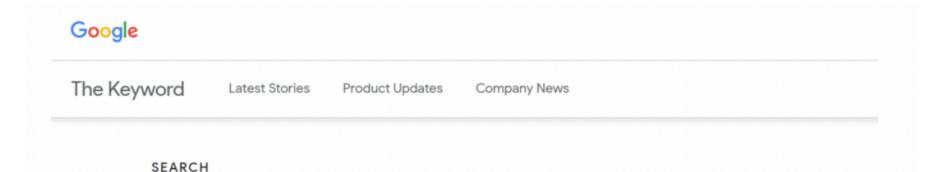
Information Retrieval applied on FAQ Chatbot

Summary

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

Information Retrieval

Information Retrieval



Understanding searches better than ever before

Pandu Nayak
Google Fellow and
Vice President, Search

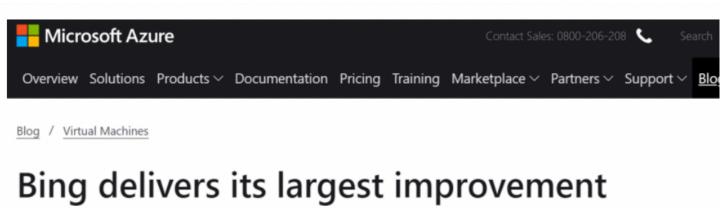
Published Oct 25, 2019

If there's one thing I've learned over the 15 years working on Google Search, it's that people's curiosity is endless. We see billions of searches every day, and 15 percent of those queries are ones we haven't seen before—so we've built ways to return results for queries we can't anticipate.

When people like you or I come to Search, we aren't always quite sure about the best way to formulate a query. We might not know the right words to use, or how to spell something, because often times, we come to Search looking to learn--we don't necessarily have the knowledge to begin with.

At its core, Search is about understanding language. It's our job to figure out what you're searching for and surface helpful information from the web, no matter how you spell or combine the words in your query. While we've continued to improve our language understanding capabilities over the years, we

Google (October 2019)



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 algorithm's understanding of a user's search intent and the related webpages so that Bing can deliver the most relevant search results to our users. We rely on deep learning computer vision techniques to enhance the discoverability of billions of images even if they don't have accompanying text descriptions or summary metadata. We leverage machine-based reading comprehension models to retrieve captions within larger text bodies that directly answer the specific questions users have. All these enhancements lead toward more relevant, contextual results for web search queries.

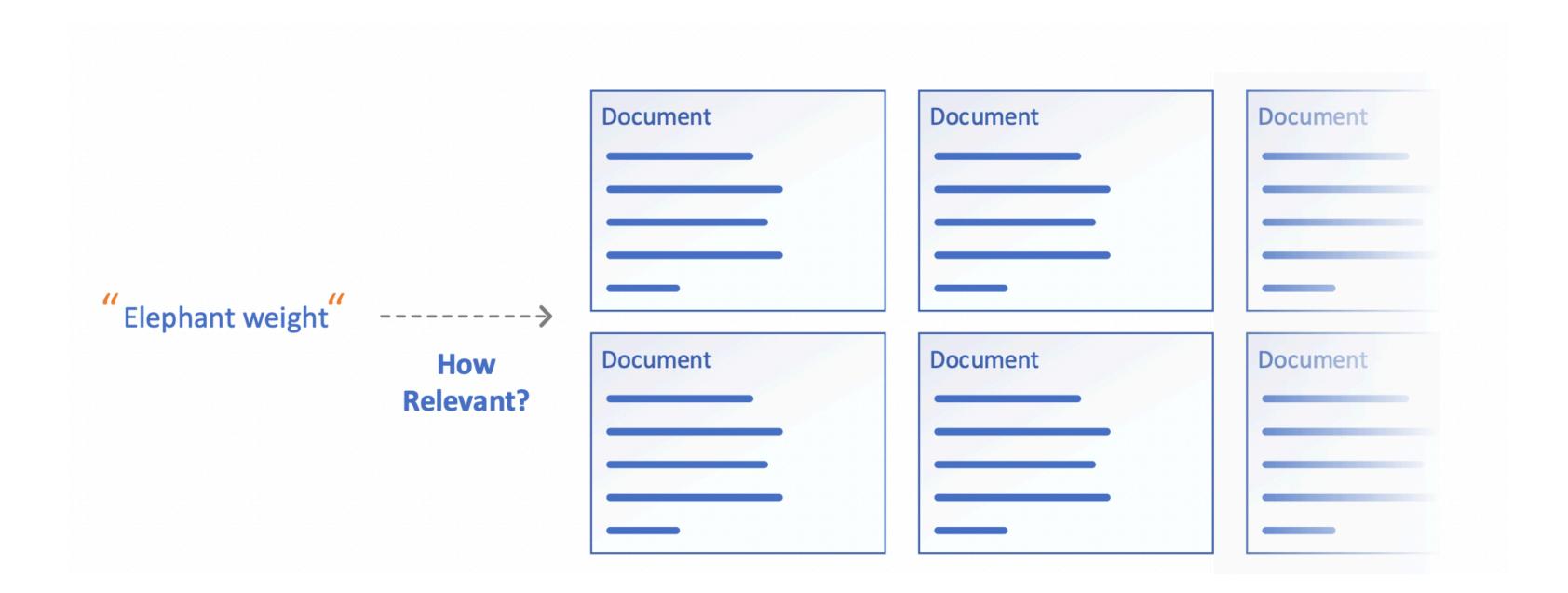
Recently, there was a breakthrough in natural language understanding with a type of model called transformers (as popularized by Bidirectional Encoder Representations from Transformers, BERT). Unlike previous deep neural network (DNN) architectures that processed words individually in order, transformers understand the context and relationship between each word and all the words around it in a sentence. Starting from April of this year, we used large transformer models to deliver the largest quality improvements to our Bing customers in the past year. For example, in the query "what can aggravate a concussion", the word "aggravate" indicates the user wants to learn about actions to be taken after a concussion and not about causes or symptoms. Our search

