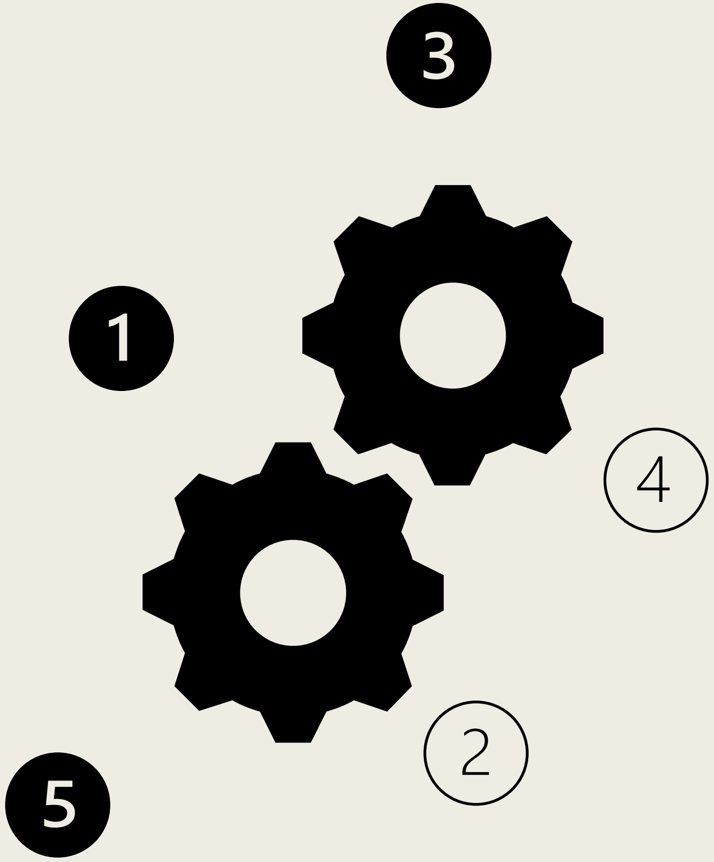


NLU & IR: NEURAL IR (II)

