Information Retrieval applied on FAQ Chatbot

Max Sobroza, April 2022

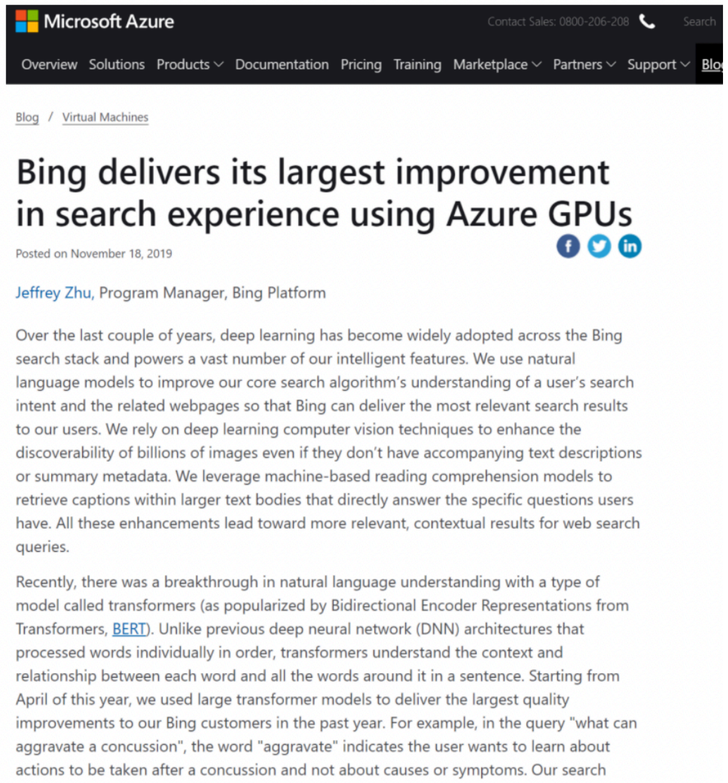
Summary

1. Information Retrieval
2. MS-MARCO
3. Data augmentation

Information Retrieval

Information Retrieval

Google (October 2019)

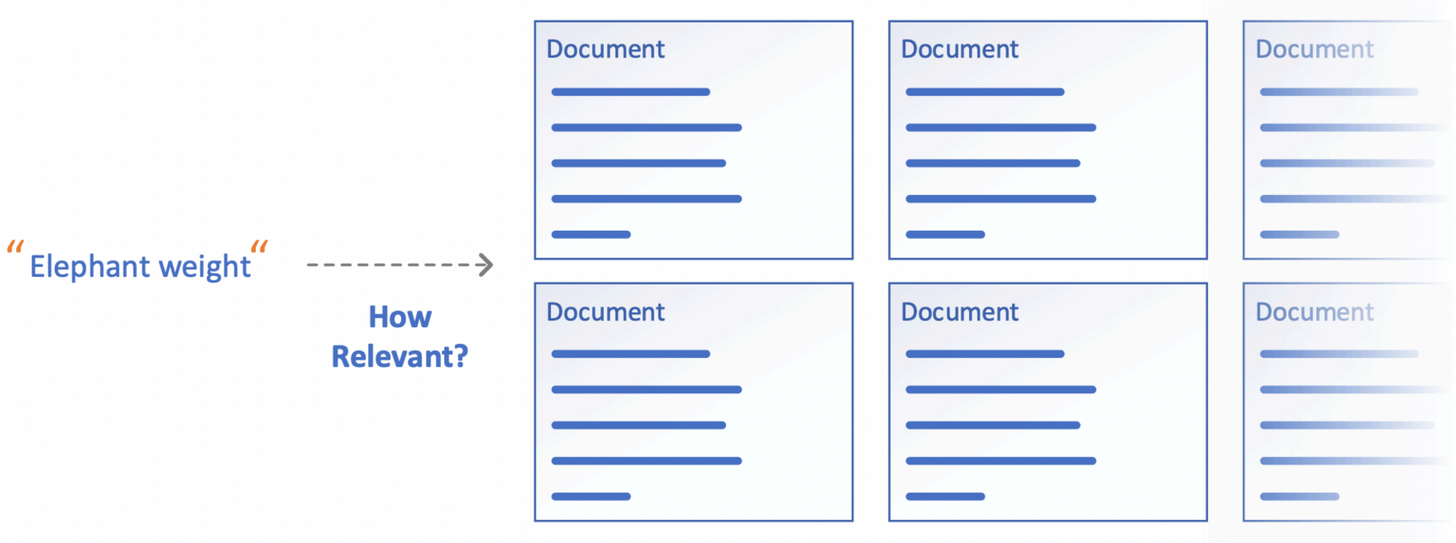


Microsoft (November 2019)

Information Retrieval

Finding the needle in the haystack

- Obtain relevant information from a collection of documents
  - **Documents:** can be anything (web pages, text, article, answer, …)
  - **Collection:** A set of documents
  - **Relevance:** Does the document satisfy the information need of the user
  - **Query:** question, set of words, sentence or even a document…



MS MARCO

- Microsoft MAchine Reading COmprehension Dataset
- IR Dataset with **a lot** of training data
  - Most of neural IR research make use of MS MARCO (including Google Research)
  - More than 8 millions of training samples
- Web search queries and passage-level answers extracted from Bing
- Negatives samples are sampled using BM25 then annotated

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, Tong Wang. MS MARCO : A Human Generated MAchine Reading COmprehension Dataset (2018).