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.

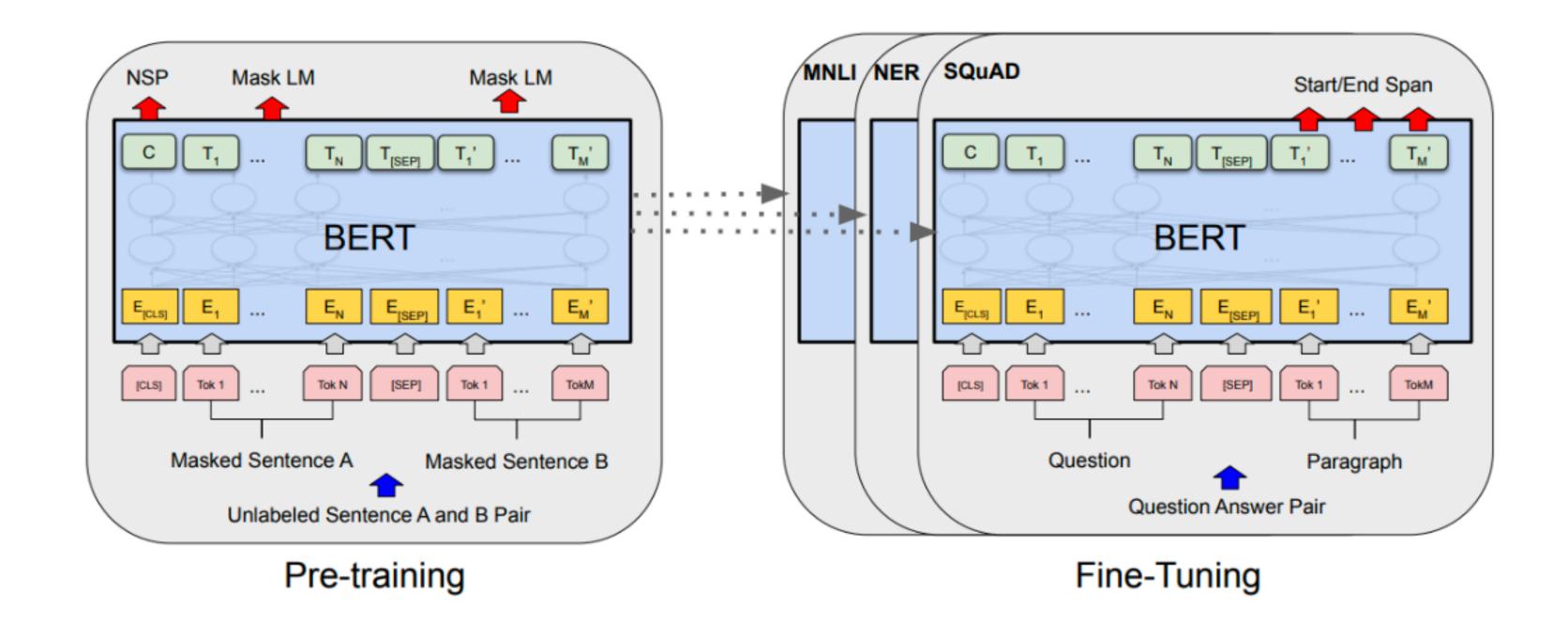
Main limitations

- Only in English language
- Sparse annotation (only 1 judgement per query)

