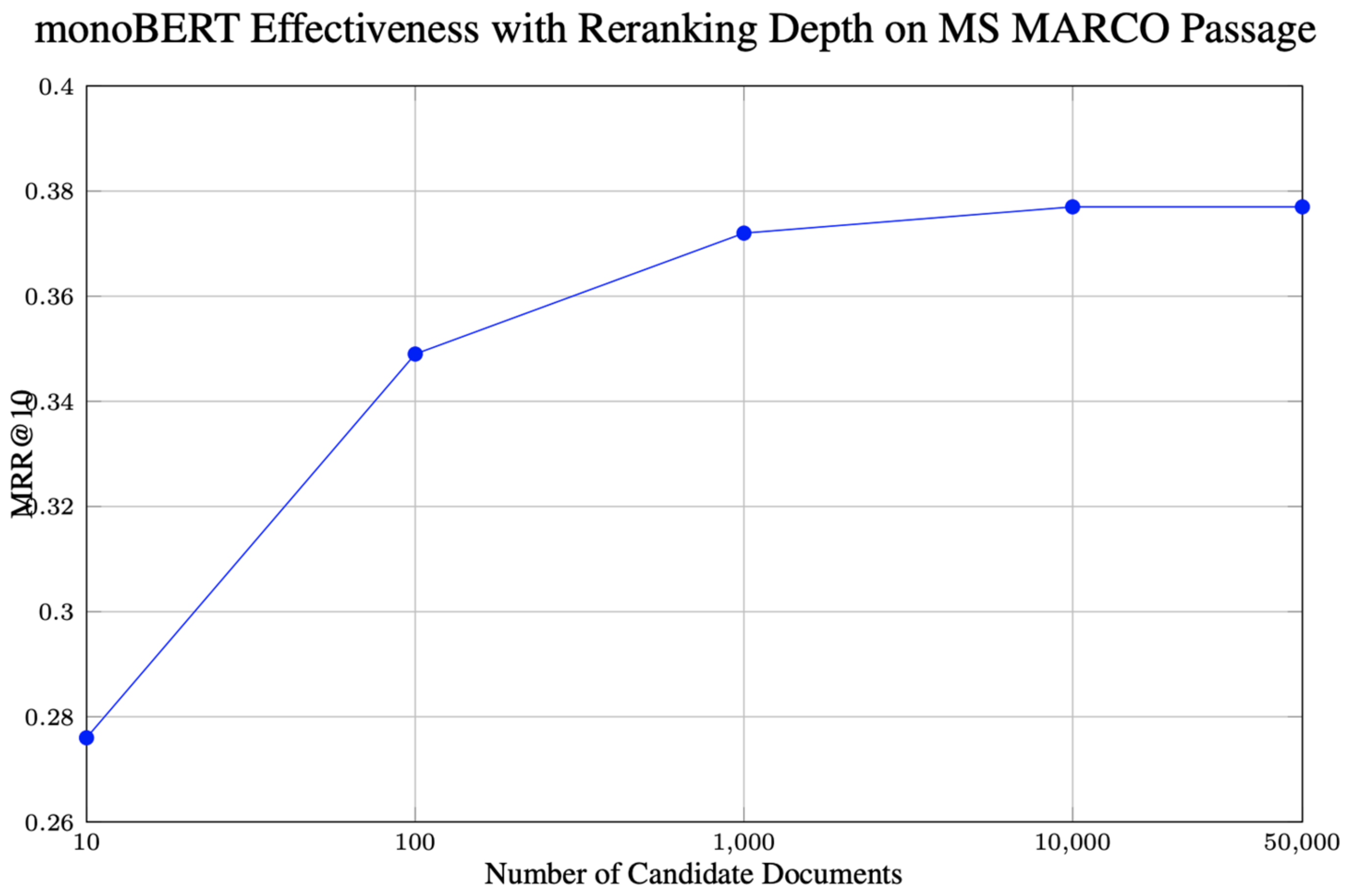
Once monoBERT is trained...



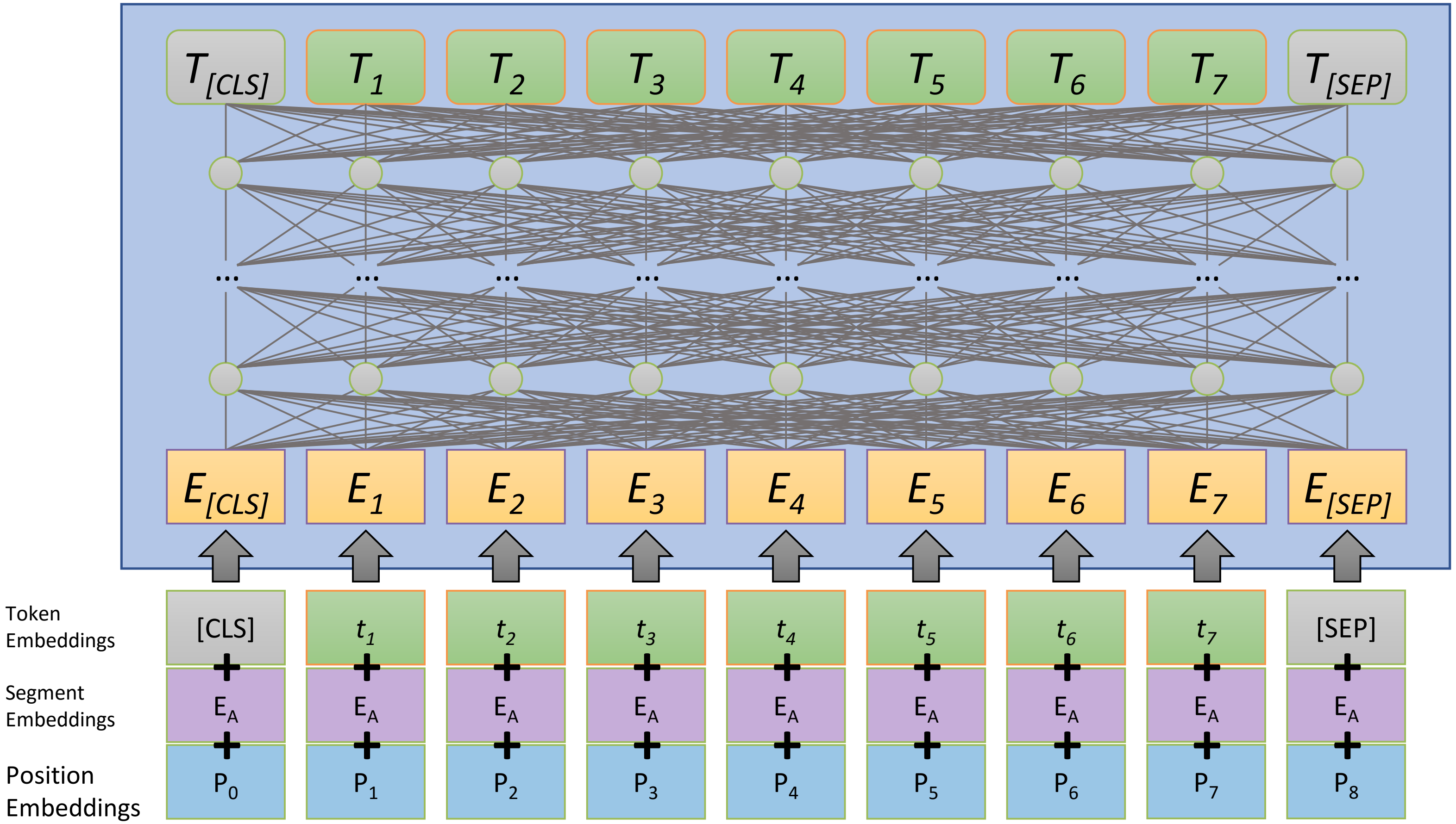
TREC 2019 - Deep Learning Track - Passage

	nDCG@10	MAP	Recall@1k
BM25	0.506	0.377	0.739
+ monoBERT	**0.738**	**0.506**	0.739
BM25 + RM3	0.518	0.427	0.788
+ monoBERT	**0.742**	**0.529**	0.788

How does retrieval depth affect performance?



