Pretrained Transformers for Text Ranking:
BERT and Beyond

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin

@andrewyates @rodrigfnogueira @lintool

max planck institut informatik

UNIVERSITY OF WATERLOO

Based on the survey:

Pretrained Transformers for Text Ranking:
BERT and Beyond

by Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
https://arxiv.org/abs/2010.06467

Tutorial organization:
- Recorded tutorial
- Live sessions: hands-on component and Q&A

Outline

- Part 1: Background
  (text ranking, IR, ML)
- Part 2: Ranking with relevance classification
- Part 3: Ranking with dense representations
- Part 4: Conclusion & future directions

Text Ranking

Text ranking problems
Transformers

Definition

Given: a piece of text
(keywords, question, news article, …)

Rank: other pieces of text
(passages, documents, queries, …)

Ordered by: their similarity

e.g., **Web search**

Yandex Google Bing Baidu百度

Focus: Ad hoc Retrieval

Given: query *q*

collection of texts

Return: a ranked list of *k* texts $d_1 \dots d_k$

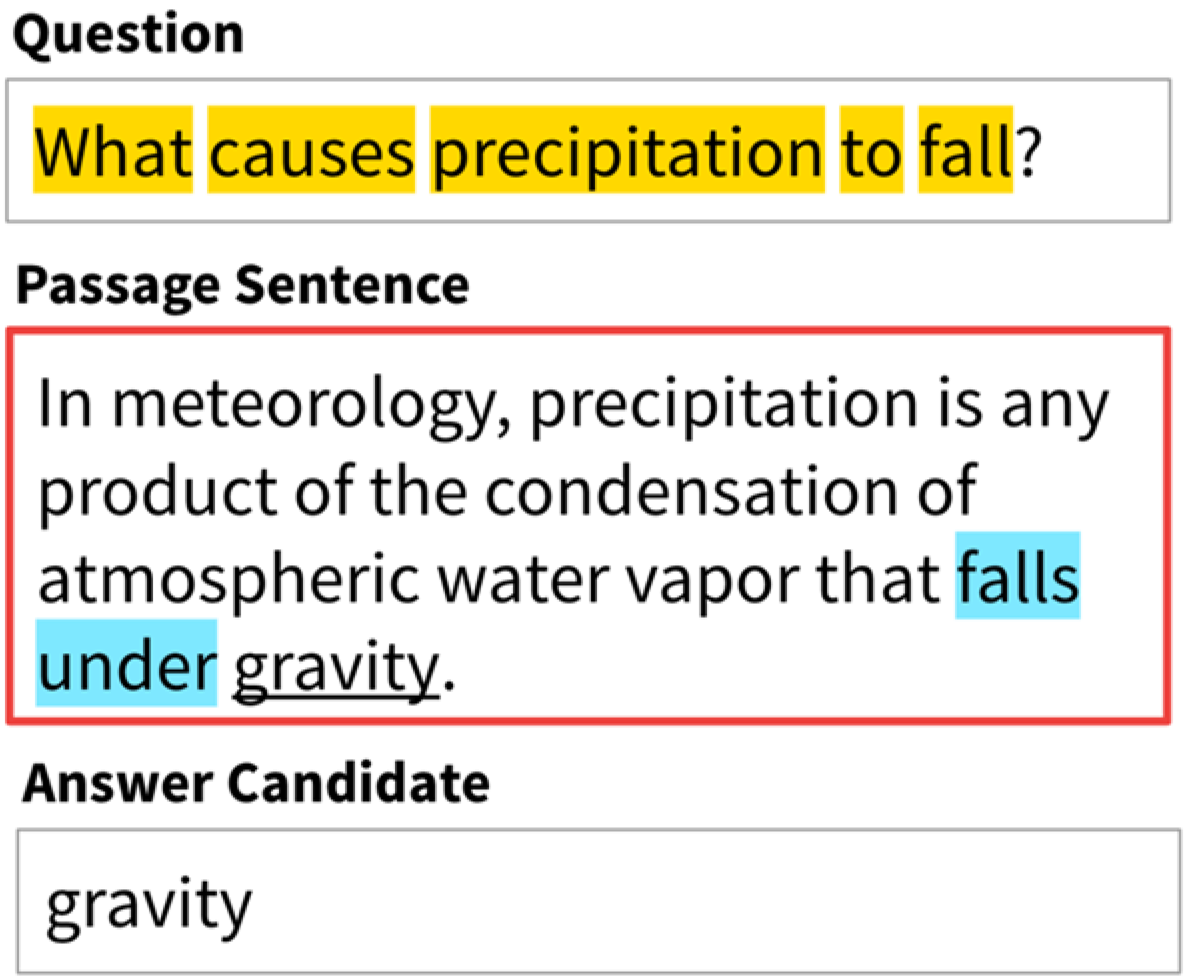
Maximizing: a metric of interest

metric: 0.66

Other Problems: Question Answering

Approach:

- Rank passages
- Rank answer spans



Source: [SQuAD]

Other Problems: Community Question Answering

New question: **What is the longest airline flight?**

Related Questions

What is the longest airline flight physically possible? Not flying around...

Once aircraft range is maxed out, what will eventually be the longest non-stop...

What is the longest commercial flight?

What is the world's shortest daily airline flight?

What is the longest airline flight you have been on? Where were you going?

What's the longest flight from New York?

Quora

Search Quora

Commercial Flights | World Records | Flights | Airlines | Aviation | Air Travel

What is the longest commercial flight?

Answer · Follow · 5 · Request

6 Answers

Anshul Choudhary, senior director in apple store at Apple (2016-present)

Answered January 12, 2019

Originally Answered: What's the longest commercial flight between two locations, in hours?

The world's longest commercial flight has left Singapore for New York, beginning a journey scheduled to cover more than 15,000km in almost 19 hours.

Singapore Airlines is relaunching the service five years after it was cut because it had become too expensive.

Source: Quora

Other Problems: Text Recommendation

NEWS ASTRONOMY

Andromeda’s and the Milky Way’s black holes will collide. Here’s how it may play out

Supermassive black holes will merge less than 17 million years after galaxy merger

By Sid Perkins

MARCH 5, 2021 AT 8:00 AM

The supermassive black holes at the centers of the Milky Way and Andromeda galaxies are doomed to engulf each other in an ill-fated cosmological dance.

Related Stories

ASTRONOMY

Supermassive black hole gets kicked to the galactic curb

By Ashley Yeager • March 28, 2017

SPACE

A newfound black hole in the Milky Way is weirdly heavy

By Christopher Crockett • November 27, 2019

PHYSICS

Big black holes can settle in the outskirts of small galaxies

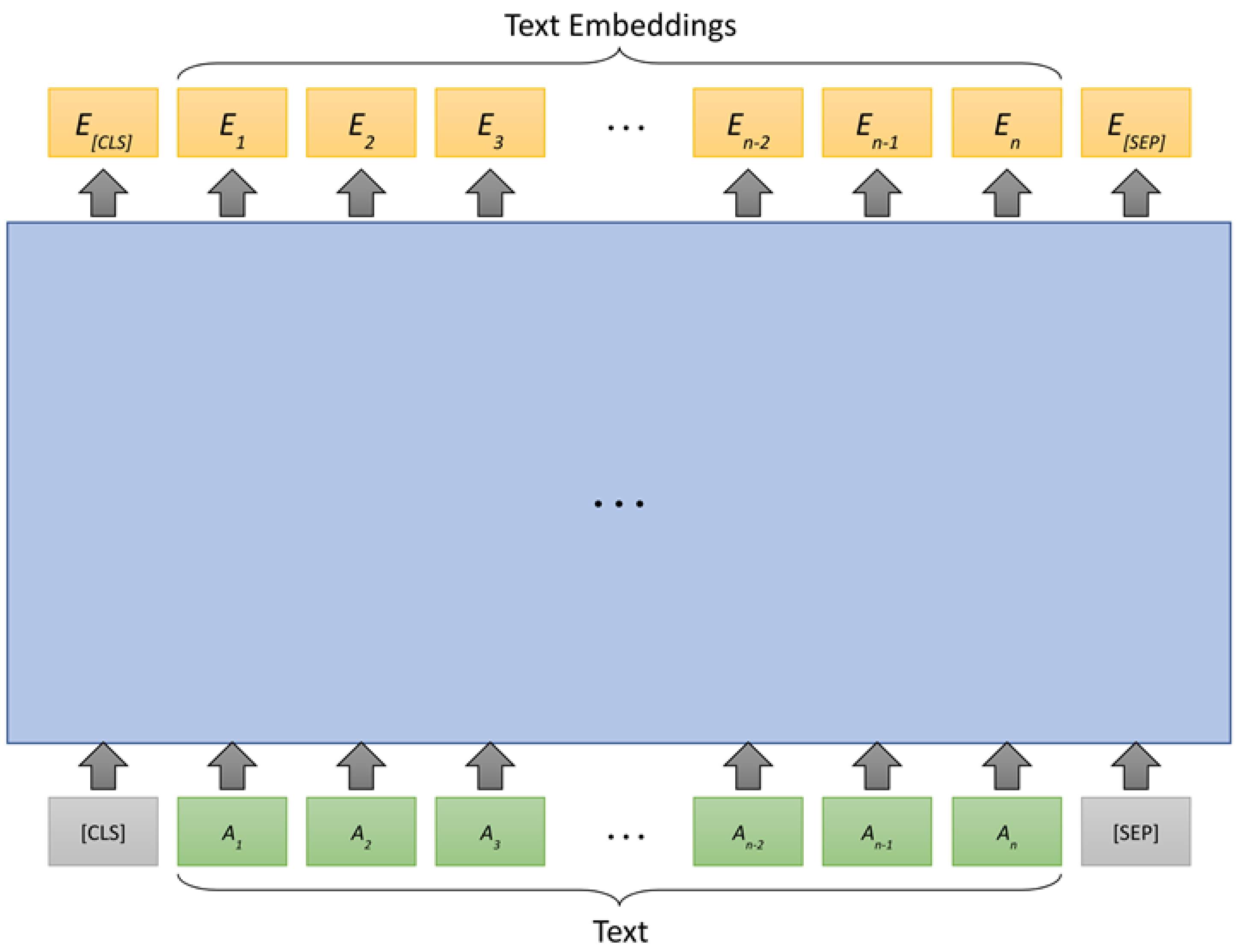
By Lisa Grossman • May 23, 2019

Source: [Science News]

Focus: Content-based Similarity

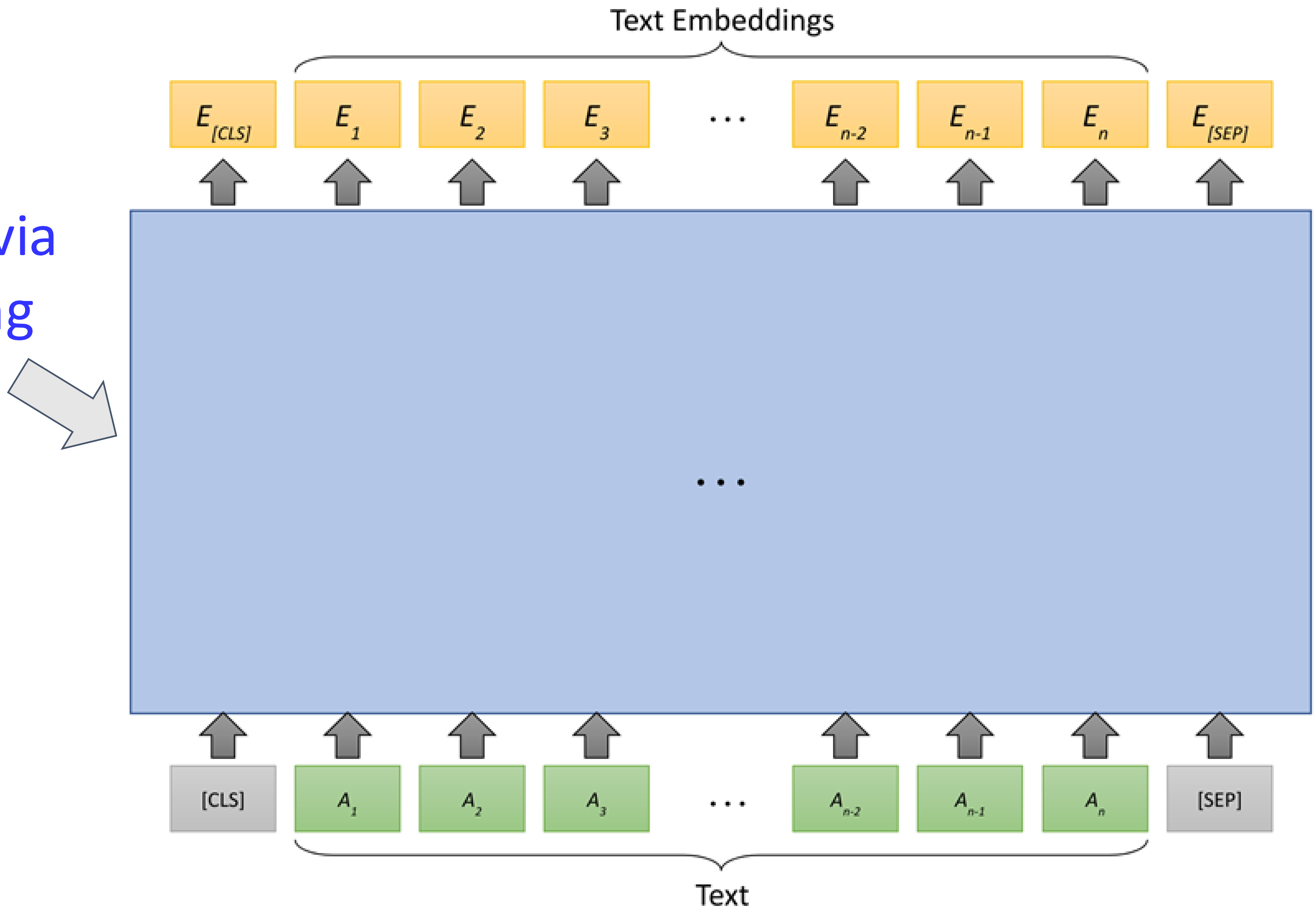
Agreement between query and a piece of text

