Since 2004
UET
ĐẠI HỌC CÔNG NGHỆ, ĐHQGHN
VNU-University of Engineering and Technology

KEPCO KDN

Since 1906
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi

Large Language Model and Its Applications

LLM05: Retrieval and Question Answering

Thi-Hai-Yen Vuong

sponsored by **KEPCO KDN Co., Ltd.**
Eco-friendly & Digital Centered Energy ICT Platform Leader

Hanoi, 10/2023

Outline

- Retrieval and Question Answering Introduction
- Retrieval-based Question Answering
  - Traditional Method
  - Deep Learning Method
  - Hybrid
- Question Answering
  - Answer Sentence Selection
  - Reading Comprehension
  - Generative Model

1. Introduction

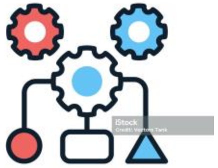
Question (Q) →

→ Answer (A)

The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language.**

Question Answering Taxonomy

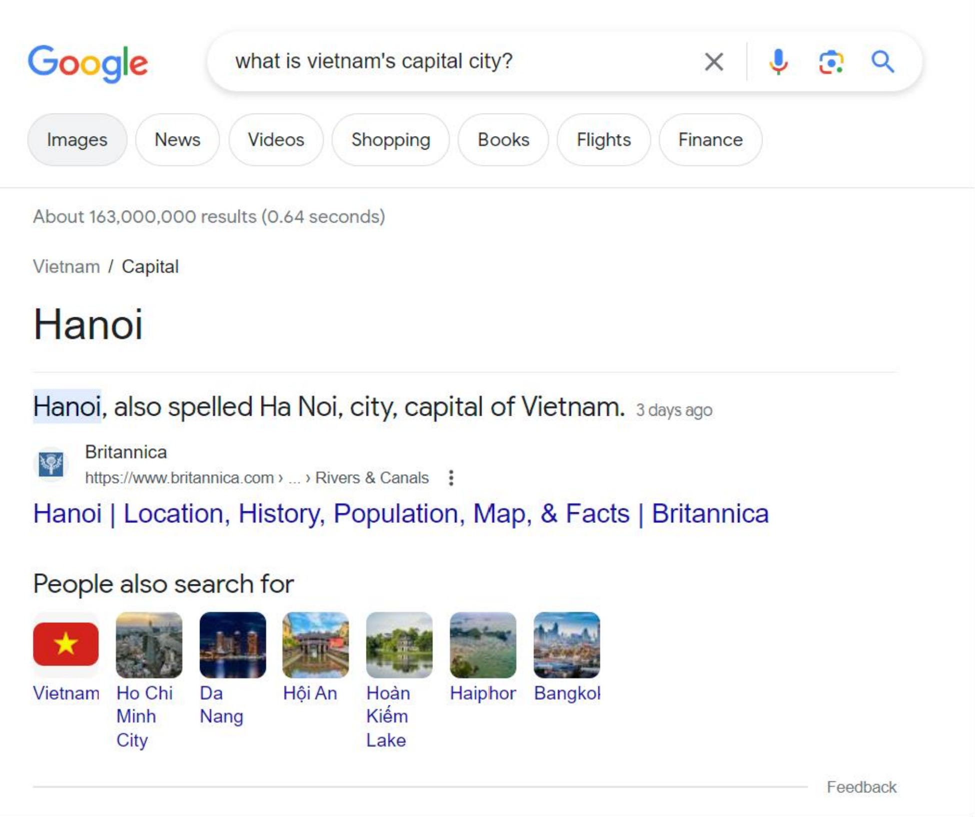
- **Source type:**
  - A text passage
  - Web documents
  - Knowledge bases
  - Tables
  - Images
  - …



- **Answer type:**
  - Sentences
  - Tokens
  - List
  - Yes/no
  - …

- **Question type:**
  - Factoid & non-factoid
  - Open-domain & closed-domain

Google Search Engine



About 163,000,000 results (0.64 seconds)

Vietnam / Capital

Hanoi

Hanoi, also spelled Ha Noi, city, capital of Vietnam. 3 days ago

Britannica
https://www.britannica.com › ... › Rivers & Canals

Hanoi | Location, History, Population, Map, & Facts | Britannica

People also search for

Vietnam

Ho Chi Minh City

Da Nang

Hội An

Hoàn Kiếm Lake

Haiphor

Bangkol

Wikipedia

WIKIPEDIA

English
The Free Encyclopedia
3 543 000+ articles

日本語
フリー百科事典
730 000+ 記事

Deutsch
Die freie Enzyklopädie
1 181 000+ Artikel

Español
La enciclopedia libre
710 000+ artículos

Fr
L'encyc
1 061

Type your search keyword here

лопедия
ей

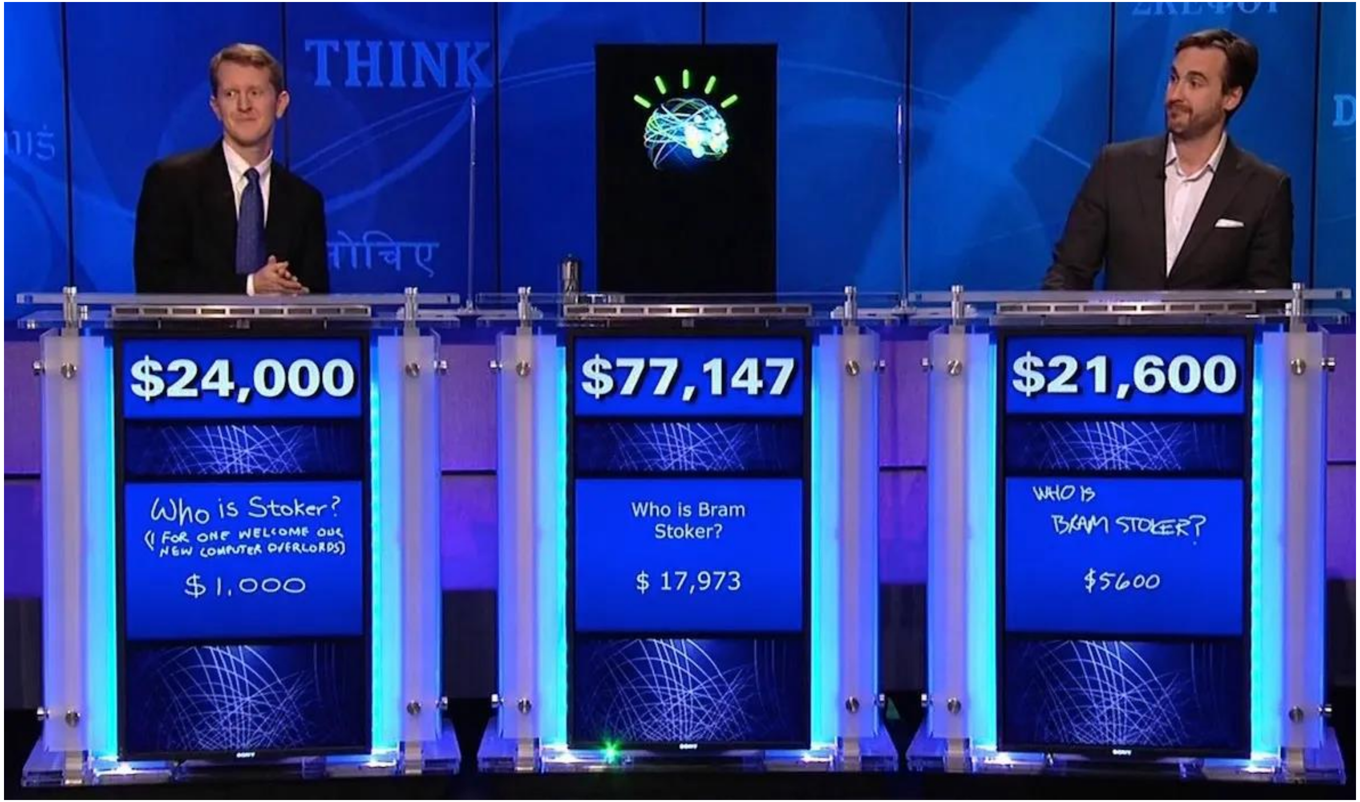
Italiano
L'enciclopedia libera
768 000+ voci

Português
A enciclopédia livre
669 000+ artigos

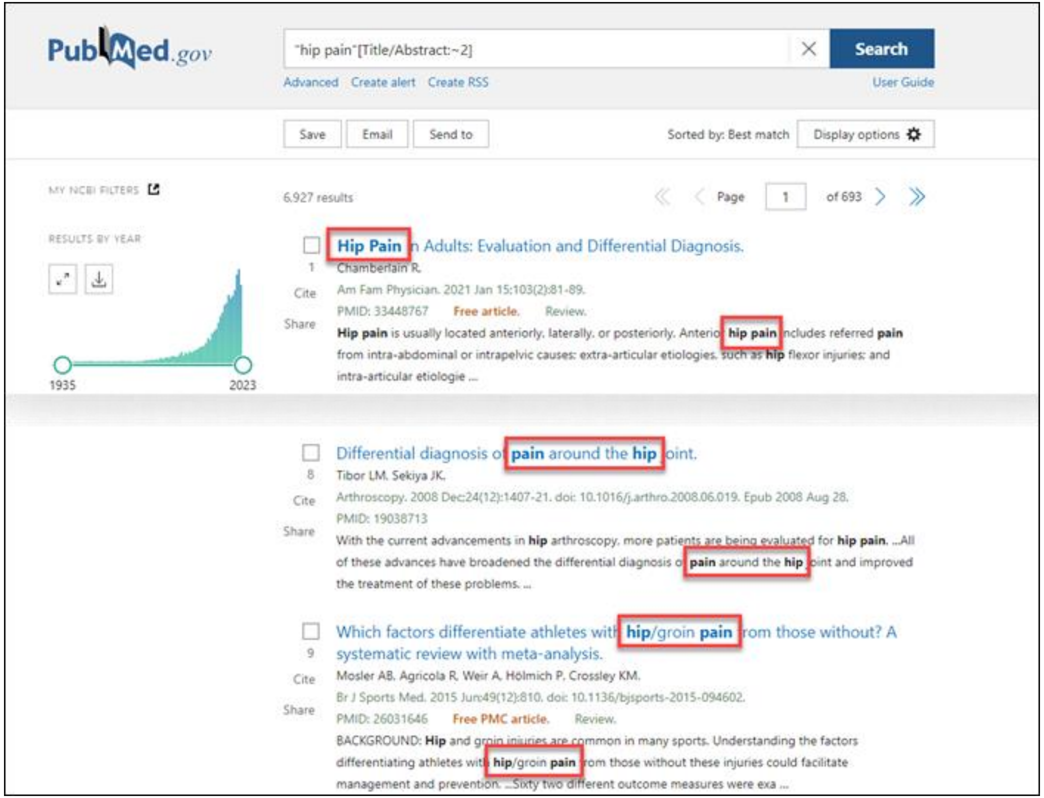
Polski
Wolna encyklopedia
769 000+ haseł

Nederlands
De vrije encyclopedie
668 000+ artikelen

2011: IBM Watson beat Jeopardy champions



Pubmed



"hip pain"[Title/Abstract:~2]