BERT's Limitations



Cannot input entire documents

what do we input?

How do we label it?


27
POLL

27

With monoBERT, given a query, we score all documents in the collection.

- Yes
- No

monBERT takes 1 document and a query and determines if the document is relevant to the query or not.

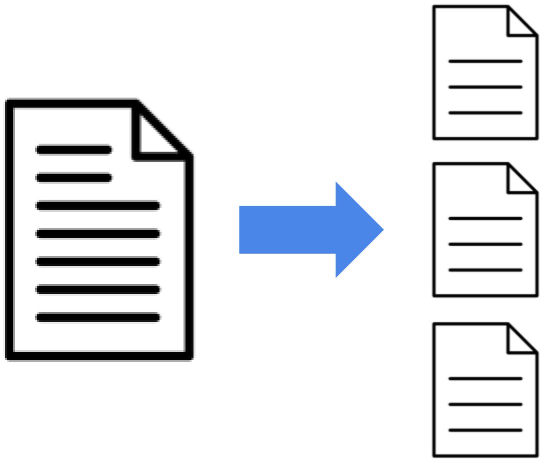
- Yes
- No

With monoBERT, increasing the depth of initial ranking (up to 10,000) increases the retrieval performance.

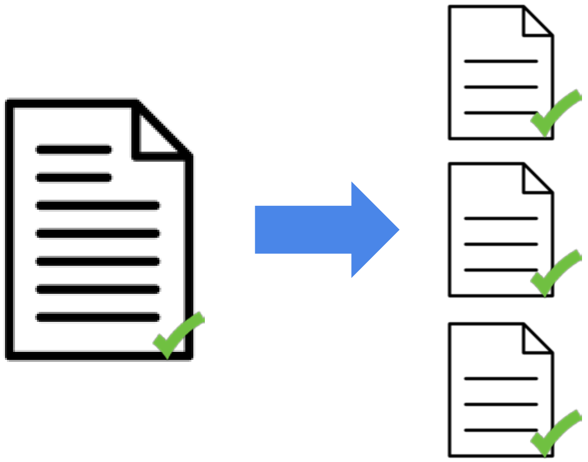
- Yes
- No

FROM PASSAGES TO DOCUMENTS

Handling Length Limitation: Training



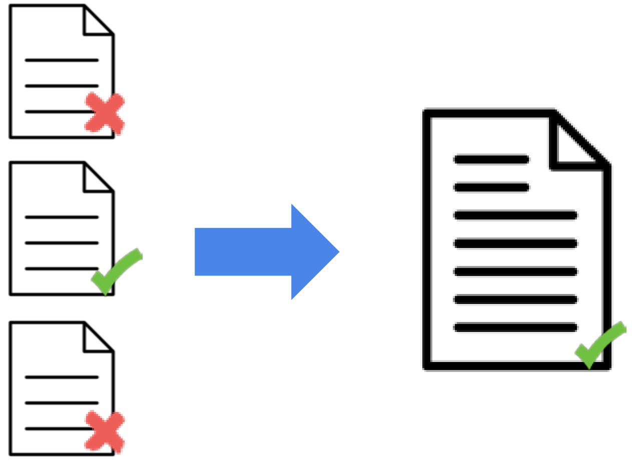
Chunk documents



Transfer labels (approximation)

Issues?

Handling Length Limitation: Inference



Aggregate Evidence

Issues?

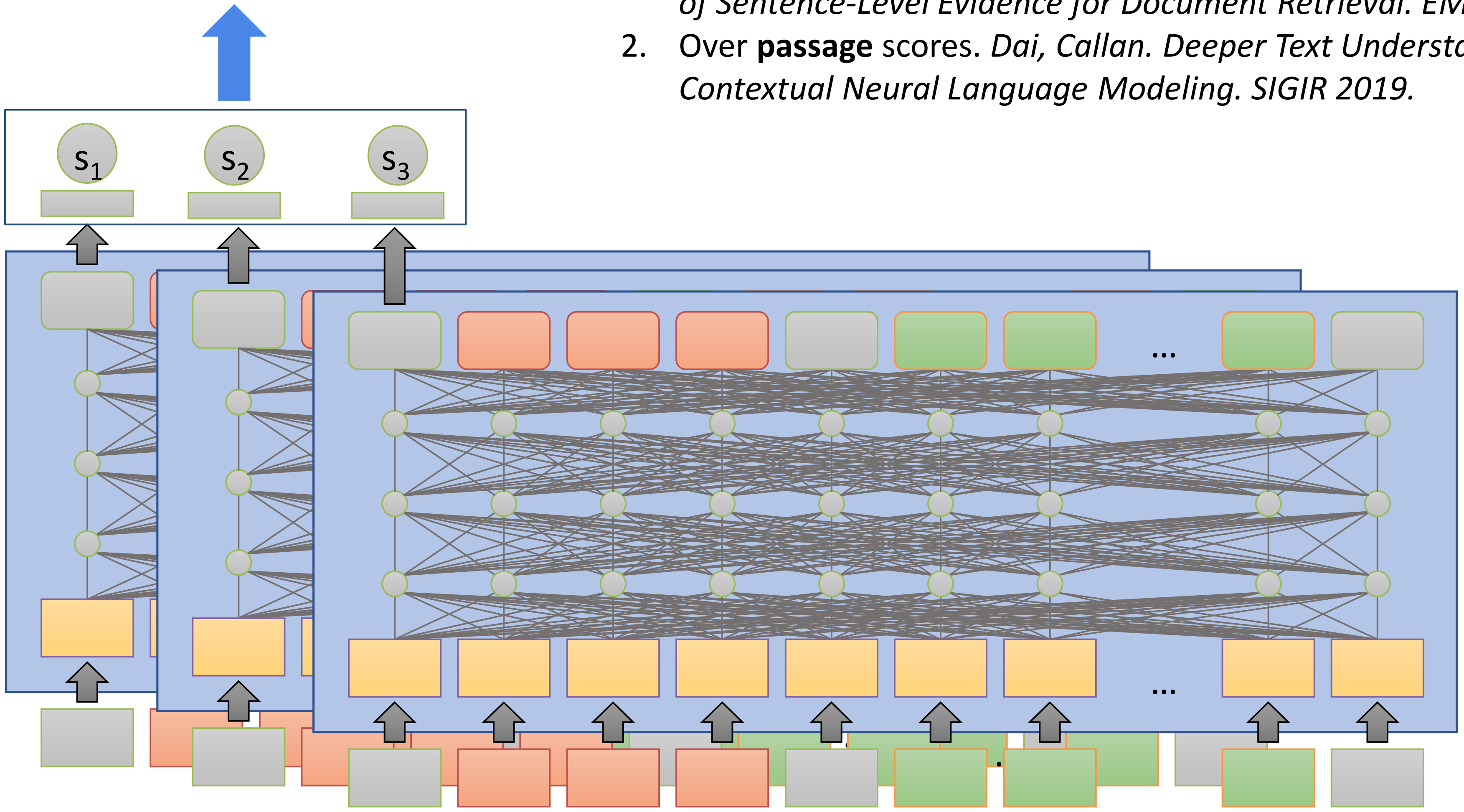
Handling Passages

Score Aggregation

Representation Aggregation

Score Aggregation

Document Score



1. Over **sentence** scores. *Yilmaz, Yang, Zhang, Lin. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. EMNLP '19.*
2. Over **passage** scores. *Dai, Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR 2019.*

Over Sentence Scores: Birch

- Trained on sentence-level judgments like tweets
- Inference on every sentence. Take top n.

$$s_f \triangleq \alpha \cdot s_d + (1 - \alpha) \cdot \sum_{i=1}^{n} w_i \cdot s_i$$

First-stage retrieval score

Sentence scores

- Interpolation weights are tuned on target dataset

Yilmaz, Yang, Zhang, Lin. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. EMNLP '19.