Transformers



Pretrained Transformers

Initialized via pretraining



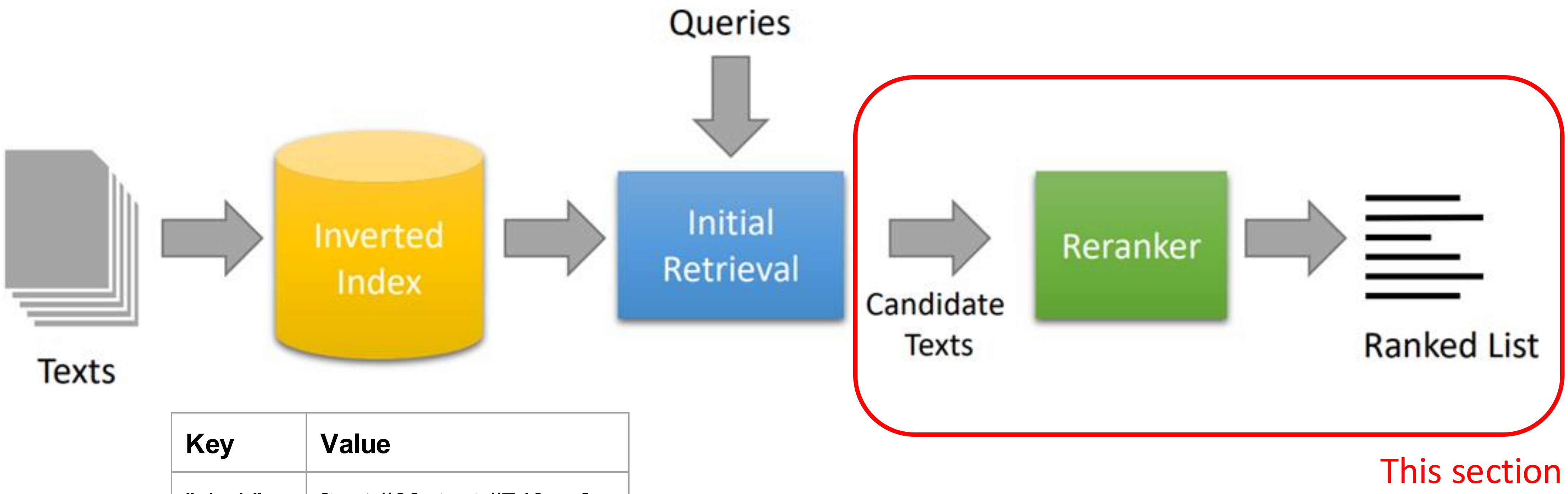
Machine Learning Background

Learning to rank
Deep learning for ranking
BERT

Machine Learning Background

Learning to rank
Deep learning for ranking
BERT

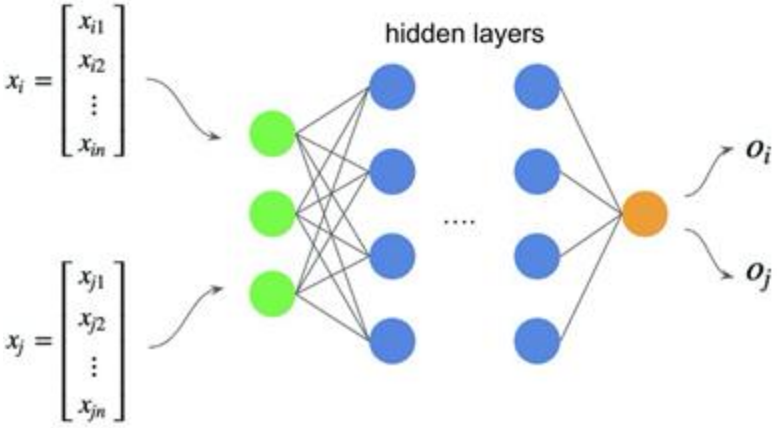
A Simple Search Engine



Key	Value
"chair"	[text #83, text #743, ...]
"store"	[text #1003, text #50, ...]
...	...

Learning to Rank (> 1990)

- Supervised machine learning techniques
- Typically based on hand-crafted features:
  - Content (e.g. term frequencies, document lengths)
  - Meta-data (e.g.: PageRank scores)
- RankNet (Burges et al., 2005): a neural net
  - Different from DL models because **they require hand-crafted features**



- Gained popularity with user click data (Burges., 2010)

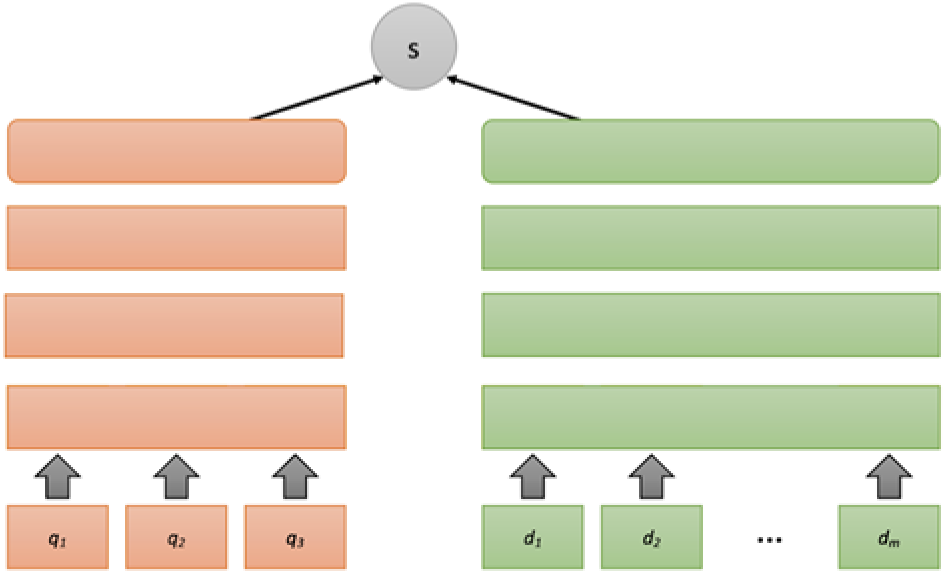
Machine Learning Background

Learning to rank
Deep learning for ranking
BERT

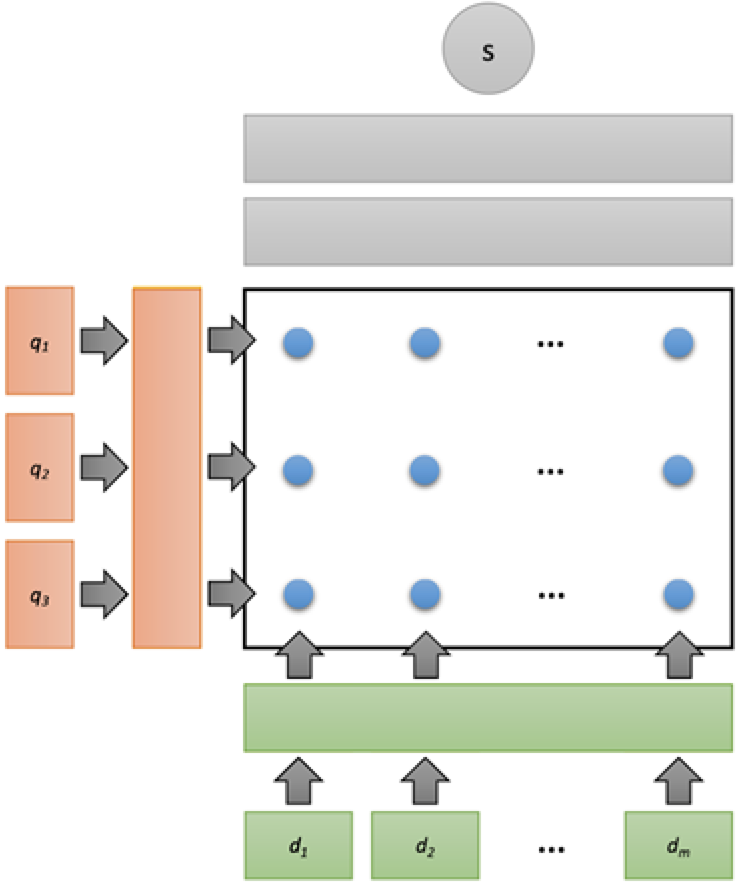
Neural Ranking Models (> 2016)

We will revisit these architectures in Dense Retrieval Section

Representation-based

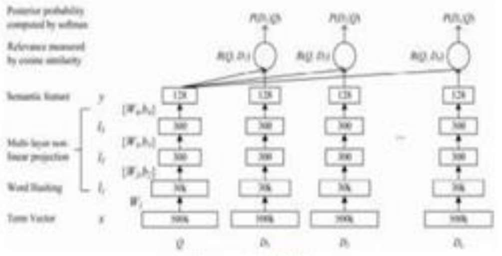


Interaction-based

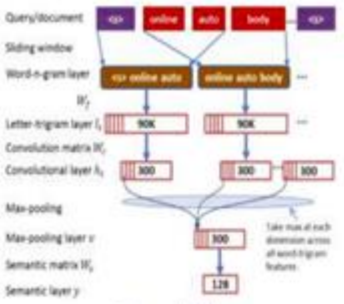


Popular Neural Ranking Models

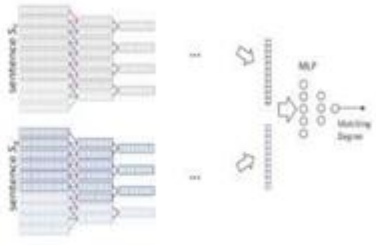
- DESM (Nalisnick et al., 2016)
- MatchPyramid (Pang et al., 2016)
- DUET (Mitra et al., 2017)
- PACRR (Hui et al., 2017)
- Co-PACRR (Hui et al., 2018)
- ConvKNRM (Dai et al., 2018)
- Query Expansion w/ Embeddings
  - (Diaz et al., 2016, Roy et al., 2016)
- ….



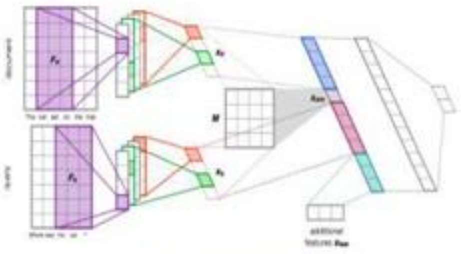
(Huang et al., 2013)



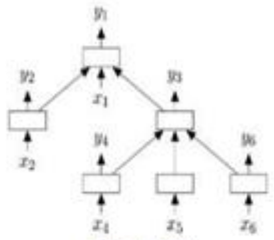
(Shen et al., 2014)



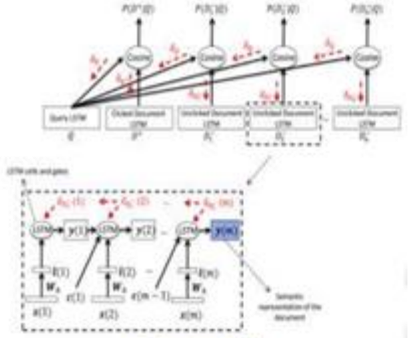
(Hu et al., 2014)



(Severyn and Moschitti, 2015)



(Tai et al., 2015)



(Palangi et al., 2015)

Microsoft

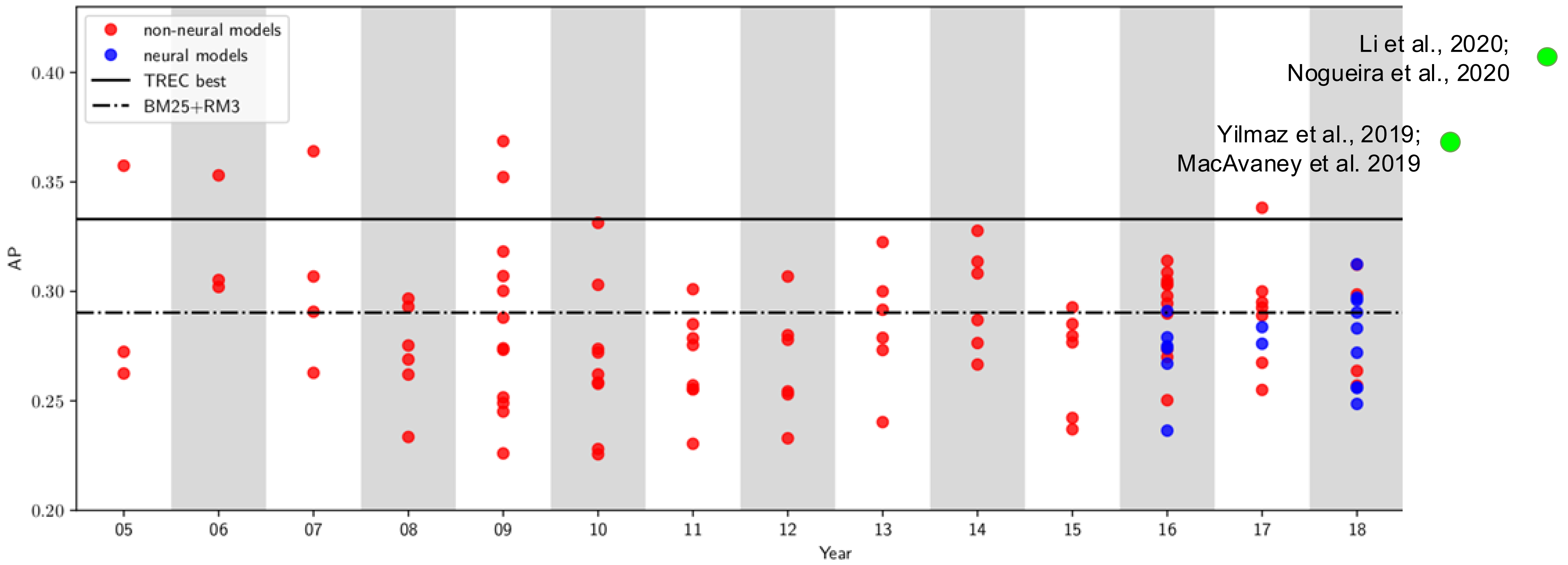
- Check Mitra and Craswell, (2017) for an excellent survey of these methods

Machine Learning Background

Learning to rank
Deep learning for ranking
BERT

Progress in Information Retrieval - Robust04

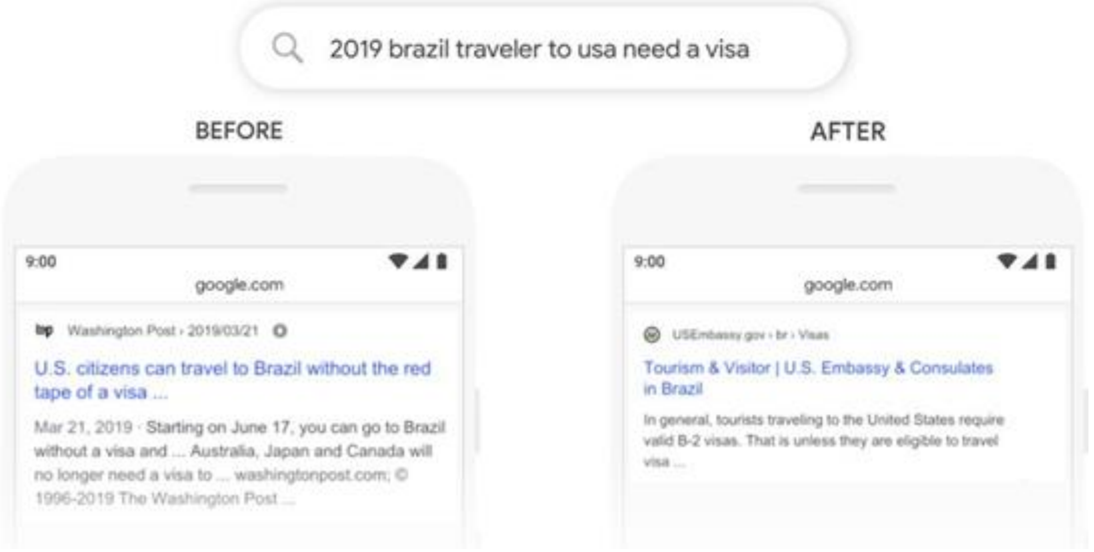
Some of them are zero-shot!