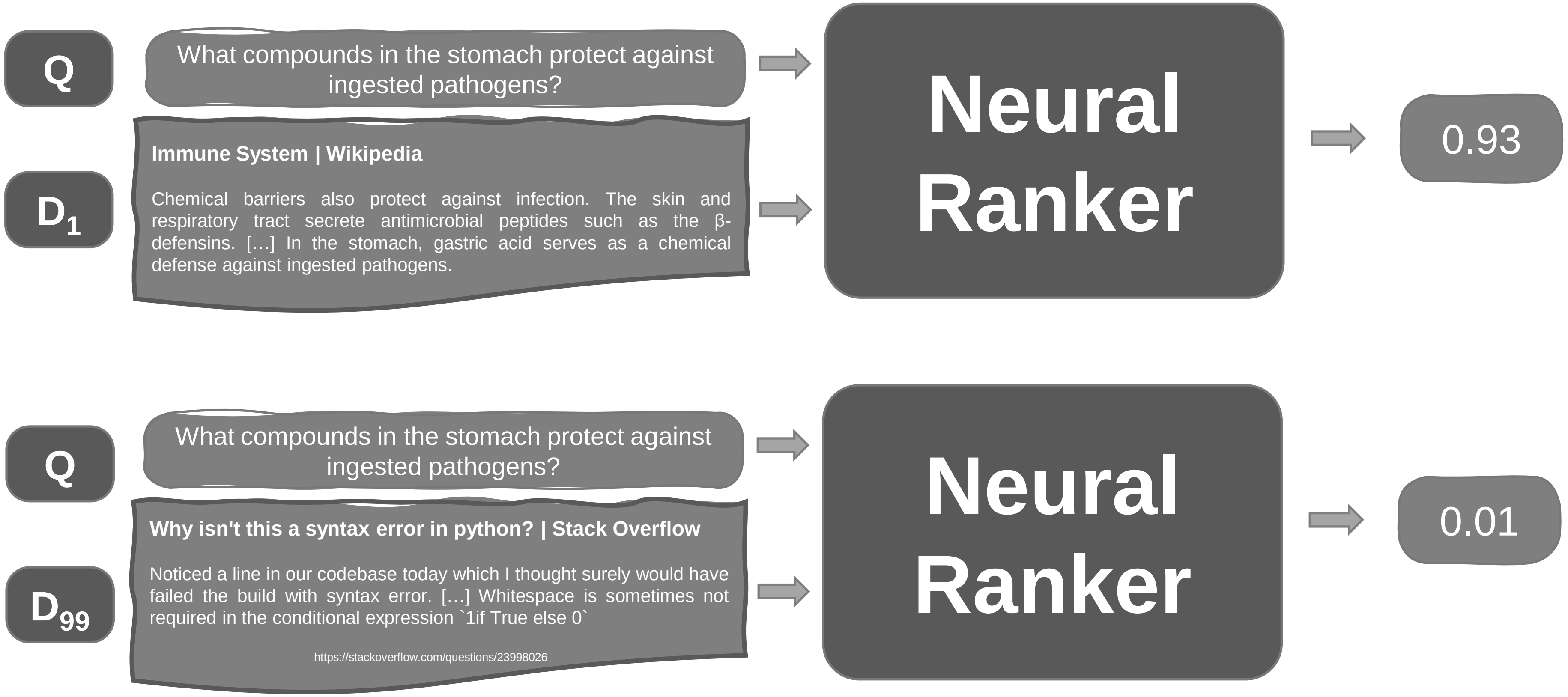
Omar Khattab

CS224U: Natural Language Understanding

Spring 2021

Neural Ranking: Functional View

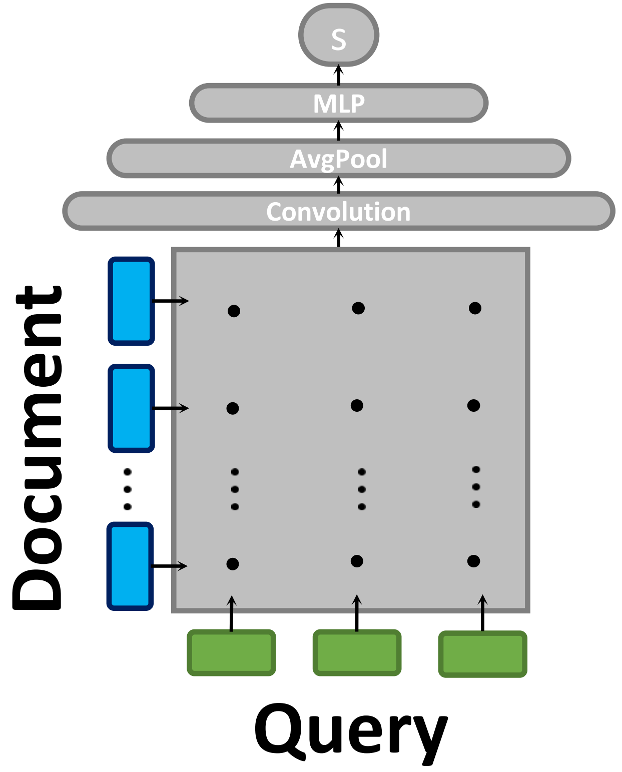
- All we need is a score for every query–document pair
  - We’ll sort the results by decreasing score



Query–Document Interaction Models

1. Tokenize the query and the document
2. Embed all the tokens of each
3. Build a query–document interaction matrix
   - Most commonly: store the cos similarity of each pair of words
4. Reduce this dense matrix to a score
   - Learn neural layers (e.g., convolution, linear layers)