Recent Neural IR models

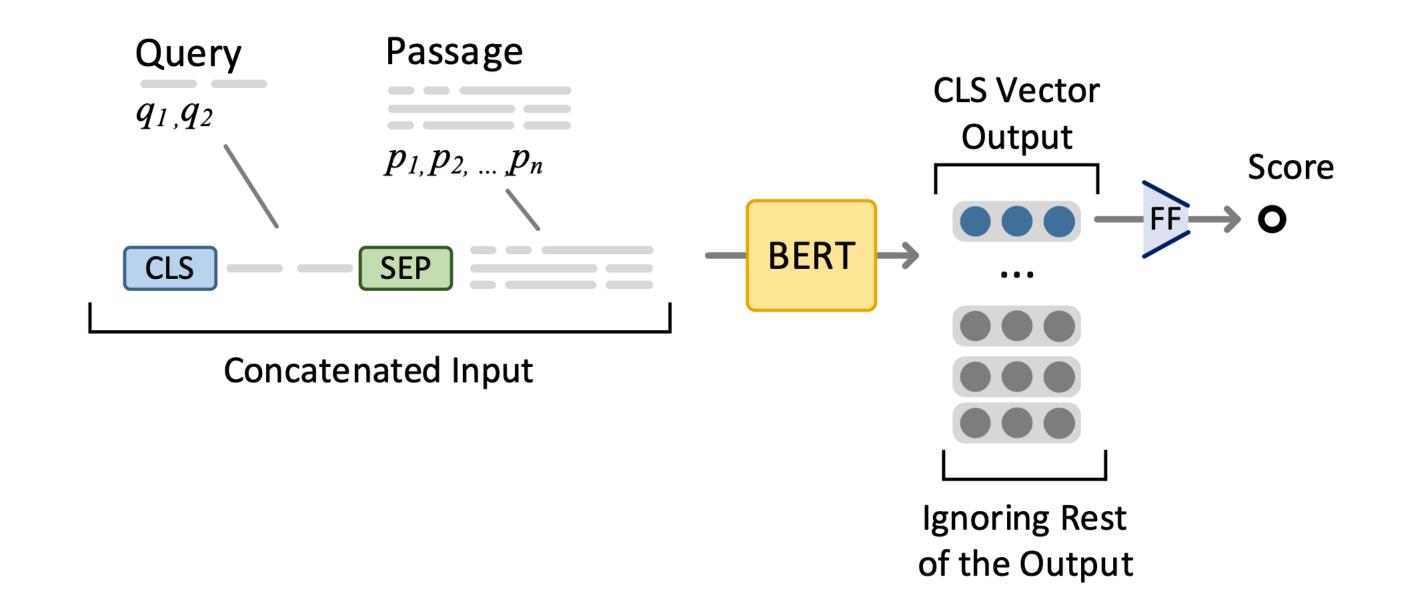
Recall: BERT Workflow

- Bidirectional Encoder Representations for Transformers
- 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

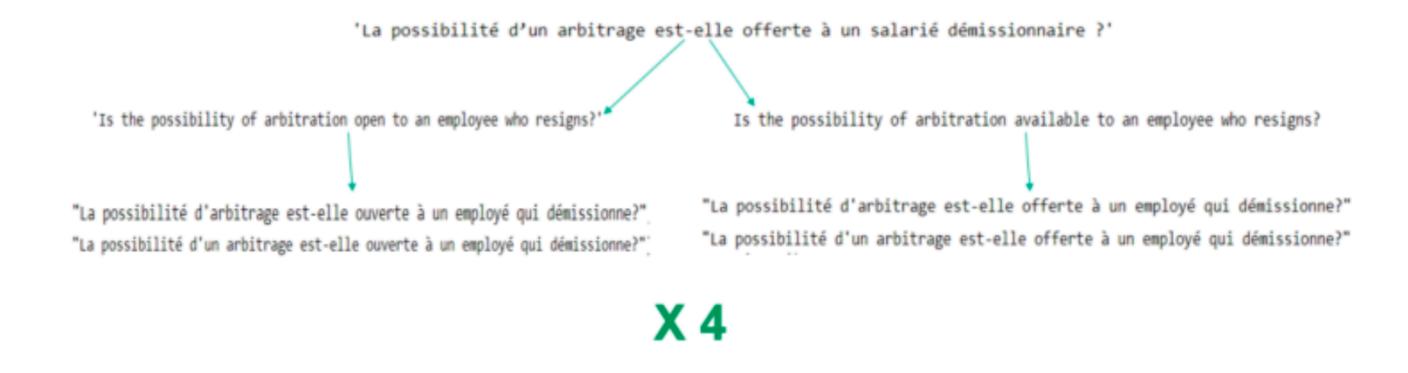


Figure 27 : illustration de la double traduction pour l'augmentation des données textuelles