Source: [Yang et al., (2019)]

Adoption by Commercial Search Engines

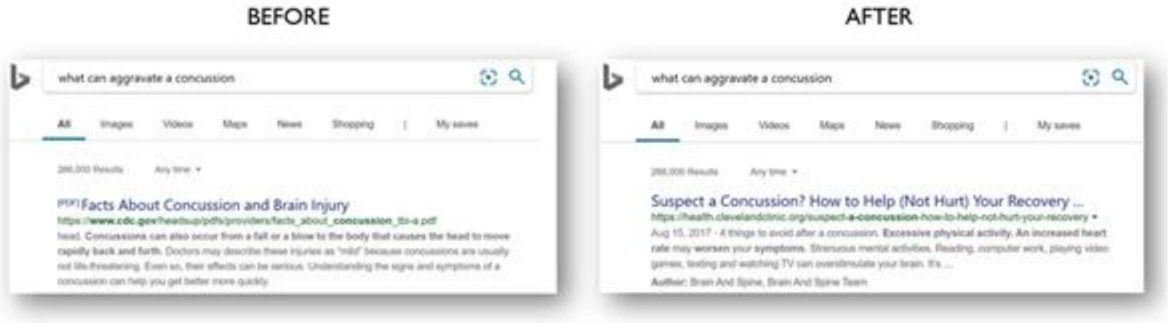
Google Search



We're making a significant improvement to how we understand queries, representing the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search.

source

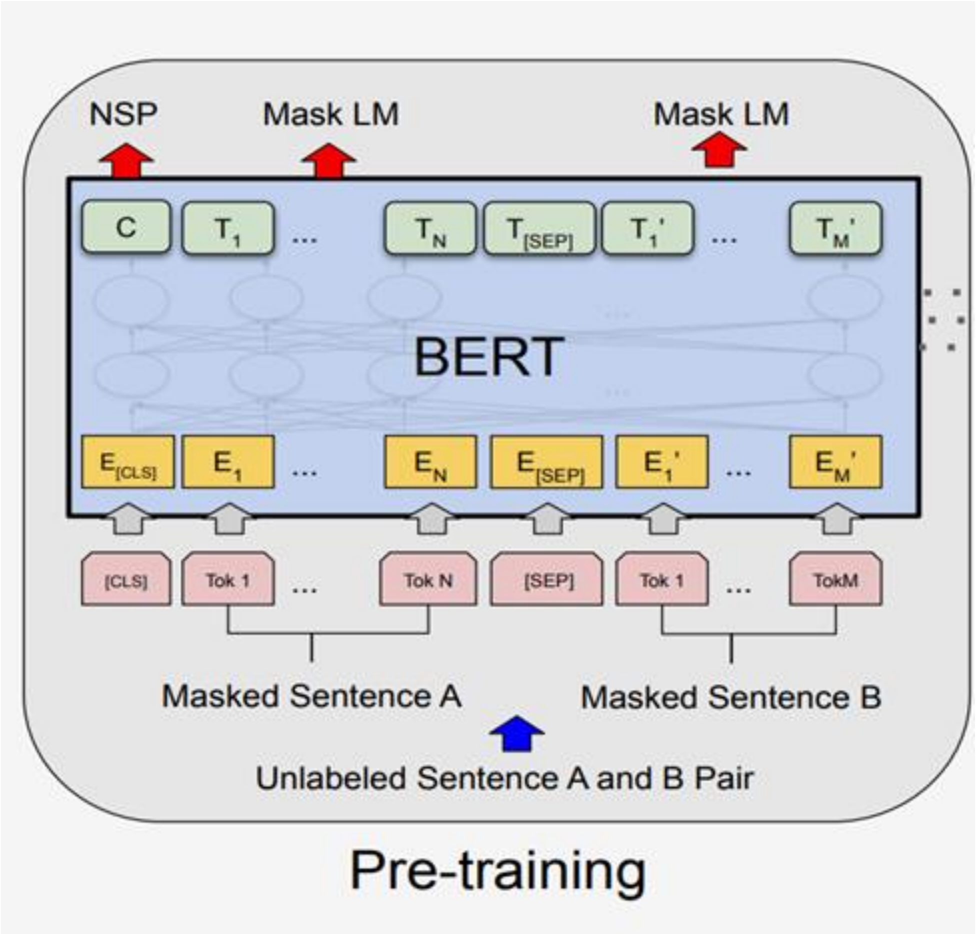
MS Bing



Starting from April of this year (2019), we used large transformer models to deliver the largest quality improvements to our Bing customers in the past year.

source

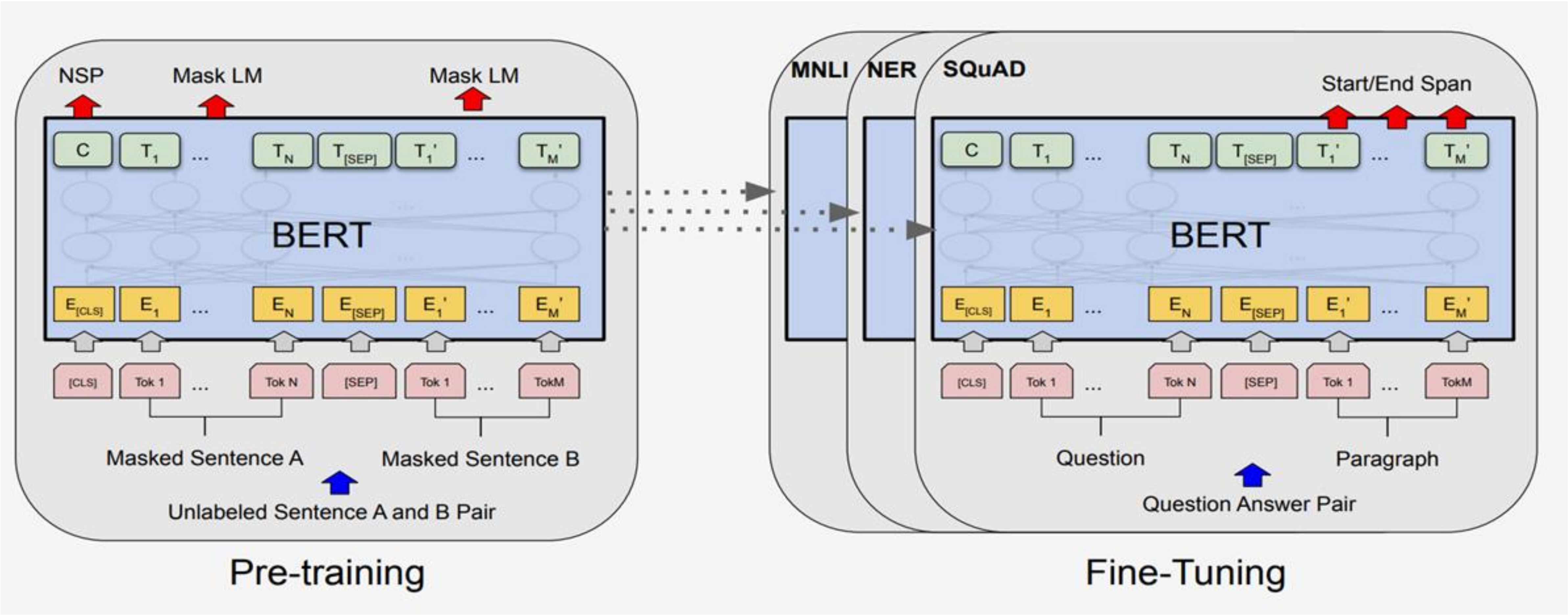
What is BERT?



Self-supervised: ∞ training data

Devlin, Chang, Lee, Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

What is BERT?