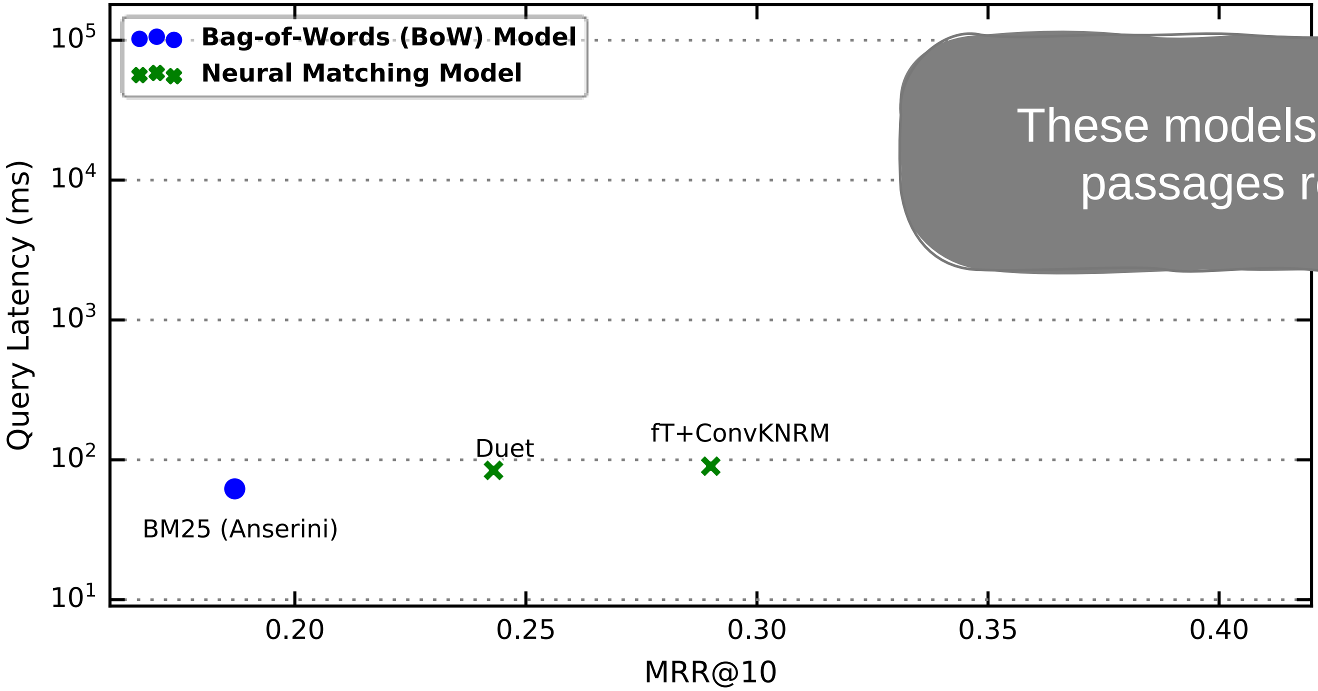
Models in this category include KNRM, Conv-KNRM, and Duet.



Chenyan Xiong, et al. End-to-end neural ad-hoc ranking with kernel pooling. SIGIR'17
Zhuyun Dai, et al. Convolutional neural networks for soft-matching n-grams in ad-hoc search. WSDM'18
Bhaskar Mitra, et al. Learning to match using local and distributed representations of text for web search. WWW'17

Query–Document Interaction Models: MS MARCO Results

- Considerable gains in **quality**—at a reasonable increase in computational cost!

