CS224v

Conversational Virtual Assistants with Deep Learning

Lecture 13: NLP Building Blocks

Omar Khattab and Monica Lam

Lecture Goals

- Technical details on two building blocks under the hood
- **Information Retrieval (on free-text)**
  - Find the relevant paragraph from billions of paragraphs that answers a question?
- **Named Entity Disambiguation**
  - Figure out the entity in a reference
    - E.g. How tall is Lincoln?
      Where is Lincoln Nebraska?

Lecture Goals

- **Information Retrieval (on free-text)**
  - Late Interaction Model
  - Optimization: time (Select k candidates + Late interaction)
  - Optimization: space with compression

- **Named Entity Disambiguation**
  - Use entity description: (Select k candidates + Cross-encoder)
  - Use type information

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

Omar Khattab
Stanford University
okhattab@stanford.edu

Matei Zaharia
Stanford University
matei@cs.stanford.edu

SIGIR '20 (Over 1200 citations)

ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction

Keshav Santhanam*
Stanford University

Omar Khattab*
Stanford University

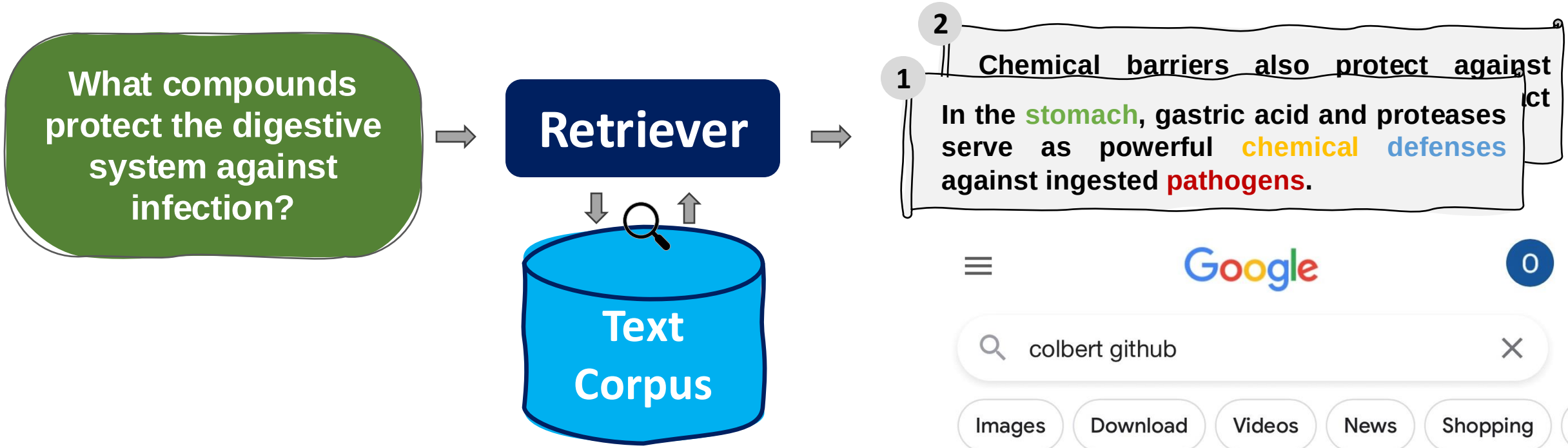
Jon Saad-Falcon
Georgia Institute of Technology

Christopher Potts
Stanford University

Matei Zaharia
Stanford University

NAACL '22

Information Retrieval at a Glance

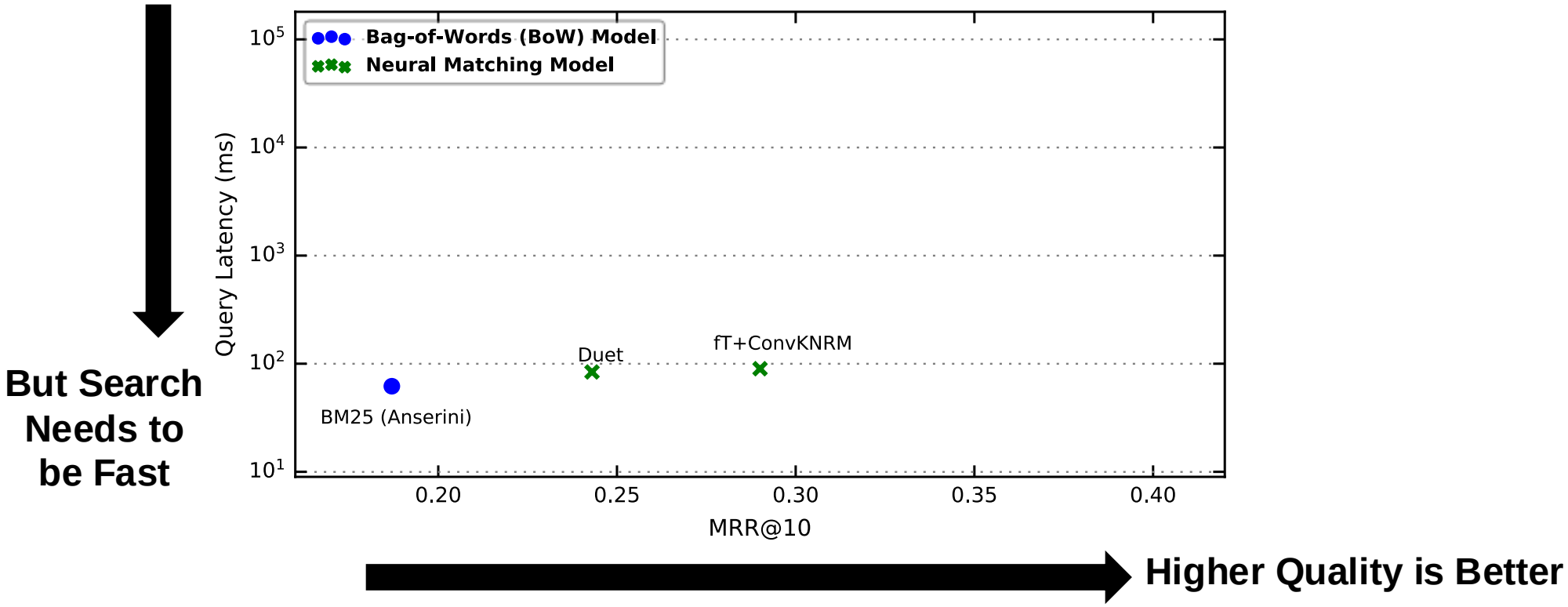


Product Search | **Question Answering**

Fact Checking | **Informative Dialogue**

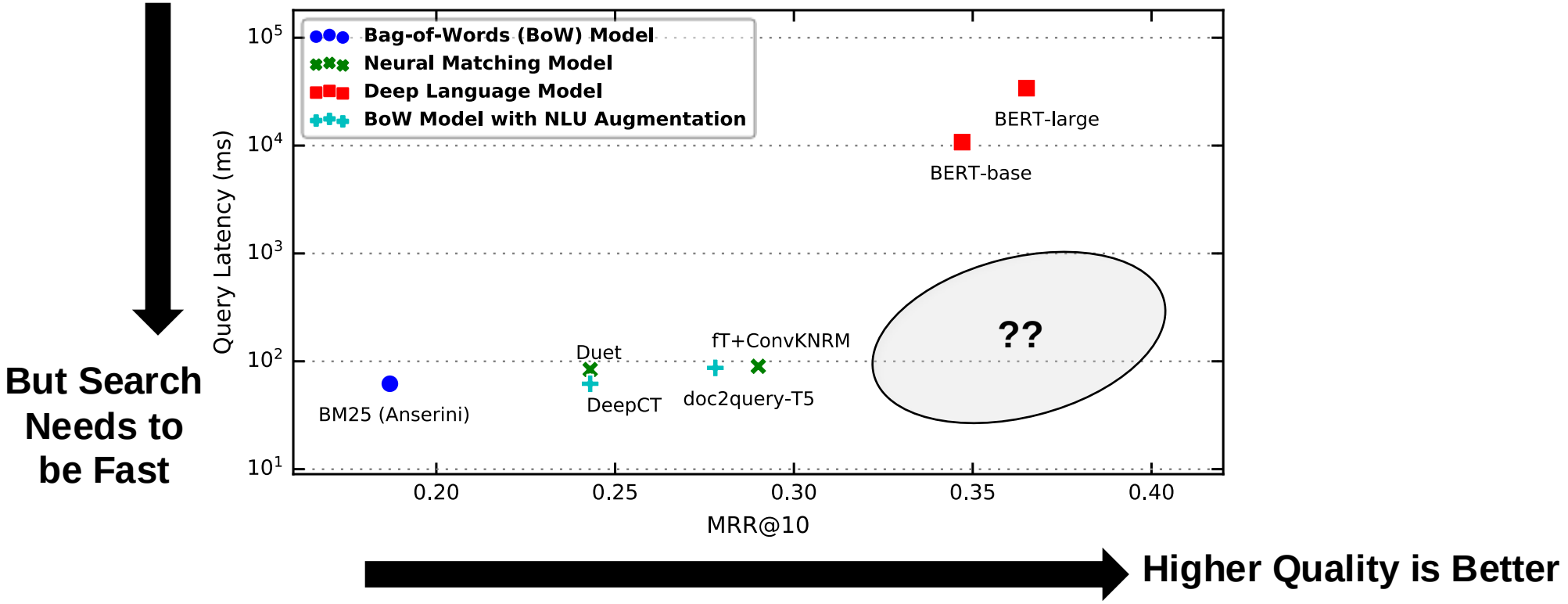
Original version of question and answer from SQuAD. All passages are short extractions from Wikipedia. Edited.
Pranav Rajpurkar, et al. 2016. SQuAD: 100,000+ questions for machine comprehension of text. EMNLP'16.

Retrievers must balance **quality** & **efficiency**

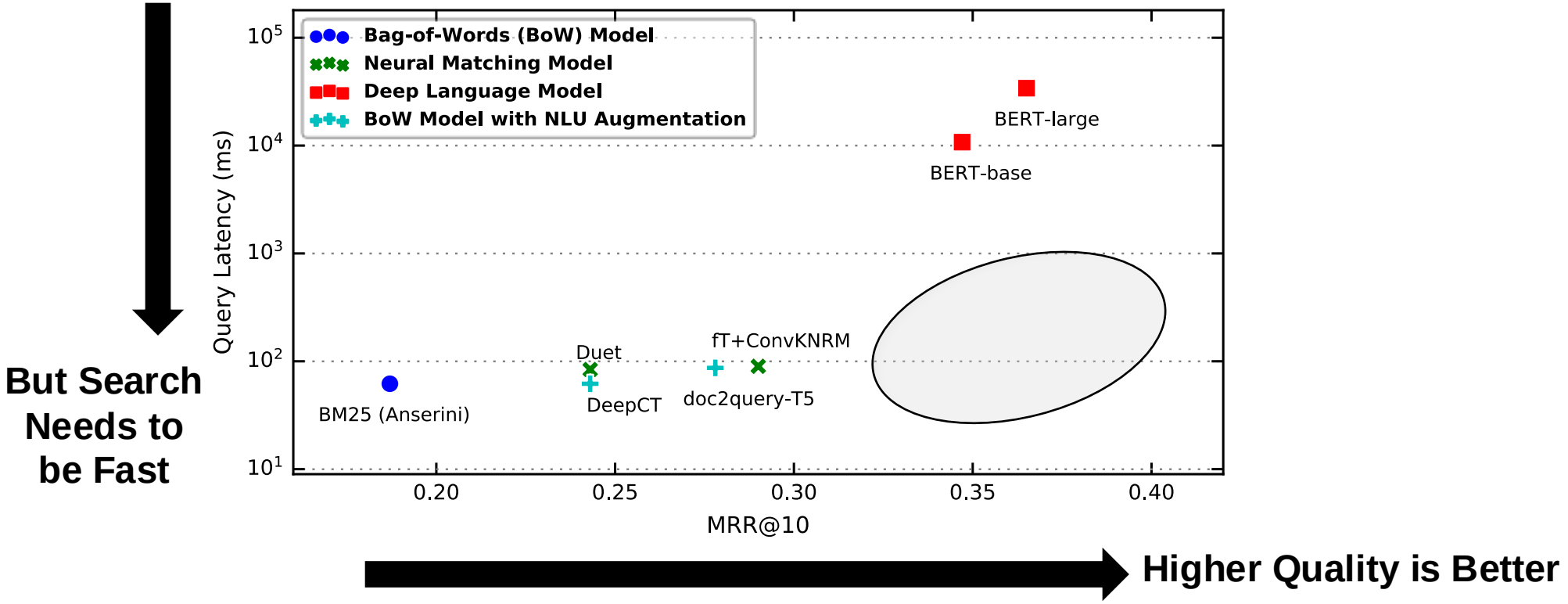


Answer **challenging queries** vs. Search over **millions** of documents in **milliseconds**!