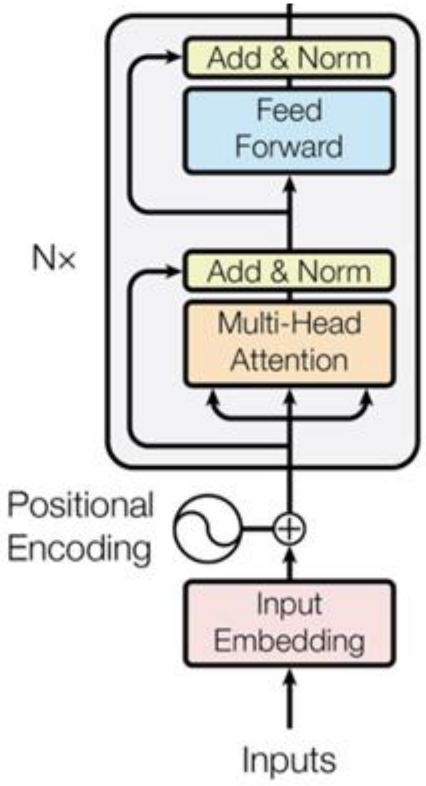


Self-supervised: ∞ training data

Supervised: (few) labeled examples

Devlin, Chang, Lee, Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

BERT's Pretraining Ingredients

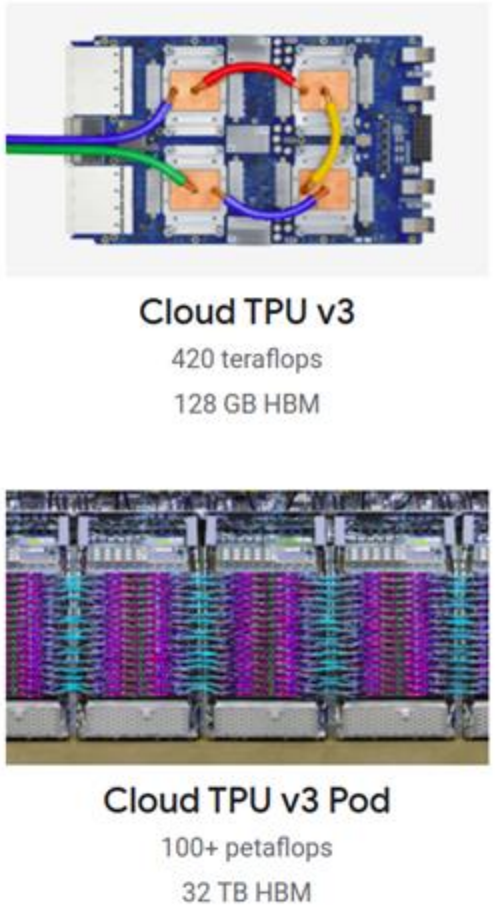


Transformer (encoder-only) with lots of parameters

+

Lots of texts

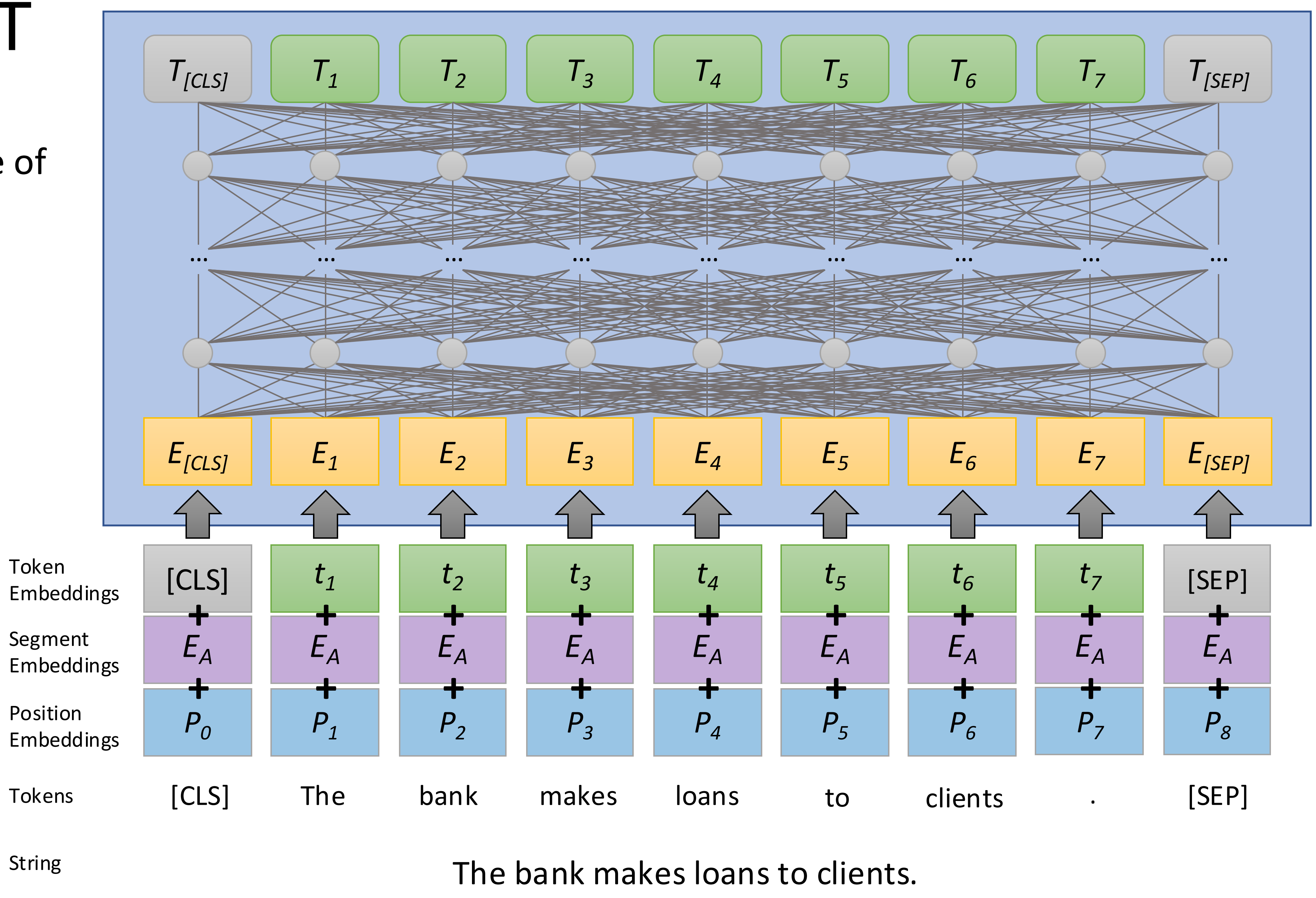
+



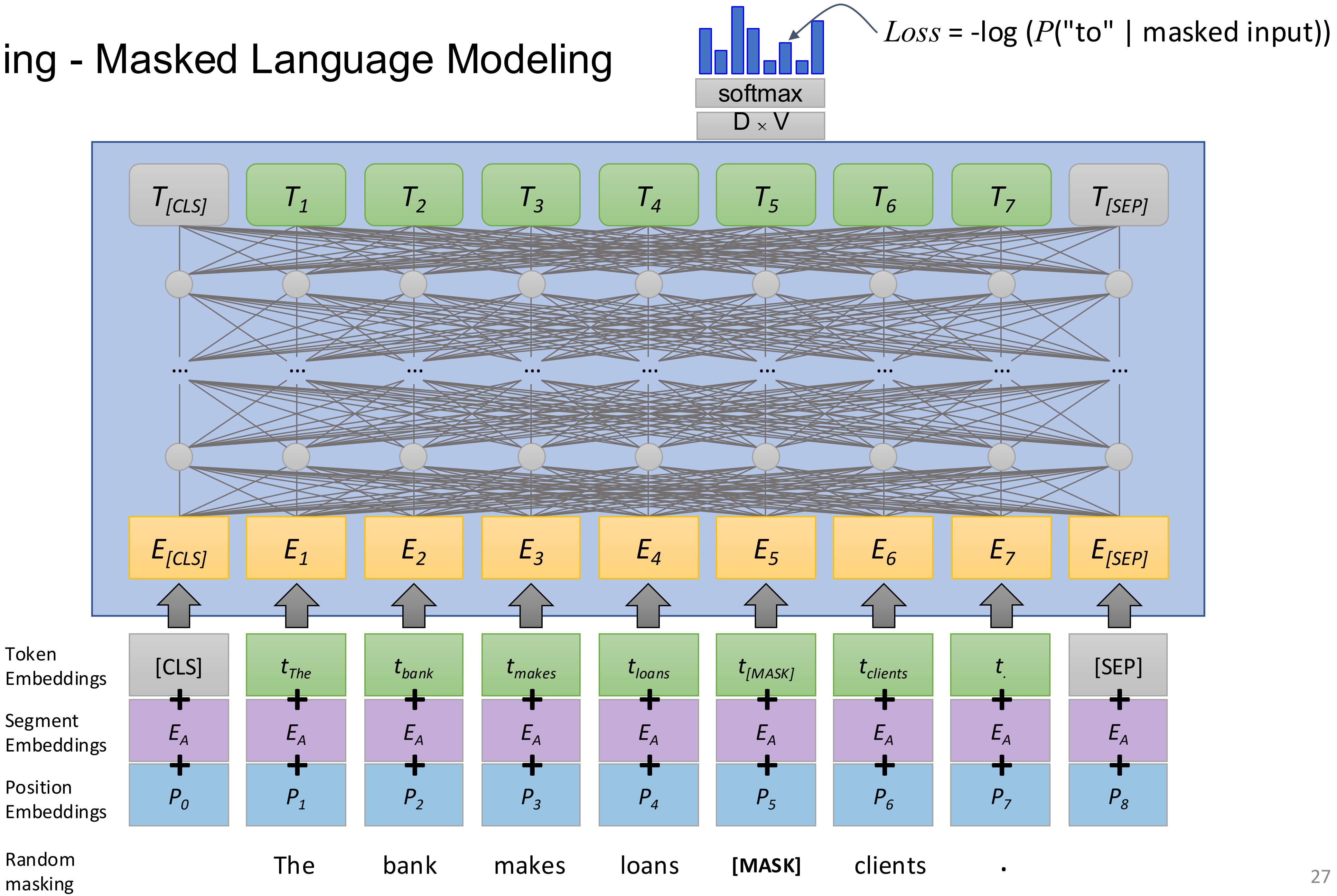
Lots of Compute

BERT

string → sequence of vectors



Pretraining - Masked Language Modeling



BERT
for Relevance Classification

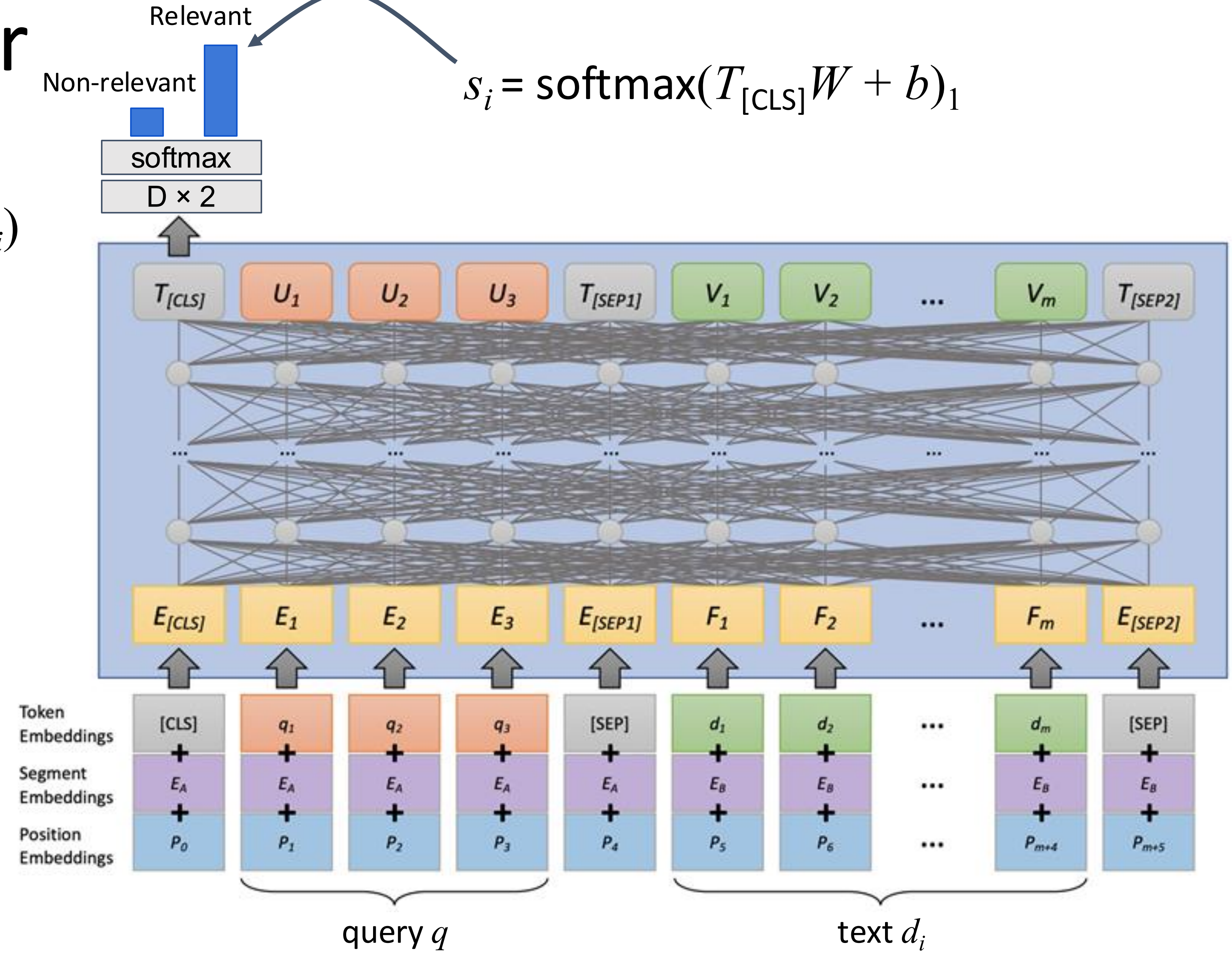
(aka monoBERT)

monoBERT: BERT reranker

We want:

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

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



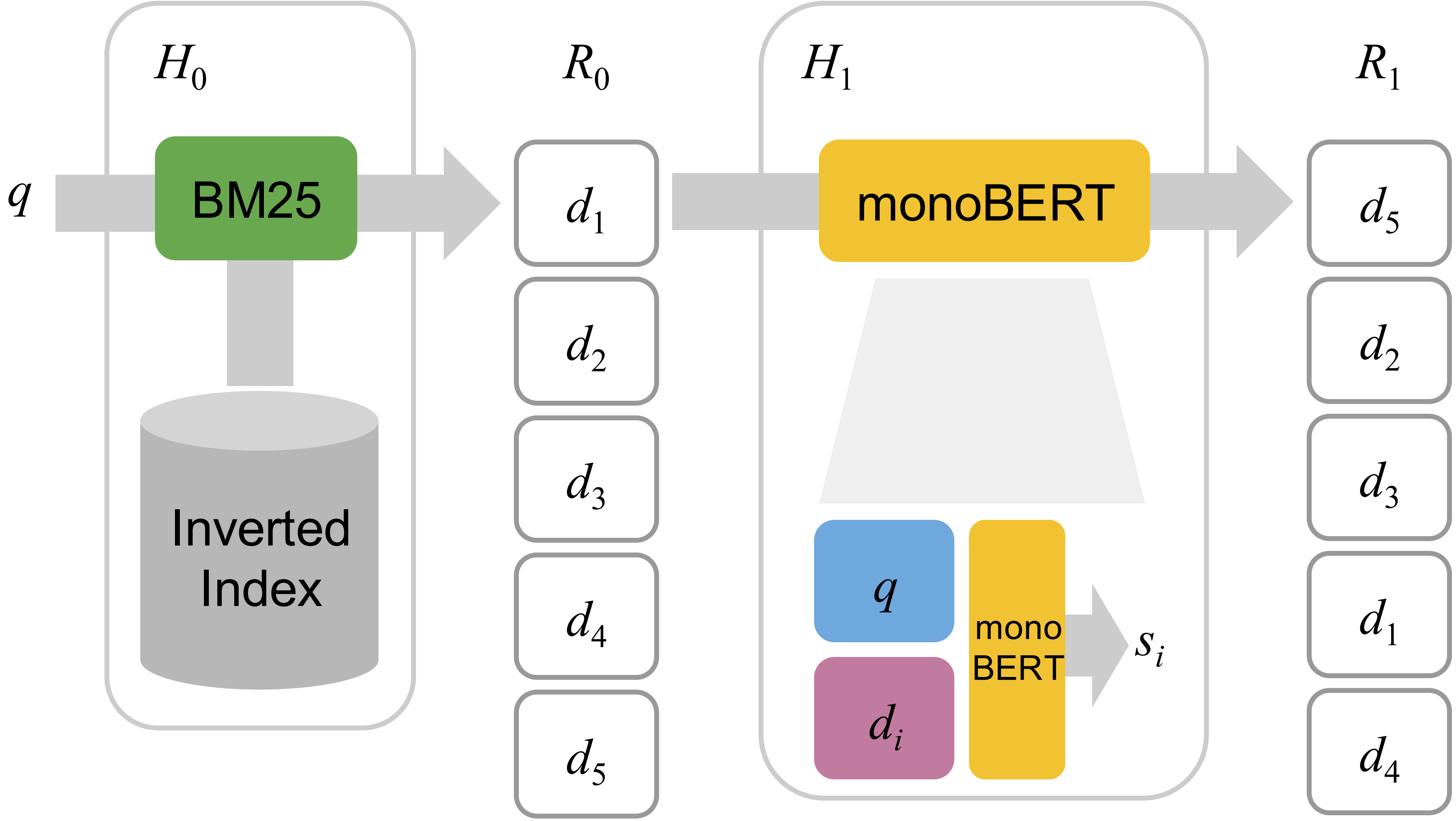
Training monoBERT

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

Humans

BM25

Once monoBERT is trained...



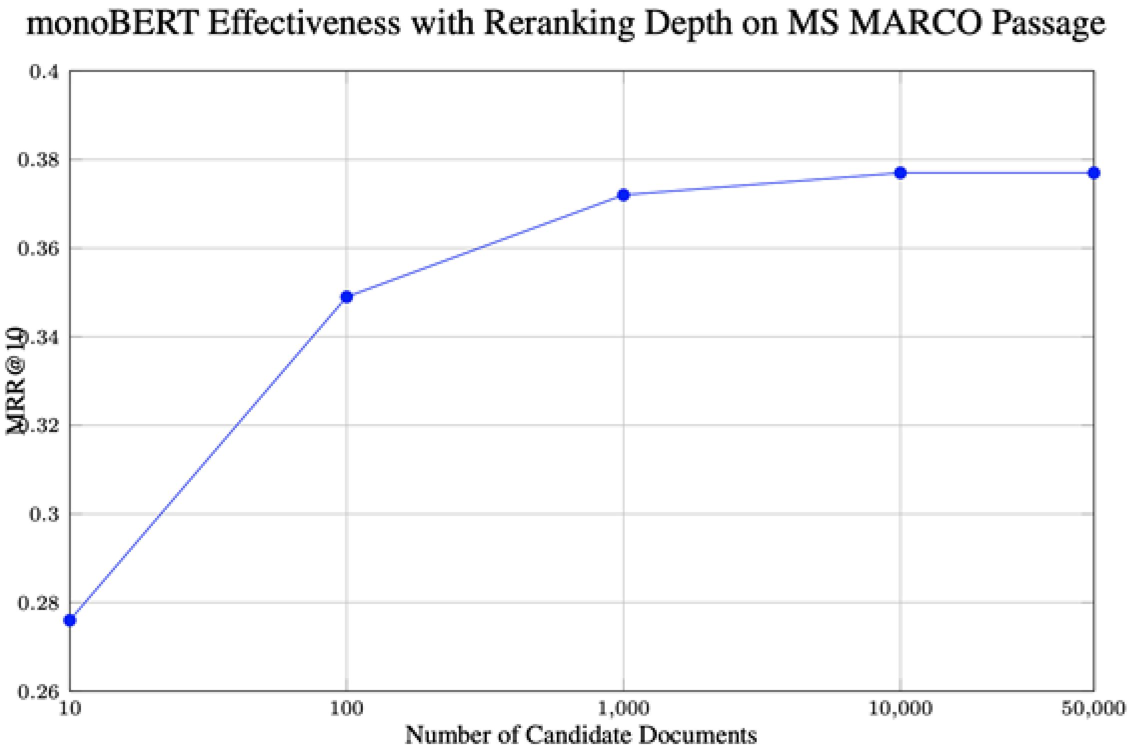
TREC 2019 - Deep Learning Track - Passage

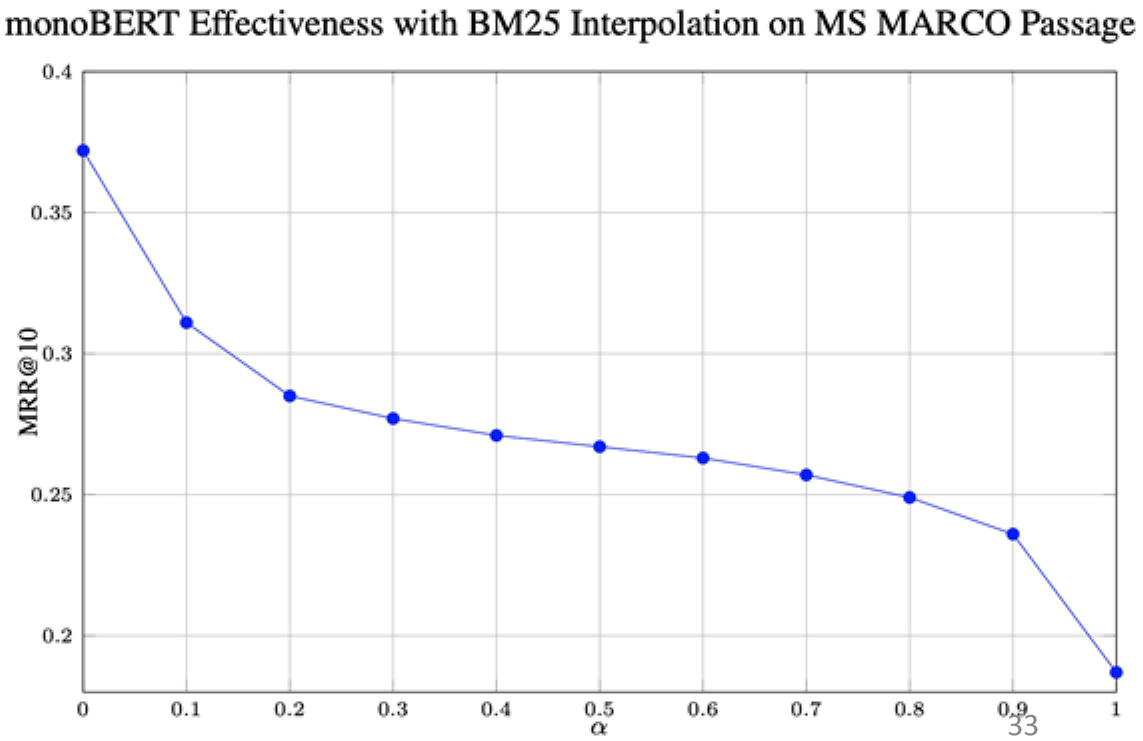
	nDCG@10	MAP	Recall@1k
BM25	0.506	0.377	0.739
+ monoBERT	**0.738**	**0.506**	0.739
BM25 + RM3	0.518	0.427	0.788
+ monoBERT	**0.742**	**0.529**	0.788

How useful is the BM25 signal?