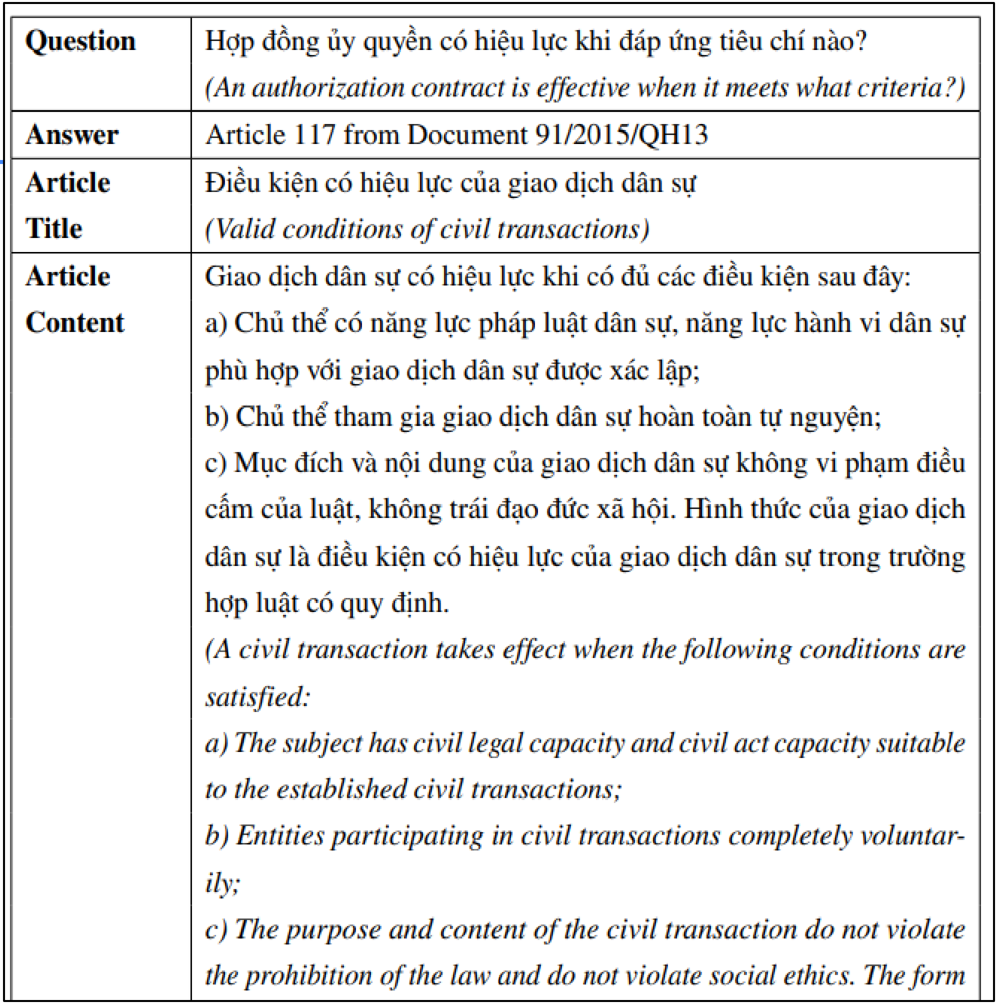
6,927 results

1. □ **Hip Pain** n Adults: Evaluation and Differential Diagnosis.
   Chamberlain R.
   Am Fam Physician. 2021 Jan 15;103(2):81-89.
   PMID: 33448767 Free article. Review.
   Hip pain is usually located anteriorly, laterally, or posteriorly. Anterio **hip pain** ncludes referred **pain** from intra-abdominal or intrapelvic causes; extra-articular etiologies, such as **hip** flexor injuries; and intra-articular etiologie ...

8. □ Differential diagnosis o **pain** around the **hip** oint.
   Tibor LM, Sekiya JK.
   Arthroscopy. 2008 Dec;24(12):1407-21. doi: 10.1016/j.arthro.2008.06.019. Epub 2008 Aug 28.
   PMID: 19038713
   With the current advancements in **hip** arthroscopy, more patients are being evaluated for **hip pain**. ...All of these advances have broadened the differential diagnosis o **pain** around the **hip** oint and improved the treatment of these problems. ...

9. □ Which factors differentiate athletes wit **hip**/groin **pain** rom those without? A systematic review with meta-analysis.
   Mosler AB, Agricola R, Weir A, Hölmich P, Crossley KM.
   Br J Sports Med. 2015 Jun;49(12):810. doi: 10.1136/bjsports-2015-094602.
   PMID: 26031646 Free PMC article. Review.
   BACKGROUND: **Hip** and groin injuries are common in many sports. Understanding the factors differentiating athletes wit **hip**/groin **pain** rom those without these injuries could facilitate management and prevention. ...Sixty two different outcome measures were exa ...

Legal Question Answering



Question	Hợp đồng ủy quyền có hiệu lực khi đáp ứng tiêu chí nào? *(An authorization contract is effective when it meets what criteria?)*
Answer	Article 117 from Document 91/2015/QH13
Article Title	Điều kiện có hiệu lực của giao dịch dân sự *(Valid conditions of civil transactions)*
Article Content	Giao dịch dân sự có hiệu lực khi có đủ các điều kiện sau đây: a) Chủ thể có năng lực pháp luật dân sự, năng lực hành vi dân sự phù hợp với giao dịch dân sự được xác lập; b) Chủ thể tham gia giao dịch dân sự hoàn toàn tự nguyện; c) Mục đích và nội dung của giao dịch dân sự không vi phạm điều cấm của luật, không trái đạo đức xã hội. Hình thức của giao dịch dân sự là điều kiện có hiệu lực của giao dịch dân sự trong trường hợp luật có quy định. *(A civil transaction takes effect when the following conditions are satisfied:* *a) The subject has civil legal capacity and civil act capacity suitable to the established civil transactions;* *b) Entities participating in civil transactions completely voluntarily;* *c) The purpose and content of the civil transaction do not violate the prohibition of the law and do not violate social ethics. The form*

QA based on Knowledge Graph

Visual Question Answering

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

2. Retrieval based Question Answering

The retrieval-based QA problem can be simply stated as follows:

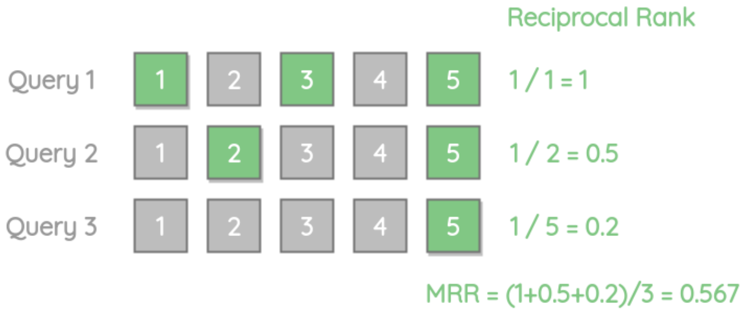
Input: given a query q and a text corpus $D = \{d_1, d_2, \ldots, d_n\}$

Ouput: the retrieval-based QA aims to find the most likely document that maximizes the relevance score: $\mathrm{d}^* = \mathrm{argmax}_{\mathrm{d} \in \mathrm{D}} R(q, d)$

where $R(q, d)$ represents the relevance score of the query q and document d.

Metric of Retrieval

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

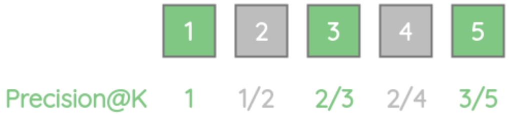


						Reciprocal Rank
Query 1	🟩 1	2	🟩 3	4	🟩 5	1 / 1 = 1
Query 2	1	🟩 2	3	4	🟩 5	1 / 2 = 0.5
Query 3	1	2	3	4	🟩 5	1 / 5 = 0.2

MRR = (1+0.5+0.2)/3 = 0.567

$$AP = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{number\ of\ relevant\ items}$$

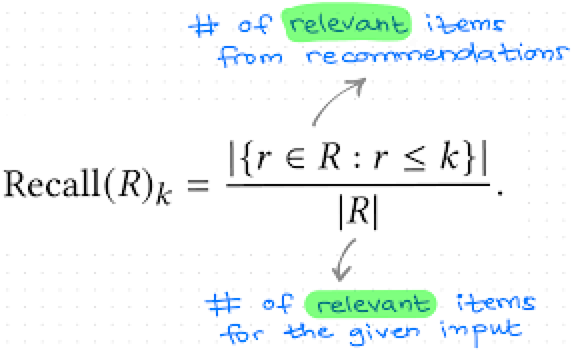
$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q)$$