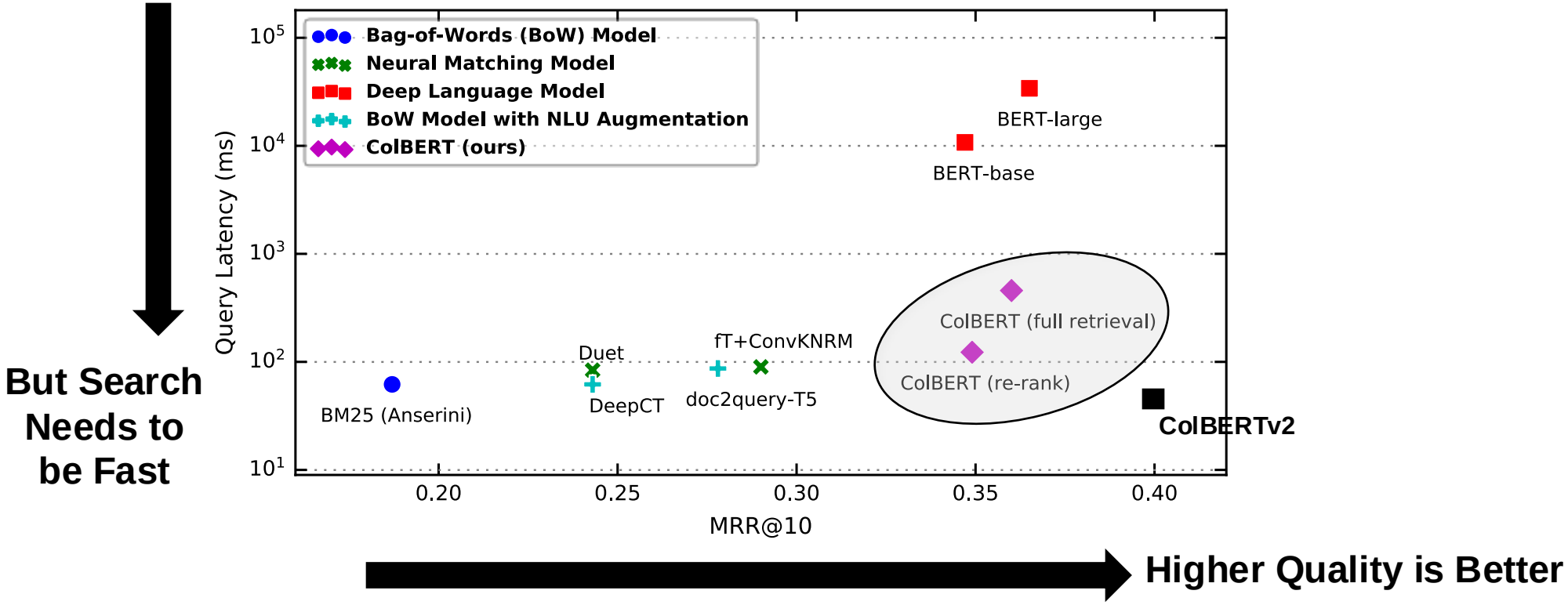
Retrievers must balance **quality** & **efficiency**



Retrievers must balance **quality** & **efficiency**

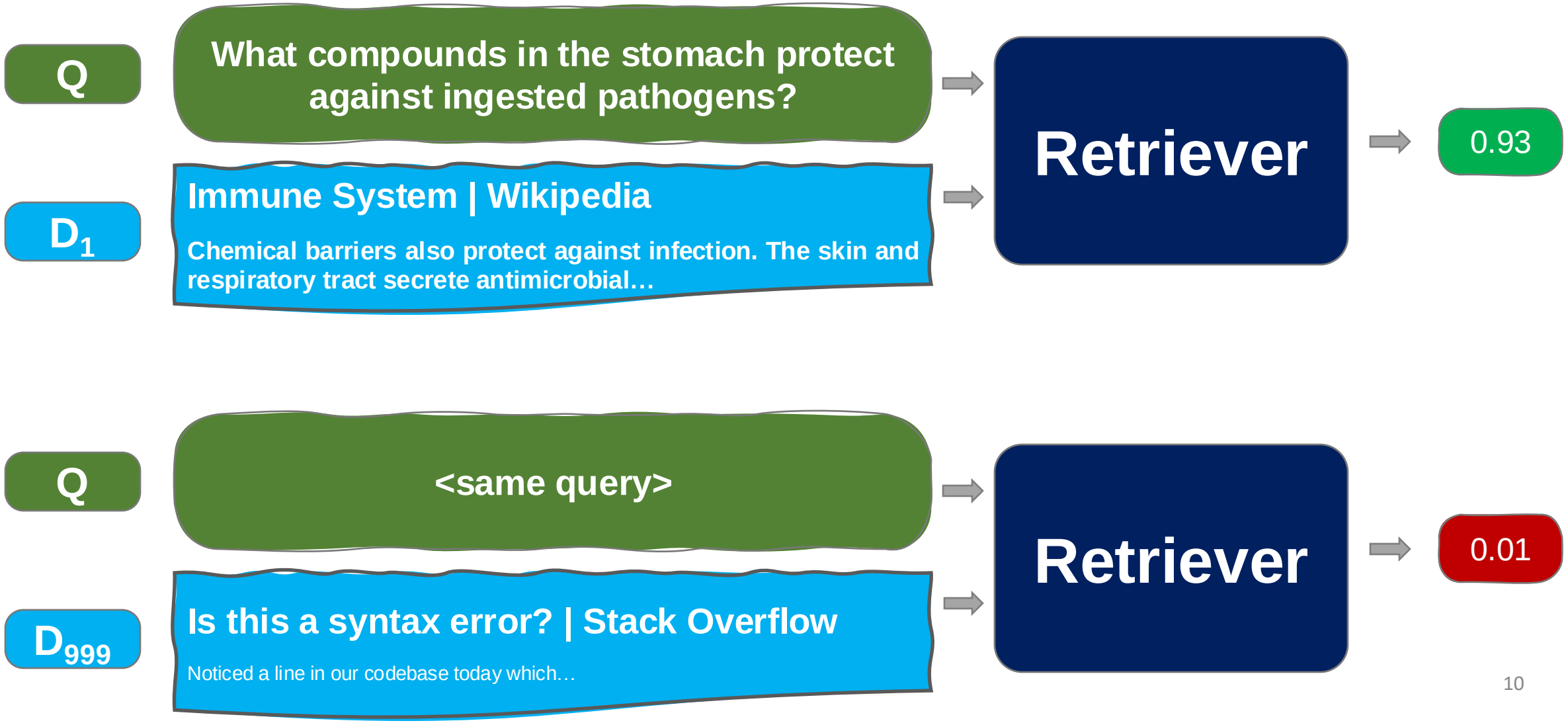


Retrievers must balance **quality** & **efficiency**



Latency (y axis) is in **log scale**. **ColBERT** can be **orders of magnitude** faster than BERT!

How Retrievers Work at a High Level



Information Retrieval Data Set

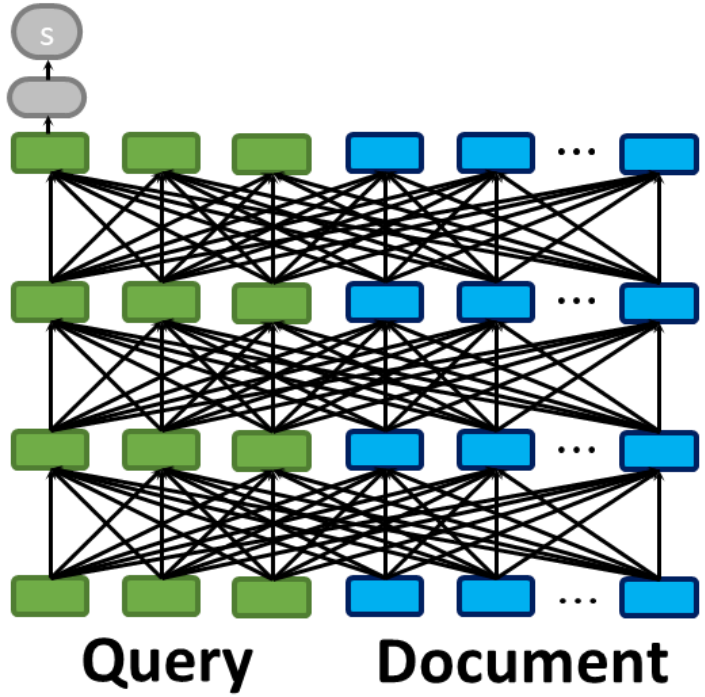
MS MARCO: A Human Generated MAchine Reading COmprehension Dataset

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang
Microsoft AI & Research

NIPS, 2016

- In 2018, it is turned into an information retrieval dataset
- Bing's 1M real-world queries
- 8.8M passages from Web pages
- Effectiveness metric: MRR-10
  - Finding the right answer within the 10 documents returned

Neural IR: Two Extreme Matching Paradigms



(a) Cross Encoder

✓ **Fine-Grained Interactions**

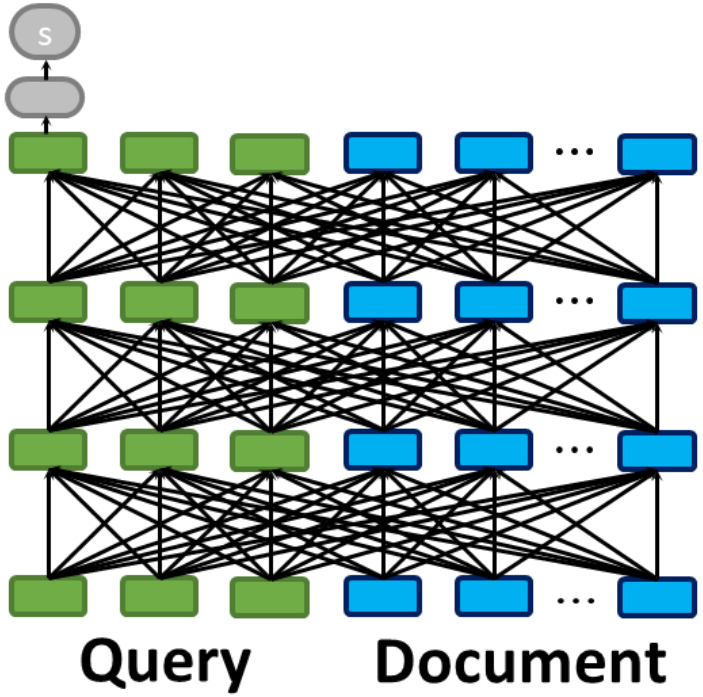
✗ **Unscalable Joint Conditioning**

Scale is a major challenge.

You might have **100 million** documents.

Even if scoring **each document** took **10 ms**, retrieval would consume **11 days** per query!

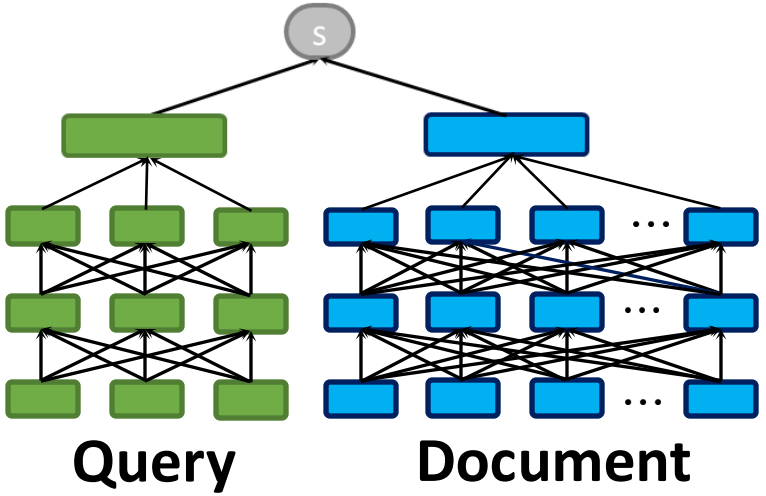
Neural IR: Two Extreme Matching Paradigms



(a) Cross Encoder

✓ Fine-Grained Interactions

✗ Unscalable Joint Conditioning

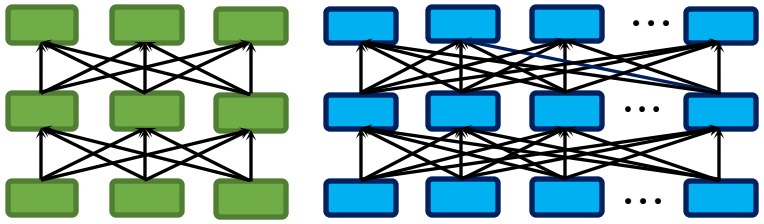


(b) Bi-Coder

✓ Independent, Dense Encoding

✗ Coarse-Grained Representation

ColBERT: **Late Interaction**

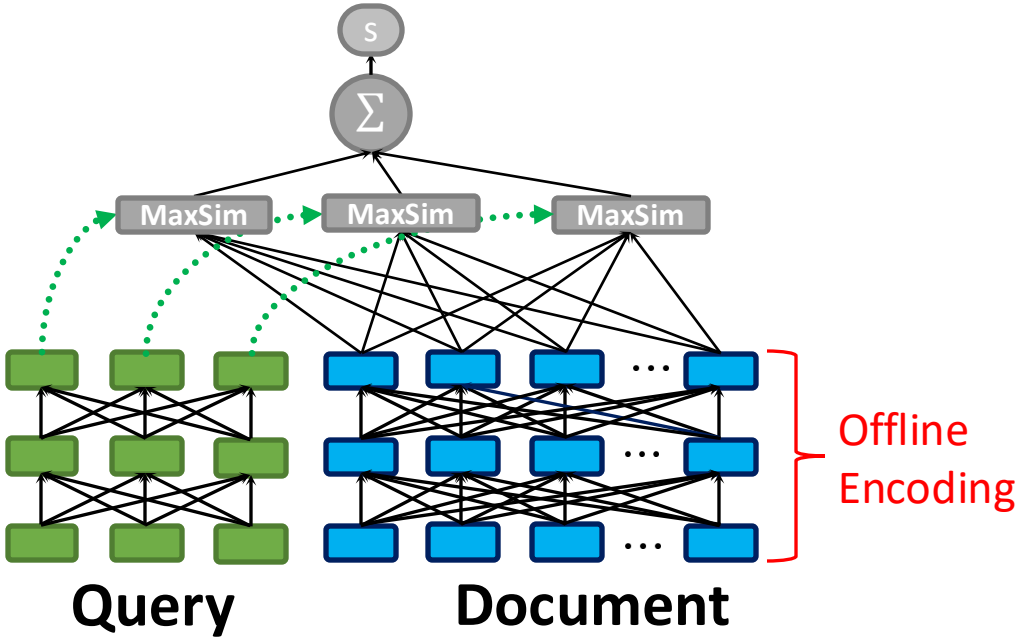


(c) Late Interaction

- ✓ Independent Encoding
- ✓ Fine-Grained Representations
- ✓ Scalable Nearest-Neighbor Search

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR 2020.