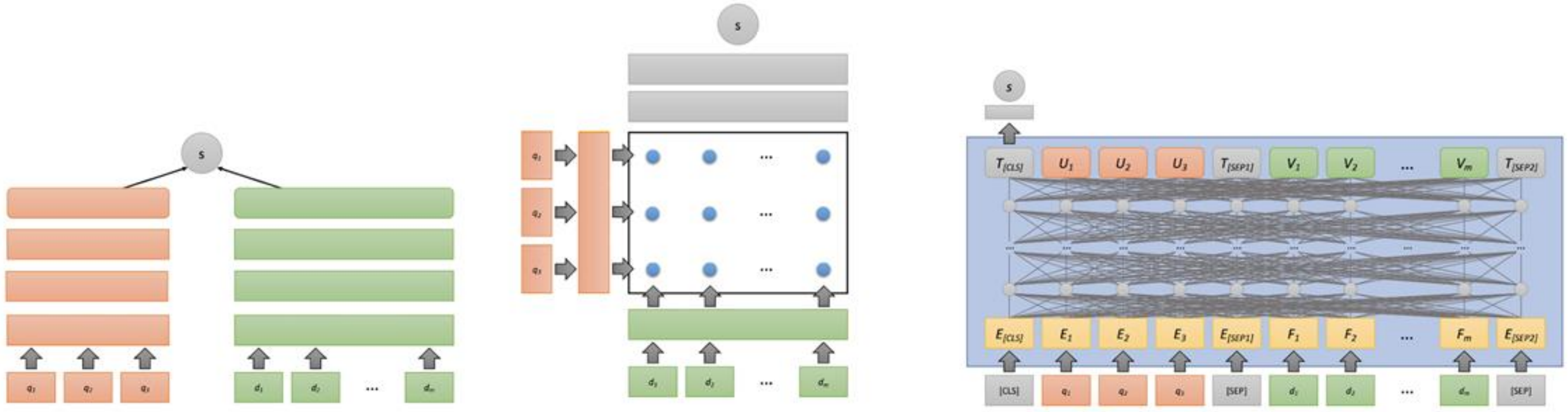
$$s_i \triangleq \alpha \cdot \hat{s}_{\text{BM25}} + (1 - \alpha) \cdot s_{\text{BERT}}$$

$$\hat{s}_{\text{BM25}} = \frac{s_{\text{BM25}} - s_{\min}}{s_{\max} - s_{\min}}$$





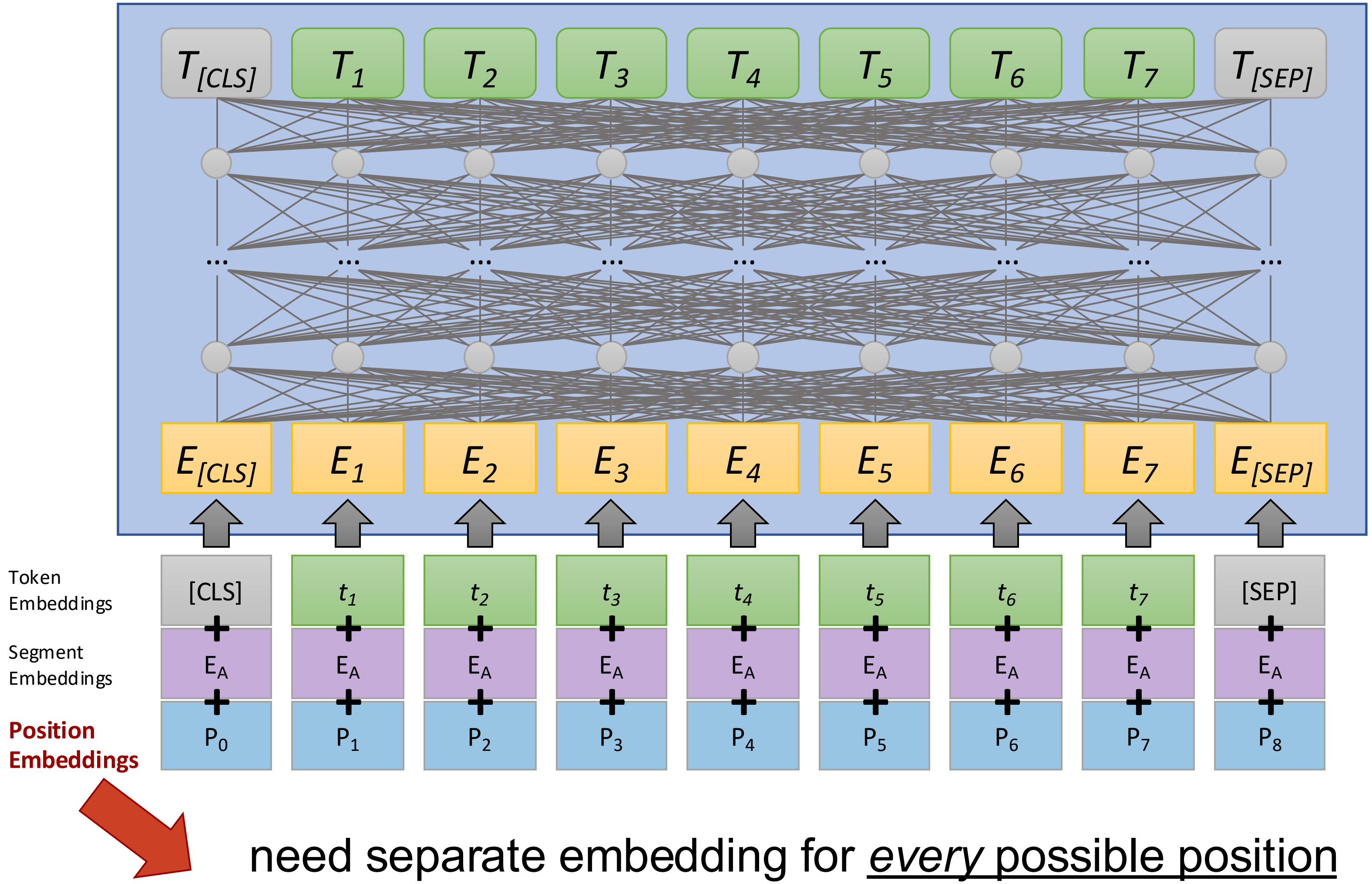
Recap: Pre-BERT vs. monoBERT



(a) Representation-Based (b) Interaction-Based (c) monoBERT

Part 2: Ranking with Relevance Classification

BERT's Limitations



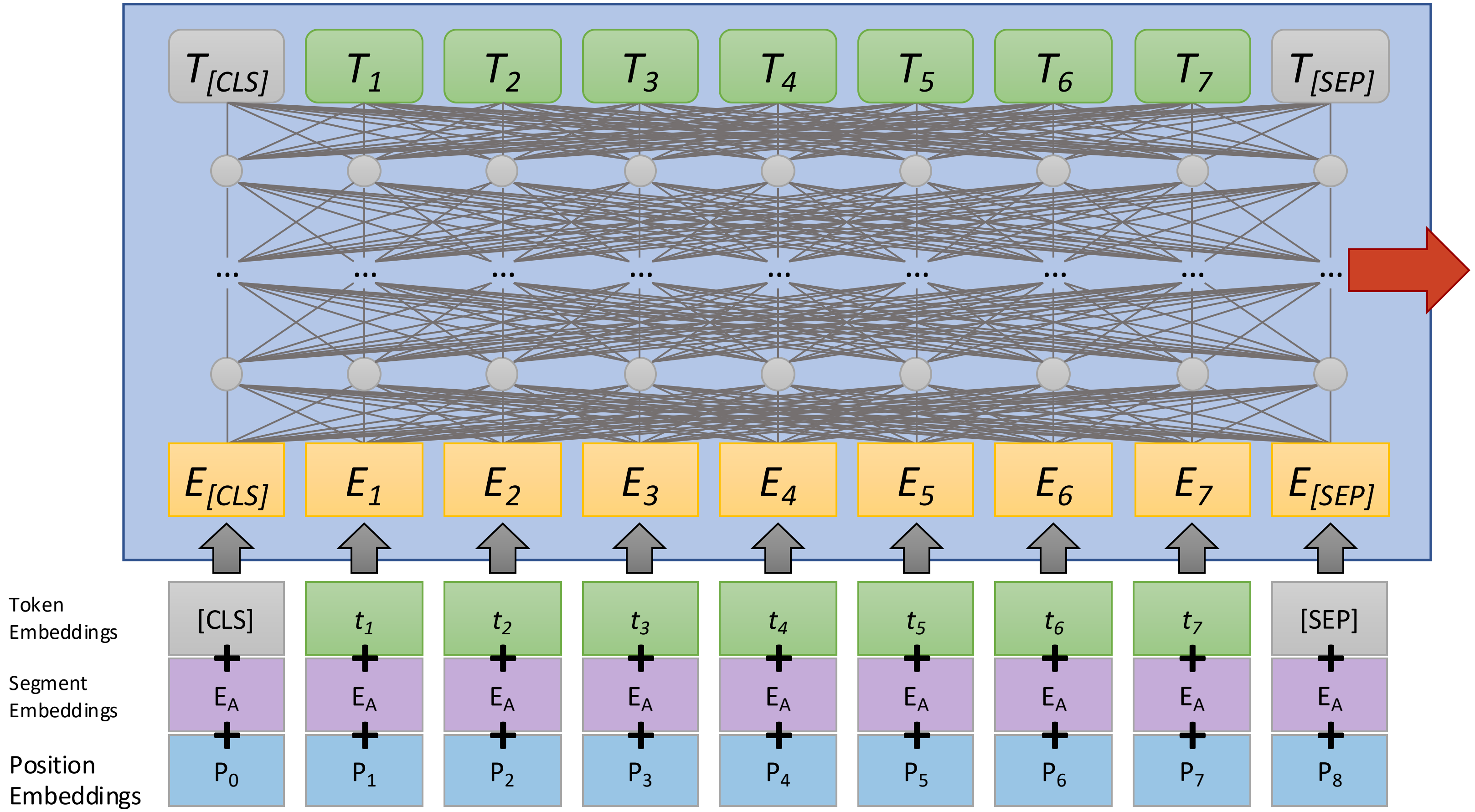
need separate embedding for <u>*every* possible position</u>

➔ restricted to indices 0-511

Cannot input entire documents

- what do we input?
- & how do we label it?

BERT's Limitations



computationally expensive layers

➔ e.g., 110+ *million* learned weights

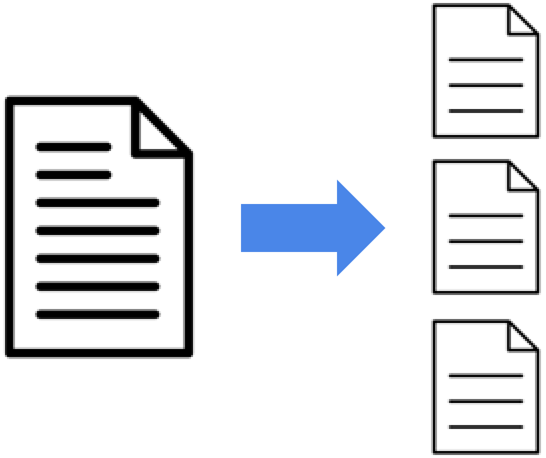
(later: *Beyond BERT* & *Dense Representations*)

Multi-stage ranking pipeline

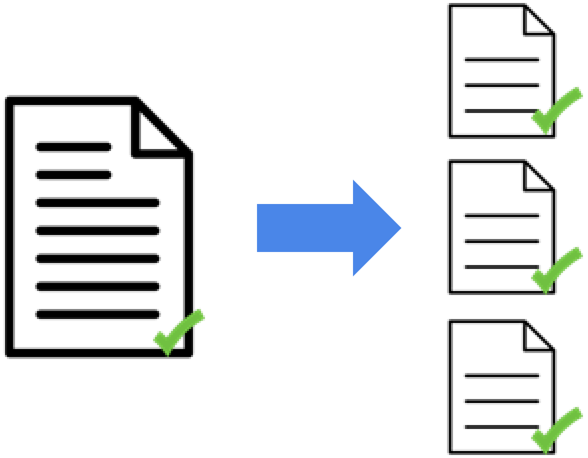
- Identify candidate documents
- Rerank

From Passages to Documents

Handling Length Limitation: Training

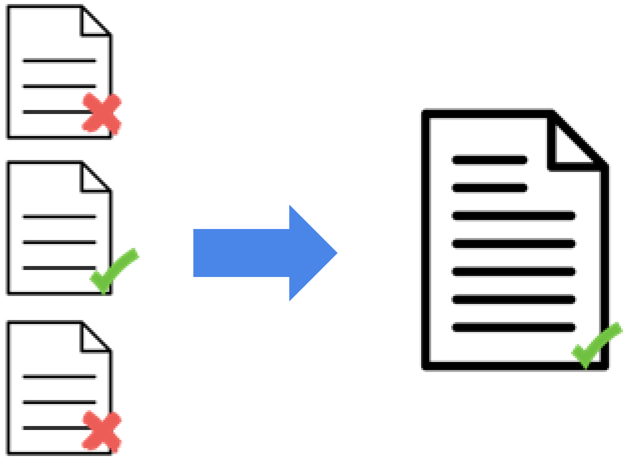


Chunk documents



Transfer labels
(approximation)