Over Sentence Scores: Results

		Robust04	
Method		MAP	nDCG@20
(1)	BM25 + RM3	0.2903	0.4407
(2a)	1S: BERT(MB)	$0.3408^{\dagger}$	$0.4900^{\dagger}$
(2b)	2S: BERT(MB)	$0.3435^{\dagger}$	$0.4964^{\dagger}$
(2c)	3S: BERT(MB)	$0.3434^{\dagger}$	$0.4998^{\dagger}$
(3a)	1S: BERT(MS MARCO)	$0.3028^{\dagger}$	0.4512
(3b)	2S: BERT(MS MARCO)	$0.3028^{\dagger}$	0.4512
(3c)	3S: BERT(MS MARCO)	$0.3028^{\dagger}$	0.4512
(4a)	1S: BERT(MS MARCO $\rightarrow$ MB)	$0.3676^{\dagger}$	$0.5239^{\dagger}$
(4b)	2S: BERT(MS MARCO $\rightarrow$ MB)	$\mathbf{0.3697}^{\dagger}$	$0.5324^{\dagger}$
(4c)	3S: BERT(MS MARCO $\rightarrow$ MB)	$0.3691^{\dagger}$	$\mathbf{0.5325}^{\dagger}$

Zero-shot Cross-domain Learning

Length mismatch!

Top sentence is good enough!

Yilmaz, Yang, Zhang, Lin. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. EMNLP '19.

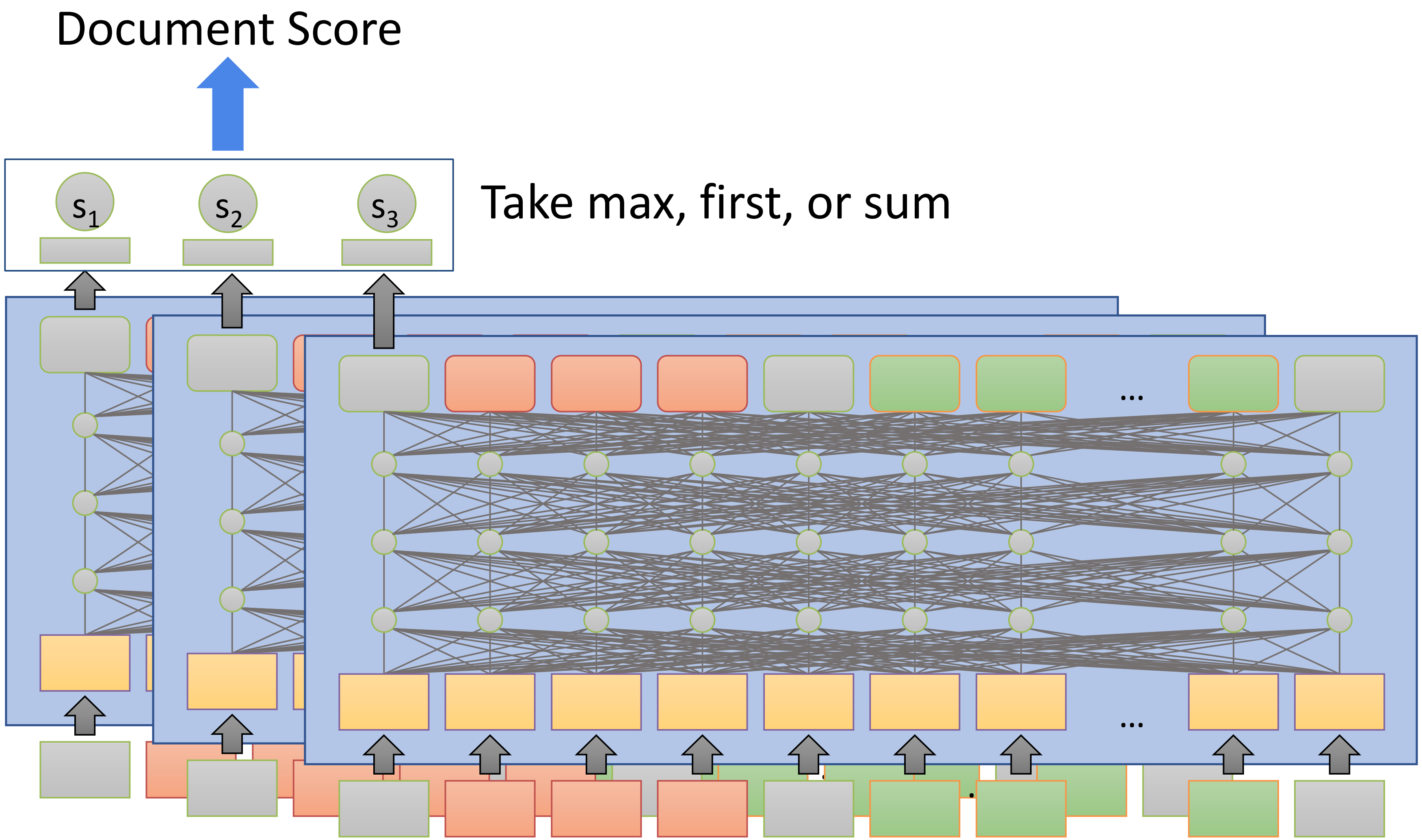
Over Passage Scores: BERT-MaxP, FirstP, SumP

- Train on overlapping passages
  - Fixed window size
  - With stride

- Aggregate passage scores: max, first, sum.

Dai, Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR 2019.

Over Passage Scores: BERT-MaxP, FirstP, SumP



Dai, Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR 2019.

Over Passage Scores: Results

		Robust04	
		nDCG@20	
Model		Title	Description
(1)	BOW	0.417	0.409
(2)	SDM	0.427	0.427
(3)	LTR	0.427	0.441
(4a)	BERT–FirstP	$0.444^{\dagger}$	$0.491^{\dagger}$
(4b)	BERT–MaxP	$\mathbf{0.469}^{\dagger}$	$\mathbf{0.529}^{\dagger}$
(4c)	BERT–SumP	$0.467^{\dagger}$	$0.524^{\dagger}$

Dai, Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR 2019.

Would longer queries be better?

Title: air traffic controller

Description: What are working conditions and pay for U.S. air traffic controllers?

Narrative: Relevant documents tell something about working conditions or pay for American controllers. Documents about foreign controllers or individuals are not relevant.

Over Passage Scores: Results

Method		Robust04 Avg. Length	nDCG@20 SDM	MaxP
(1)	Title	3	0.427	0.469
(2a)	Description	14	0.404	0.529
(2b)	Description, keywords	7	0.427	0.503
(3a)	Narrative	40	0.278	0.487
(3b)	Narrative, keywords	18	0.332	0.471
(3c)	Narrative, negative logic removed	31	0.272	0.489

Stop words!


28
POLL

28

We can finetune BERT with multiple training datasets.

- Yes
- No

With BERT, longer queries tend to give worse performance.