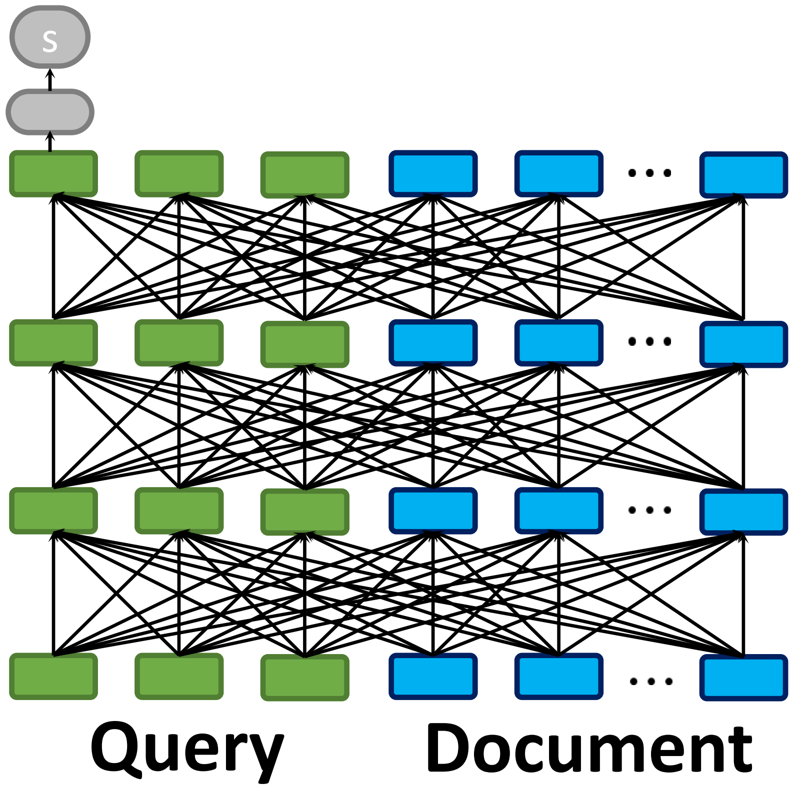


These models re-rank the top-1000 passages retrieved by BM25.

Bhaskar Mitra and Nick Craswell. An Updated Duet Model for Passage Re-ranking. arXiv:1903.07666 (2019)
Sebastian Hofstätter, et al. On the effect of low-frequency terms on neural-IR models. SIGIR'19

All-to-all Interaction with BERT

1. Feed BERT “[CLS] Query [SEP] Document [SEP]”
2. Run this through all the BERT layers
3. Extract the final [CLS] output embedding
   - Reduce to a single score through a linear layer



This is essentially a standard BERT classifier, used for ranking passages.

Of course, we must fine-tune BERT for this task with positives and negatives to be effective.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. arXiv:1901.04085 (2019)
Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR’19

BERT Rankers: SOTA 2019 (in quality)

Rank	Model	Submission Date	MRR@10 On Eval
1	**BERT + Small Training** Rodrigo Nogueira and Kyunghyun Cho - New York University	January 7th, 2019	35.87
2	**IRNet (Deep CNN/IR Hybrid Network)** Dave DeBarr, Navendu Jain, Robert Sim, Justin Wang, Nirupama Chandrasekaran – Microsoft	January 2nd, 2019	28.061


Google The Keyword

SEARCH

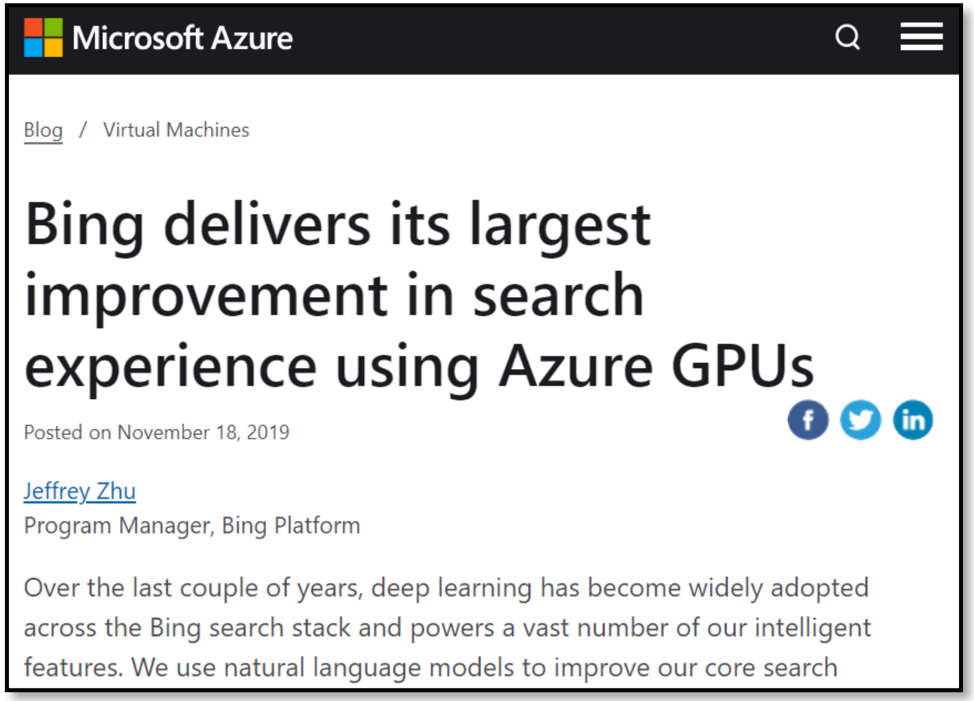
Understanding searches better than ever before