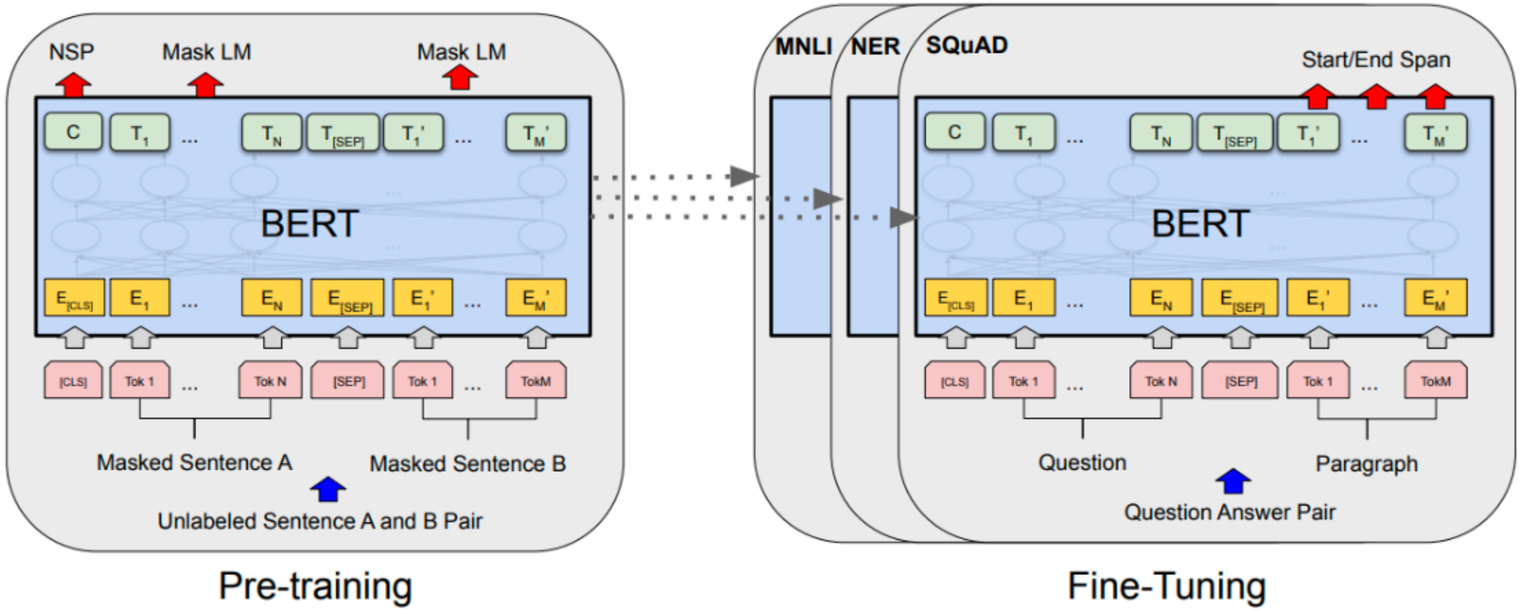
MS MARCO

- Training triples
  - **Query:** *what fruit is native to Australia*
  - **Relevant:** *Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, ...*
  - **Non-Relevant:** *The kola nut is the fruit of the kola tree, a genus (Cola) of trees that are native to the tropical rainforests of Africa.*
- **Main limitations**
  - Only in English language
  - Sparse annotation (only 1 judgement per query)