ColBERT: **Late Interaction**



(c) Late Interaction

MaxSim: Maximum Similarity
e.g. cosine similarity

- ✓ Independent Encoding
- ✓ Fine-Grained Representations
- ✓ Scalable Nearest-Neighbor Search

End-to-End Retrieval over Wikipedia (21M passages) takes **70ms**.

Omar Khattab and Matei Zaharia. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR 2020.

Late Interaction: Real Example of Matching

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] **on** August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The **Transformers** […] […] It was released […] on August 8, 1986

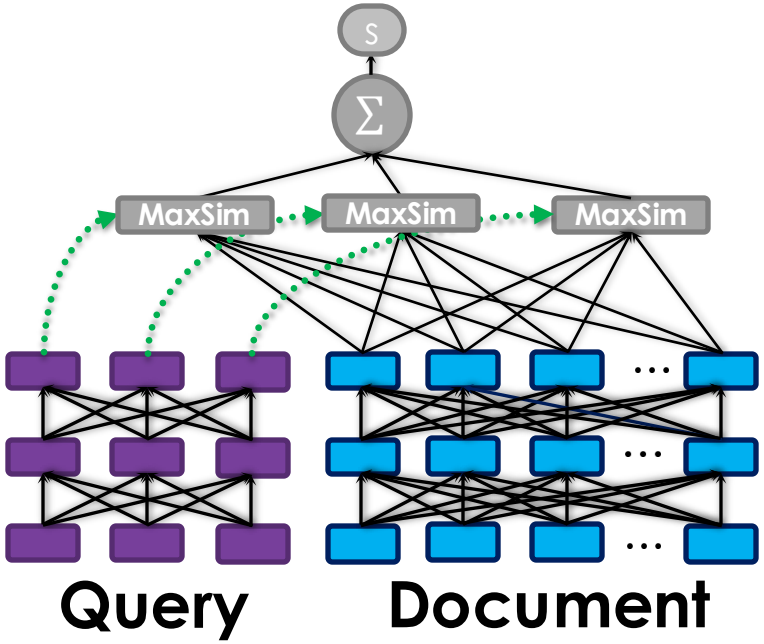
when did the transformers cartoon series come out?

[…] the **animated** […] The Transformers […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was **released** […] on August 8, 1986

Colbert Inference Time Optimization



Documents are all encoded offline

Given a query, find the top k matching docs (from 100M docs)

- Ranking every document is too costly
  1. Narrow down candidates w/o late interaction (**100M** docs to **10K**)
  2. Use late interaction on **10K** docs

- Space: The embeddings are too large
  - Compress the storage (Colbertv2)

POSTED ON MARCH 29, 2017 TO AI RESEARCH, DATA INFRASTRUCTURE, ML APPLICATIONS

Faiss: A library for efficient similarity search

- Faiss: Facebook AI Similarity Search
- Returns k-nearest neighbors constructed on billions of high-dimensional vectors (embeddings).

Given a set of vectors x_i in dimension d, Faiss builds a data structure in RAM from it. After the structure is constructed, when given a new vector x in dimension d it performs efficiently the operation:

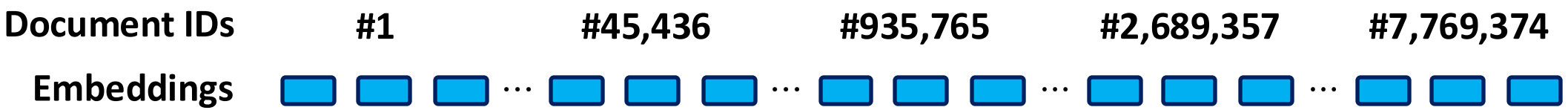
$$j = argmin_i \|x - x_i\|$$

where $\|\cdot\|$ is the Euclidean distance (L^2).

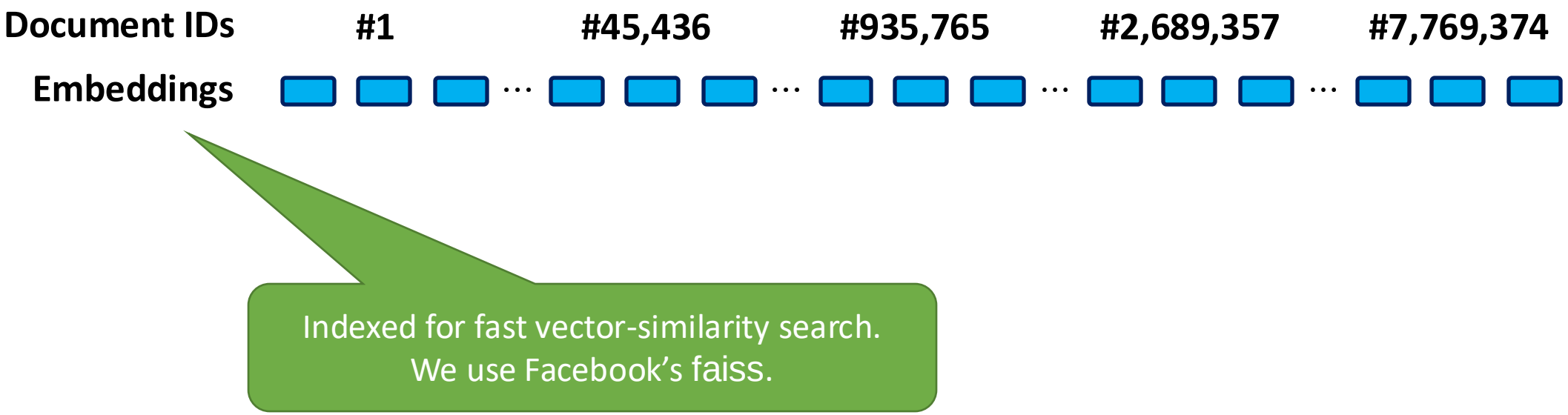
ColBERT: Step 1 (No Late Interaction)

Document IDs	**#1**	**#45,436**	**#935,765**	**#2,689,357**	**#7,769,374**

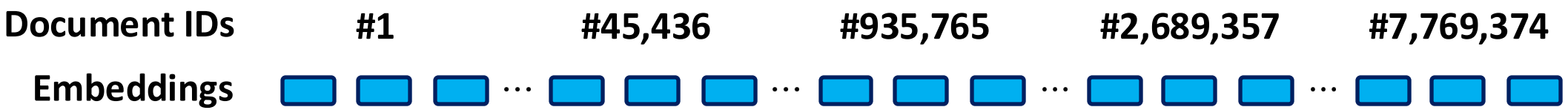
ColBERT: Step 1 (No Late Interaction)



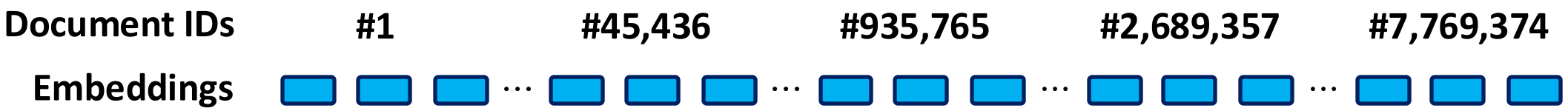
ColBERT: Step 1 (No Late Interaction)

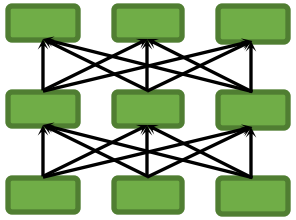


ColBERT: Step 1 (No Late Interaction)

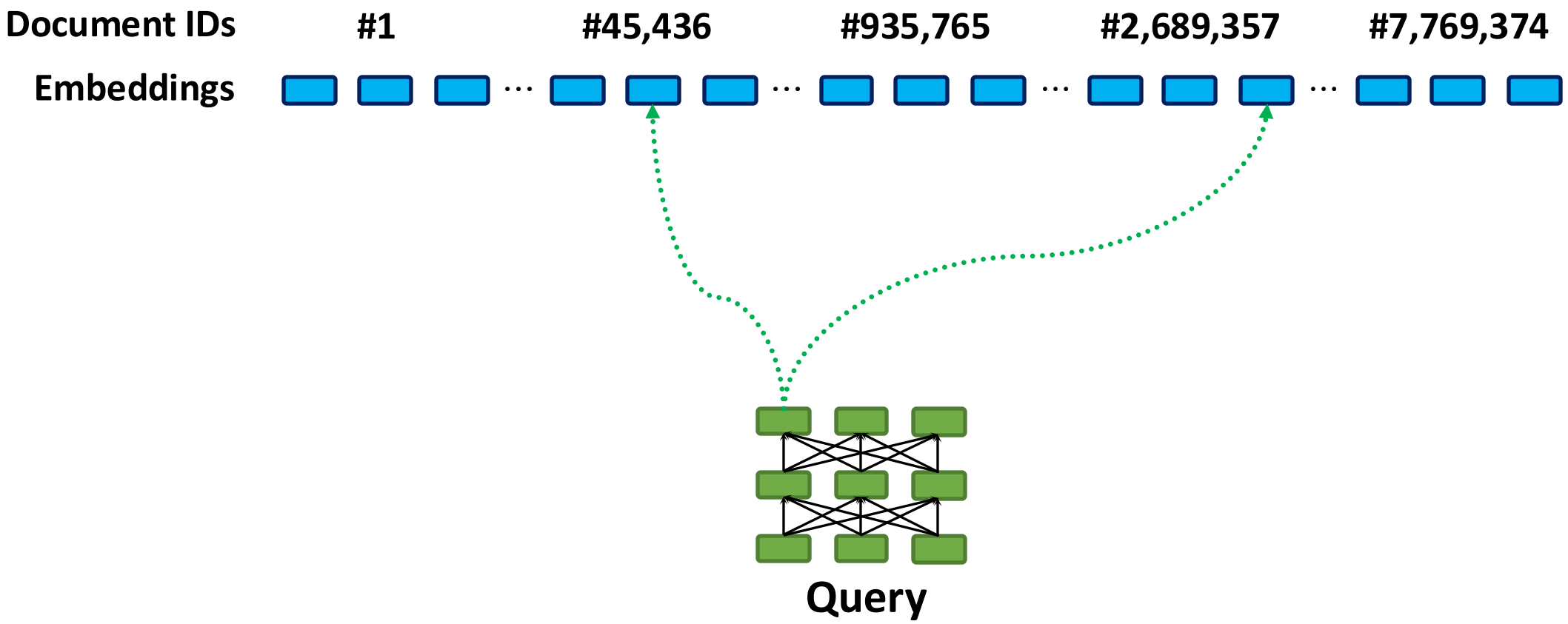


ColBERT: Step 1 (No Late Interaction)