	🟩 1	2	🟩 3	4	🟩 5
Precision@K	1	1/2	2/3	2/4	3/5

$$AP = \frac{(1 + 2/3 + 3/5)}{3} = 0.7555$$

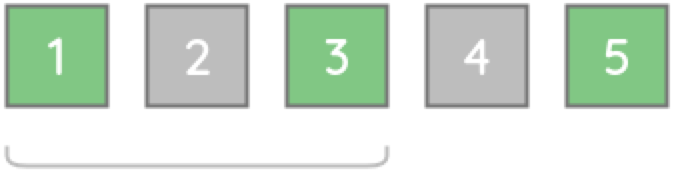
Metric of Retrieval



# of relevant items from recommendations

$$\mathrm{Recall}(R)_k = \frac{|\{r \in R : r \leq k\}|}{|R|}.$$

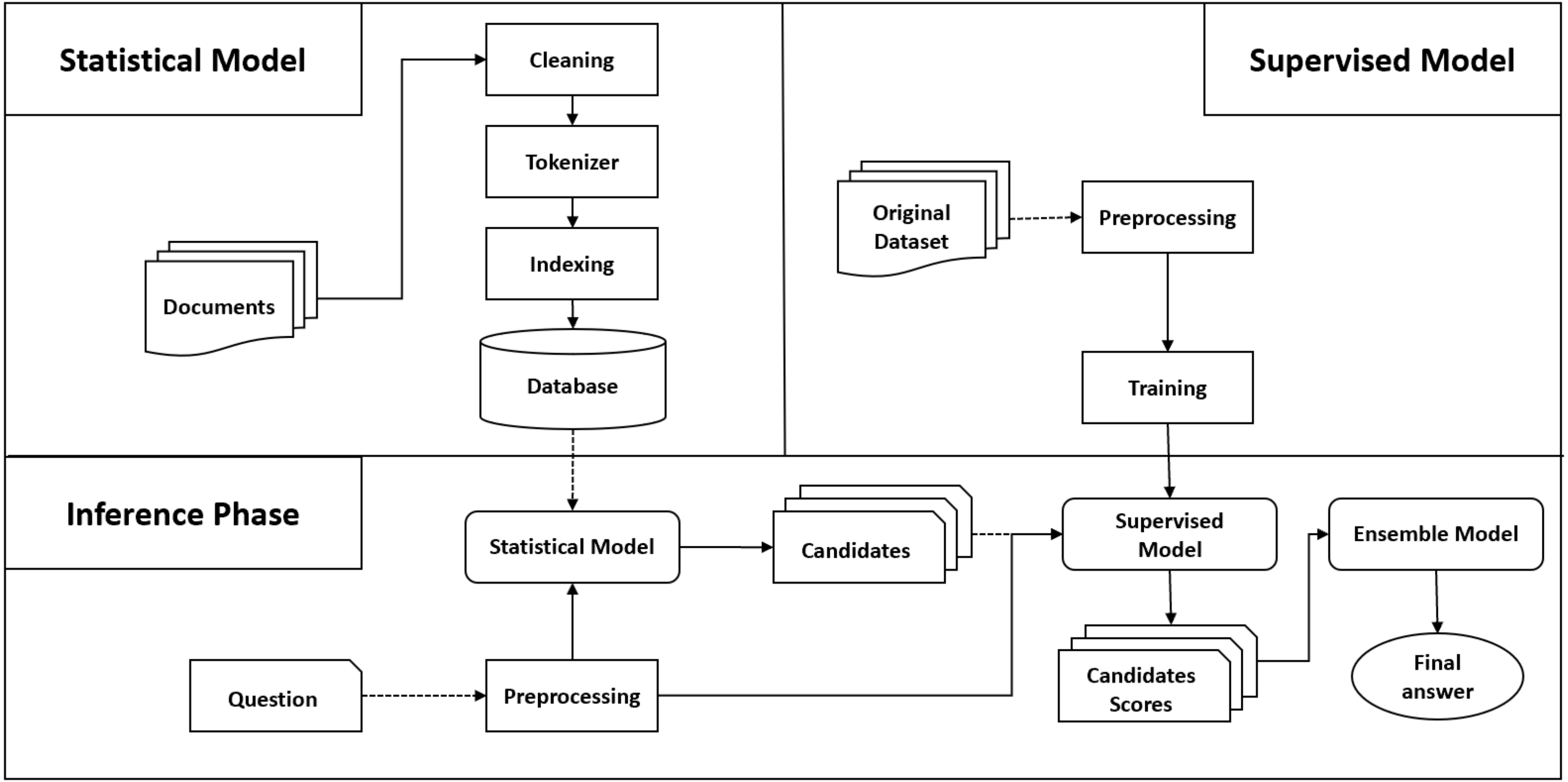
# of relevant items for the given input



Recall@3 = 2/(2+1) = 2/3 = 0.67

$$F_2 = \frac{5P \times R}{4P + R}$$

Retrieval based Question Answering



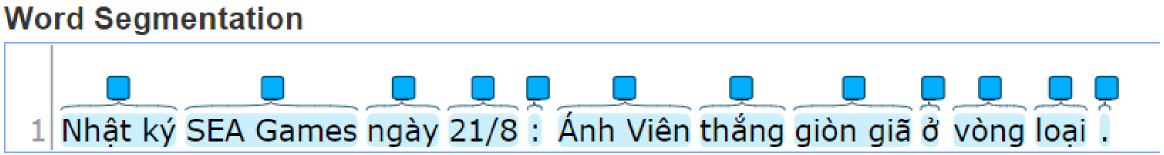
Pre-processing

- **Normalizing Text:** remove capitalization that would confuse a computer model:
  - *'Hey'* becomes *'hey'*.
  - *'Amazon'* becomes *'amazon'*.
  - *'PLEASE FIX'* becomes *'please fix'*.
  - *'@AmazonHelp'* becomes *'@amazonhelp'*.
- **Removing Unicode Characters**: punctuation, Emoji's, URL's and @'s confuse AI models because they are uniques signatures that either end up being translated unhelpfully into Unicode
- **Removing stopwords:** some words that don't directly apply to interpretation.
- **Stemming and Lemmatization:**
  - Stemming, the simpler of the two, groups words by their root stem.
  - Lemmatization, on the other hand, groups words based on root definition, and allows us to differentiate between present, past, and indefinite.

Pre-processing

Part of Speech (POS) Tagging: There are eight main parts of speech.

Tokenizer



Word Segmentation

Nhật ký SEA Games ngày 21/8 : Ánh Viên thắng giòn giã ở vòng loại .

Noun	N
Verb	V
Adjective	ADJ
Adverb	ADV
Preposition	P
Conjunction	CON
Pronoun	PRO
Interjection	INT

Further Sorting: Text cleaning has three further sorting functions that may be of use:

- Translation
- Typo Correction
- Number Unification

Term frequency matching

TF-IDF

(Term frequency-inverse document frequency)

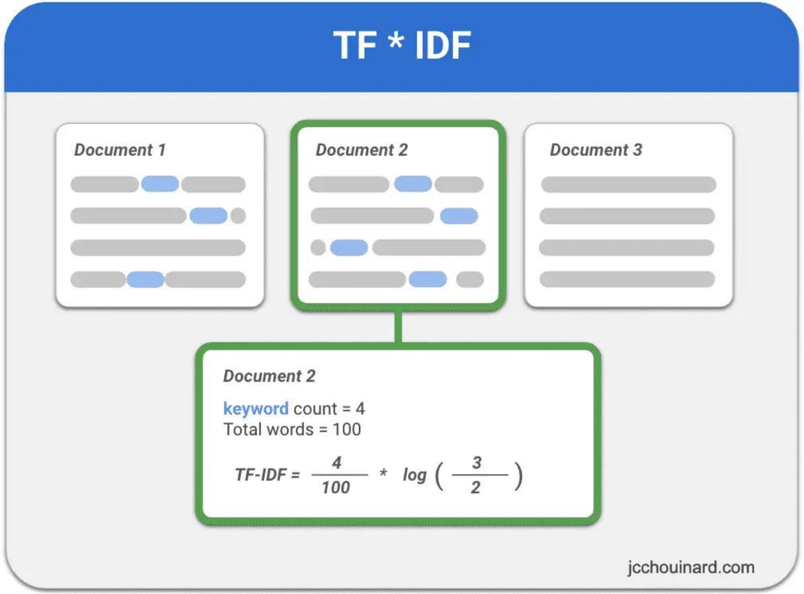
$$TF(t,A) = \frac{f(t,A)}{|A|}$$

$$IDF(t) = log(\frac{n}{n(q_i)})$$

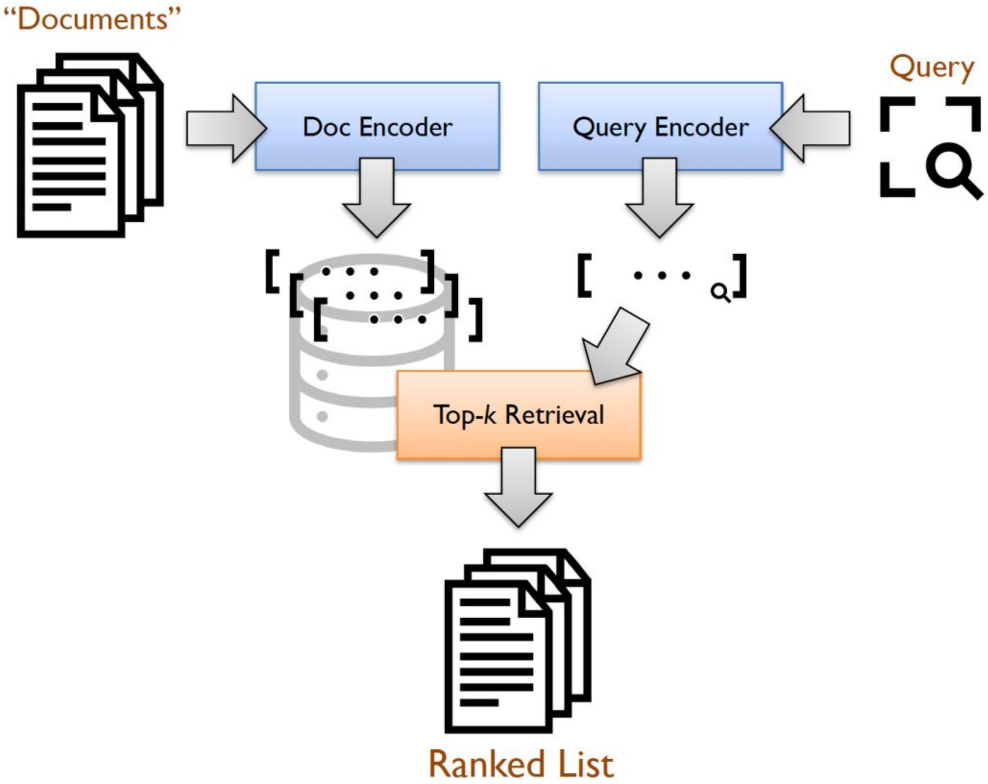
$$TF - IDF(t,A) = TF * IDF = \frac{f(t,A)}{|A|} * log(\frac{n}{n(q)})$$

The BM25

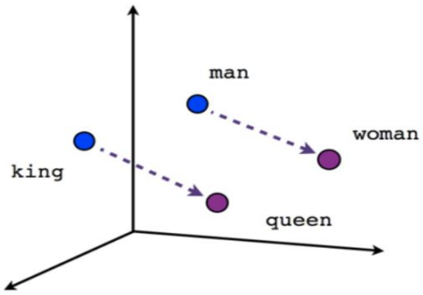
$$BM25S(Q,A) = \sum_{i=1}^{n} IDF(t_i) \cdot \frac{f(t_i,A) \cdot (k_1 + 1)}{f(t_i,A) + k_1 \cdot (1 - b + b \cdot \frac{|A|}{avgdl})}$$



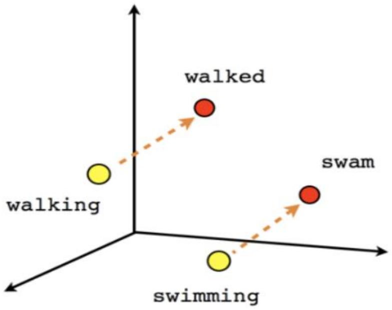
Semantic Search



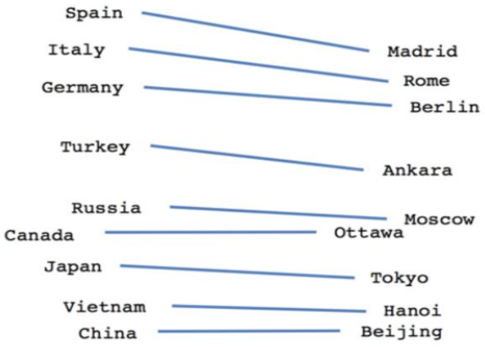
Word2vec



Male-Female



Verb tense



Country-Capital

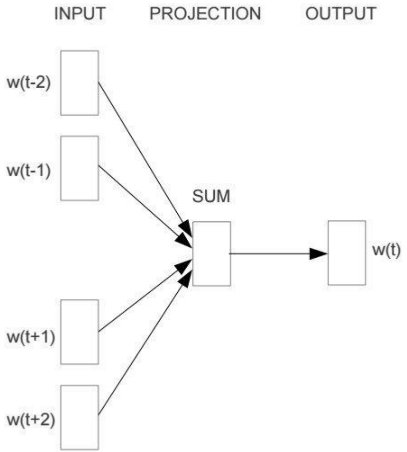
Best word2vec method:

- Word2vec (Milokov el al., 2013)
- Glove (Pennington el al., 2014)
- FastText (Bojanowski et al., 2017)

Word2vec

Cosine similarity

$$\cos(\boldsymbol{u}_i, \boldsymbol{u}_j) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{u}_j}{||\boldsymbol{u}_i||_2 \times ||\boldsymbol{u}_j||_2}.$$



BOW

Advantages:

- Short vectors are easier to use as features in ML systems
- Dense vectors may generalize better than storing explicit counts
- They do better at capturing synonymy

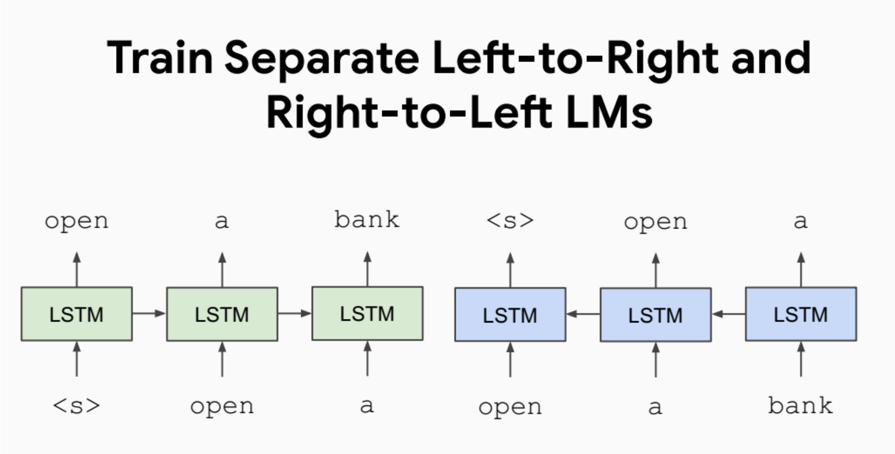
Average

$$\begin{array}{ccccccccc} W_1 & & W_2 & & & & W_n & & D \\ \begin{bmatrix} W_{11} \\ W_{12} \\ \\ \\ W_{1n} \end{bmatrix} & + & \begin{bmatrix} W_{21} \\ W_{22} \\ \\ \\ W_{2n} \end{bmatrix} & + & \ldots. & + & \begin{bmatrix} W_{n1} \\ W_{n2} \\ \\ \\ W_{nn} \end{bmatrix} & = & \begin{bmatrix} \frac{W_{11}+W_{21}+\ldots+W_{n1}}{n} \\ \\ \\ \frac{W_{1n}+W_{2n}+\ldots+W_{nn}}{n} \end{bmatrix} \end{array}$$

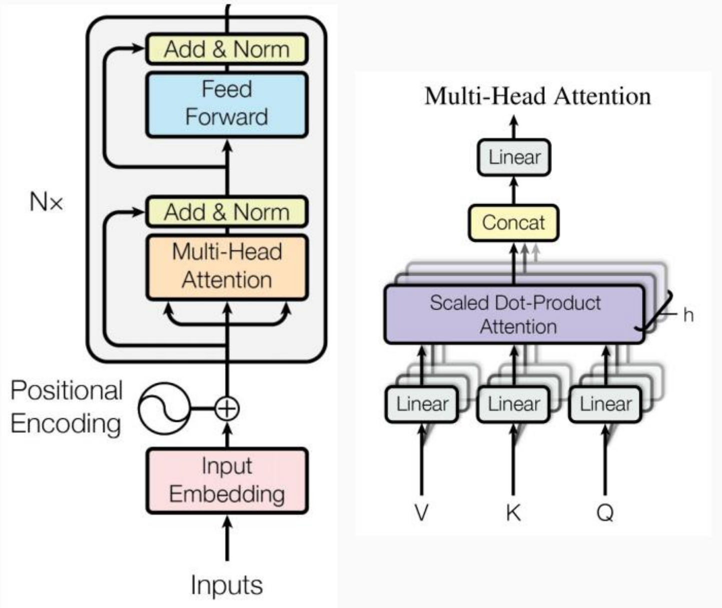
Contextual embedding

Contextual embedding refers to the representation of words or phrases in a text that takes into account their surrounding context

ELMo: Deep Contextual Word Embeddings, AI2 & University of Washington, 2017

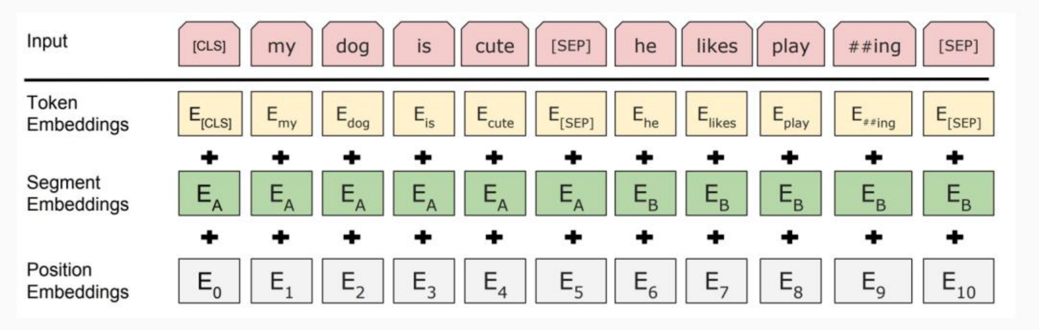


Improving Language Understanding by Generative Pre-Training, OpenAI, 2018

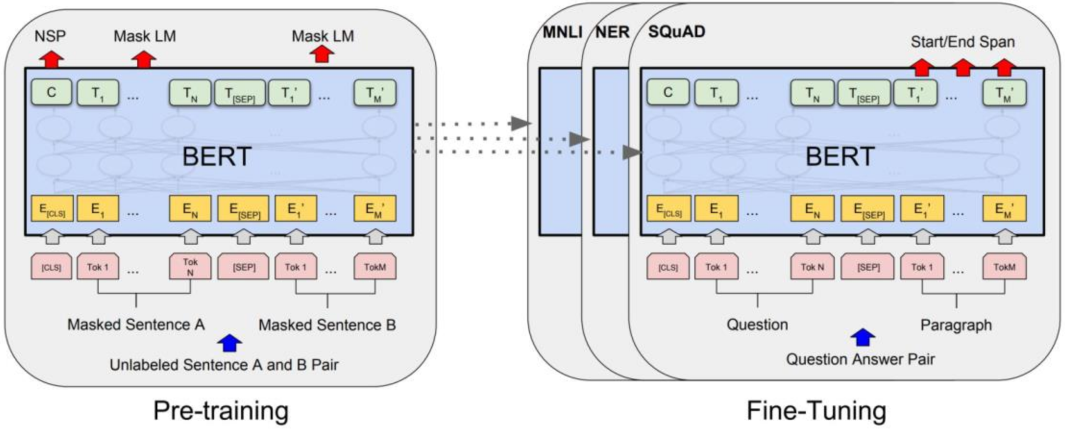


Contextual embedding

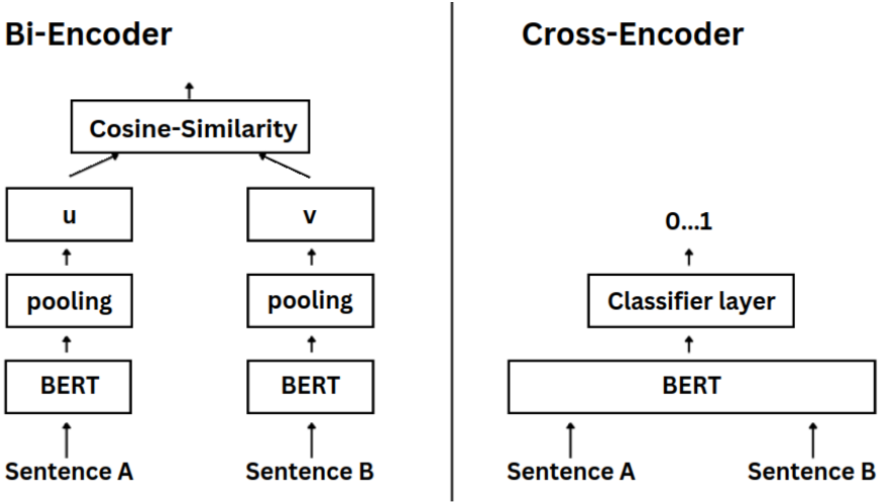
- Use 30,000 Word Piece vocabulary on input.
- Each token is sum of three embeddings
- Single sequence is much more efficient.