Pandu Nayak

Google Fellow and Vice President, Search

Published Oct 25, 2019

If there's one thing I've learned over the 15 years working on Google Search, it's that people's curiosity is endless.


Microsoft Azure

Blog / Virtual Machines

Bing delivers its largest improvement in search experience using Azure GPUs

Posted on November 18, 2019

Jeffrey Zhu

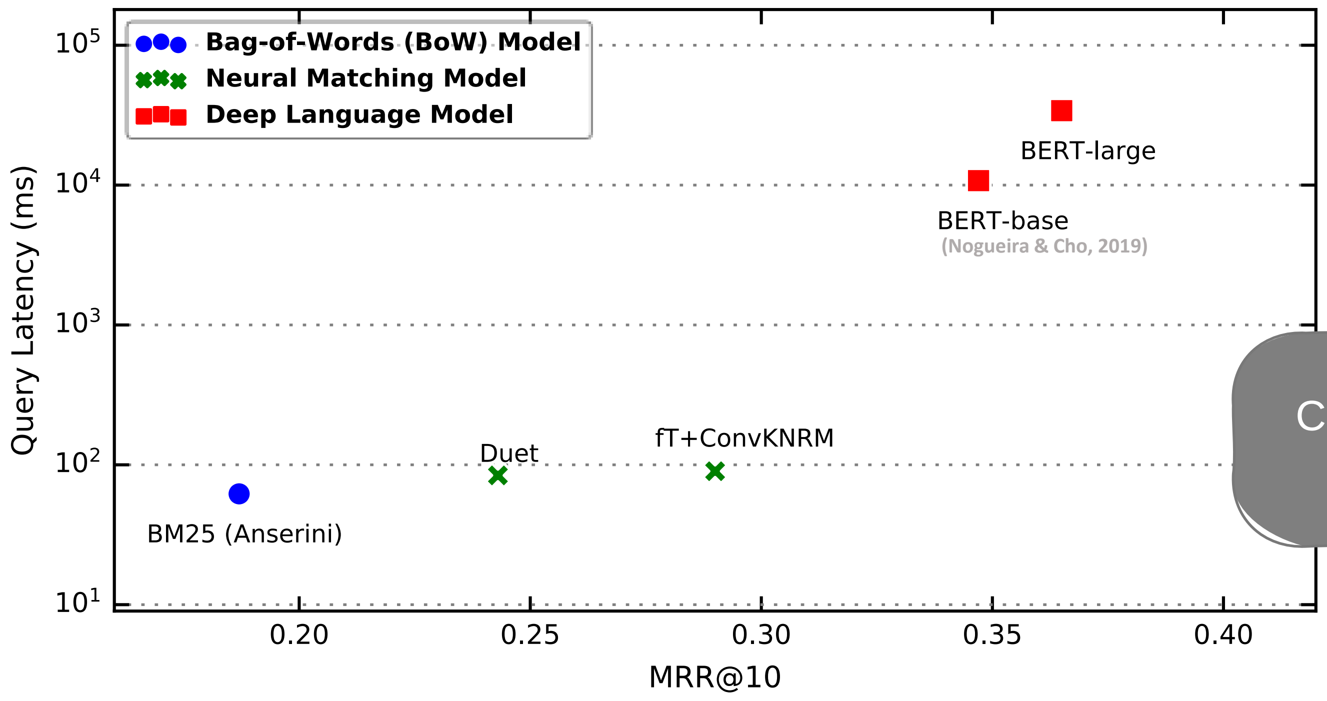
Program Manager, Bing Platform

Over the last couple of years, deep learning has become widely adopted across the Bing search stack and powers a vast number of our intelligent features. We use natural language models to improve our core search

MS MARCO Ranking screenshot as of Jan 2019. From Rodrigo Nogueira's Brief History of DL applied to IR (UoG talk).
https://blog.google/products/search/search-language-understanding-bert/
https://azure.microsoft.com/en-us/blog/bing-delivers-its-largest-improvementin-search-experience-using-azure-gpus/

BERT Rankers: Efficiency–Effectiveness Tradeoff

- Dramatic gains in **quality**—but also a dramatic increase in **computational cost**!