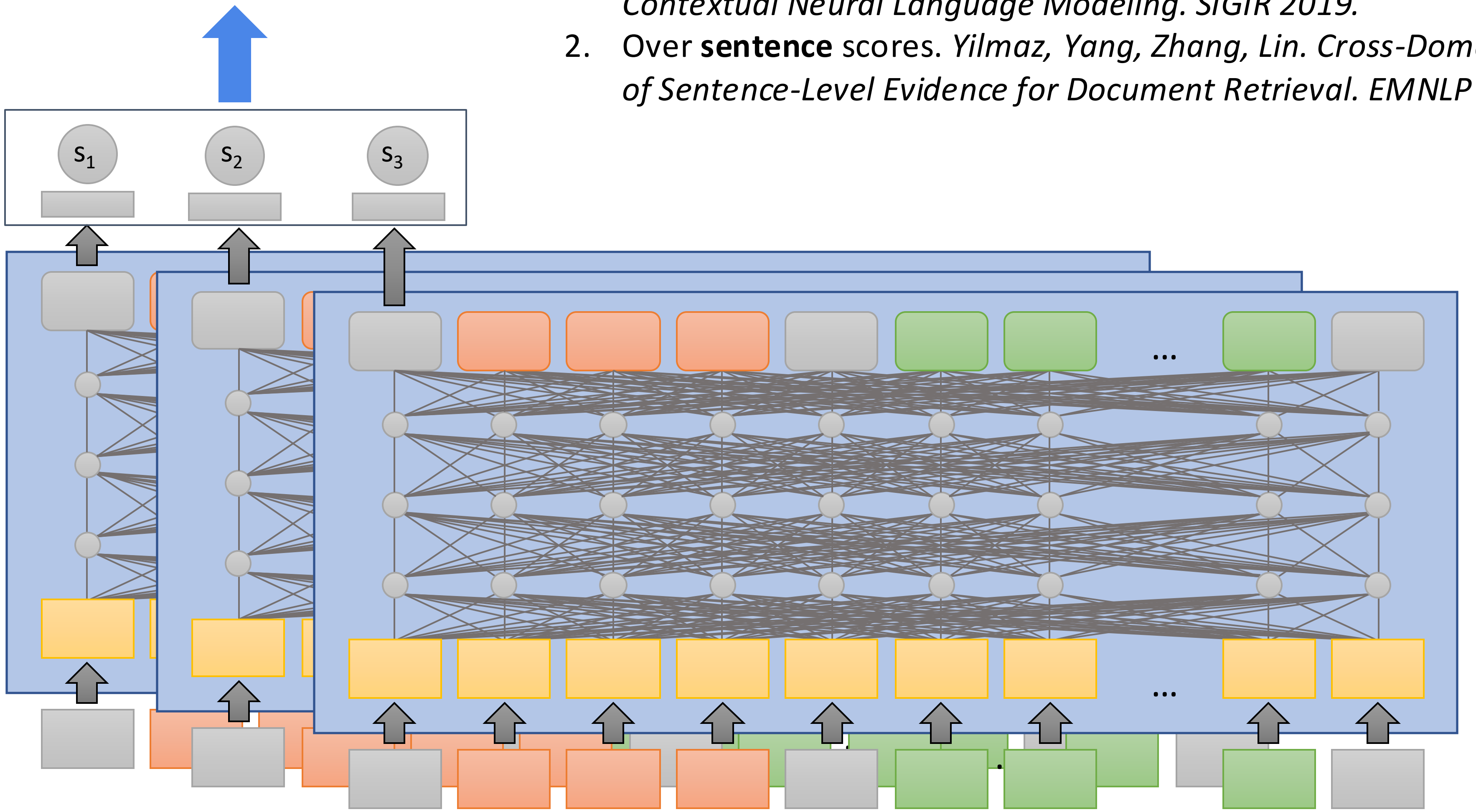
Handling Length Limitation: Inference



Aggregate Evidence

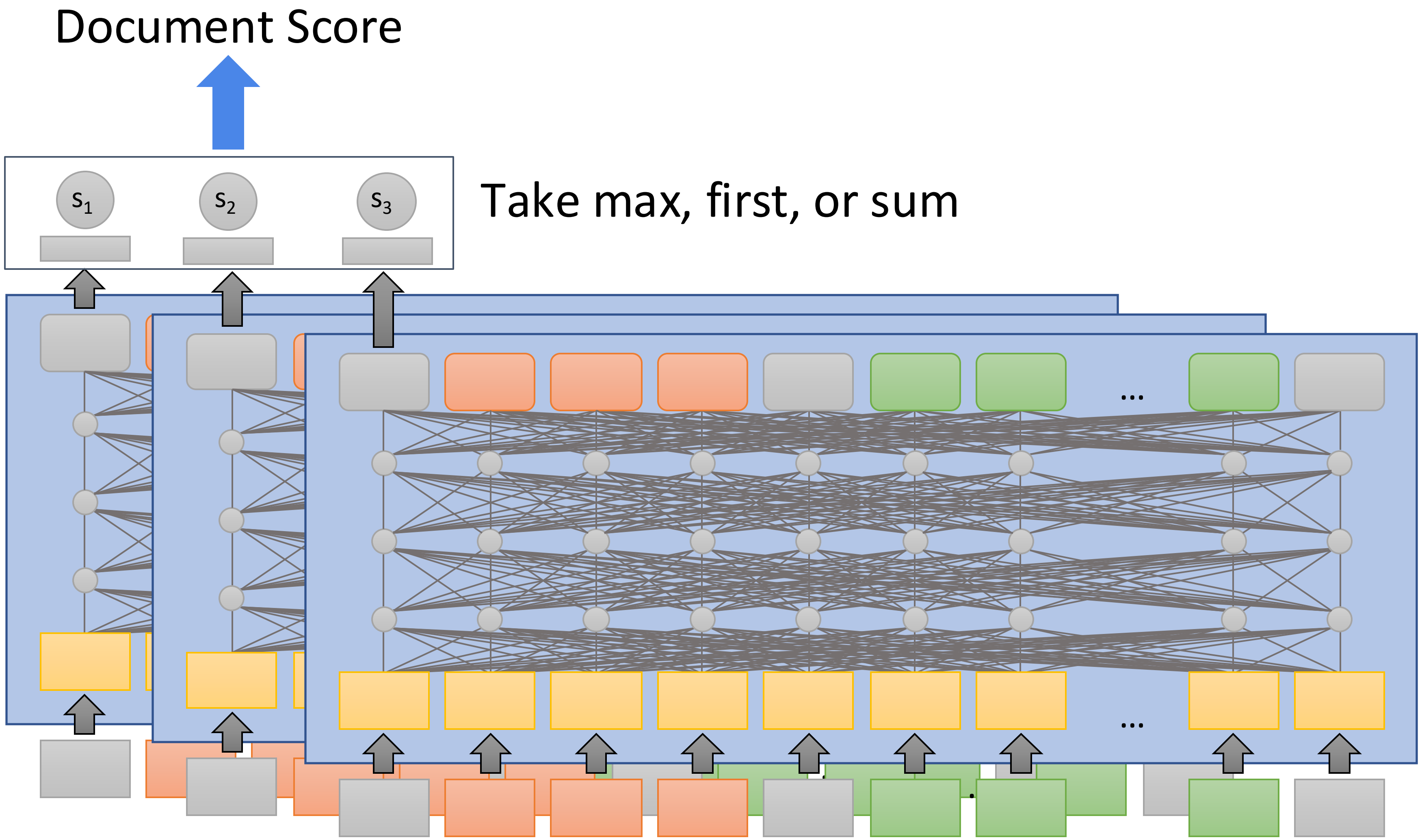
Approach #1: Score Aggregation

Document Score



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

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

Model		Robust04 nDCG@20 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†	0.491†
(4b)	BERT–MaxP	**0.469**†	**0.529**†
(4c)	BERT–SumP	0.467†	0.524†

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

Over Sentence Scores: Birch

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

First-stage retrieval score

Sentence scores

- Trained on sentence-level judgments like tweets
- Interpolation weights are tuned on target dataset

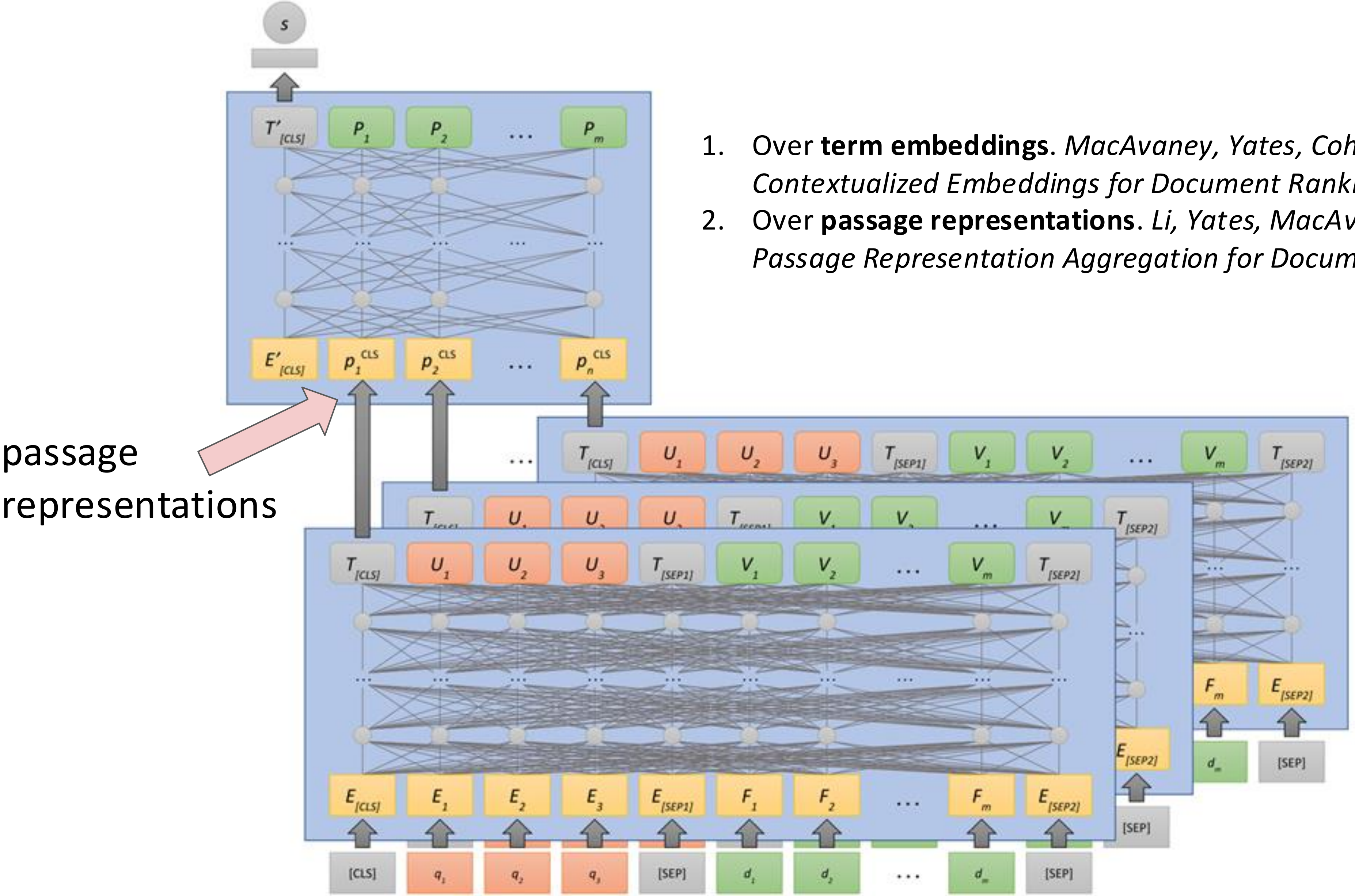
Yilmaz, Yang, Zhang, Lin. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. EMNLP '19.

Over Sentence Scores: Results

	Method	Robust04	
		MAP	nDCG@20
(1)	BM25 + RM3	0.2903	0.4407
(2a)	1S: BERT(MB)	0.3408†	0.4900†
(2b)	2S: BERT(MB)	0.3435†	0.4964†
(2c)	3S: BERT(MB)	0.3434†	0.4998†
(3a)	1S: BERT(MS MARCO)	0.3028†	0.4512
(3b)	2S: BERT(MS MARCO)	0.3028†	0.4512
(3c)	3S: BERT(MS MARCO)	0.3028†	0.4512
(4a)	1S: BERT(MS MARCO → MB)	0.3676†	0.5239†
(4b)	2S: BERT(MS MARCO → MB)	**0.3697**†	0.5324†
(4c)	3S: BERT(MS MARCO → MB)	0.3691†	**0.5325**†

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

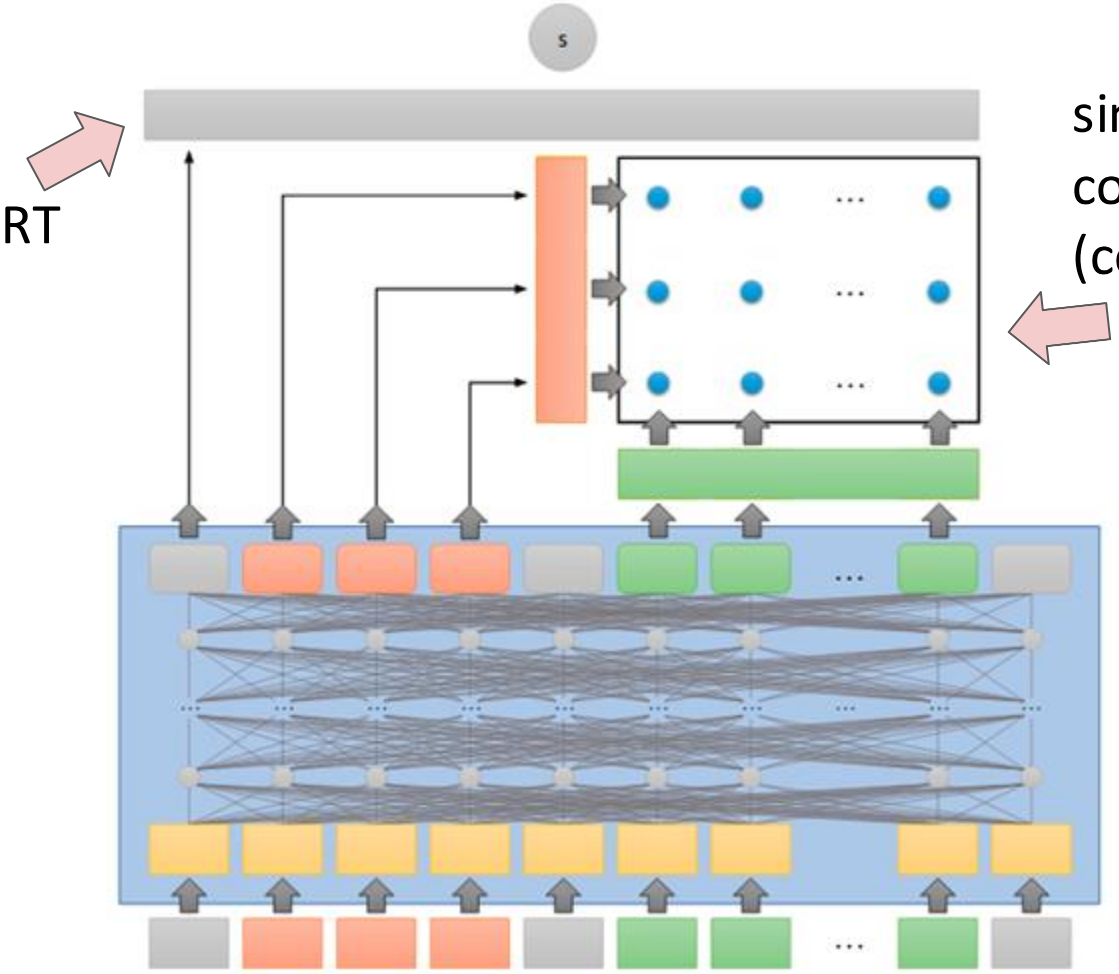
Approach #2: Representation Aggregation



1. Over **term embeddings**. *MacAvaney, Yates, Cohan, Goharian. CEDR: Contextualized Embeddings for Document Ranking. SIGIR 2019.*
2. Over **passage representations**. *Li, Yates, MacAvaney, He, Sun. PARADE: Passage Representation Aggregation for Document Reranking. arXiv 2020.*

Over Term Embeddings: CEDR

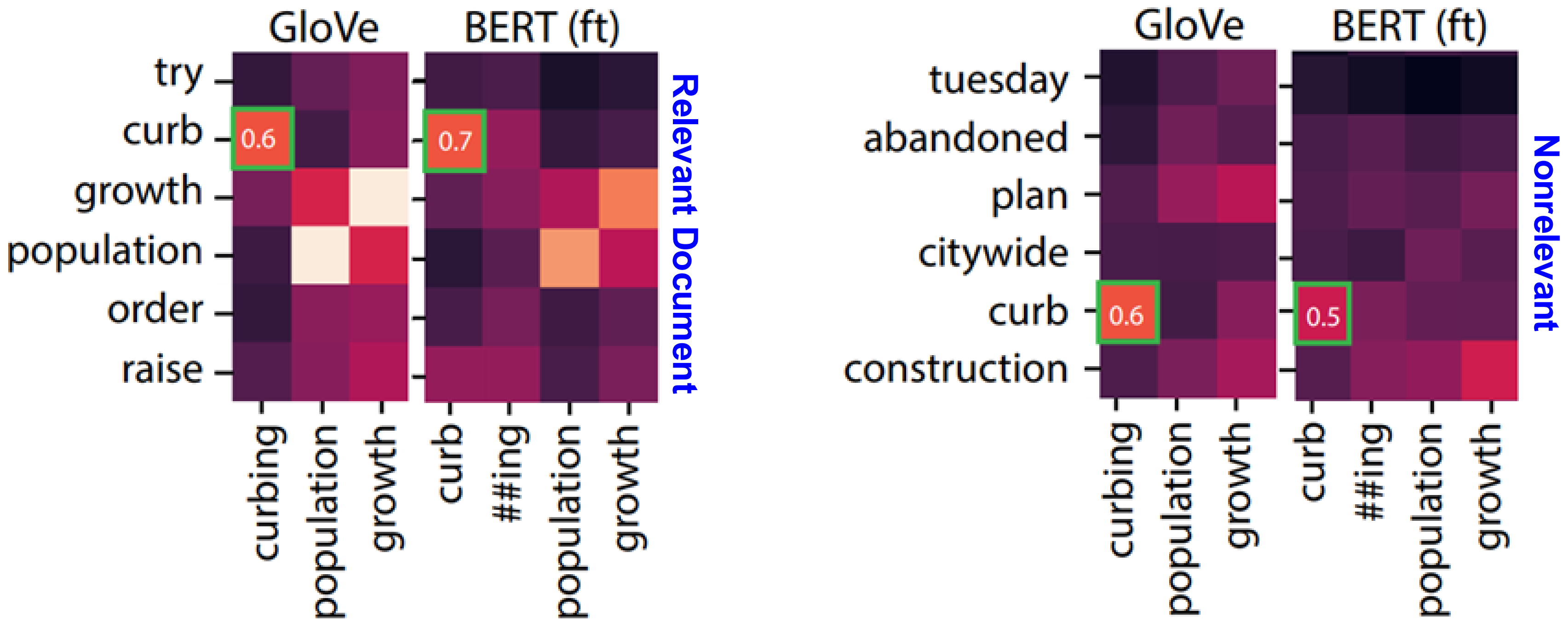
interaction-based pre-BERT model (PACRR, KNRM)



similarity matrix using contextualized embeddings (concatenate passages)

MacAvaney, Yates, Cohan, Goharian. CEDR: Contextualized Embeddings for Document Ranking. SIGIR 2019.