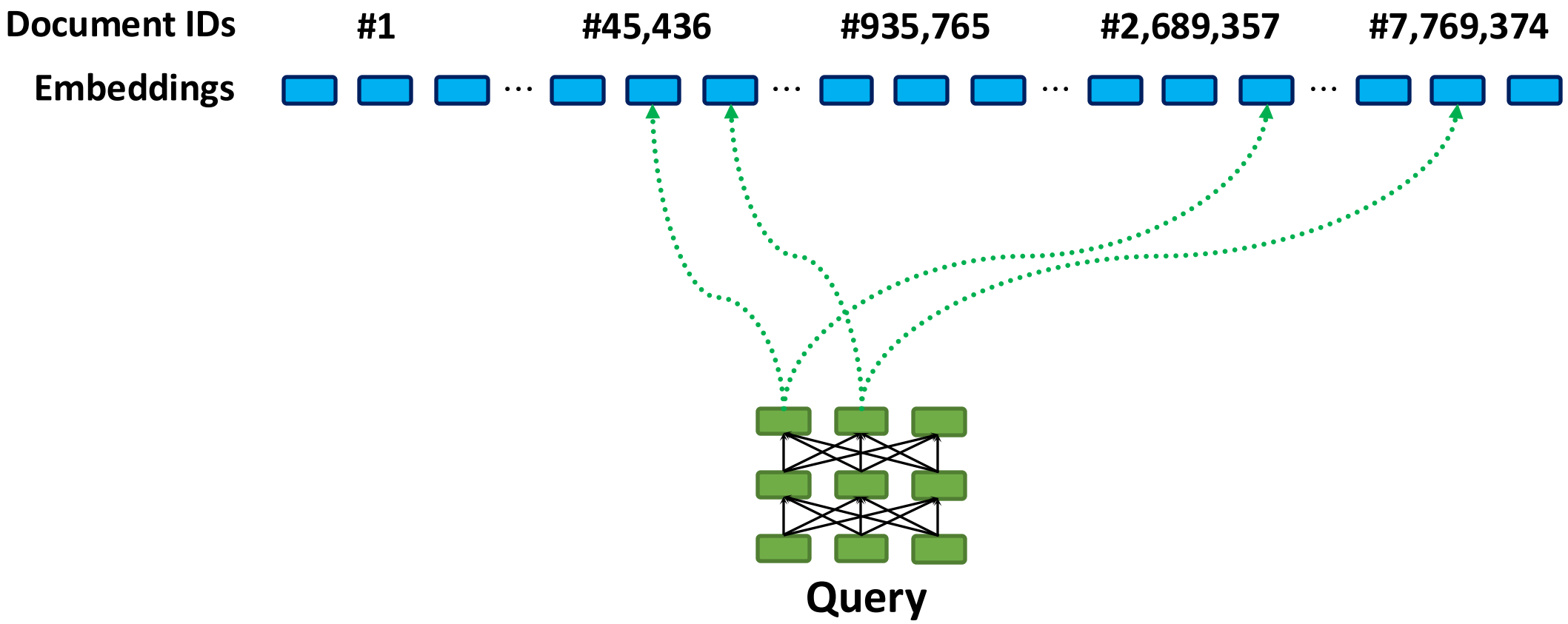




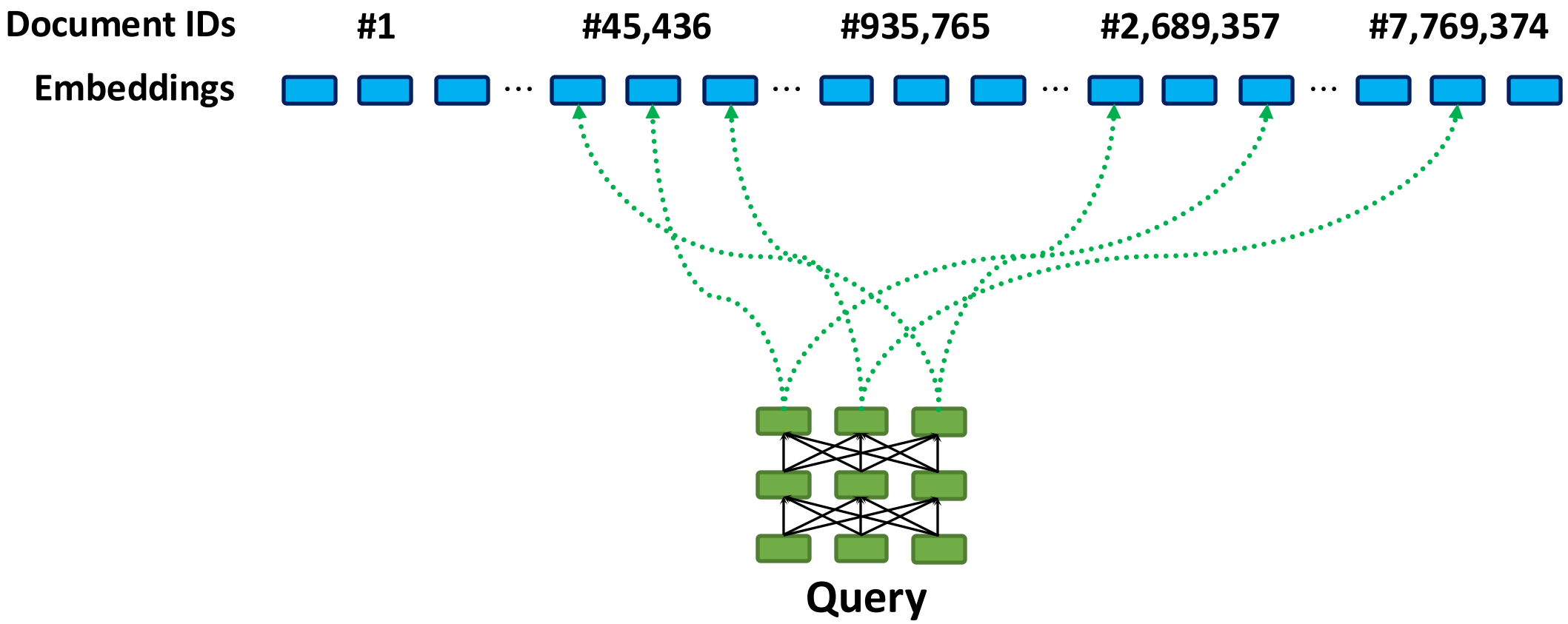
ColBERT: Step 1 (No Late Interaction)



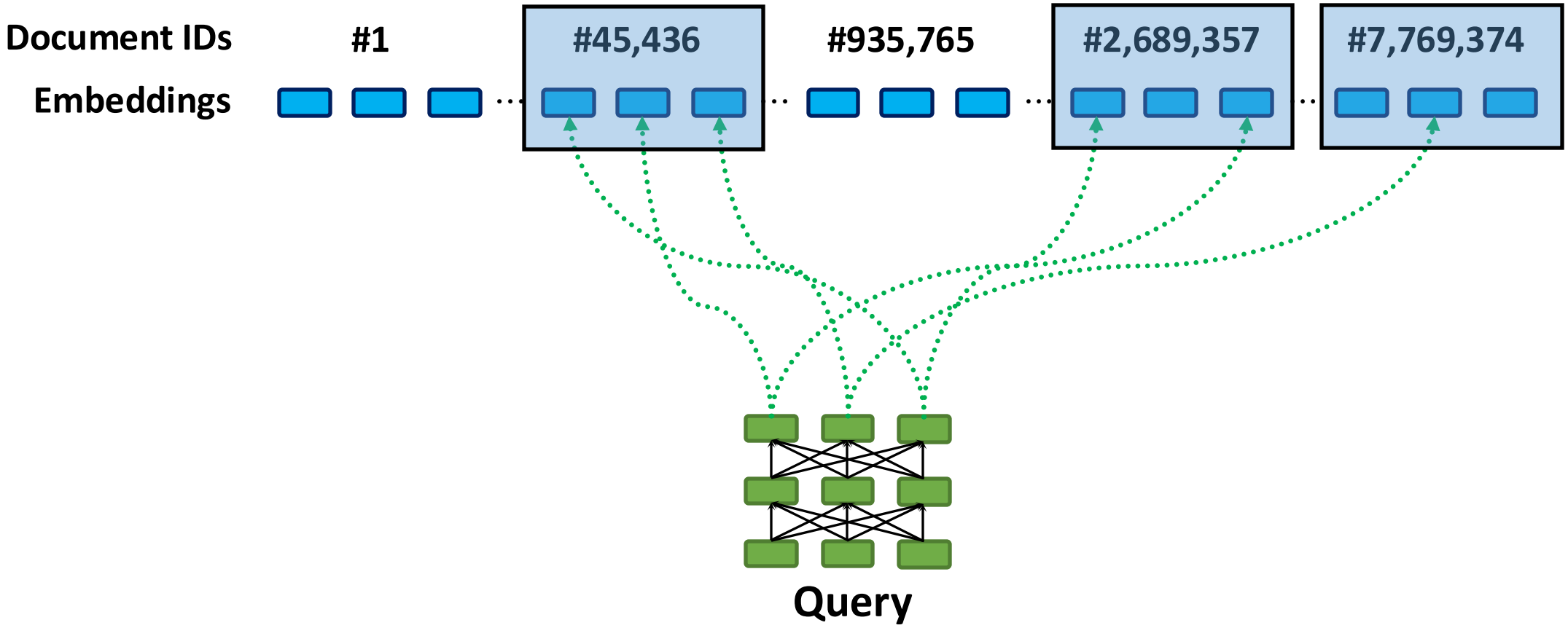
ColBERT: Step 1 (No Late Interaction)



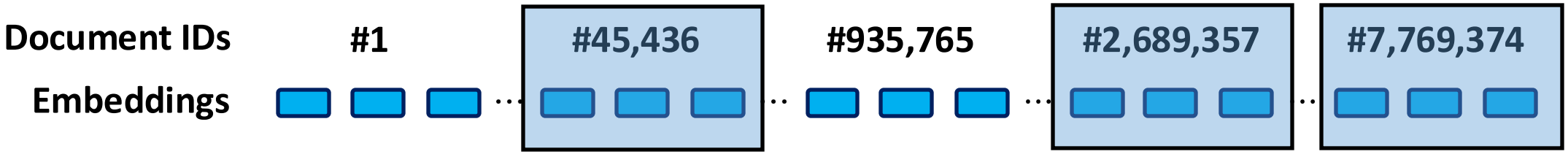
ColBERT: Step 1 (No Late Interaction)



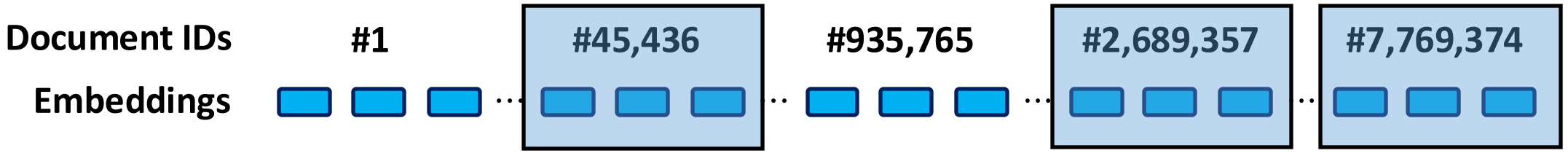
ColBERT: Step 1 (No Late Interaction)



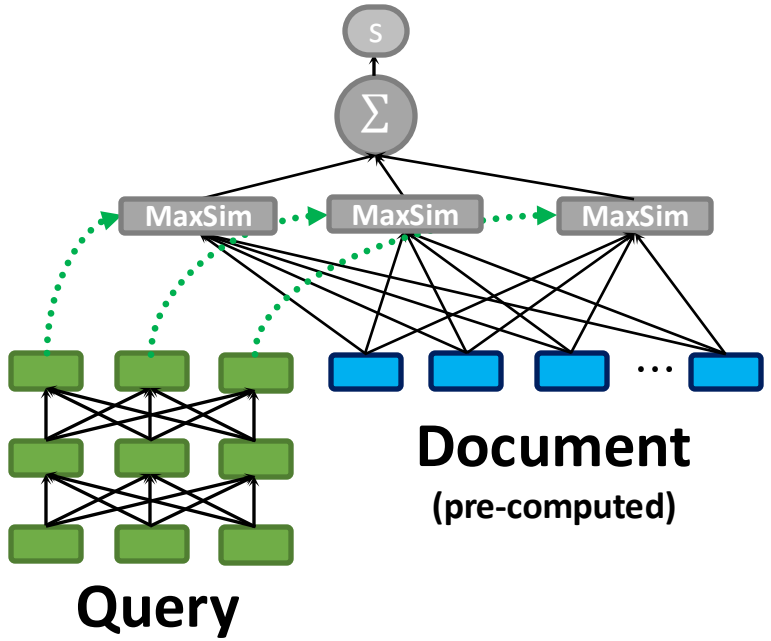
ColBERT: Step 1 (No Late Interaction)

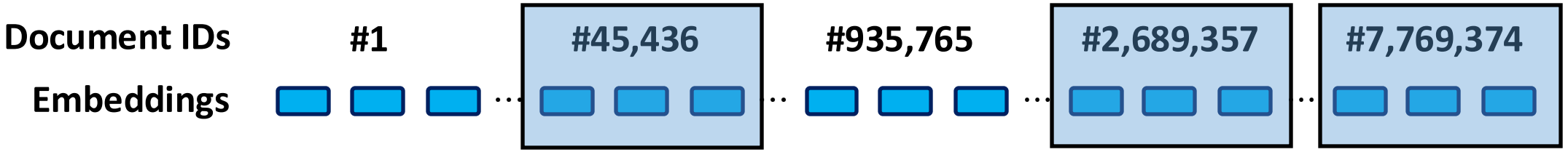


ColBERT: Step 2 (Late Interaction)