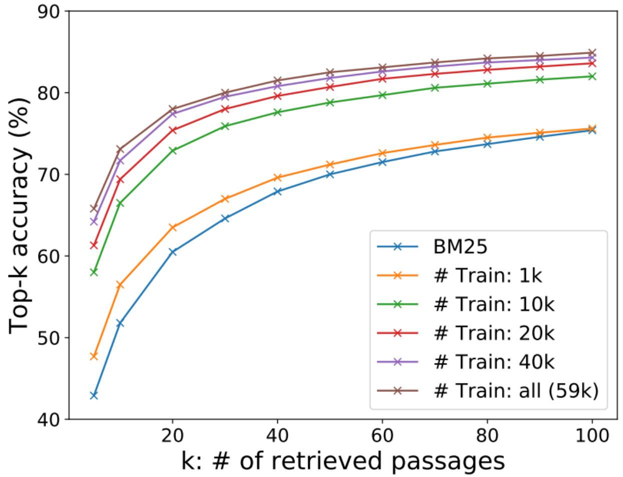


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}



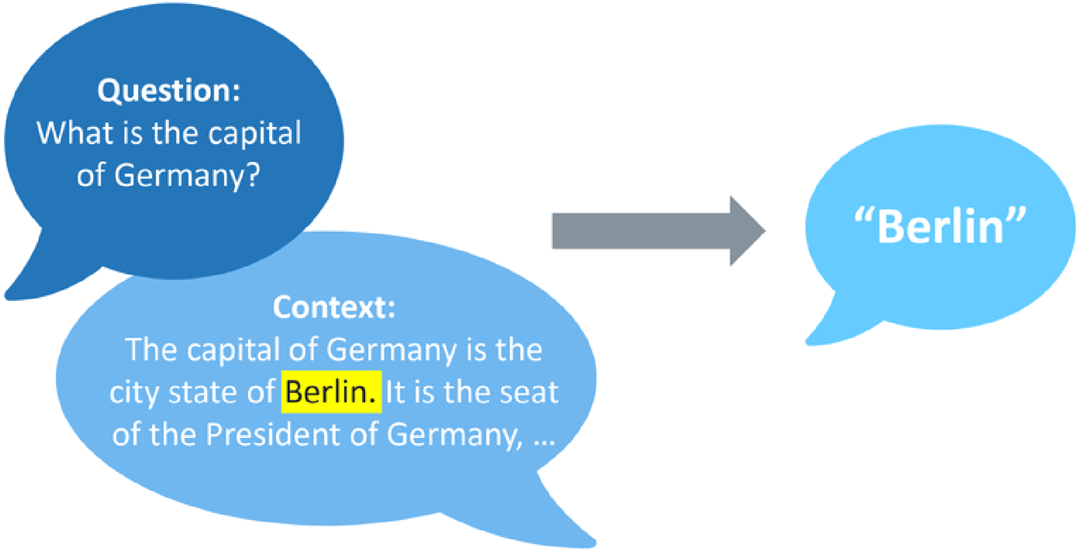
Applying Language Model in Retrieval





Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

3. Question Answering

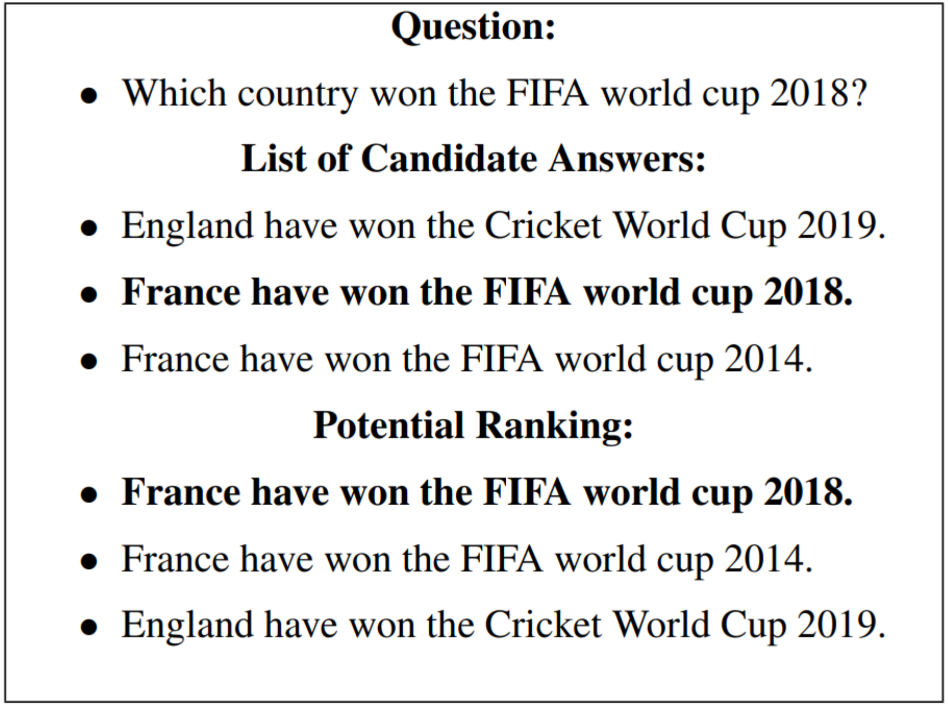


Question:
What is the capital of Germany?

Context:
The capital of Germany is the city state of Berlin. It is the seat of the President of Germany, …

“Berlin”

3. Answer Sentence Selection



Question:

- Which country won the FIFA world cup 2018?

List of Candidate Answers:

- England have won the Cricket World Cup 2019.
- **France have won the FIFA world cup 2018.**
- France have won the FIFA world cup 2014.

Potential Ranking:

- **France have won the FIFA world cup 2018.**
- France have won the FIFA world cup 2014.
- England have won the Cricket World Cup 2019.

3. Answer Sentence Selection

Input: given a question q and a text corpus $S = \{s_1, s_2, \ldots, s_n\}$

Ouput: $Answer = \{s_i | s_i \in S\}$

- A family of LSTM-based models with attention (2016–2018):

Zhiguo Wang, Wael Hamza and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. IJCAI 2017.

Yi Tay, Minh C. Phan, Luu Anh Tuan and Siu Cheung Hui. 2017 Learning to Rank Question Answer Pairs with Holographic Dual LSTM Architecture. In SIGIR 2017.

- Fine-tuning BERT-like models for reading comprehension (2019+)

Siddhant Garg, Thuy Vu, Alessandro Moschitti, TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection, in AAAI 2020

Md Tahmid Rahman Laskar, Jimmy Huang, Enamul Hoque, Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task, In LREC 2020

Bilateral Multi-Perspective Matching

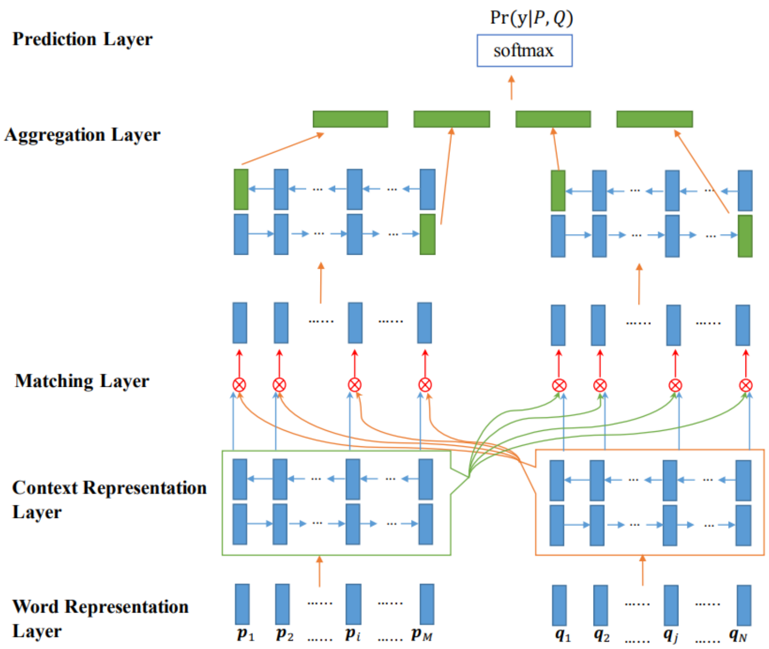


Figure 1: Architecture for Bilateral Multi-Perspective Matching (BiMPM) Model, where ⊗ is the multi-perspective matching operation described in sub-section 3.2.

Zhiguo Wang, Wael Hamza and Radu Florian. Bilateral Multi-Perspective Matching for Natural Language Sentences. IJCAI 2017.

Holographic Dual LSTM

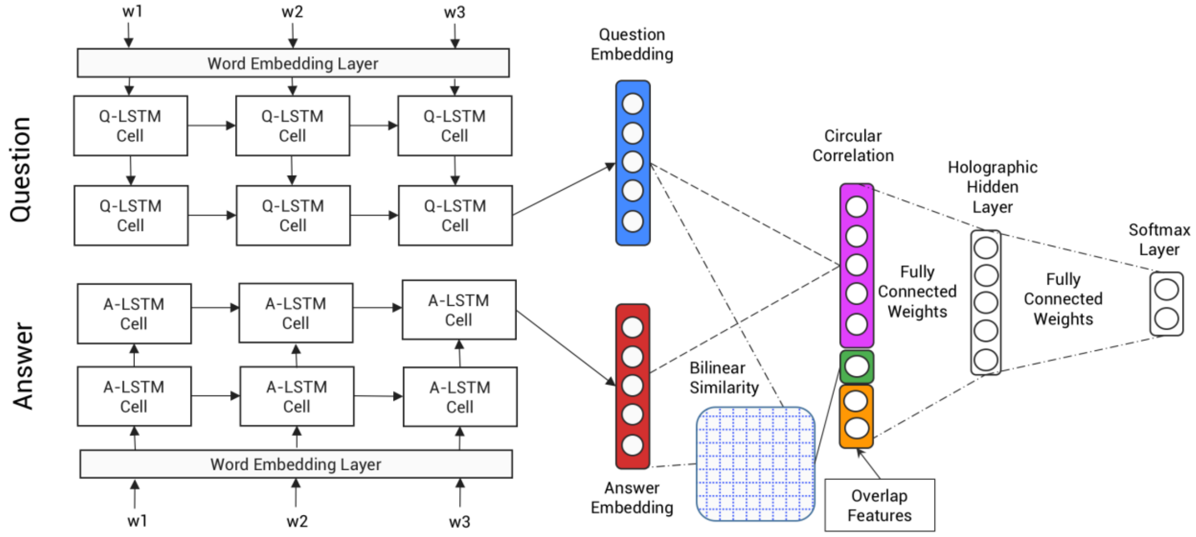


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Yi Tay, Minh C. Phan, Luu Anh Tuan and Siu Cheung Hui. 2017 Learning to Rank Question Answer Pairs with Holographic Dual LSTM Architecture. In SIGIR 2017.

Contextualized Embeddings based Transformer Encoder

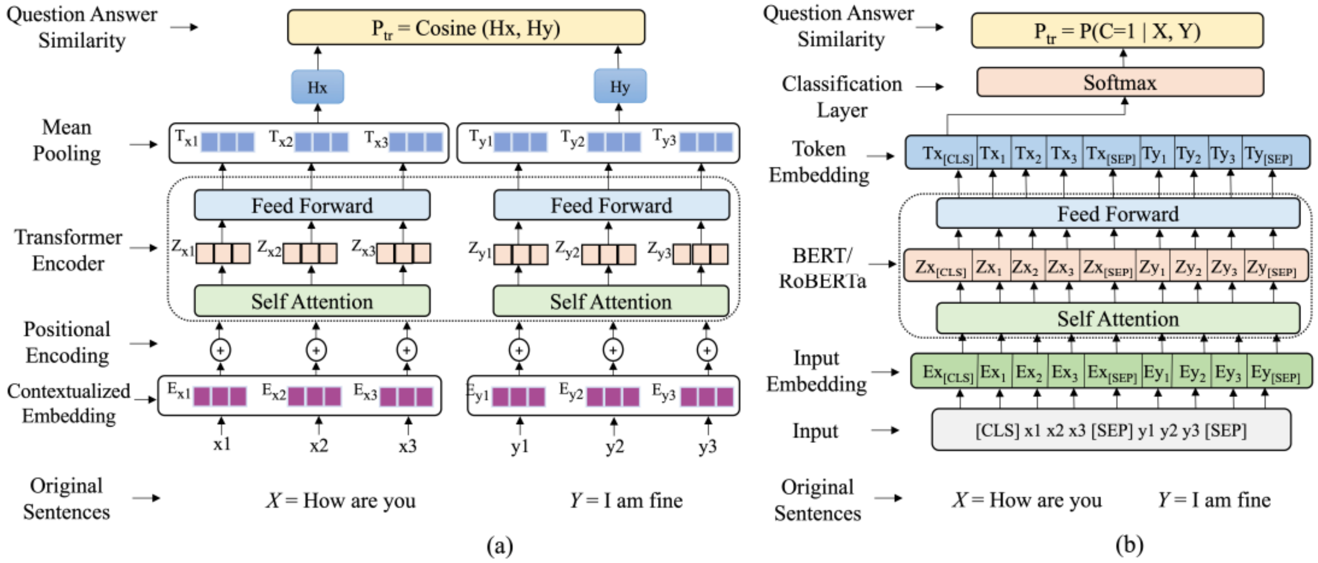
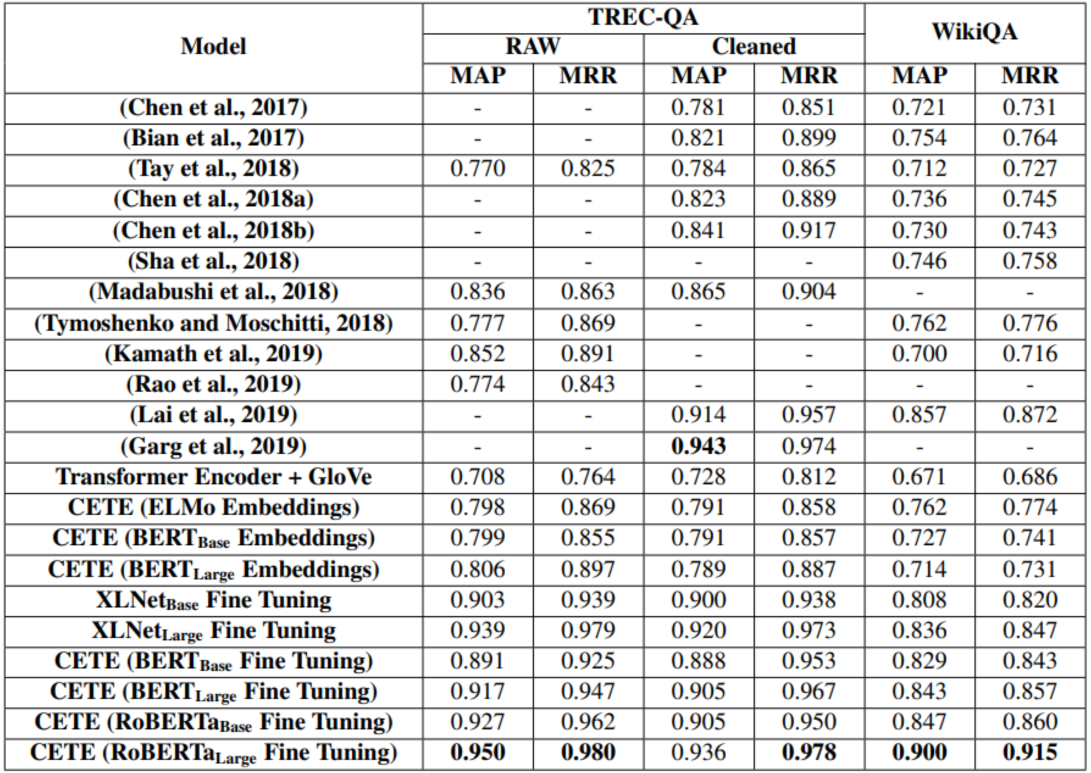