Over Term Embeddings: CEDR



MacAvaney, Yates, Cohan, Goharian. CEDR: Contextualized Embeddings for Document Ranking. SIGIR 2019.

Over Term Embeddings: Results

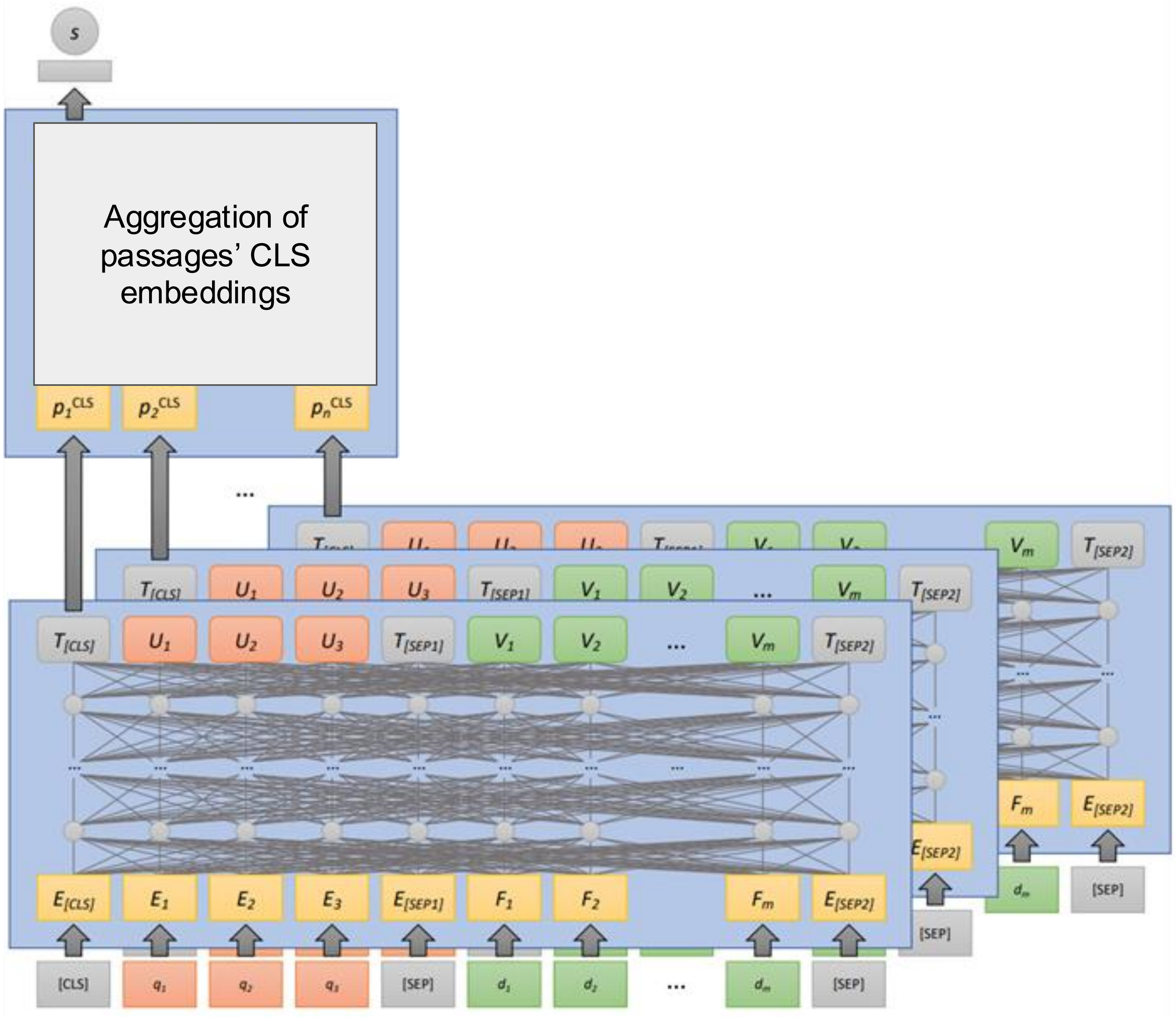
	Method	Input Representation	Robust04 nDCG@20
(1)	BM25	n/a	0.4140
(2)	Vanilla BERT	BERT (fine-tuned)	[B] 0.4541
(3a)	PACRR	GloVe	0.4043
(3b)	PACRR	BERT	0.4200
(3c)	PACRR	BERT (fine-tuned)	[BVG] 0.5135
(3d)	CEDR–PACRR	BERT (fine-tuned)	**[BVG] 0.5150**
(4a)	KNRM	GloVe	0.3871
(4b)	KNRM	BERT	[G] 0.4318
(4c)	KNRM	BERT (fine-tuned)	[BVG] 0.4858
(4d)	CEDR–KNRM	BERT (fine-tuned)	**[BVGN] 0.5381**
(5a)	DRMM	GloVe	0.3040
(5b)	DRMM	BERT	0.3194
(5c)	DRMM	BERT (fine-tuned)	[G] 0.4135
(5d)	CEDR–DRMM	BERT (fine-tuned)	**[BVGN] 0.5259**

MacAvaney, Yates, Cohan, Goharian. CEDR: Contextualized Embeddings for Document Ranking. SIGIR 2019.

Over Passage Representations: PARADE

Aggregation approaches:
(increasing complexity)

- Average feature value
- Max feature value
- Attn-weighted average
- Two Transformer layers



Li, Yates, MacAvaney, He, Sun. PARADE: Passage Representation Aggregation for Document Reranking. arXiv 2020.

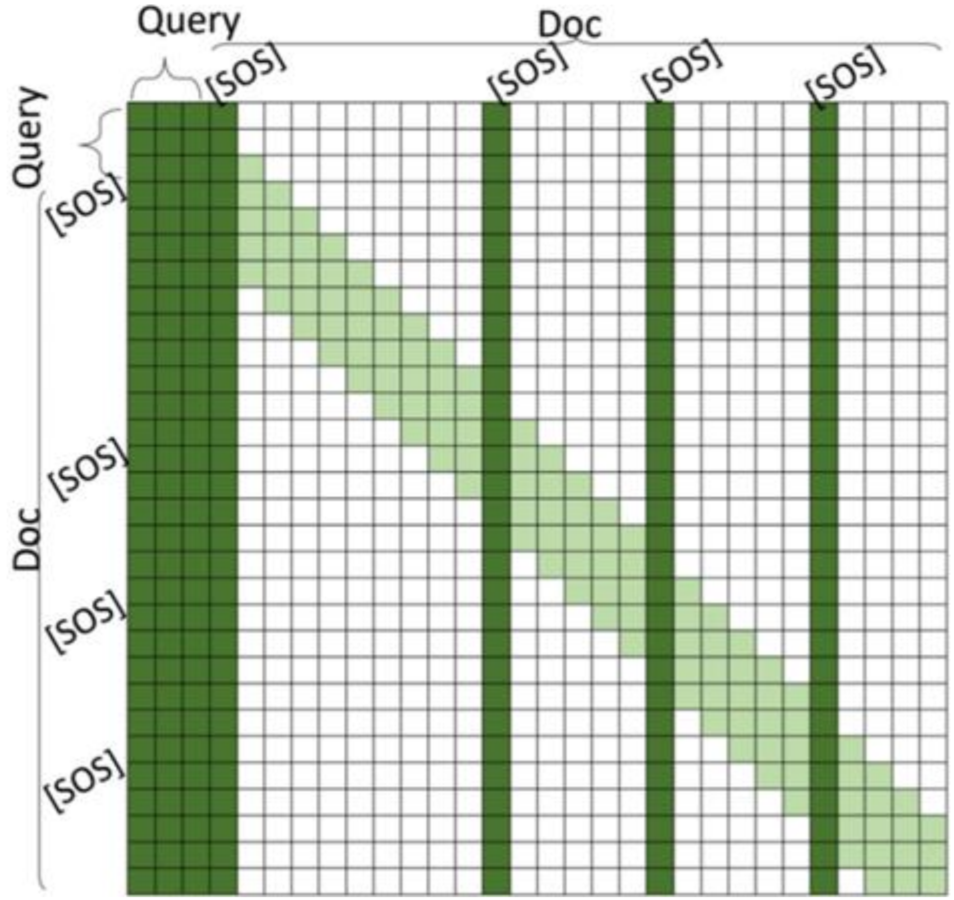
Over Passage Representations: Results

Method		Robust04 nDCG@20 Title	Robust04 nDCG@20 Description
(1)	BM25	0.4240	0.4058
(2)	BM25 + RM3	0.4407	0.4255
(3a)	Birch (MS)	0.4227	0.4053
(3b)	Birch (MS→MB)	0.5137	0.5069
(4)	BERT–MaxP (MS)	0.4931	0.5453
(5a)	$\text{PARADE}_{\text{Avg}}$	$0.4917^{\dagger}$	$0.5324^{\dagger\ddagger}$
(5b)	$\text{PARADE}_{\text{Max}}$	$0.5115^{\dagger\S}$	$0.5487^{\dagger\ddagger}$
(5c)	$\text{PARADE}_{\text{Attn}}$	$0.5134^{\dagger\S}$	$0.5517^{\dagger\ddagger}$
(5d)	PARADE	**0.5252**†§	**0.5605**†‡§
(6)	PARADE (with $\text{BERT}_{\text{Large}}$)	0.5243	-

Li, Yates, MacAvaney, He, Sun. PARADE: Passage Representation Aggregation for Document Reranking. arXiv 2020.

Enlarge Passage Representations: Longformer, QDS

Longformer: sparse attention
QDS-Transformer: specialize to IR



	Method	MS MARCO Doc	TREC 2019 DL Doc	
		MRR@10	nDCG@10	MAP
(1)	Birch (BM25+RM3)	-	0.640	0.328
(2)	Sparse-Transformer	0.328	0.634	0.257
(3)	Longformer-QA	0.326	0.627	0.255
(4)	QDS-Transformer	0.360	0.667	0.278

Beltagy, Peters, Cohan. Longformer: The Long-Document Transformer. arXiv 2020.
Jiang, Xiong, Lee, Wang. Long Document Ranking with Query-Directed Sparse Transformer. Findings of EMNLP 2020.

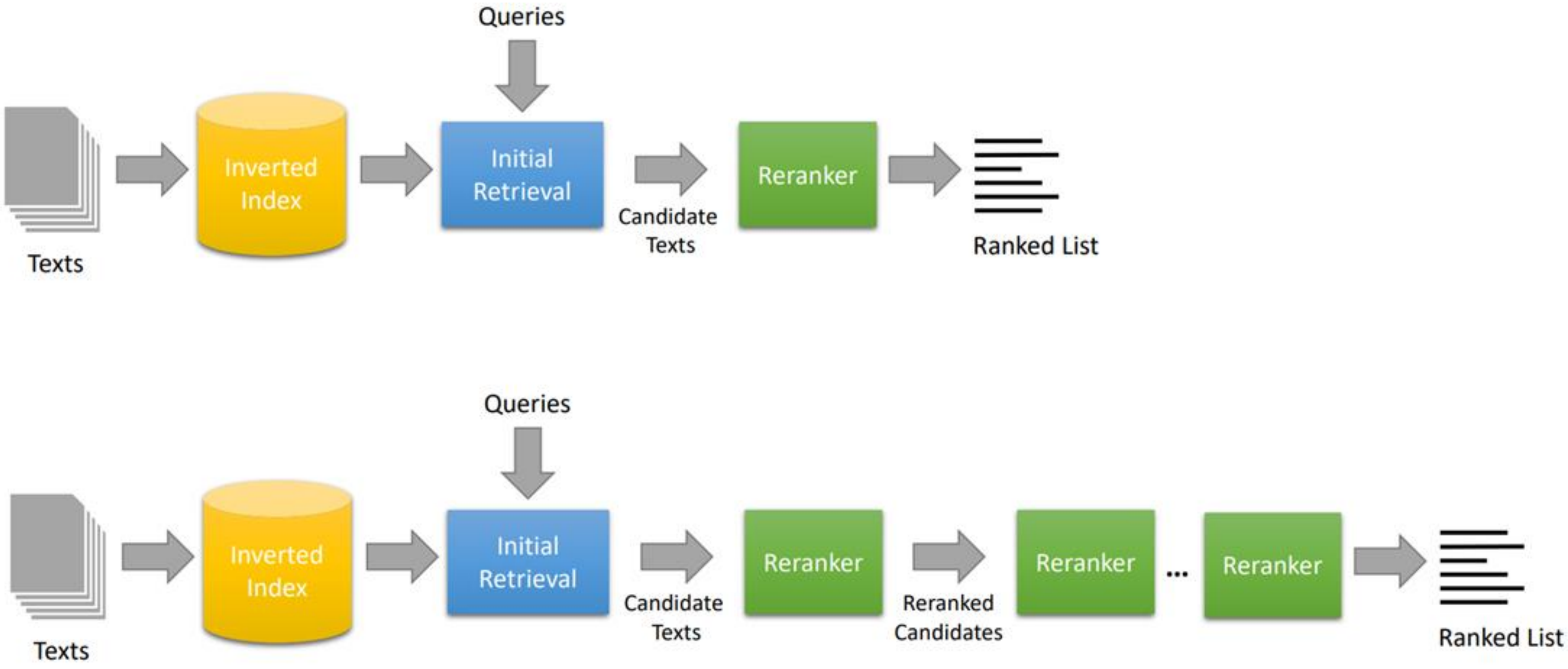
Multi-stage rerankers

why multi-stage?
duoBERT

Multi-stage rerankers

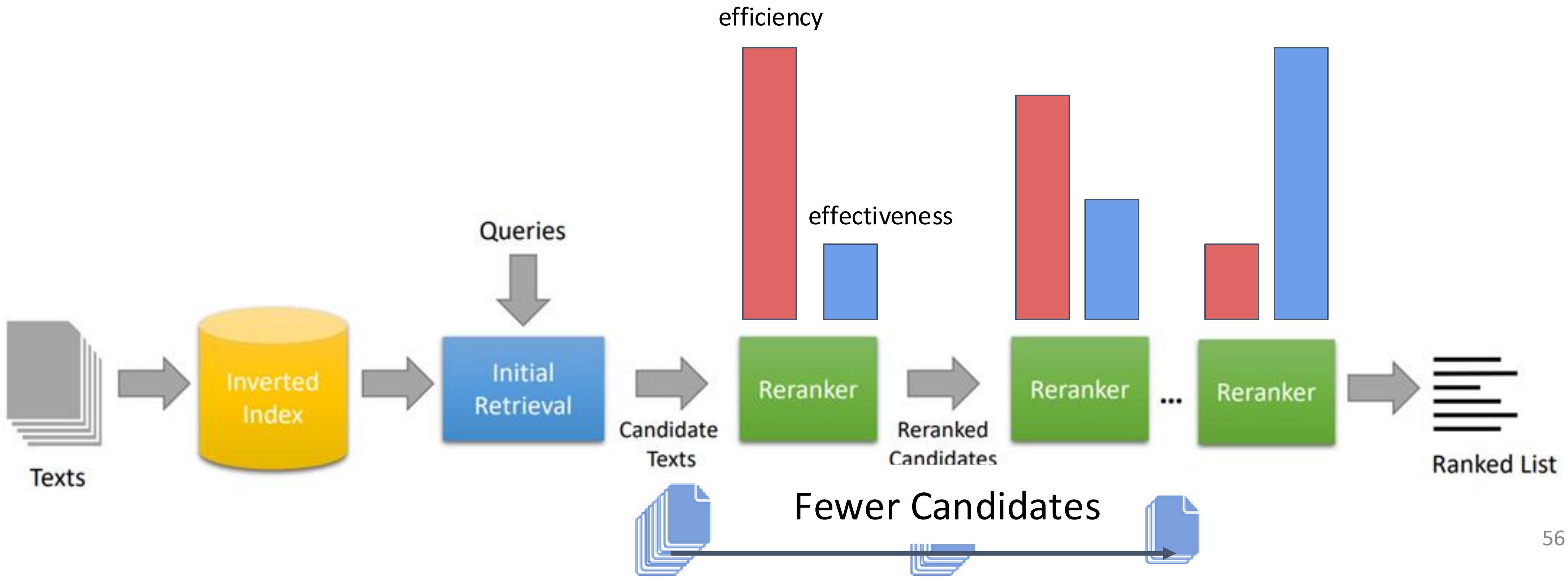
why multi-stage?
duoBERT

From Single to Multiple Rerankers



Why Multi-stage?