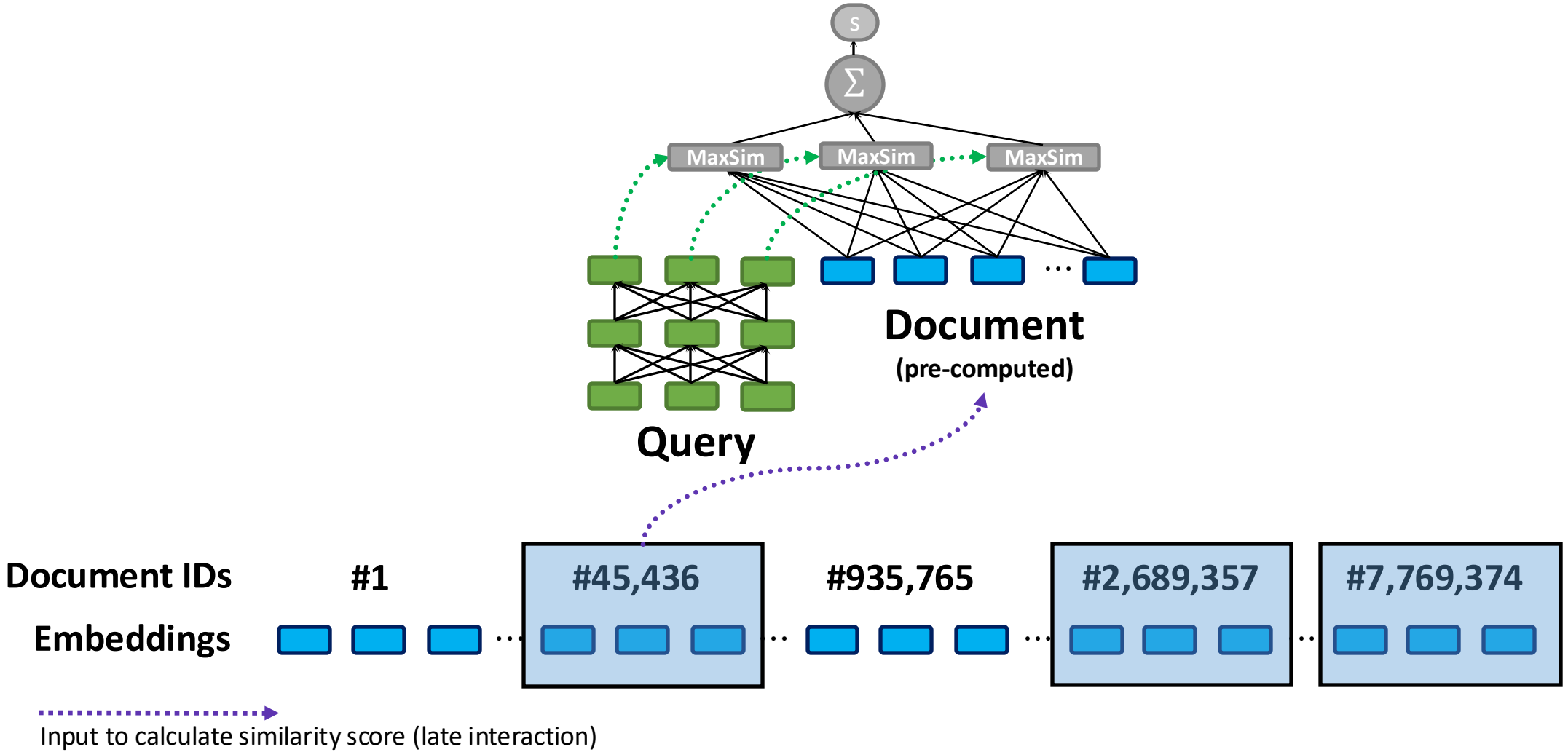


ColBERT: Step 2 (Late Interaction)

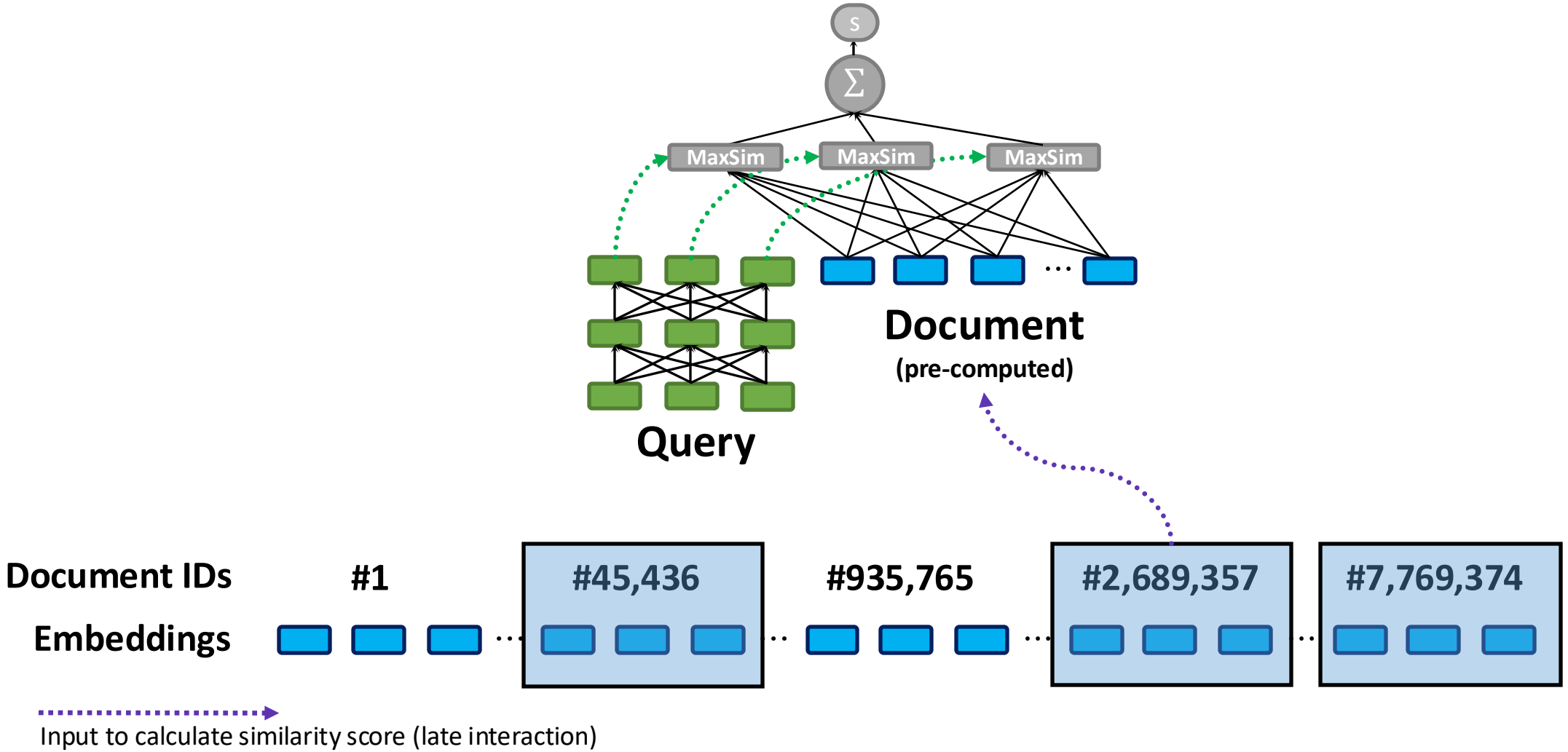




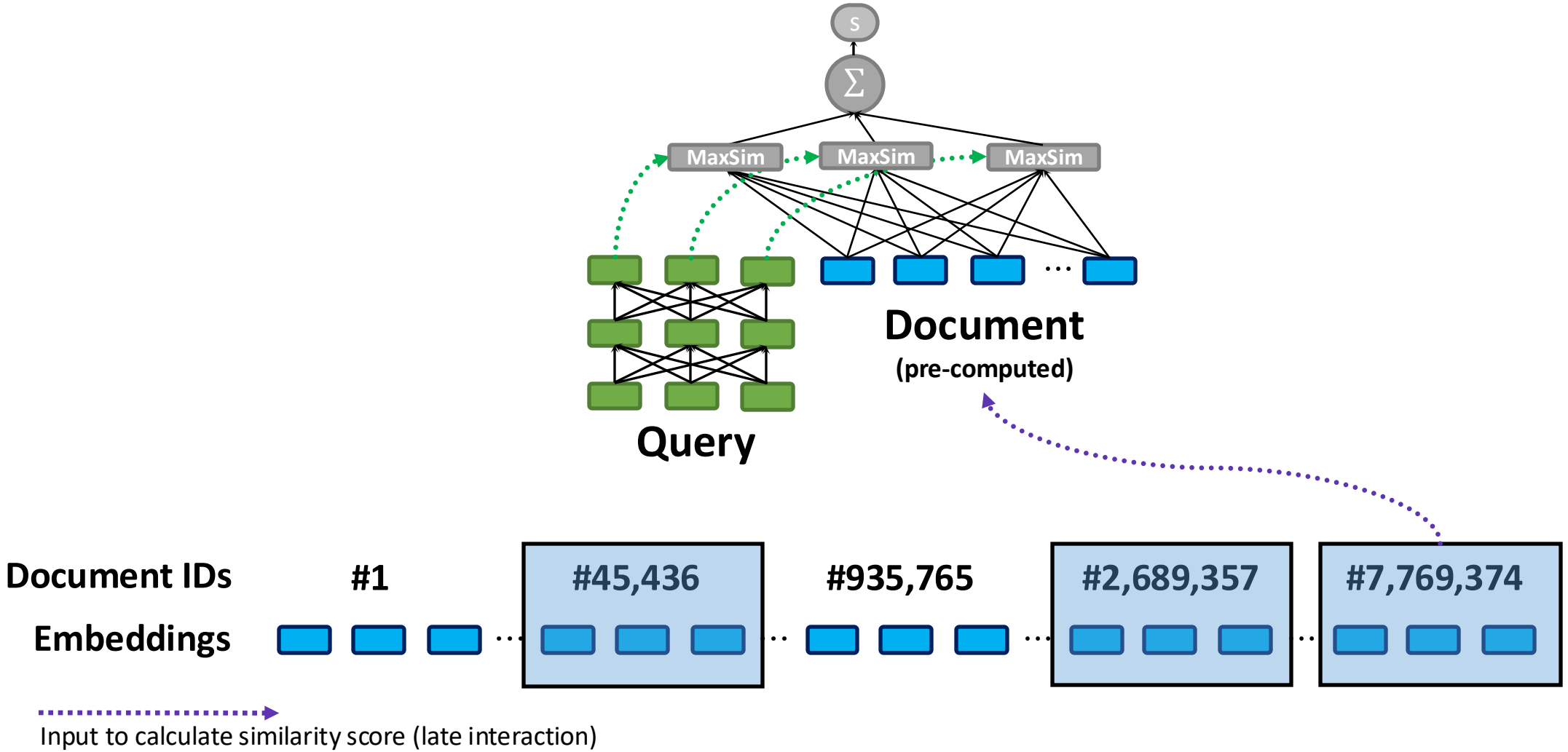
ColBERT: Step 2 (Late Interaction)



ColBERT: Step 2 (Late Interaction)



ColBERT: Step 2 (Late Interaction)



ColBERT: Step 2 (Late Interaction)

Sort Similarity Scores to Find The Top K

Late interaction delivers large gains…

Late interaction delivers large gains…

	Passage Ranking
Model	**MS MARCO** MRR@10
BM25	18.7
DPR	31.1
ANCE	33.0
ColBERT[QA]	**36.0 / 37.5**

DPR and ANCE result figures from the respective papers. DPR on MS MARCO was evaluated by the ANCE authors.

Late interaction delivers large gains...

	Passage Ranking	Open-Domain QA Retrieval over Wikipedia'18		
Model	**MS MARCO** MRR@10	**NaturalQs** Success@20	**TriviaQA** Success@20	**Open-SQuAD** Success@20
BM25	18.7	64.0	77.3	71.4
DPR	31.1	79.4	79.9	71.5
ANCE	33.0	81.9	80.3	-
ColBERT[QA]	**36.0 / 37.5**	**85.3**	**85.6**	**83.7**

DPR and ANCE result figures from the respective papers. DPR on MS MARCO was evaluated by the ANCE authors.

Late interaction delivers large gains…

	Passage Ranking	Open-Domain QA Retrieval over Wikipedia'18		
Model	**MS MARCO** MRR@10	**NaturalQs** Success@20	**TriviaQA** Success@20	**Open-SQuAD** Success@20
BM25	18.7			

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych

Evaluating Extrapolation Performance of Dense Retrieval

Jingtao Zhan[1], Xiaohui Xie[1], Jiaxin Mao[2], Yiqun Liu[1]*, Min Zhang[1], Shaoping Ma[1]

Toward A Fine-Grained Analysis of Distribution Shifts in MSMARCO

Simon Lupart Stéphane Clinchant

RELIC: Retrieving Evidence for Literary Claims

Katherine Thai Yapei Chang Kalpesh Krishna Mohit Iyyer

And the gaps are often larger when there's a **domain shift** or a **challenging downstream task**!

