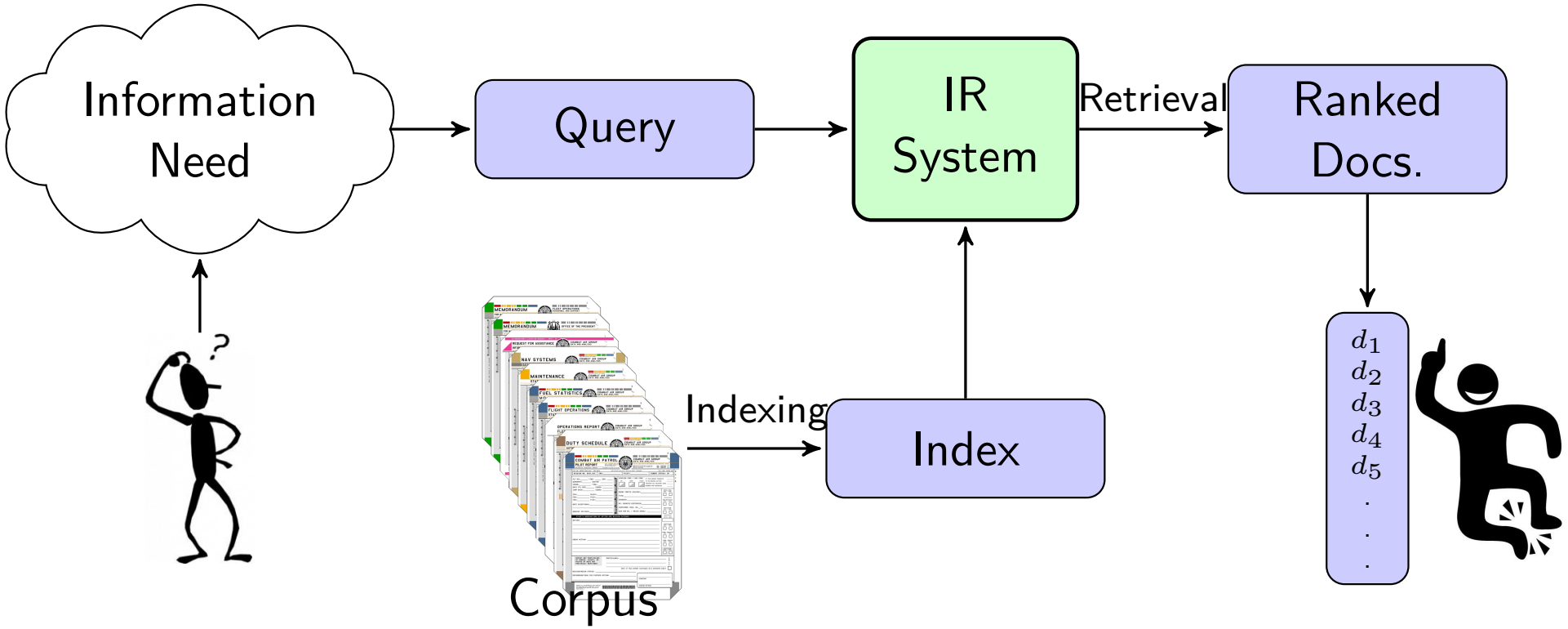
Neural IR

Sourav Saha

Indian Statistical Institute, Kolkata

July 17, 2021

Information Retrieval: A Graphical Representation



BERT - Reranker

Architecture

- vanilla BERT, mono BERT, simply BERT
- Input : [CLS] Query [SEP] Document
- Output from [CLS] token
- Predict score with a single neural layer
- r = $BERT([CLS]; q_{1...n}; [SEP]; d_{1...m})_{CLS}$
- choice of any BERT model
- s = $r \times W$
- cross entropy loss

Reference: Nogueira and Cho `https://arxiv.org/abs/1901.04085`

Problem of vanilla BERT Ranker

Longer Documents

- max 512 BERT Tokens

Problem of vanilla BERT Ranker

Longer Documents

- max 512 BERT Tokens

Time Latency

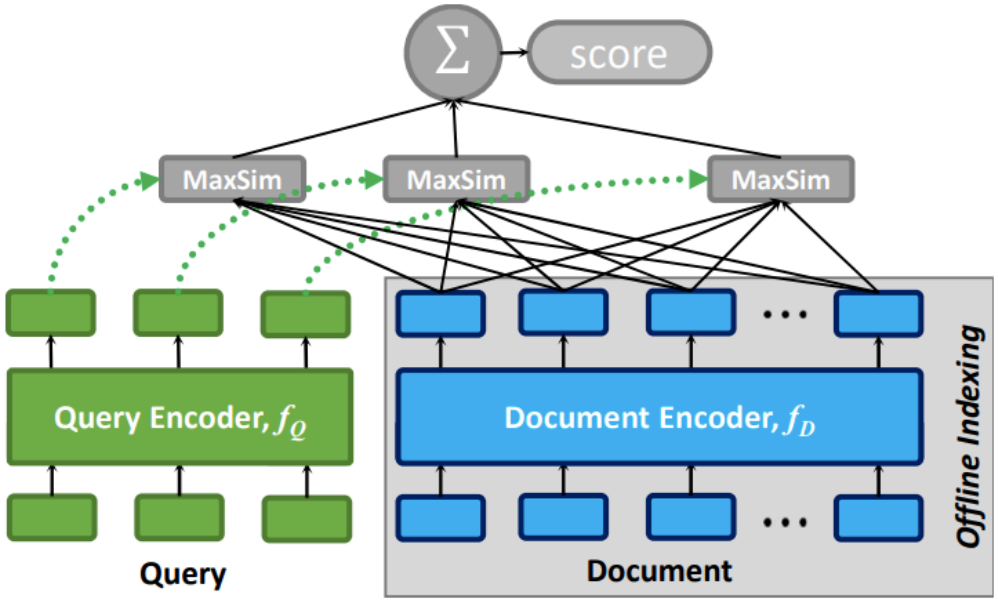
- Effective but very slow!
- Solution 1: reduce model size
- Solution 2: Precompute documents