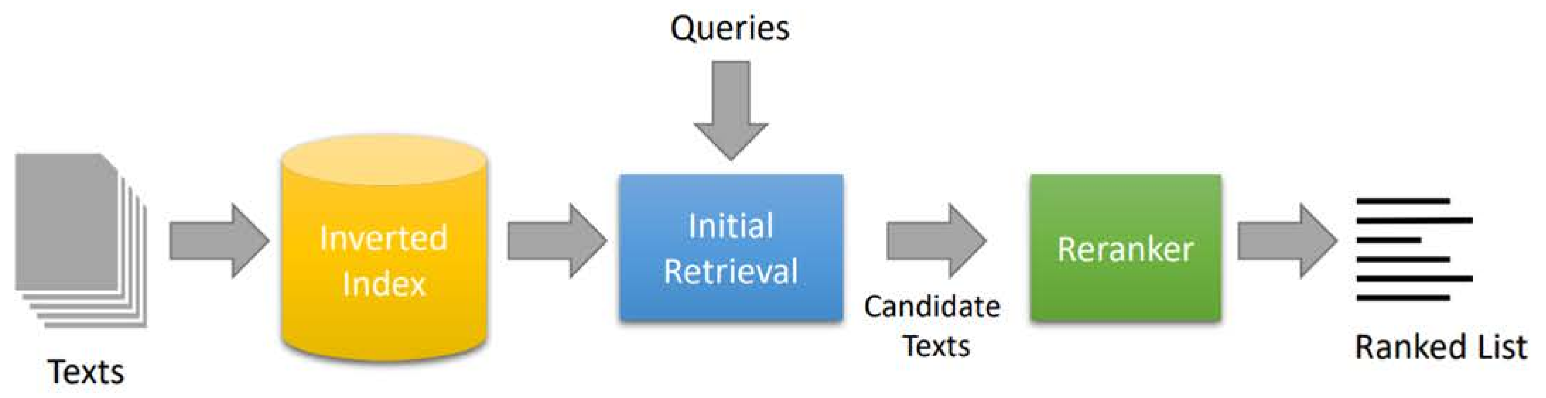
- Yes
- No

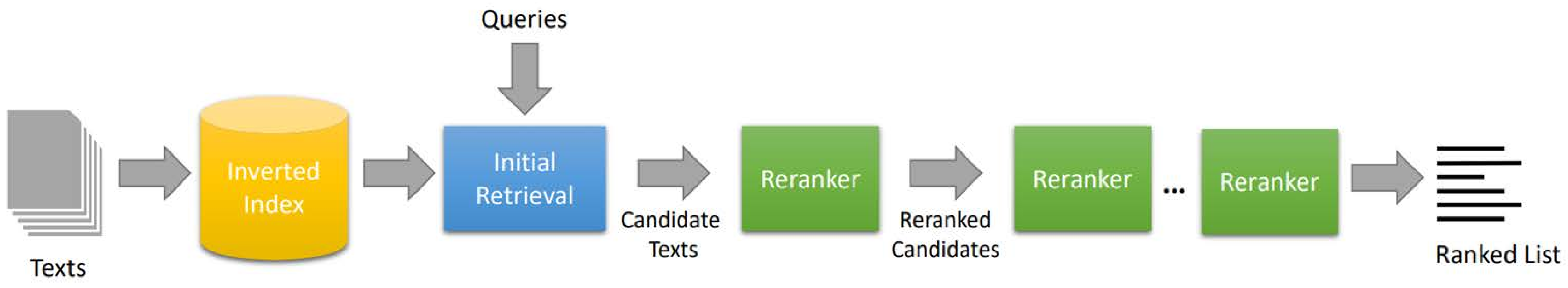
Stop words are of little use for BERT.