Recent Neural IR models

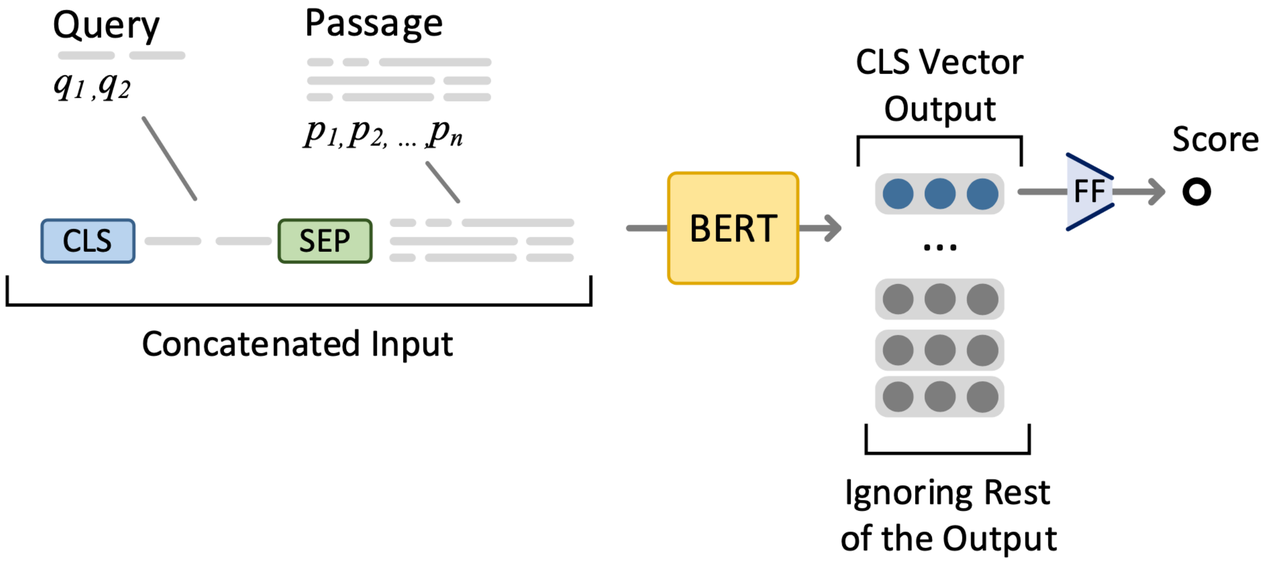
Recall: BERT Workflow

- **B**idirectional **E**ncoder **R**epresentations for **T**ransformers
- Someone with lots of computer or time pre-trains a large model
  - BERT uses Masked Language Modelling (MLM) and Next Sentence Prediction (NSP)
- We download it and fine-tune on our task



BERT Re-ranking IR model

- monoBERT
- Concatenating the two sequences to fit in BERT workflow
  - [CLS] query [SEP] passage
  - Pool [CLS] token
  - Predict score with a single linear layer (binary cross entropy loss)
- Needs to be repeated for every passage



Passage Re-ranking with BERT. Rodrigo Nogueira, Kyunghyun Cho. 2019

https://arxiv.org/abs/1901.04085

BERT Re-ranking IR model

- SOTA of Neural IR models in several IR Datasets
- MSMARCO-Passage ranking
  - MRR@10 from 0.194 (BM25) to 0.385 (ALBERT-Large)
  - Doubles the result quality
- Longer Documents
  - Works well (MS MARCO-Document ranking)
  - Sliding window over the document
  - Take max window score as document score
- **Main limitations**
  - May not work very well for domain specific tasks
  - Needs a large supervised IR Dataset
  - Inference computation time for a large collection of documents

Deeper Text Understanding for IR with Contextual Neural Language Modeling. Zhuyun Dai, Jamie Callan

https://arxiv.org/abs/1905.09217

Fine-tuning model on MSMARCO

Re-ranking task IR: MS MARCO

- Translated MS MARCO Dataset
- Zero-Shot Learning case (not fine-tuned on target data)
- Same labels to target task: relevant or not relevant
  - This passage is relevant to answer the question
- Mitigates two drawbacks of BERT modelling
  - BERT was not Pre-trained on target text
  - Need of a large supervised IR dataset

Wenpeng Yin, Jamaal Hay, Dan Roth. Benchmarking Zero-shot Text Classification : Datasets, Evaluation and Entailment Approach, 2019

Matthew E. Peters, Sebastian Ruder, Noah A. Smith To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, 2019

Rosa, Guilherme Moraes; Rodrigues, Ruan Chaves; Lotufo, Roberto; Nogueira, Rodrigo. To tune or not to tune?: zero-shot models for legal case entailment, ICAIL 2021

Data augmentation

Data augmentation

Back translation

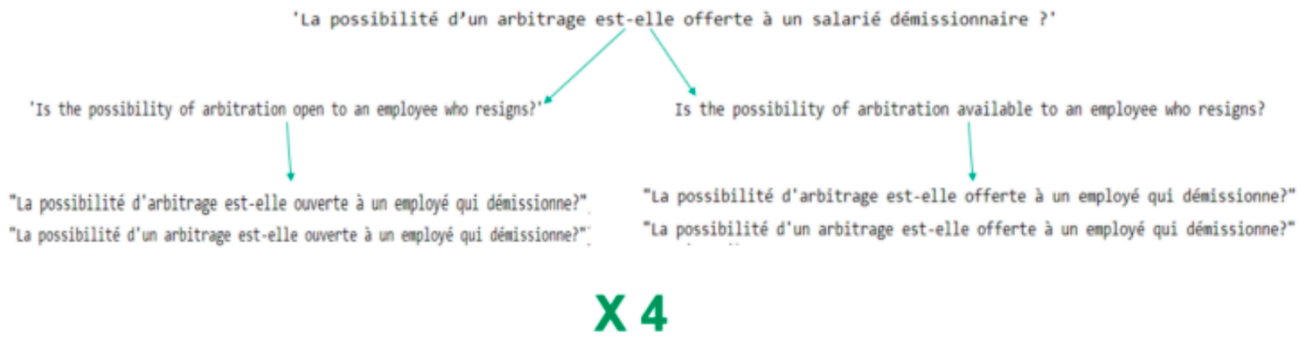


Figure 27 : illustration de la double traduction pour l'augmentation des données textuelles