Can we achieve high MRR
and low latency?

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. arXiv:1901.04085 (2019)

Toward Faster Ranking: **Pre-computation**

- BERT rankers are slow because their computations be **redundant**:
  - **Represent the query** (1000 times for 1000 documents)
  - **Represent the document** (once for every query!)
  - Conduct matching between the query and the document

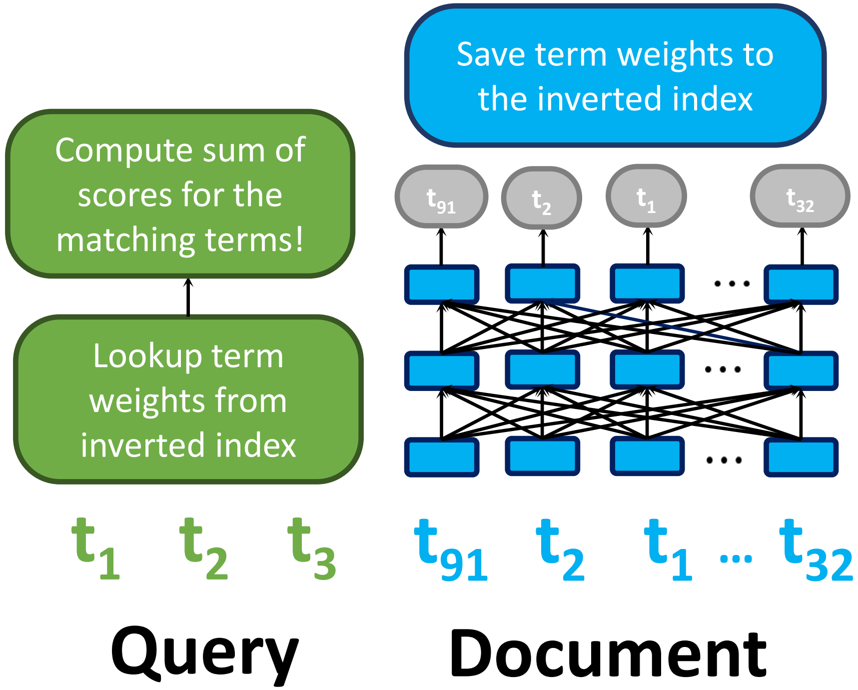
- We have the documents in advance.
  - Can we **pre-compute** the document representations?
  - And “cache” these representations for use across queries

Is there a unique value in **jointly** representing queries and documents?

Neural IR Paradigms: **Learning term weights**

- BM25 decomposed a document's score into a summation over term–document weights. **Can we learn term weights with BERT?**