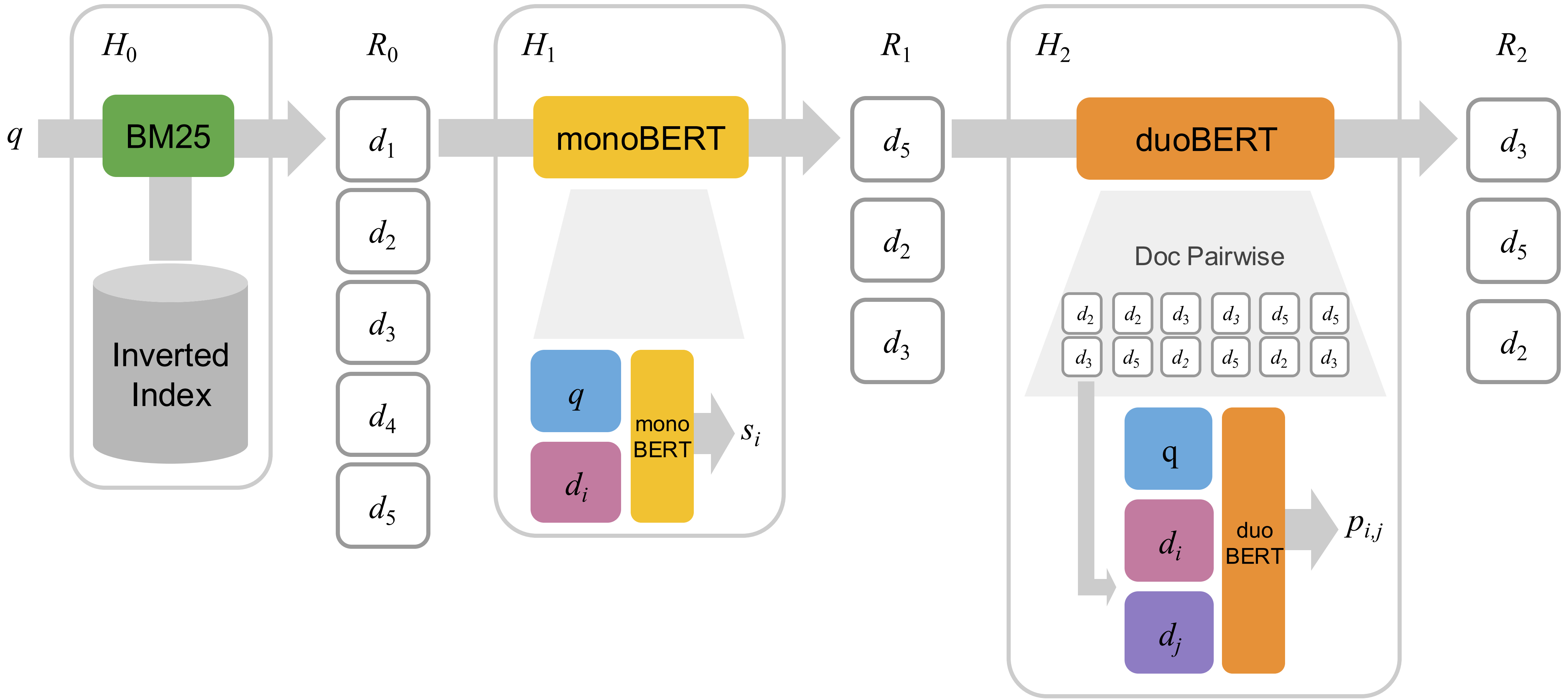
- Trade-off between effectiveness (quality of the ranked lists) and efficiency (retrieval latency)



Multi-stage Rerankers

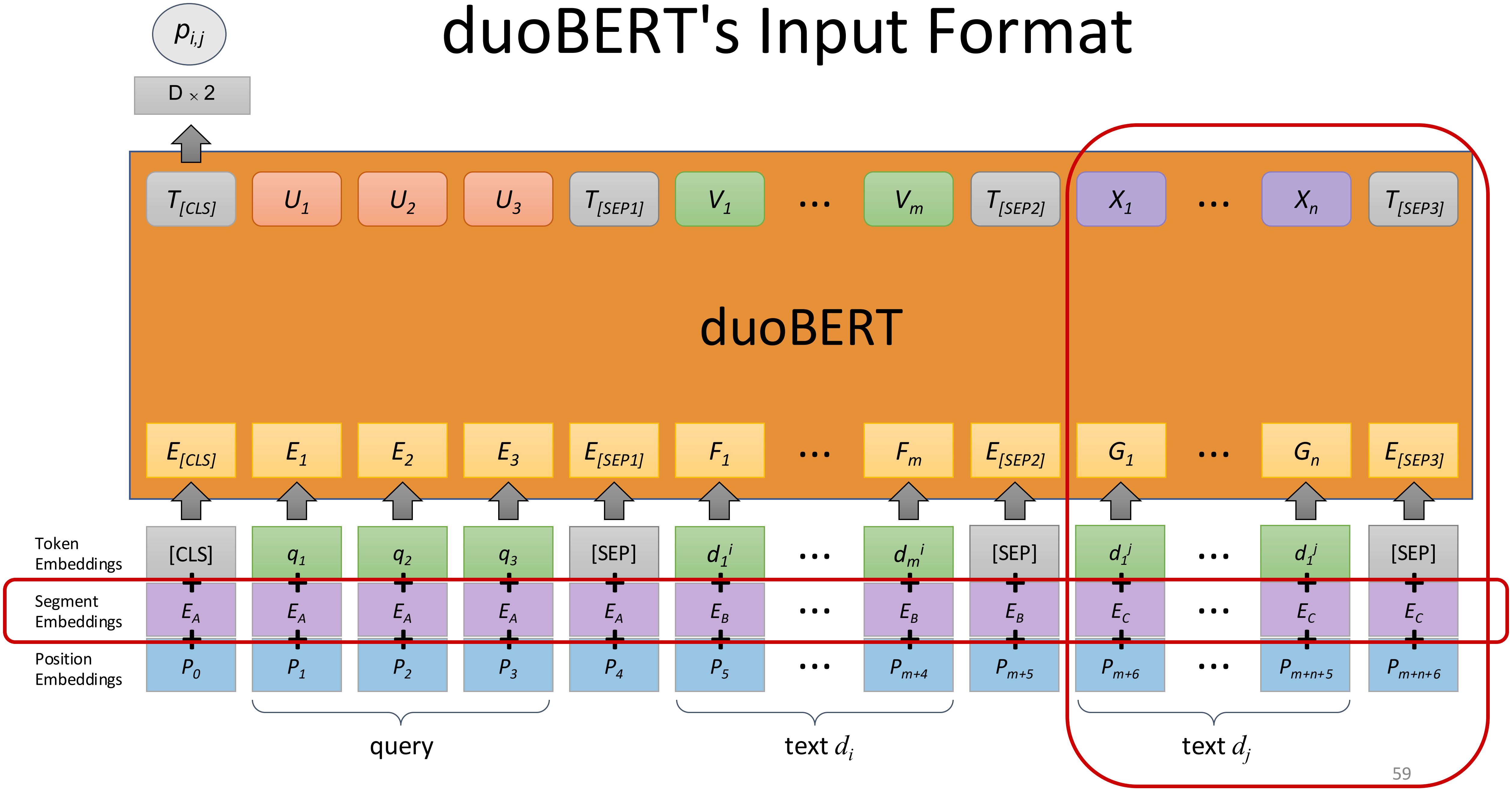
why multi-stage?
duoBERT

Multi-stage with duoBERT



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

duoBERT's Input Format



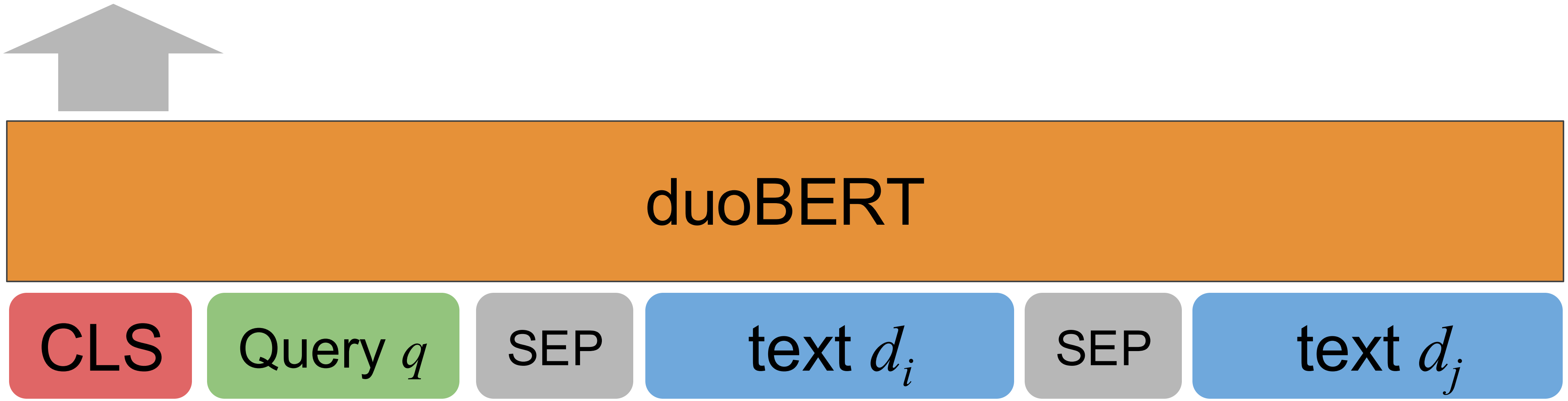
Training duoBERT

Loss:

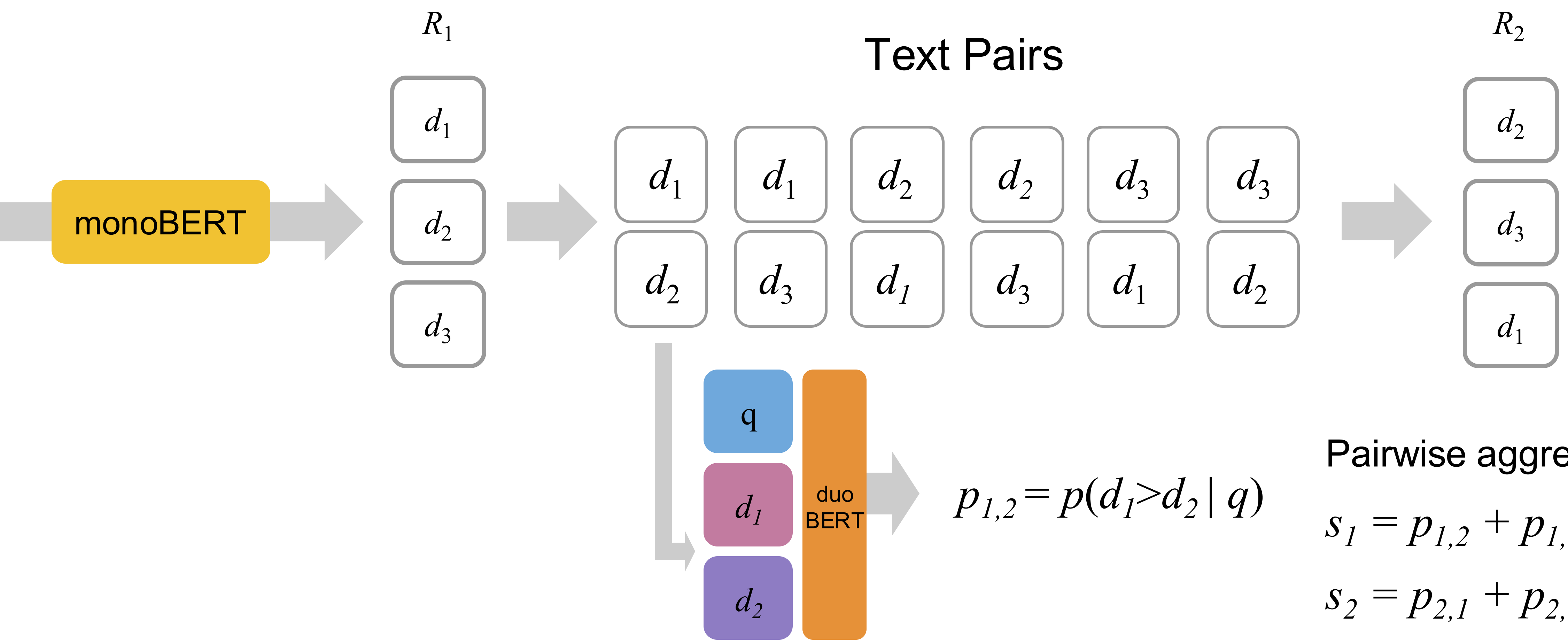
$$L_{\text{duo}} = -\sum_{i \in J_{\text{pos}}, j \in J_{\text{neg}}} \log(p_{i,j}) - \sum_{i \in J_{\text{neg}}, j \in J_{\text{pos}}} \log(1 - p_{i,j})$$

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



$p_{1,2} = p(d_1 > d_2 \mid q)$

Pairwise aggregation:

$s_1 = p_{1,2} + p_{1,3}$

$s_2 = p_{2,1} + p_{2,3}$

$s_3 = p_{3,1} + p_{3,2}$

Takeaways of Multi-stage Rerankers

Advantage:

- more tuning knobs → more flexibility in effectiveness/efficiency tradeoff space

Disadvantage:

- more tuning knobs → more complexity

We are only starting exploring the design space for multi-stage reranking pipelines with Transformers