Figure 1: Our similarity modeling framework that applies contextualized embeddings: (a) Feature-based approach using the transformer encoder. (b) Fine-tuning-based approach using the BERT/RoBERTa model.

Md Tahmid Rahman Laskar, Jimmy Huang, Enamul Hoque, Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task, In LREC 2020

Performance comparisons with recent progress on TREC-QA and WikiQA datasets



Model	TREC-QA				WikiQA	
	RAW		Cleaned			
	MAP	MRR	MAP	MRR	MAP	MRR
(Chen et al., 2017)	-	-	0.781	0.851	0.721	0.731
(Bian et al., 2017)	-	-	0.821	0.899	0.754	0.764
(Tay et al., 2018)	0.770	0.825	0.784	0.865	0.712	0.727
(Chen et al., 2018a)	-	-	0.823	0.889	0.736	0.745
(Chen et al., 2018b)	-	-	0.841	0.917	0.730	0.743
(Sha et al., 2018)	-	-	-	-	0.746	0.758
(Madabushi et al., 2018)	0.836	0.863	0.865	0.904	-	-
(Tymoshenko and Moschitti, 2018)	0.777	0.869	-	-	0.762	0.776
(Kamath et al., 2019)	0.852	0.891	-	-	0.700	0.716
(Rao et al., 2019)	0.774	0.843	-	-	-	-
(Lai et al., 2019)	-	-	0.914	0.957	0.857	0.872
(Garg et al., 2019)	-	-	**0.943**	0.974	-	-
Transformer Encoder + GloVe	0.708	0.764	0.728	0.812	0.671	0.686
CETE (ELMo Embeddings)	0.798	0.869	0.791	0.858	0.762	0.774
CETE (BERT$_{Base}$ Embeddings)	0.799	0.855	0.791	0.857	0.727	0.741
CETE (BERT$_{Large}$ Embeddings)	0.806	0.897	0.789	0.887	0.714	0.731
XLNet$_{Base}$ Fine Tuning	0.903	0.939	0.900	0.938	0.808	0.820
XLNet$_{Large}$ Fine Tuning	0.939	0.979	0.920	0.973	0.836	0.847
CETE (BERT$_{Base}$ Fine Tuning)	0.891	0.925	0.888	0.953	0.829	0.843
CETE (BERT$_{Large}$ Fine Tuning)	0.917	0.947	0.905	0.967	0.843	0.857
CETE (RoBERTa$_{Base}$ Fine Tuning)	0.927	0.962	0.905	0.950	0.847	0.860
CETE (RoBERTa$_{Large}$ Fine Tuning)	**0.950**	**0.980**	0.936	**0.978**	**0.900**	**0.915**

Performance comparisons

Advantages of LSTM-based model:

Resource Efficiency: less computationally intensive, making them more accessible for smaller-scale projects or when computational resources are limited.

Customizability: can be tailored to specific tasks by adjusting the number of layers, hidden units, and other hyperparameters.

Disadvantages of LSTM-based Models:

Limited Context Understanding: difficulty capturing long-range dependencies and complex contextual relationships in text, which is a limitation for many natural language processing tasks.

Data Dependency: require a substantial amount of labeled training data for each specific task, which may not be available or practical for all applications.

Advantages of BERT:

Contextual Understanding: excels at capturing contextual information by considering the entire input text bidirectionally, enabling it to understand nuances and complex language relationships.

Transfer Learning: is pre-trained on a vast amount of text data, making it highly effective for transfer learning across a wide range of natural language processing tasks. Fine-tuning BERT models often requires less task-specific labeled data.

Large-scale Pre-training: BERT benefits from extensive pre-training on large corpora, giving it a strong understanding of language.

Disadvantages of BERT:

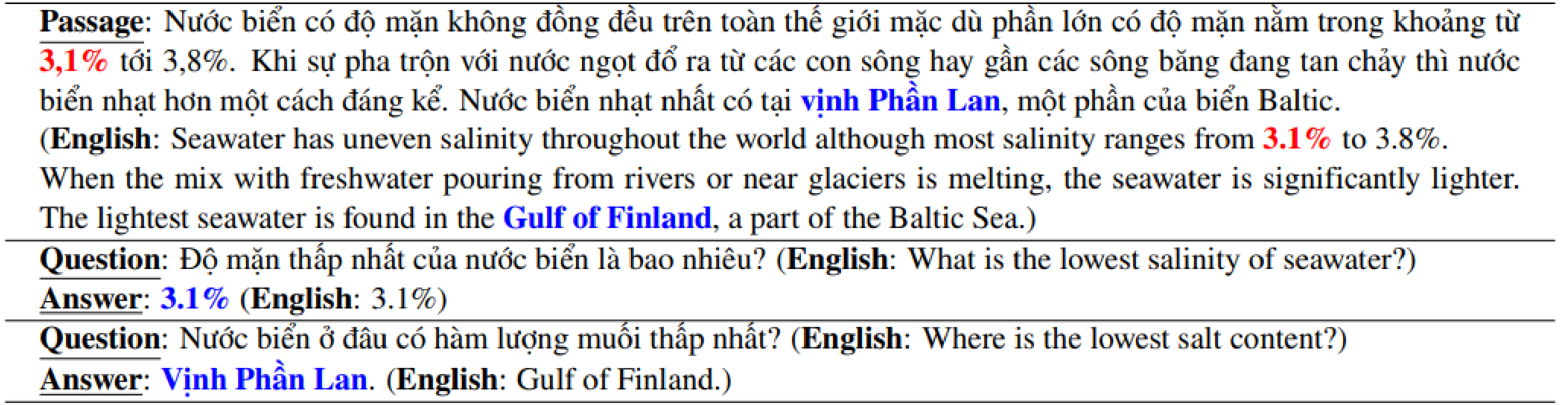
Computational Intensity: BERT models are computationally demanding, requiring powerful hardware and significant training time. This can be a drawback for smaller projects or those with limited resources.

Complexity: BERT's architecture is more complex than LSTMs, making it less interpretable and challenging to fine-tune for specific domain or task requirements.

Need for Pre-trained Models: BERT models are most effective when fine-tuned from pre-trained models, which means they may not be as suitable for tasks that lack access to relevant pre-trained models or a substantial amount of task-specific data.

Reading Comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) ⟶ A



Passage: Nước biển có độ mặn không đồng đều trên toàn thế giới mặc dù phần lớn có độ mặn nằm trong khoảng từ **3,1%** tới 3,8%. Khi sự pha trộn với nước ngọt đổ ra từ các con sông hay gần các sông băng đang tan chảy thì nước biển nhạt hơn một cách đáng kể. Nước biển nhạt nhất có tại **vịnh Phần Lan**, một phần của biển Baltic.
(**English**: Seawater has uneven salinity throughout the world although most salinity ranges from **3.1%** to 3.8%. When the mix with freshwater pouring from rivers or near glaciers is melting, the seawater is significantly lighter. The lightest seawater is found in the **Gulf of Finland**, a part of the Baltic Sea.)

Question: Độ mặn thấp nhất của nước biển là bao nhiêu? (**English**: What is the lowest salinity of seawater?)
Answer: **3.1%** (**English**: 3.1%)

Question: Nước biển ở đâu có hàm lượng muối thấp nhất? (**English**: Where is the lowest salt content?)
Answer: **Vịnh Phần Lan**. (**English**: Gulf of Finland.)

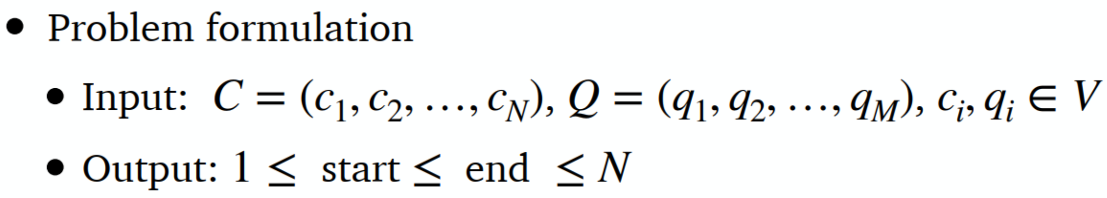
Metric

Evaluation: exact match (0 or 1) and F1 (partial credit).

- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to *each* gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- Estimated human performance: EM = 82.3, F1 = 91.2

Q: What did Tesla do in December 1878?
A: {left Graz, left Graz, left Graz and severed all relations with his family}
Prediction: {left Graz and served}
Exact match: max{0, 0, 0} = 0 F1: max{0.67, 0.67, 0.61} = 0.67

Reading Comprehension



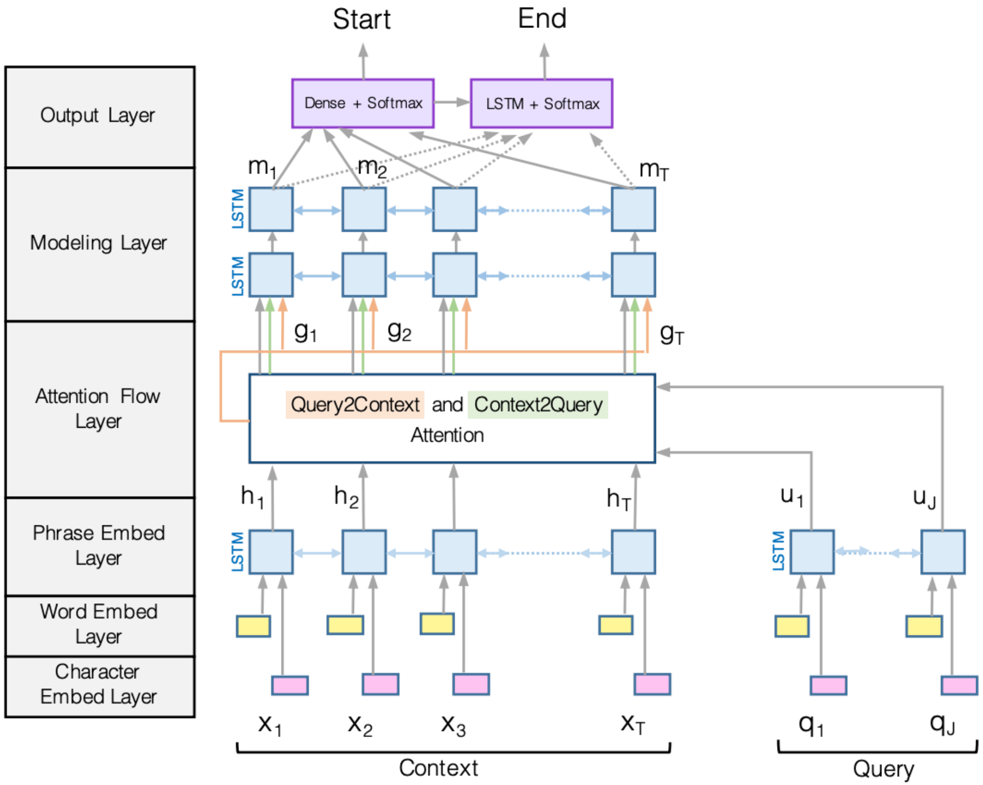
- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N), Q = (q_1, q_2, \ldots, q_M), c_i, q_i \in V$
  - Output: $1 \leq$ start $\leq$ end $\leq N$

A family of LSTM-based models with attention (2016–2018)

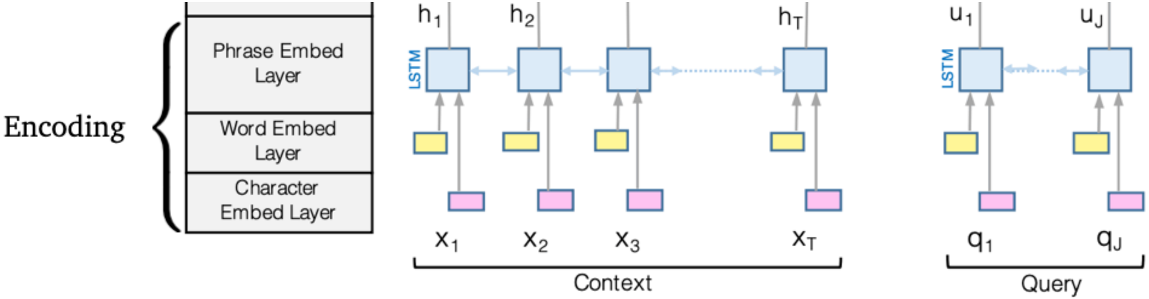
Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), MatchLSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..

- Fine-tuning BERT-like models for reading comprehension (2019+)

BiDAF: the Bidirectional Attention Flow



BiDAF: the Bidirectional Attention Flow



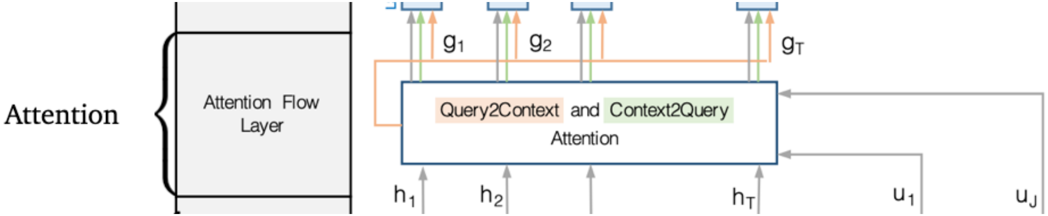
- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query.

$$e(c_i) = f([\text{GloVe}(c_i); \text{charEmb}(c_i)]) \qquad e(q_i) = f([\text{GloVe}(q_i); \text{charEmb}(q_i)])$$

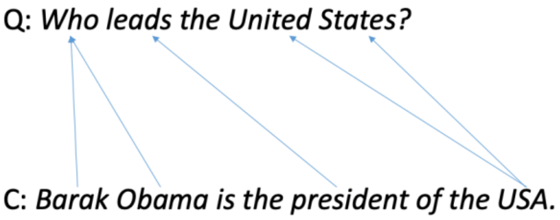
- Then, use two **bidirectional** LSTMs separately to produce contextual embeddings for both context and query.

$$\begin{aligned}
\overrightarrow{\mathbf{c}}_i &= \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \\
\overleftarrow{\mathbf{c}}_i &= \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \\
\mathbf{c}_i &= [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}
\end{aligned}
\qquad
\begin{aligned}
\overrightarrow{\mathbf{q}}_i &= \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H \\
\overleftarrow{\mathbf{q}}_i &= \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H \\
\mathbf{q}_i &= [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}
\end{aligned}$$

BiDAF: the Bidirectional Attention Flow



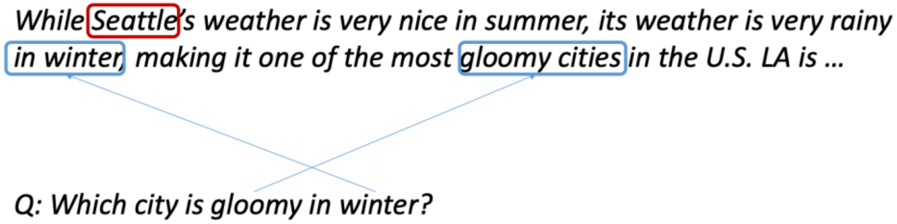
- Context-to-query attention: For each context word, choose the most relevant words from the query words.
- Query-to-context attention: choose the context words that are most relevant to one of query words.



Q: *Who leads the United States?*

C: *Barak Obama is the president of the USA.*

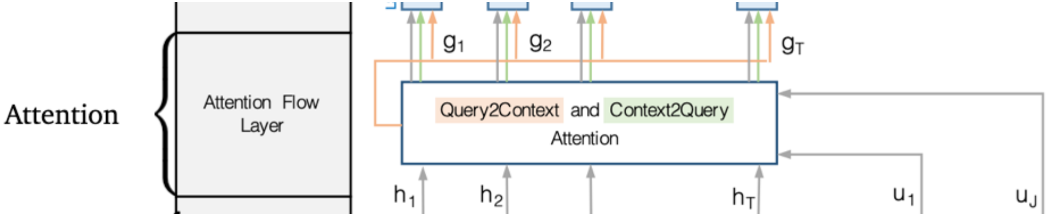
For each context word, find the most relevant query word.



While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is …

Q: *Which city is gloomy in winter?*

BiDAF: the Bidirectional Attention Flow



- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\mathsf{T}}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

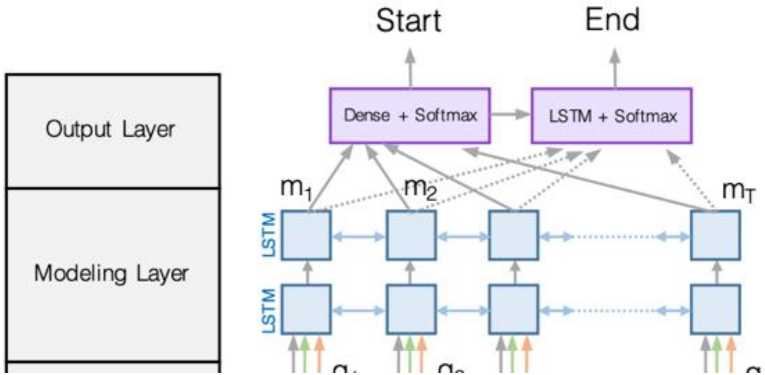
- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query-to-context attention (which context words are relevant to some question words):

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

BiDAF: the Bidirectional Attention Flow

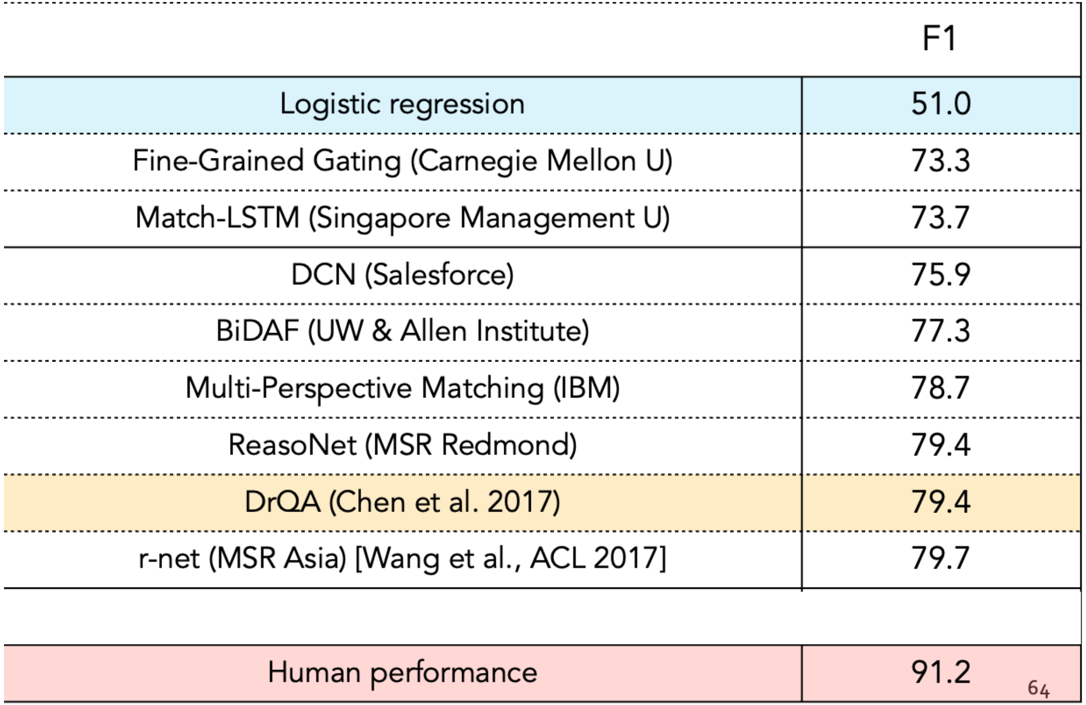


Modeling layer: pass to another two layers of **bi-directional** LSTMs.