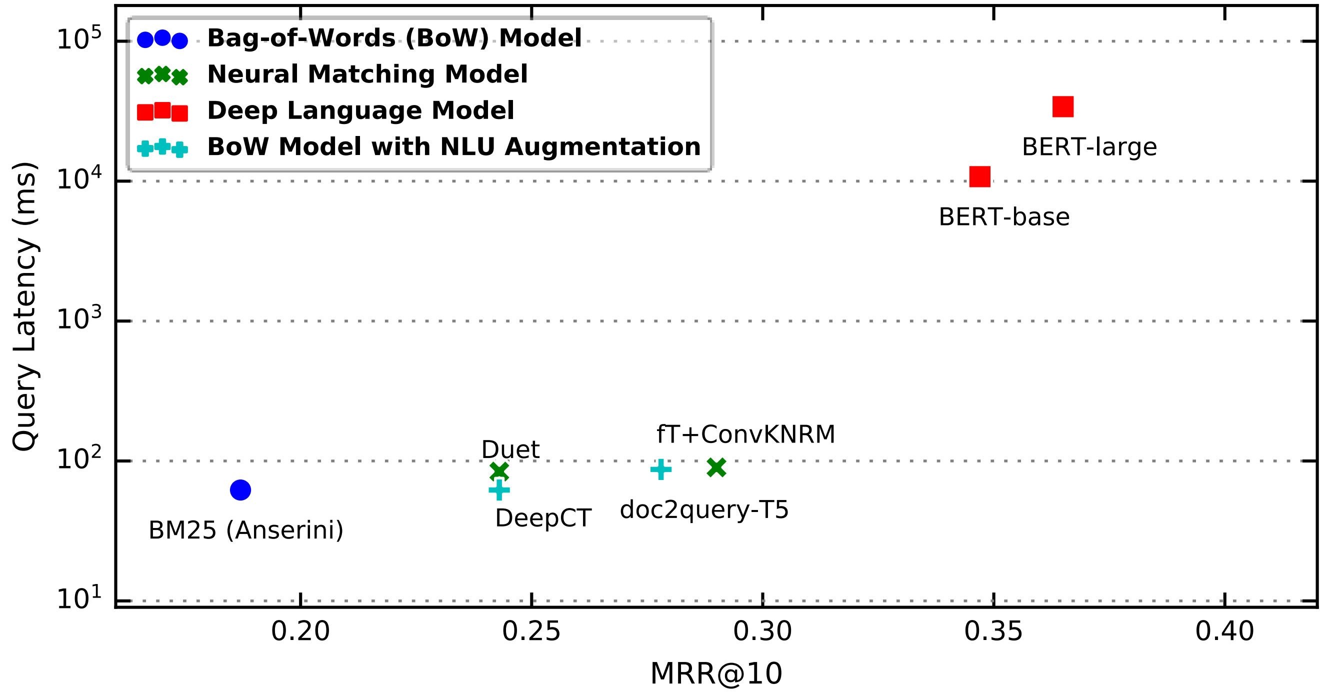
- Tokenize the query/document
- Use BERT to produce a score for each token in the document
- Add the scores of the tokens that also appear in the query



Dai, Zhuyun, and Jamie Callan. "Context-aware term weighting for first stage passage retrieval." SIGIR'20
Nogueira, Rodrigo and Jimmy Lin. "From doc2query to docTTTTTquery." Online preprint (2019).
Mallia, Antonio, et al. "Learning Passage Impacts for Inverted Indexes." SIGIR'21.

Learning term weights

- We get to learn the term weights with BERT and to **re-use** them!
- But our query is back to being a "bag of words".



DeepCT and **doc2query** are two major models under this paradigm.

Can we do better?

Next: Can we achieve high MRR *and* low latency?

- Yes! We’ll discuss two rich neural IR paradigms:
  - **Representation Similarity**
  - **Late Interaction**

References

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT.“ SIGIR’20

Chenyan Xiong, et al. End-to-end neural ad-hoc ranking with kernel pooling. SIGIR’17

Zhuyun Dai, et al. Convolutional neural networks for soft-matching n-grams in ad-hoc search. WSDM’18

Bhaskar Mitra, et al. Learning to match using local and distributed representations of text for web search. WWW’17

Bhaskar Mitra and Nick Craswell. An Updated Duet Model for Passage Re-ranking. arXiv:1903.07666 (2019)

Sebastian Hofstätter, et al. On the effect of low-frequency terms on neural-IR models. SIGIR’19

Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. SIGIR’19

Rodrigo Nogueira. “A Brief History of Deep Learning applied to Information Retrieval” (UoG talk). Retrieved from https://docs.google.com/presentation/d/1_mlvmyev0pjdG0OcfbEWManRREC0jCdjD3b1tPPvcbk

Zhuyun Dai, and Jamie Callan. "Context-aware term weighting for first stage passage retrieval.” SIGIR’20

Rodrigo Nogueira and Jimmy Lin. "From doc2query to docTTTTTquery." Online preprint (2019).

Antonio Mallia, et al. "Learning Passage Impacts for Inverted Indexes.“ SIGIR’21.