Relevance-guided Supervision for OpenQA with ColBERT

Omar Khattab Christopher Potts Matei Zaharia

Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval

Omar Khattab Christopher Potts Matei Zaharia

Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval

Wei Zhong, Jheng-Hong Yang, and Jimmy Lin

Learning Cross-Lingual IR from an English Retriever

Yulong Li[†*], Martin Franz[‡*], Md Arafat Sultan[‡*], Bhavani Iyer[‡], Young-Suk Lee[‡] and Avirup Sil[‡]

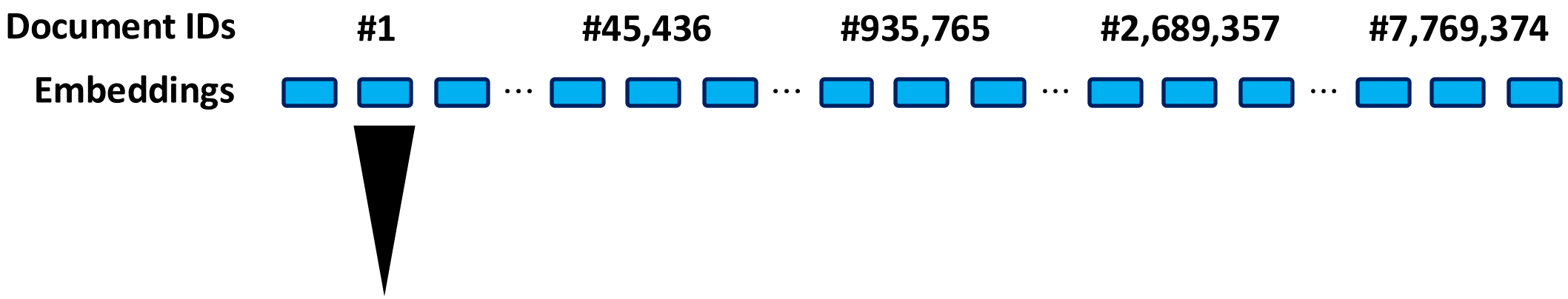
But in the first version of ColBERT, this came at a cost!

Model	Passage Ra… MS MAR… MRR@10	… Success@20	… Success@20	…D Success@20
BM25	18.7	64.0	77.3	71.4
DPR	31.1	79.4	79.9	71.5
ANCE	33.0	81.9	81.9	-
ColBERT[QA]	34.9	83.9	83.1	**83.7**

However, ColBERT's index is an order of magnitude larger than baselines, at **650 GB** for Wikipedia!

Can we advance ColBERT's large quality advantage and **reduce its footprint by an order of magnitude**?

ColBERTv2: **Can we reduce the storage requirements?**



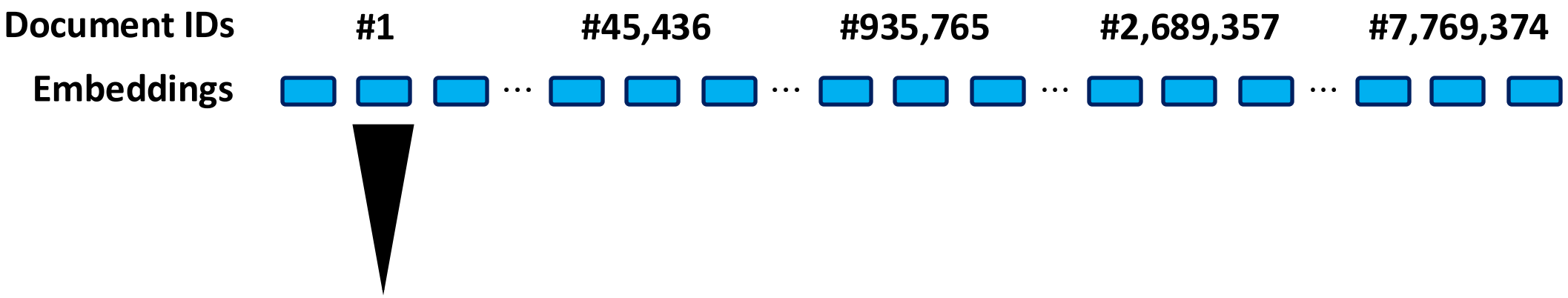
In **ColBERTv1**, each vector needs **128 x float** (4 bytes) **= 512 bytes**

In **ColBERTv2**, each vector consumes just **20 bytes** – How?

Residual Compression

- Given a set of centroids C
- Encode each vector v
  - the index of its closest centroid C_t and a *quantized* vector $\tilde{r}$
  - that approximates the residual $r = v - C_t$.
- At search time, use the centroid index t and residual $\tilde{r}$ recover an approximate $\tilde{v} = C_t + r$

ColBERTv2: **Residual** Compression



In **ColBERTv1**, each vector needs
128 x float (4 bytes) **= 512 bytes**

In **ColBERTv2**, each vector encodes
cluster ID (4 bytes)
+ **128 x bit** (16 bytes)
= **20 bytes** only

ColBERTv2: **Residual** Compression

Indexing clusters a sample of token vectors.

Represent each vector as a **cluster ID** and **a 1-bit delta per dimension**.
This can consume as little **20 bytes** per vector.

	MS MARCO Passage Ranking		
Model	Storage	MRR@10	Recall@50
ColBERT **v1**	154 GB	**36.2**	**82.1**
+ 2 bit residual compression (6x)	25 GB	**36.2**	**82.3**
+ 1 bit residual compression (10x)	16 GB	35.5	81.6

ColBERT has been deeply **influential** in IR and NLP

The ColBERT line of work has been cited by over 1,000 papers

Best Paper Awards (analyses & extensions of ColBERT)
- A White Box Analysis of ColBERT
- SparseEmbed: Learning Sparse Lexical Representations with Contextual Embeddings for Retrieval

Advanced ColBERT-based architectures
- ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval
- ED2LM: Encoder-Decoder to Language Model for Faster Document Re-ranking Inference
- Effective Contrastive Weighting for Dense Query Expansion
- AligneR from Google
- LAIT, LUMEN, GLIMMER from Google

Optimizations for ColBERT
- XTR from Google
- A Study on Token Pruning for ColBERT
- On Approximate Nearest Neighbour Selection for Multi-Stage Dense Retrieval
- Query Embedding Pruning for Dense Retrieval
- Static Pruning for Multi-Representation Dense Retrieval

Extensions
- Distilling Dense Representations for Ranking using Tightly-Coupled Teachers
- Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation
- VIRT: Improving Representation-based Models for Text Matching through Virtual Interaction
- I^3 Retriever: Incorporating Implicit Interaction in Pre-trained Language Models for Passage Retrieval
- SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes
- Reproducibility, Replicability, and Insights into Dense Multi-Representation Retrieval Models: from ColBERT to Col*

Applications
- FILIP: Fine-grained Interactive Language-Image Pre-Training (+ 2-3 other key ones for multi-modal models)
- LI-RAGE: Late Interaction Retrieval Augmented Generation with Explicit Signals for Open-Domain Table Question Answering
- IRLab-Amsterdam at TREC 2021 Conversational Assistant Track
- Soft Prompt Tuning for Augmenting Dense Retrieval with Large Language Models
- Beyond Two-Tower Matching: Learning Sparse Retrievable Cross-Interactions for Recommendation
- Too Few Bug Reports? Exploring Data Augmentation for Improved Changeset-based Bug Localization

Out of Domain Generalization
- BEIR, RELIC, Token-Level Math Information Retrieval, Evaluating Extrapolation Performance in IR (I & II)
- NevIR: Negation in Neural Information Retrieval

Cross Lingual
- IBM's Learning Cross Lingual IR from an English Retriever
- Cross-lingual Knowledge Transfer via Distillation for Multilingual Information Retrieval
- Multilingual ColBERT-X

Our Current IR System

Text Retrieval:

mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval

Xin Zhang[1,2], **Yanzhao Zhang**[1], **Dingkun Long**[1], **Wen Xie**[1], **Ziqi Dai**[1], **Jialong Tang**[1], **Huan Lin**[1] **Baosong Yang**[1], **Pengjun Xie**[1], **Fei Huang**[1], **Meishan Zhang***, **Wenjie Li**[2], **Min Zhang**
[1]Tongyi Lab, Alibaba Group, [2]The Hong Kong Polytechnic University
{linzhang.zx,zhangyanzhao.zyz,dingkun.ldk}@alibaba-inc.com

EMNLP Nov. 2024

Primarily a bi-coder

Vector similarity search:

Qdrant Vector Search

https://qdrant.tech/

Lecture Goals

- **Information Retrieval (on free-text)**
  - Late Interaction Model
  - Optimization: time (Select k candidates + Late interaction)
  - Optimization: space with compression

- **Named Entity Disambiguation**
  - Use entity description: (Select k candidates + Cross-encoder)
  - Use type information

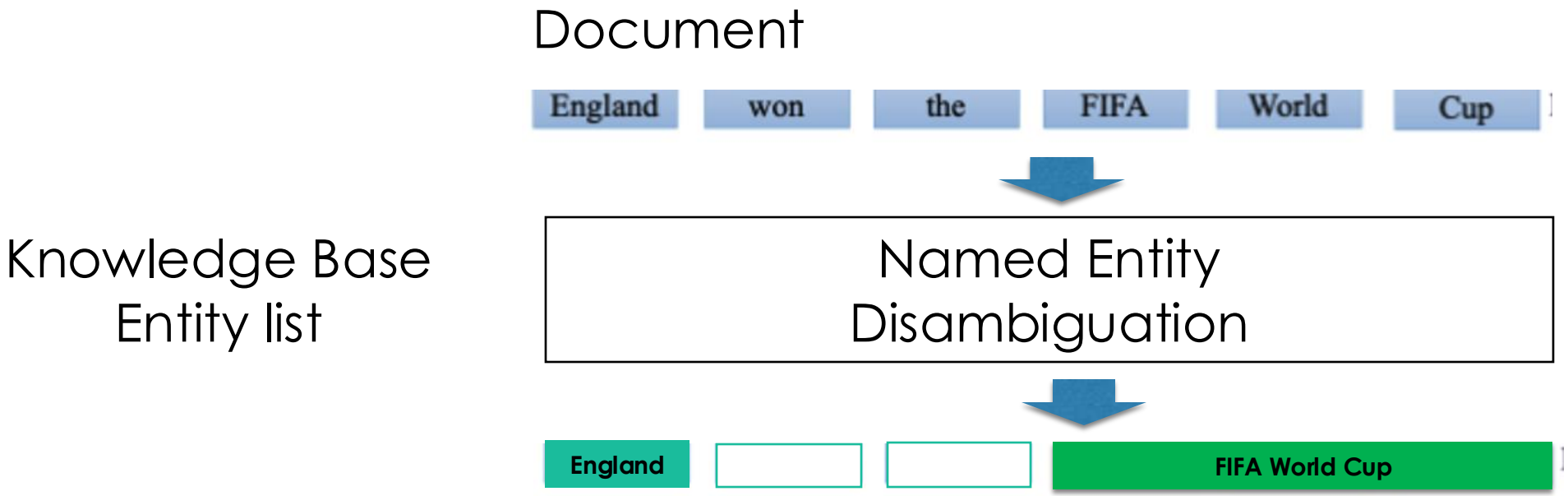
Entity Linking

"Barack Obama, born in Hawaii, served as the 44th President of the United States. He graduated from Harvard Law School and won the Nobel Peace Prize in 2009."

Entity Linking

"**Barack Obama**, born in **Hawaii**, served as the 44th President of the **United States**. He graduated from **Harvard Law School** and won the **Nobel Peace**

Named Entity Disambiguation



Tasks:

- Mention: a span that refers to a named entity
- Entity: the entity in the entity list

Entity Linking Applications

- Question Answering
- Relation Extraction
- Automated Construction of Knowledge Bases

Scalable Zero-shot Entity Linking with Dense Entity Retrieval

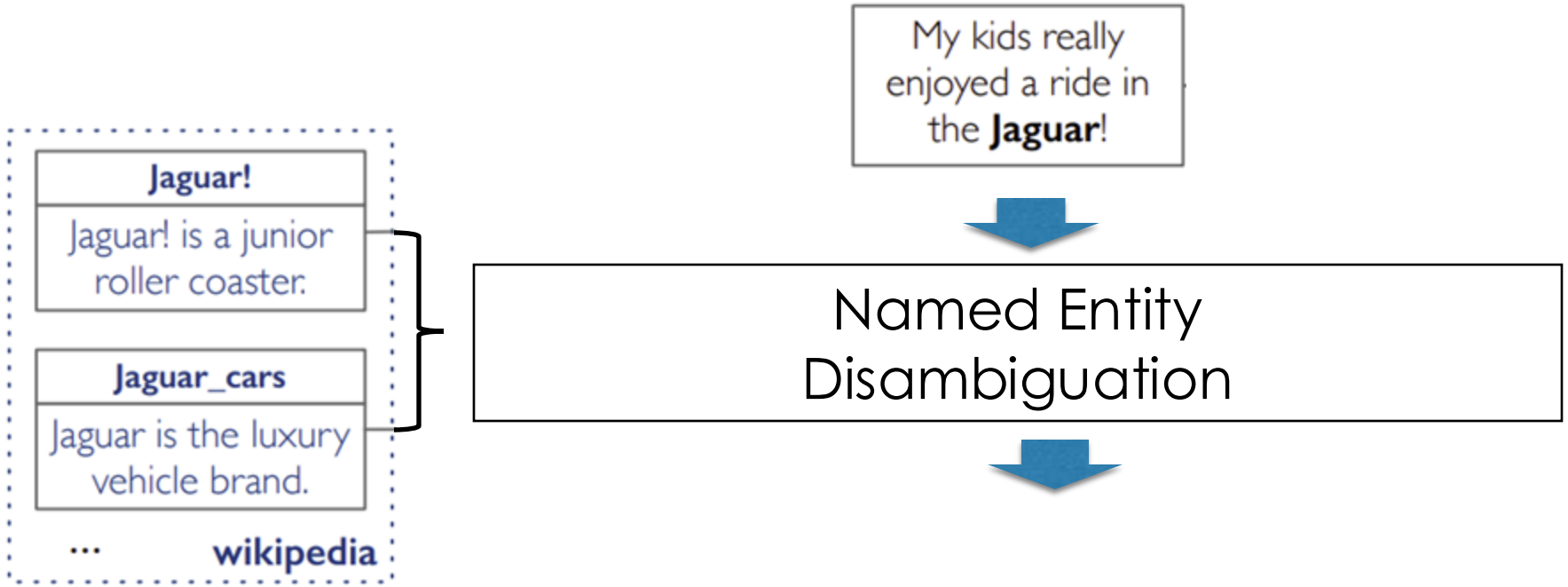
Ledell Wu,[1] **Fabio Petroni,**[1] **Martin Josifoski,**[3*] **Sebastian Riedel,**[1,2] **Luke Zettlemoyer**[1]
[1]Facebook AI Research
{ledell, fabiopetroni, sriedel, lsz}@fb.com
[2]University College London
[3]Ecole Polytechnique Federale de Lausanne
martin.josifoski@epfl.ch