ColBERT

- Encodes query and documents with BERT vectors
- Late interactions of Query terms and Docs

ColBERT

- Encodes query and documents with BERT vectors
- Late interactions of Query terms and Docs



Reference: Khattab and Zaharia SIGIR 2020

ColBERT

- $E_q = \hat{q}_{1...n} = \text{BERT}([\text{CLS}]; q_{1...n})$
- $E_d = = \hat{d}_{1...m} = \text{BERT}([\text{CLS}]; d_{1...m})$

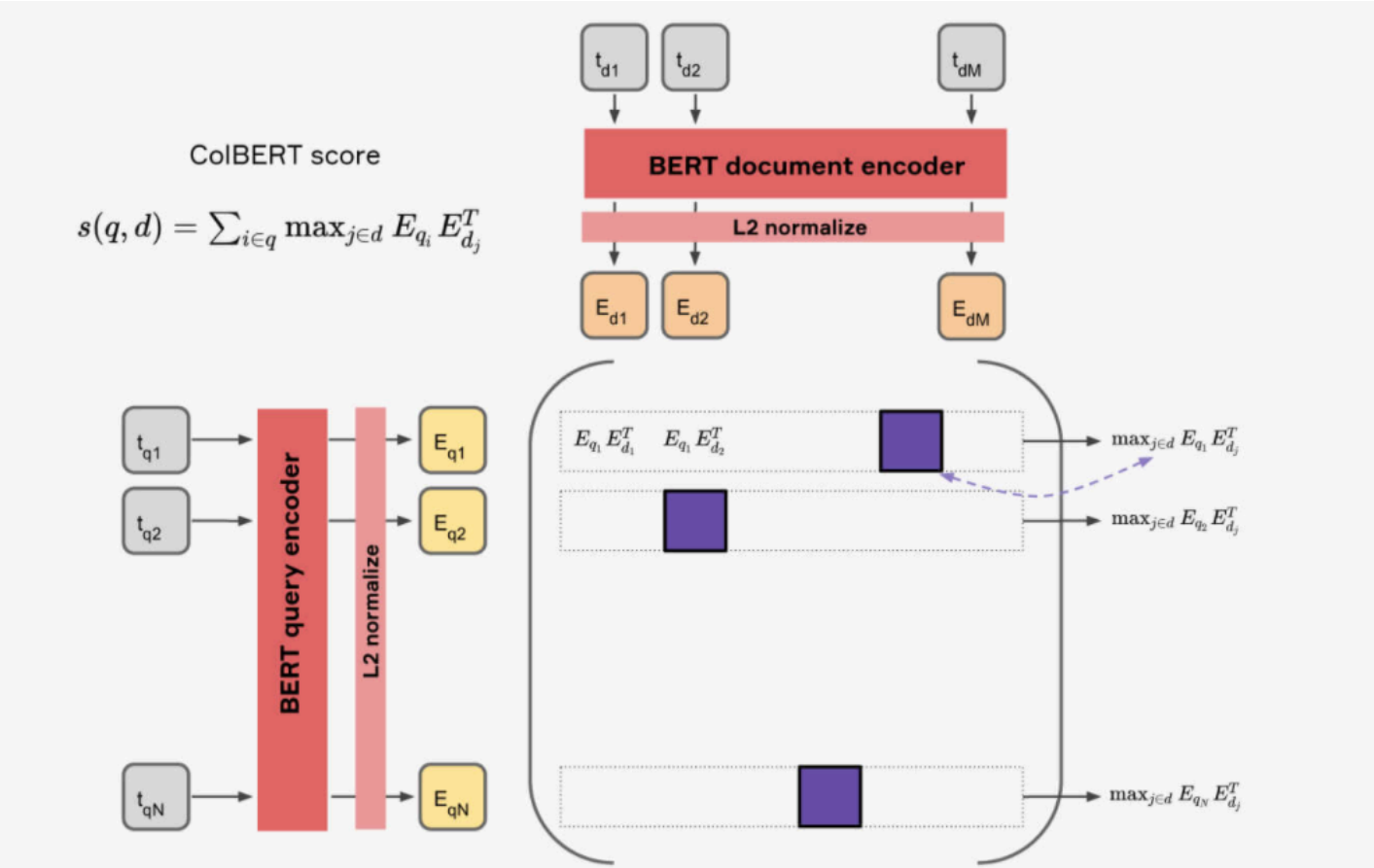


Figure 1. ColBERT model description.

Picture source: `https://europe.naverlabs.com/blog/a-white-box-analysis-of-colbert/`

ColBERT

- Optionally reduce dimensions of $\hat{q}$ and $\hat{d}$

ColBERT

- Optionally reduce dimensions of $\hat{q}$ and $\hat{d}$
- $E_q = \text{Normalize}(\text{CNN}(\text{BERT}([\text{CLS}]; q_{1...n})))$
- $E_d = \text{Filter}(\text{Normalize}(\text{CNN}(\text{BERT}([\text{CLS}]; d_{1...m}))))$

Resources

- Lin et. al `https://arxiv.org/abs/2010.06467`
- Sebastian Hofstätter course on neural IR
- `https://arxiv.org/list/cs.IR/recent`

Thank You!