- Yes
- No

MULTI-STAGE RERANKERS

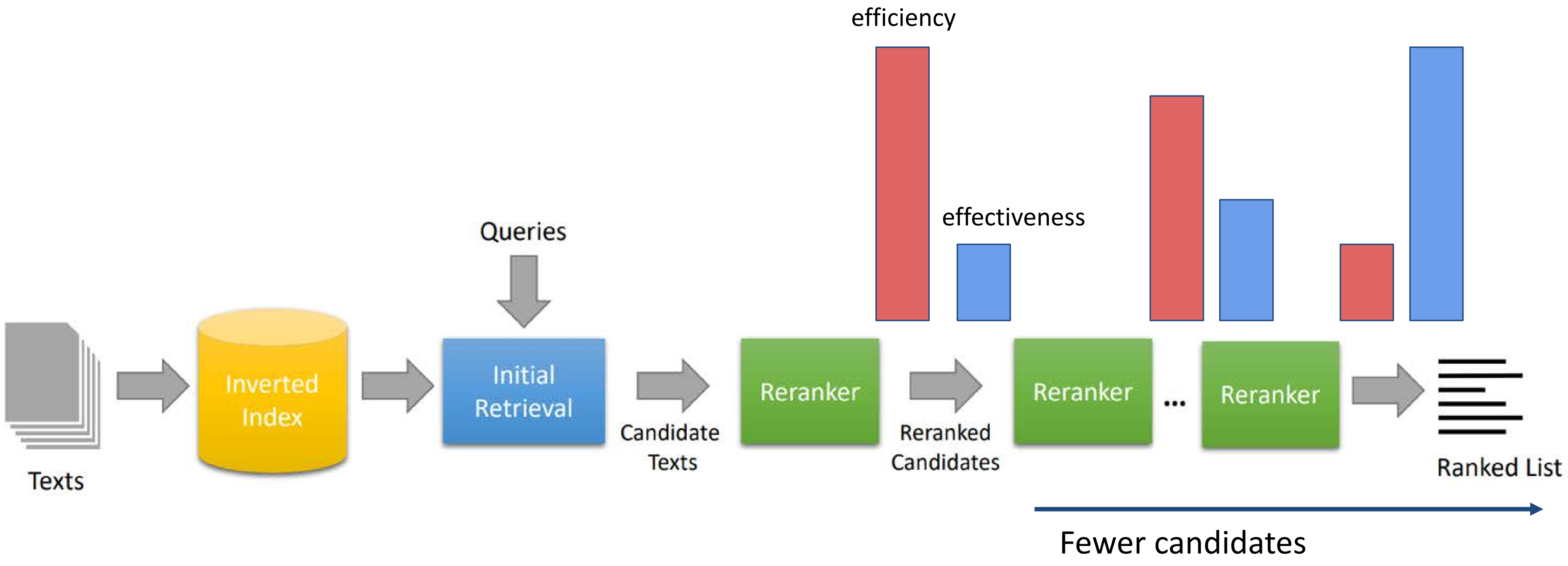
From Single to Multiple Rerankers





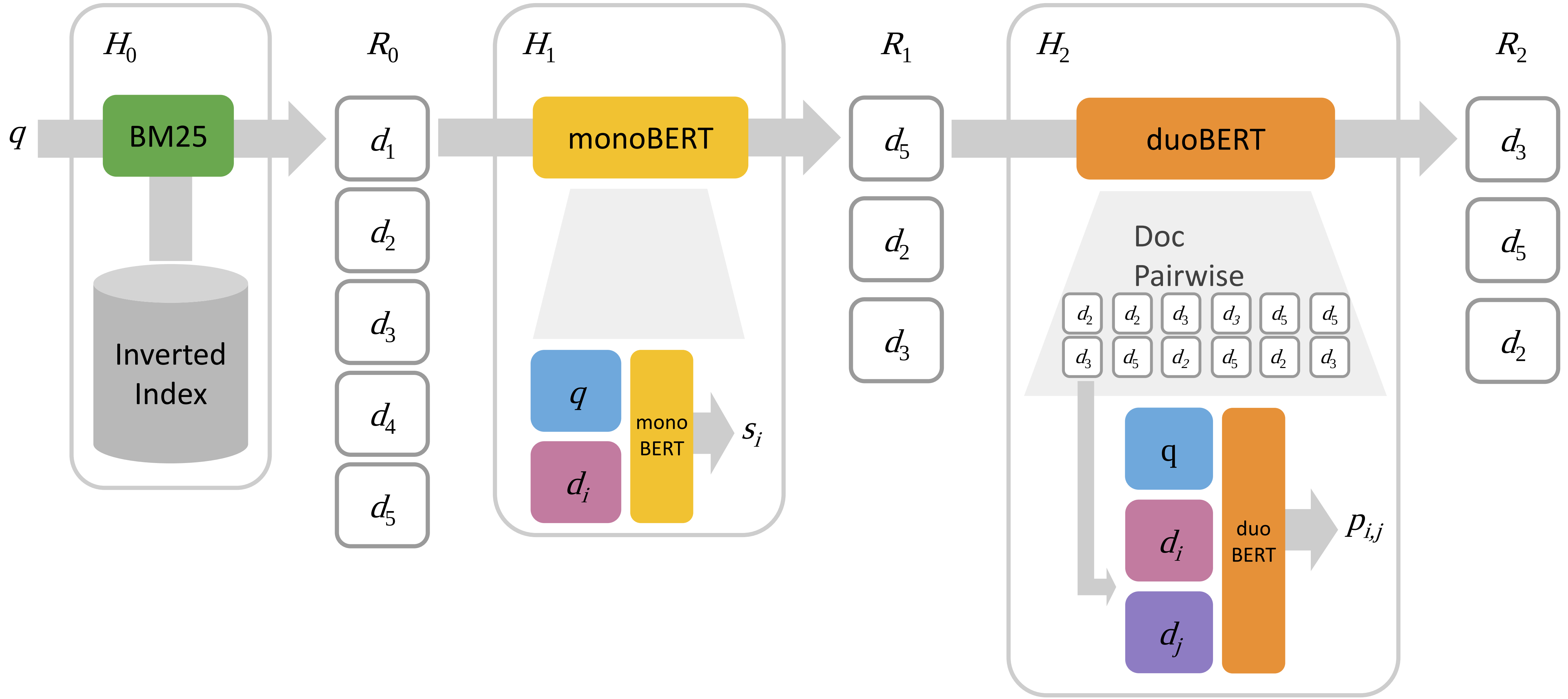
Why Multi-stage?

- Trade-off between effectiveness (quality of the ranked lists) and efficiency (retrieval latency)



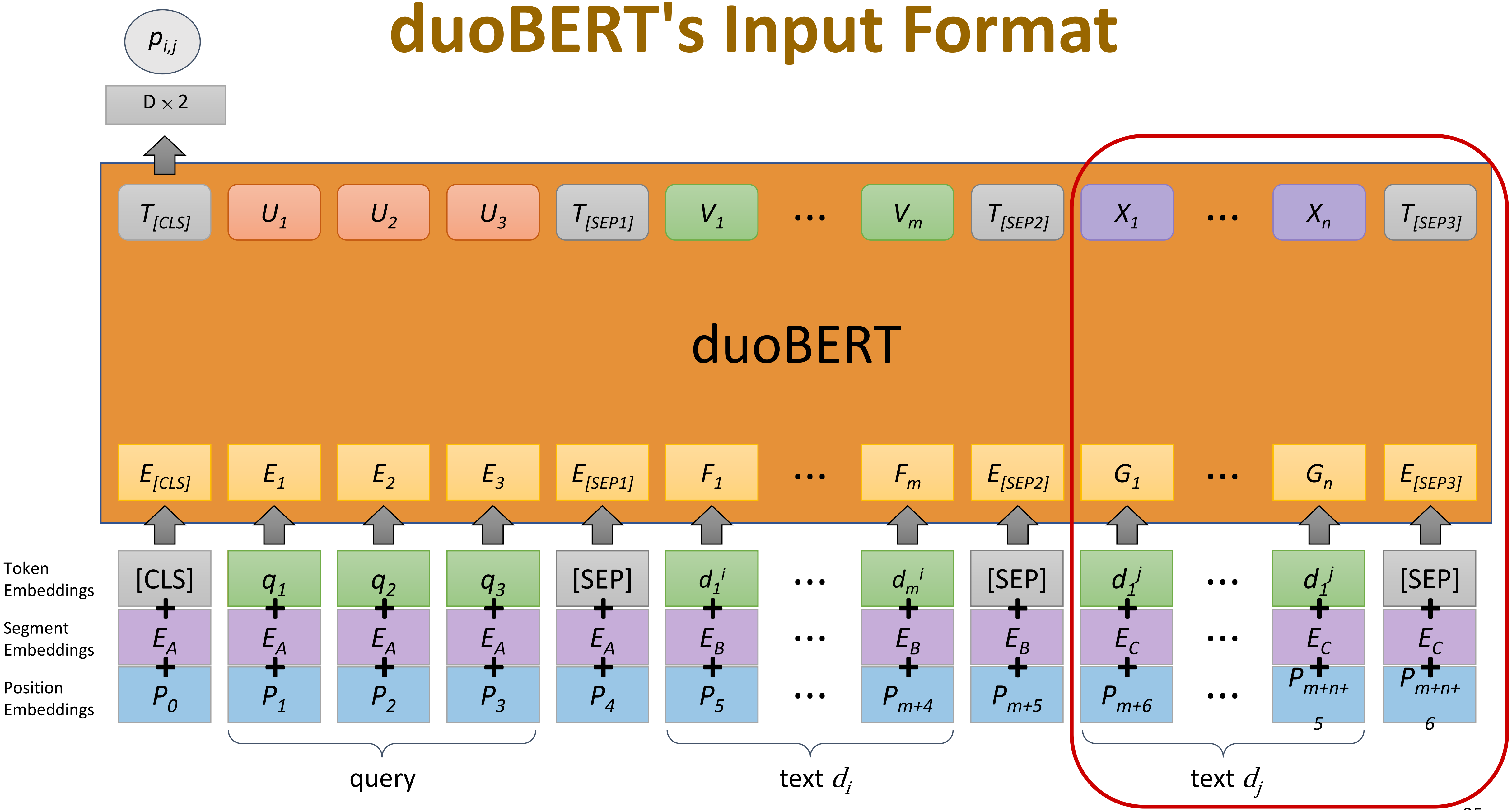
DUOBERT

Multi-stage with duoBERT



Nogueira, Yang, Cho, Lin. Multi-stage document ranking with bert. 2019.

duoBERT's Input Format



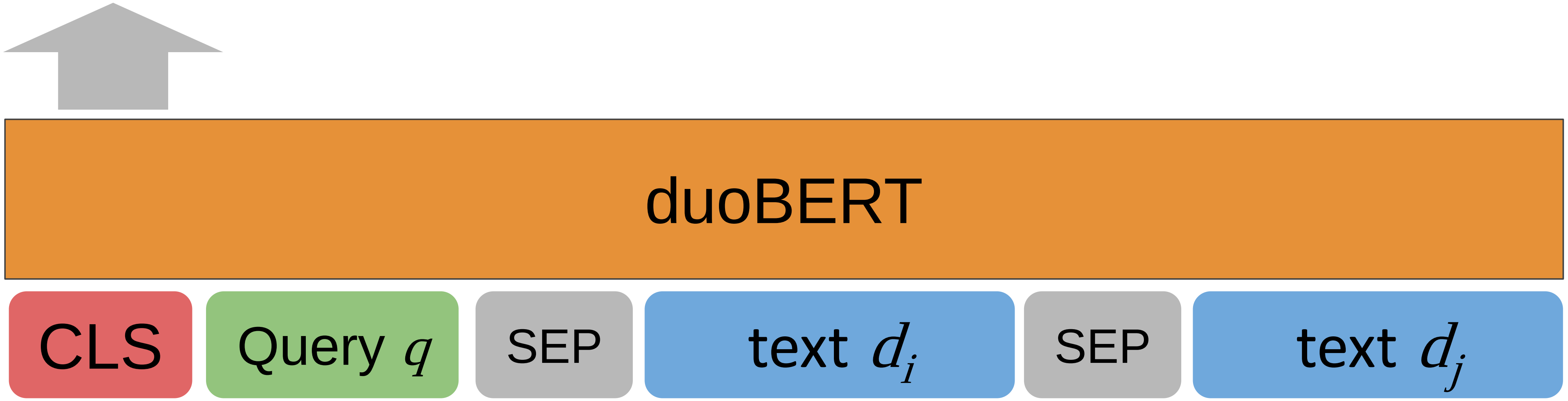
Training duoBERT

Loss:

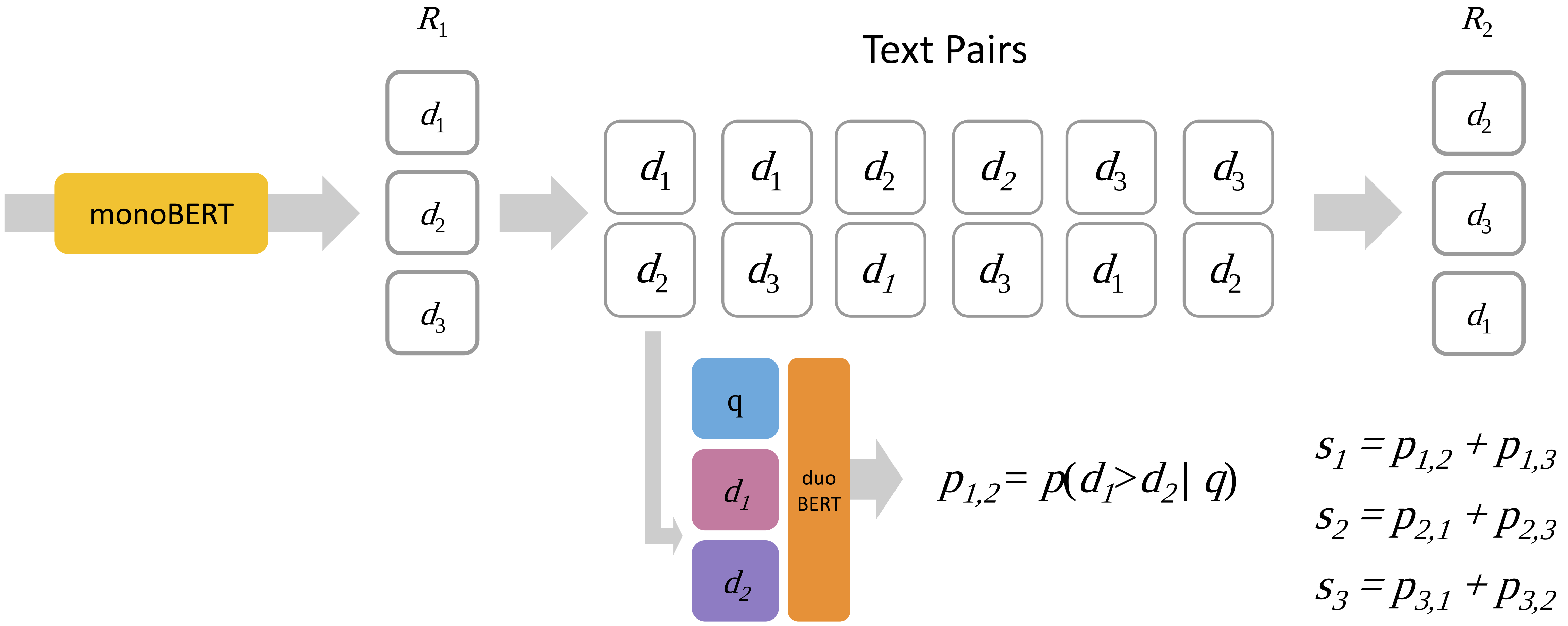
$$L_{\text{duo}} = - \sum_{i \in J_{\text{pos}}, j \in J_{\text{neg}}} \log(p_{i,j}) - \sum_{i \in J_{\text{neg}}, j \in J_{\text{pos}}} \log(1 - p_{i,j})$$

Is doc d_i more relevant than doc d_j to the query q?

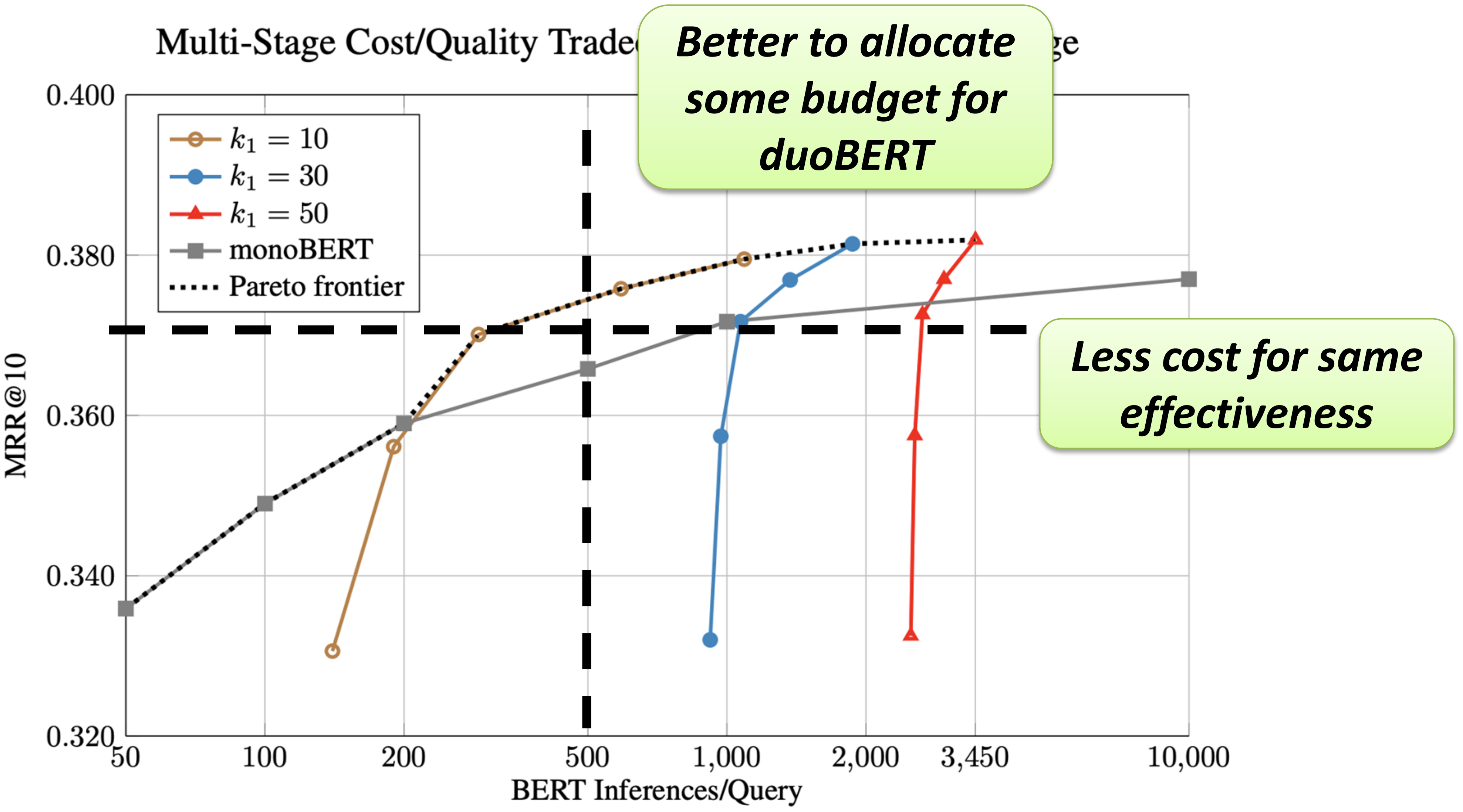
$p_{i,j} = p(d_i > d_j \mid q)$



Inference with duoBERT



monoBERT vs. duoBERT



29

29

duoBERT takes 2 documents and a query and determines if both documents are relevant to the query or not.