EMNLP 2020

Key Concept

Add Descriptions to the Entities as Input

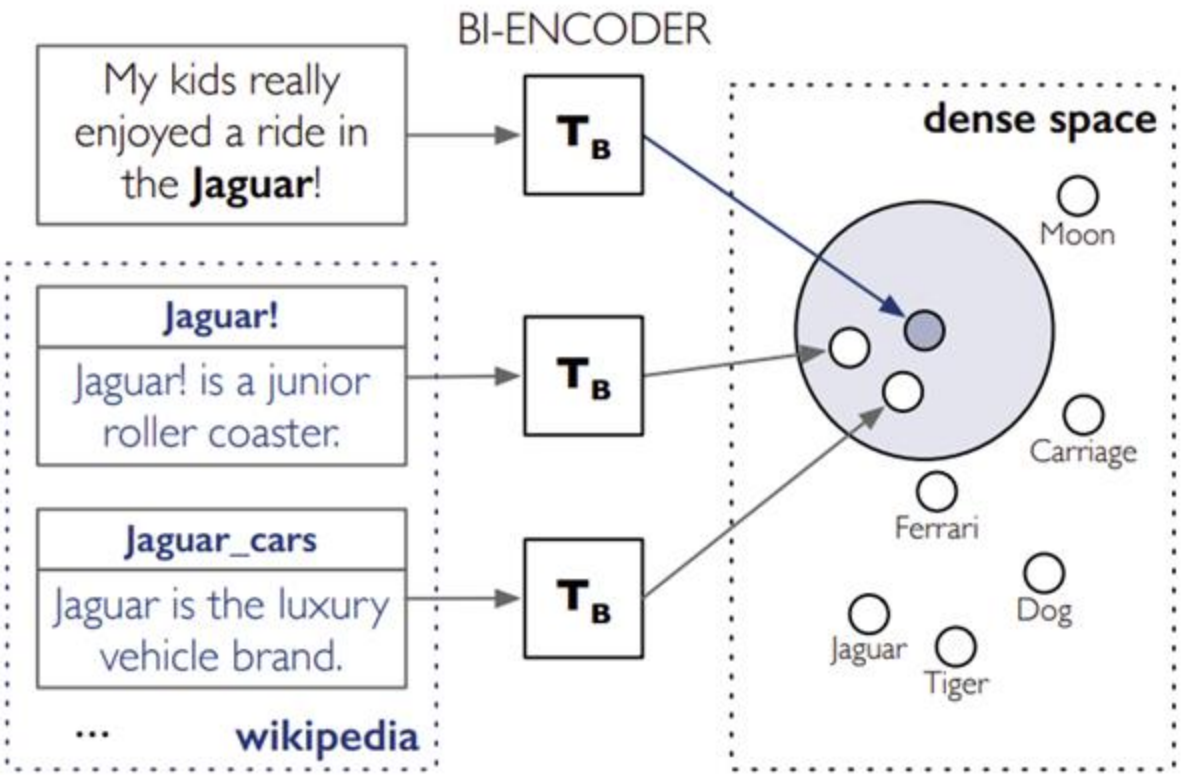
Descriptions Enable Zero-Shot NED



Tasks:

- Mention: a span that refers to a named entity
- Entity: the entity in the entity list

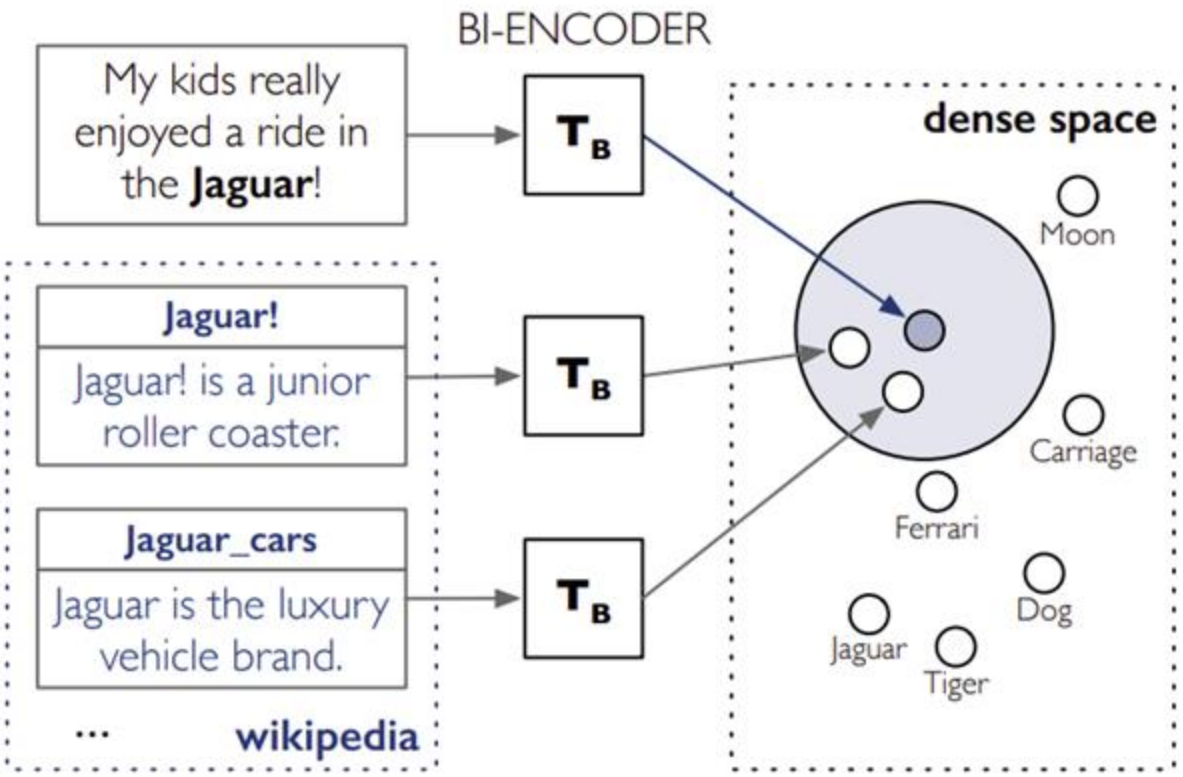
Two Step Approach

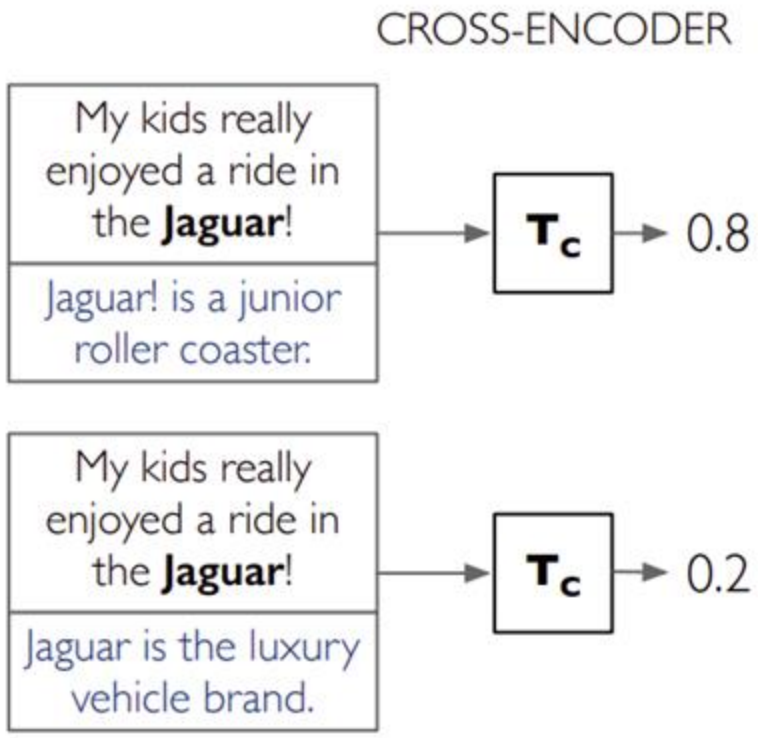


1. Candidate Generation and Ranking (Fast)

- Encode doc, entities to the same dense space
- Retrieve k closest entities

Two Step Approach





1. Candidate Generation and Ranking (Fast)
- Encode doc, entities to the same dense space
- Retrieve k closest entities

2. Rerank k candidates using a cross-encoder
- Slower, More accurate

ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, Andrea Pierleoni
Amazon Alexa AI
Cambridge, UK
{tayoola, tshubhi, fshjos, chrchrs, apierleo}@amazon.com