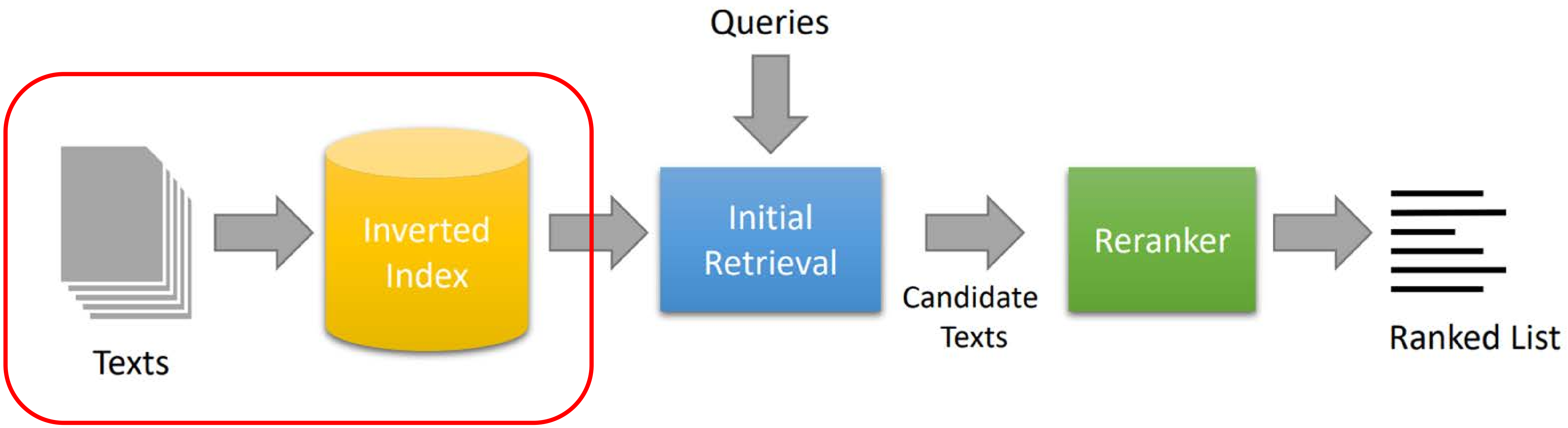
- Yes
- No

For scoring a list of 15 documents using duoBERT, we use/call BERT

- 14 times
- 15 times
- 15 x 14 times
- 15 + 14 times

DOCUMENT EXPANSION

A Simple Search Engine



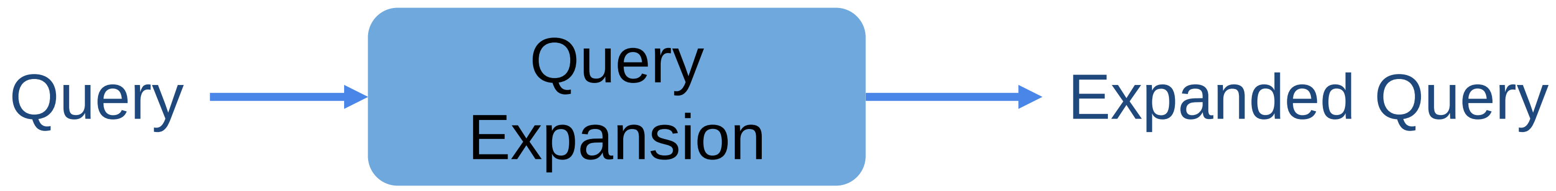
Vocabulary Mismatch

Is it still a problem?

- Initial candidate generation stage still depends on exact matching (e.g., BM25).
- Relevant text that has no overlap with query terms will not be retrieved → will never be reranked!

Expansion?

Query Expansion vs. Document Expansion



Input has little information

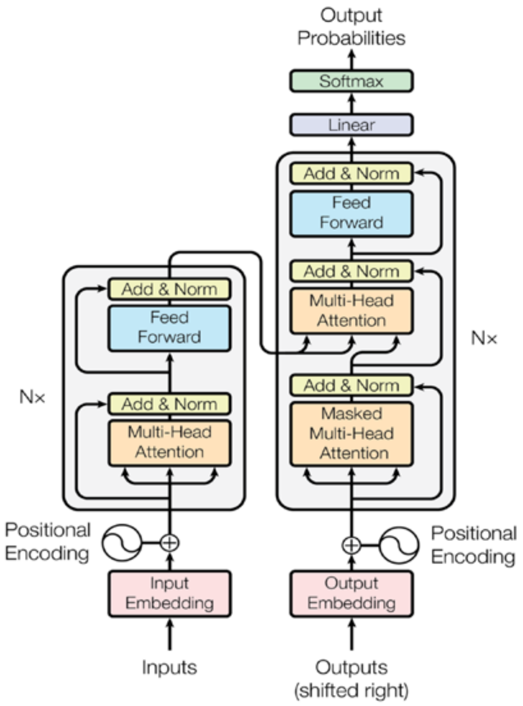
Input has a lot of information

doc2query

Document → seq2seq Transformer → Query

Supervised training:

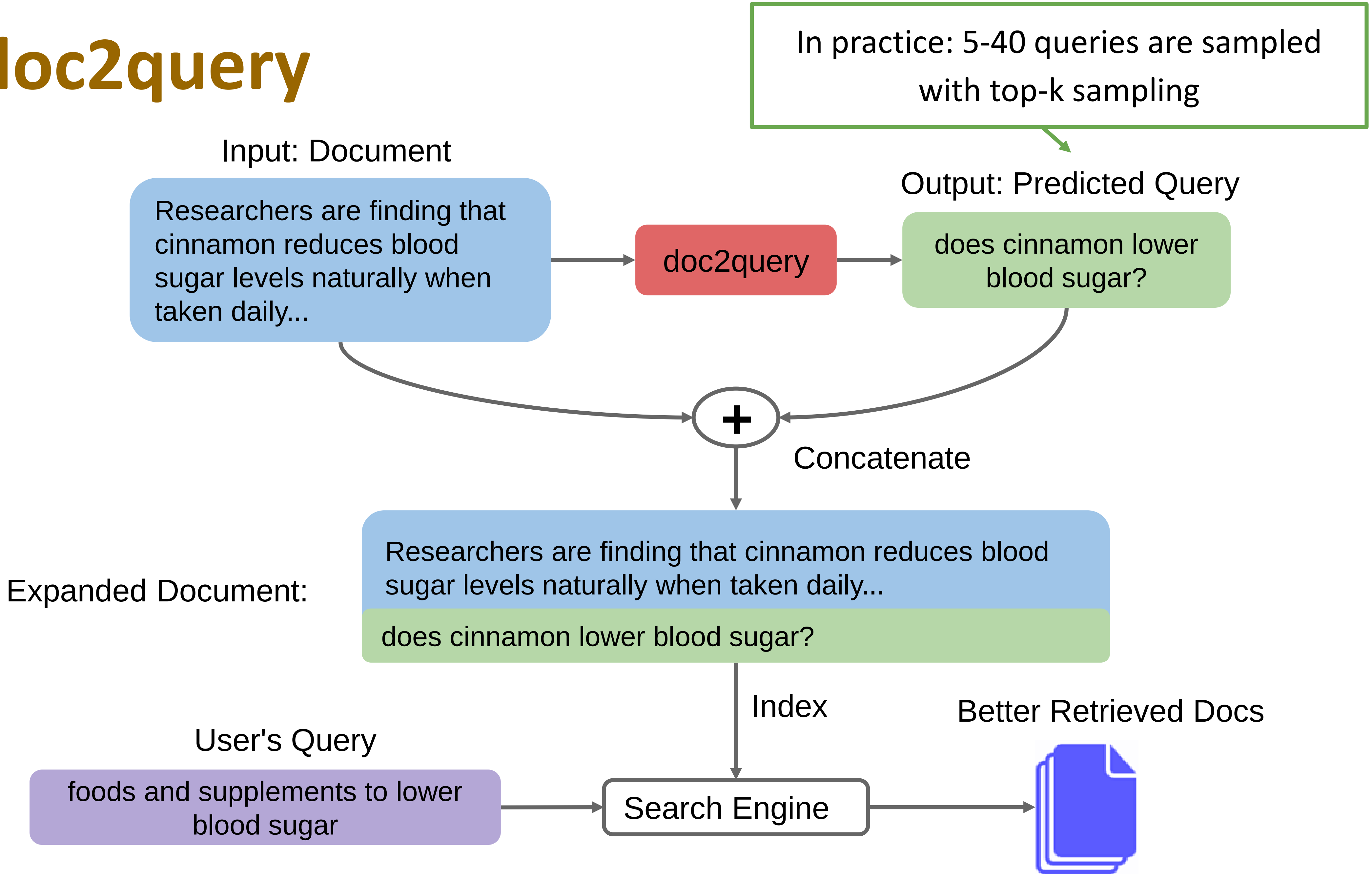
pairs of <relevant document, query>



Source: Vaswani et al., 2017

Nogueira, Yang, Lin, Cho. Document expansion by query prediction. 2019.

doc2query


In practice: 5-40 queries are sampled with top-k sampling
Input: Document
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
doc2query
Output: Predicted Query
does cinnamon lower blood sugar?
Concatenate
Expanded Document:
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
does cinnamon lower blood sugar?
Index
User's Query
foods and supplements to lower blood sugar
Search Engine
Better Retrieved Docs

Results

	MARCO Passage (MRR@10)	TREC-DL 19 (nDCG@10)
BM25	0.184	0.506
+ doc2query	**0.277**	**0.642**

Examples

Input Document:	July is the hottest month in Washington DC with an average temperature of 27C (80F) and the coldest is January at 4C (38F) with the most daily sunshine hours at 9 in July. The wettest month is May with an average of 100mm of rain.
Predicted Query:	weather in washington dc
Target query:	what is the temperature in washington
Input Document:	The Delaware River flows through Philadelphia into the Delaware Bay. It flows through and aqueduct in the Roundout Reservoir and then flows through Philadelphia and New Jersey before emptying into the Delaware Bay.
Predicted Query:	what river flows through delaware
Target Query:	where does the delaware river start and end
Input Document:	sex chromosome - (genetics) a chromosome that determines the sex of an individual; mammals normally have two sex chromosomes chromosome - a threadlike strand of DNA in the cell nucleus that carries the genes in a linear order; humans have 22 chromosome pairs plus two sex chromosomes.
Predicted Query:	what is the relationship between genes and chromosomes
Target Query:	which chromosome controls sex characteristics

Excluding stop-words:

69% copied

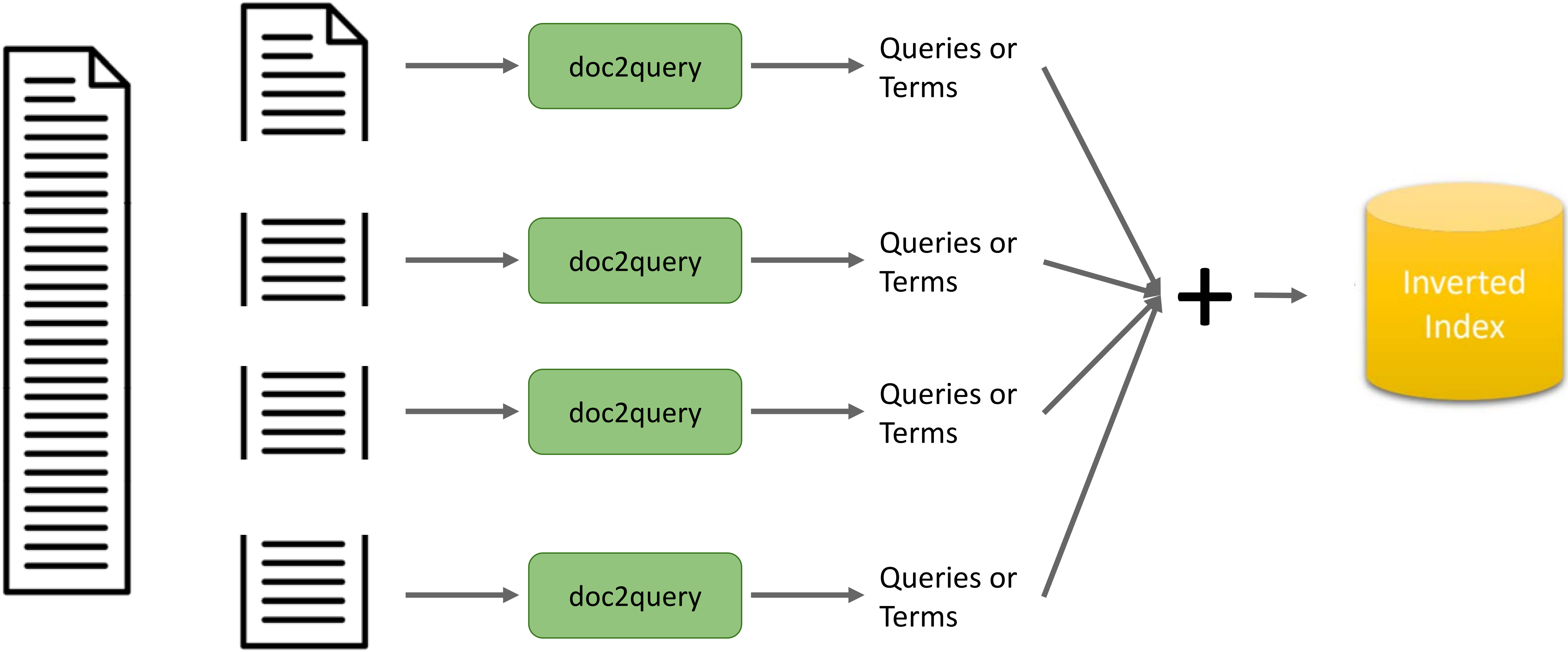
31% new

What is more important? copied or new words?

	MRR@10	R@1000
Original Document	.184	.853
+ Expansion New Words	.195	.907
+ Expansion Copied Words	.221	.893
+ Expansion Copied + New	.277	.944

Predicted queries are "better" than documents

How to Expand Long Documents?



Takeaways of Document Expansion

Advantages:

- Documents have more context than queries →easy prediction task
- Documents can be processed offline *and* in parallel

Disadvantages:

- Have to iterate over the entire collection
- Longer Documents → increase in query latency

30

POLL

30

Document expansion is easier to find relevant words than query expansion

- Yes
- No

Doc2query can be used for ... (you can check multiple)

- reweighting of original terms of the doc
- adding new terms to the doc
- removing original terms

- *My group's ongoing research*
- *Open research in IR*
- *Resources*
- *Project ideas for MSc & PhD students*
- *Where can you go from here?*