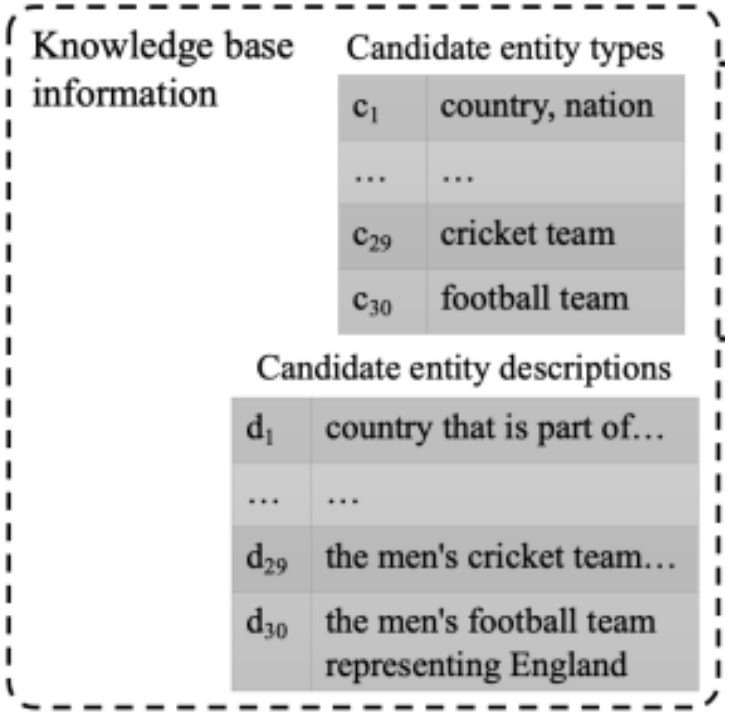
NAACL 2022

Key Concept

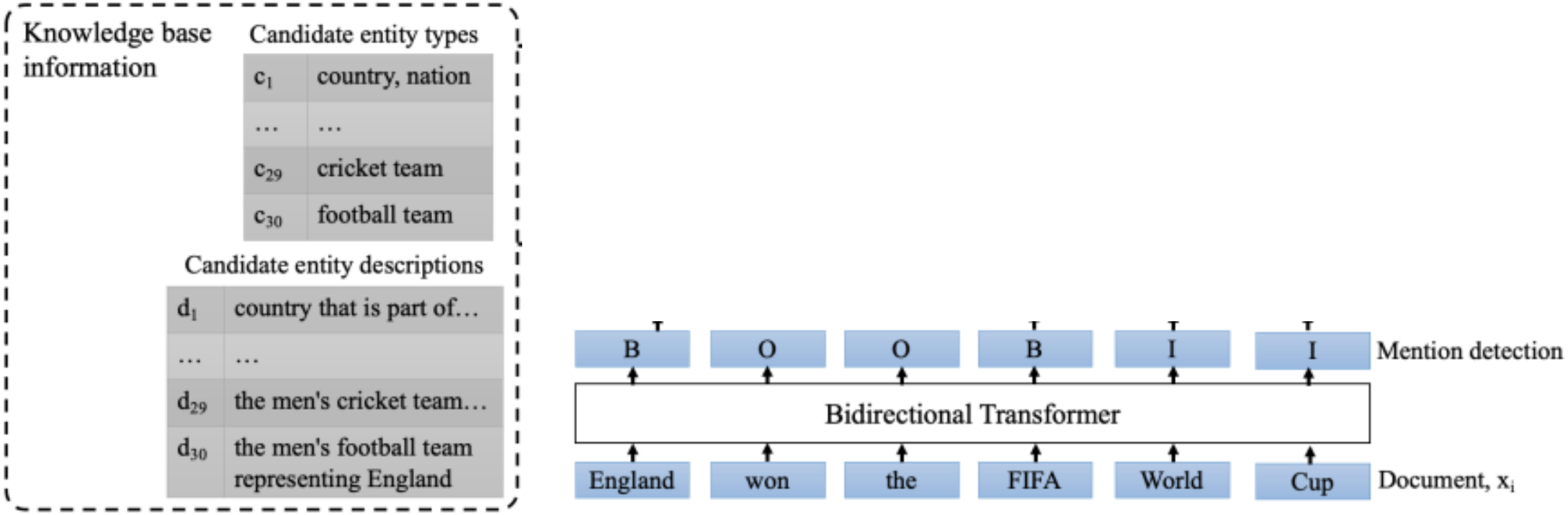
Add Type Information to Entities

ReFinED: Representation and Fine-grained typing for ED

ReFinED: Representation and Fine-grained typing for ED

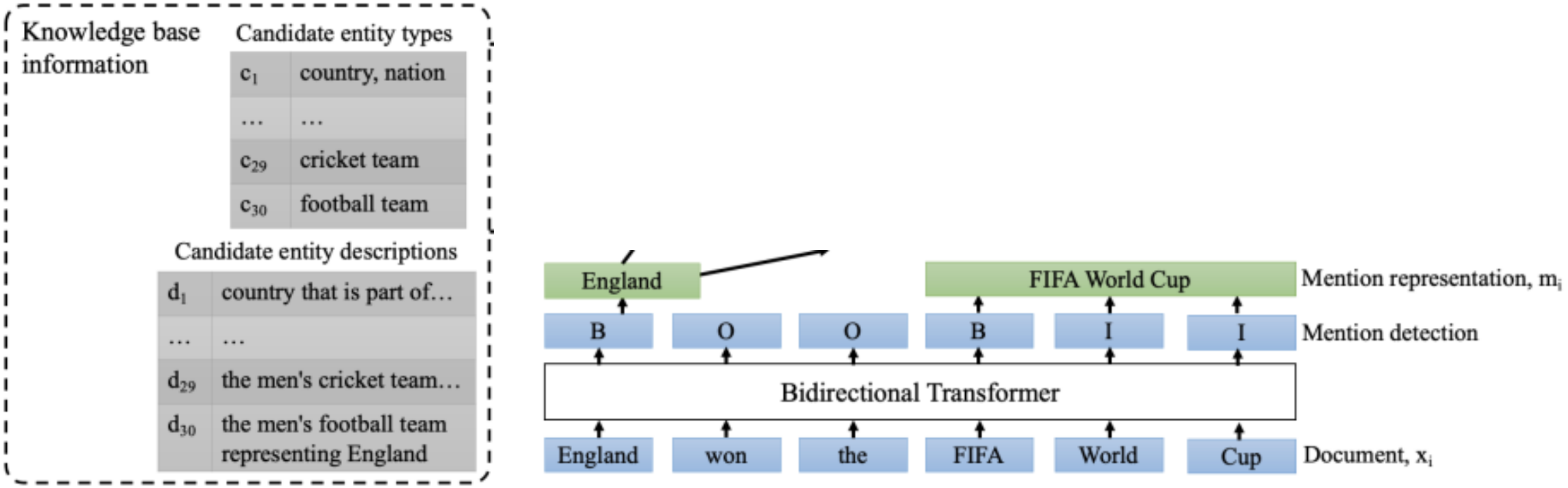


ReFinED: Representation and Fine-grained typing for ED



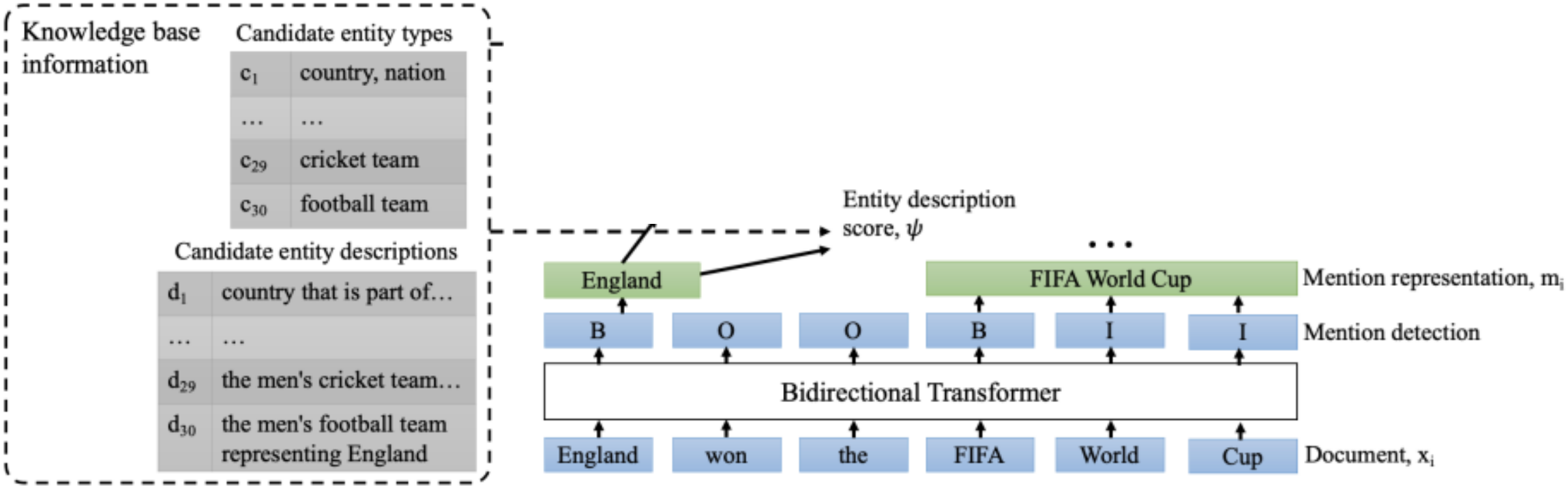
Mention Detection with Begin/Inside/Outside (BIO) Tagging.

ReFinED: Representation and Fine-grained typing for ED



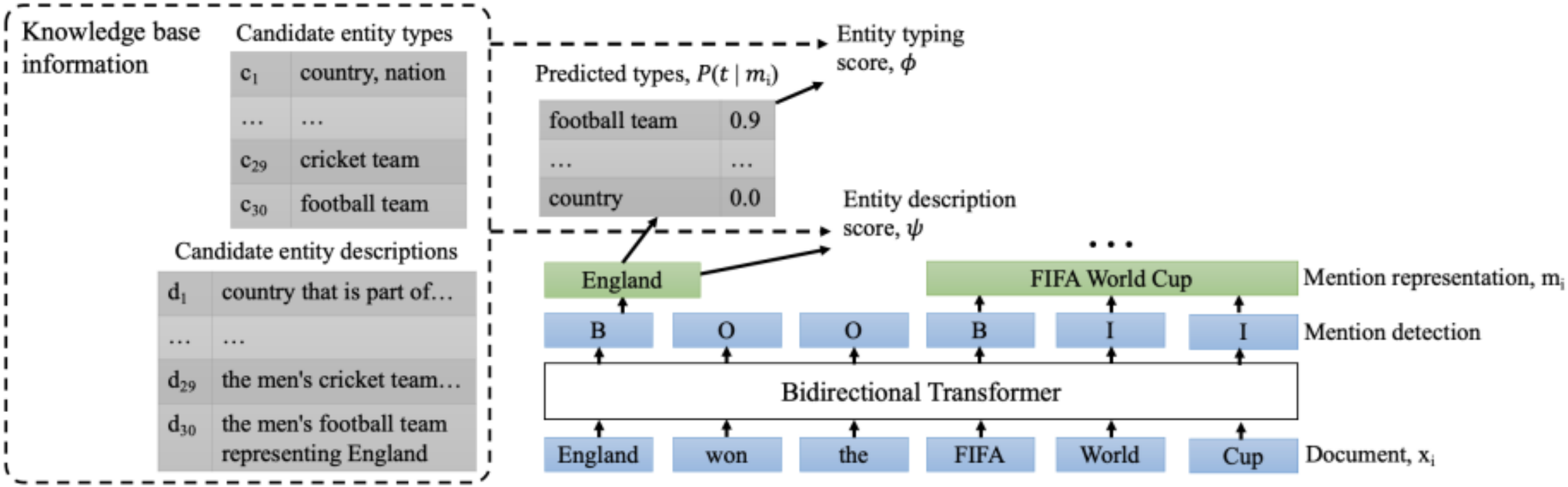
Mention Representation via mean pooling.

ReFinED: Representation and Fine-grained typing for ED



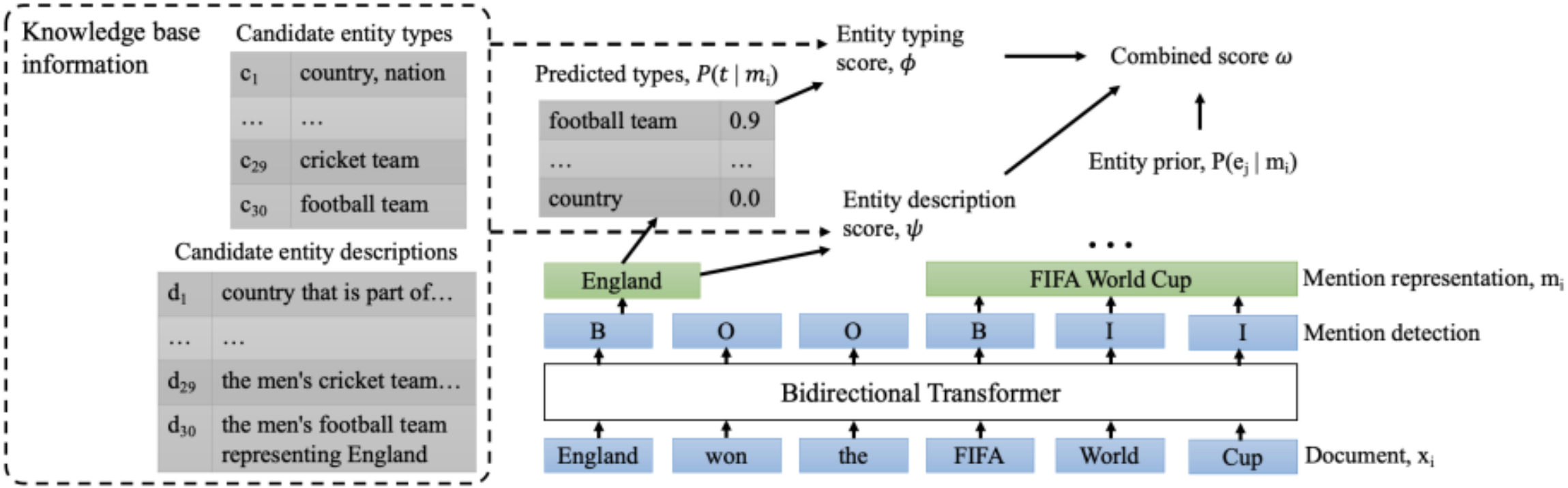
Entity Description score: multiple simultaneous bi-encoder representations

ReFinED: Representation and Fine-grained typing for ED



Entity Typing score: L2 distance between type vectors

ReFinED: Representation and Fine-grained typing for ED



Linear Combination of {Prior, Types, Description} scores

- The global entity prior is obtained from a corpus (Hoffart et al., 2011) or a popularity metric (Diefenbach and Thalhammer, 2018).
- P^(e|m) included to improve results for cases where context is limited (e.g. short question text).

ReFinED: Recap!

- Mention Detection with Begin/Inside/Outside (BIO) Tagging.
- Mention Representation via mean pooling.
- Entity Typing score.
- Entity Description score.
- Linear Combination of {Prior, Types, Description}.

ReFinED deployed by Amazon Alexa at "web scale"

- Populate a KB from a **billion web pages**, *multiple times per year*
- Requires **2 days of processing**, using **500 T4 GPUs**
- Pro: Uniform architecture is **easy to scale**!
- Pro: Generalizes well to **90M entities** at scale

Method	Time taken (s)	Avg. ED F1
Cao et al. (2020)	2100	88.7
Wu et al. (2020) bi-encoder	93	80.4
Wu et al. (2020) cross-encoder	917	87.2
Orr et al. (2021)	438	77.6
ReFinED	**15**	**89.4**

Table 5: Time taken in seconds for EL inference on AIDA-CoNLL test dataset.

Conclusions

- **Information Retrieval (on free-text)**
  - Finding a document that answers a query
    - Text retrieval: ColBERT → mGTE
    - Vector search: FAISS → Qdrant

- **Named Entity Disambiguation: ReFinED**
  - Use entity description and type information