- Attention layer is modeling interactions between query and context
- Modeling layer is modeling interactions within context words

Output layer: two classifiers predicting the start and end positions

BiDAF: the Bidirectional Attention Flow

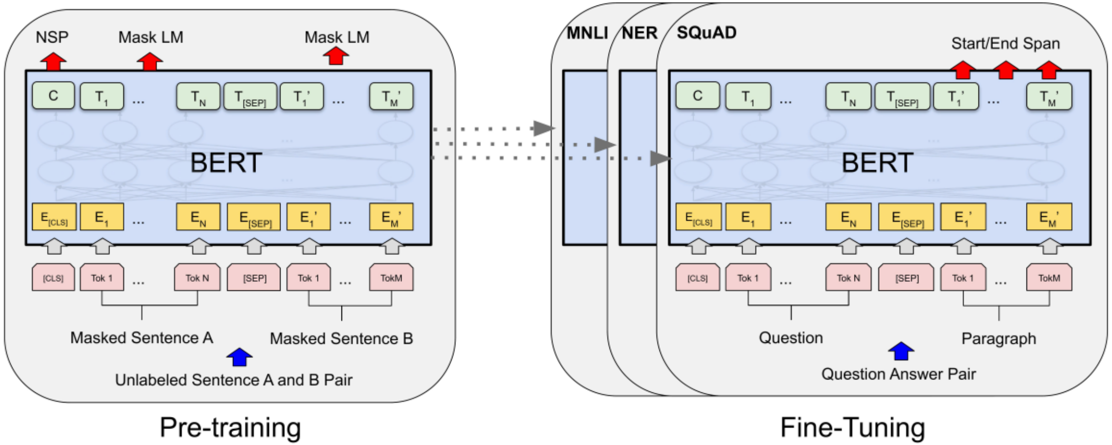


	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

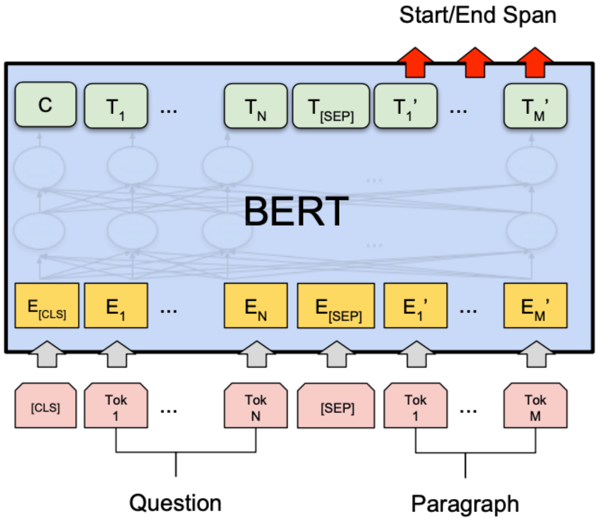
64

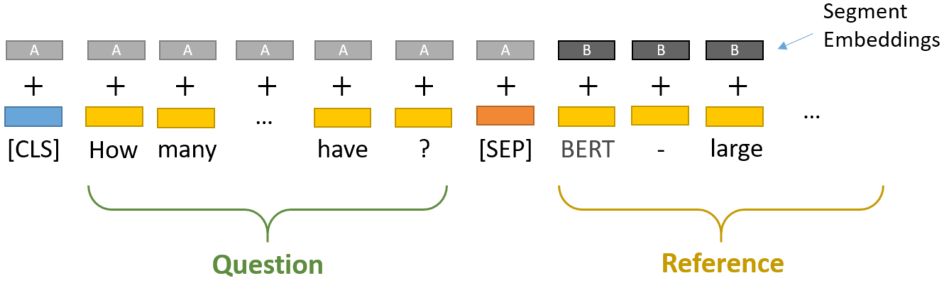
BERT for reading comprehension

- BERT is a deep bidirectional Transformer encoder pre-trained on large amounts of text (Wikipedia + BooksCorpus)
- BERT is pre-trained on two training objectives:
  - ○ Masked language model (MLM)
  - ○ Next sentence prediction (NSP)
- BERTbase has 12 layers and 110M parameters, BERTlarge has 24 layers and 330M parameters



BERT for reading comprehension





Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

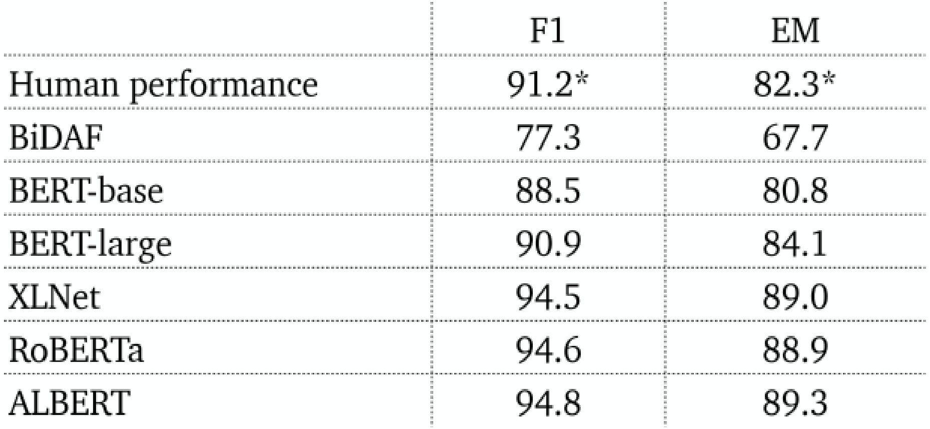
Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B

BERT for reading comprehension

It works amazing well. Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pretrained models.



	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

Compare

- BERT model has many many more parameters (110M or 330M)

BiDAF has ~2.5M parameters.

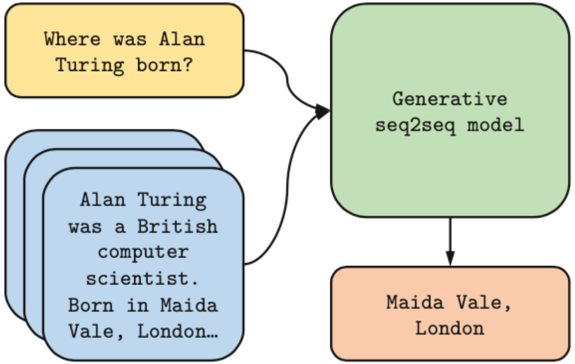
- BiDAF is built on top of several bidirectional LSTMs while BERT is

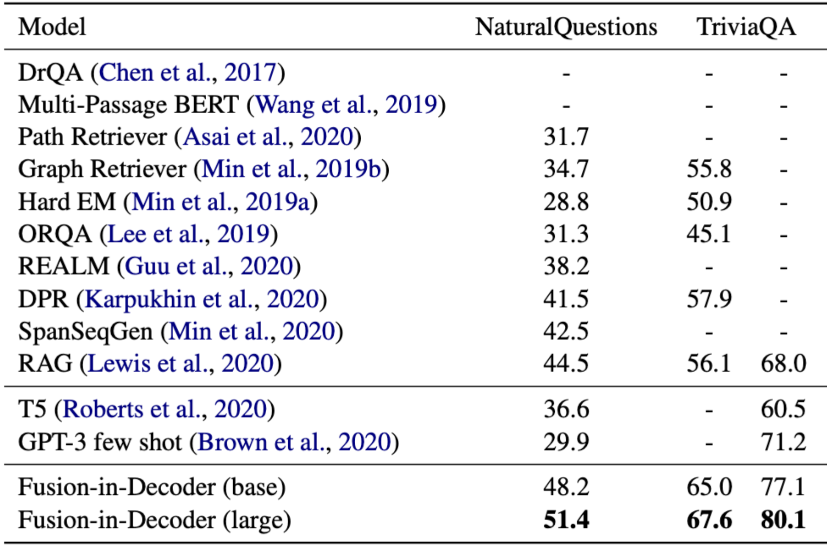
  built on top of

Transformers (no recurrence architecture and easier to parallelize).

- BERT is **pre-trained** while BiDAF is only built on top of GloVe (and all the remaining parameters need to be learned from the supervision datasets)

Generative question answering

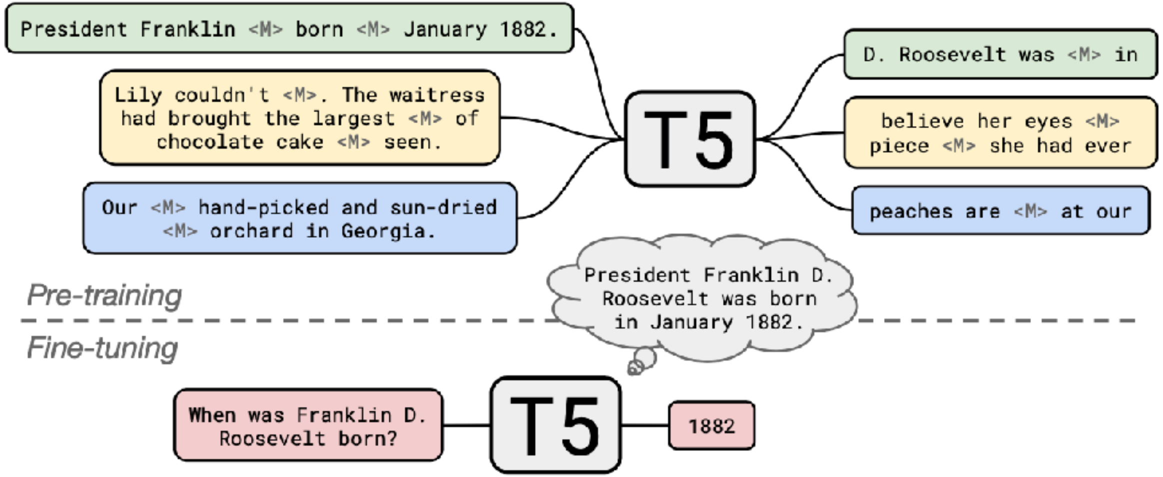

Where was Alan Turing born?
Alan Turing was a British computer scientist. Born in Maida Vale, London…
Generative seq2seq model
Maida Vale, London



Model	NaturalQuestions	TriviaQA	
DrQA (Chen et al., 2017)	-	-	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-
Path Retriever (Asai et al., 2020)	31.7	-	-
Graph Retriever (Min et al., 2019b)	34.7	55.8	-
Hard EM (Min et al., 2019a)	28.8	50.9	-
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

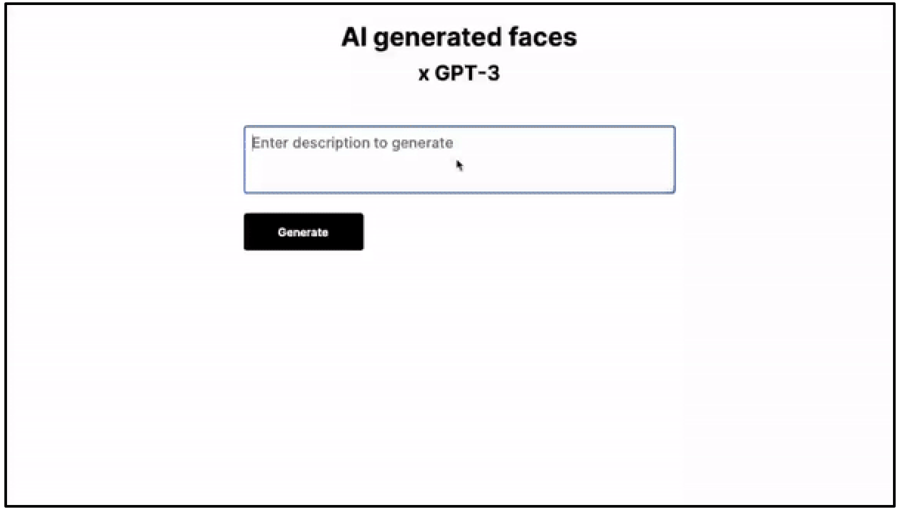
Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Generative Model


President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
President Franklin D. Roosevelt was born in January 1882.
Pre-training
Fine-tuning
When was Franklin D. Roosevelt born?
T5
1882

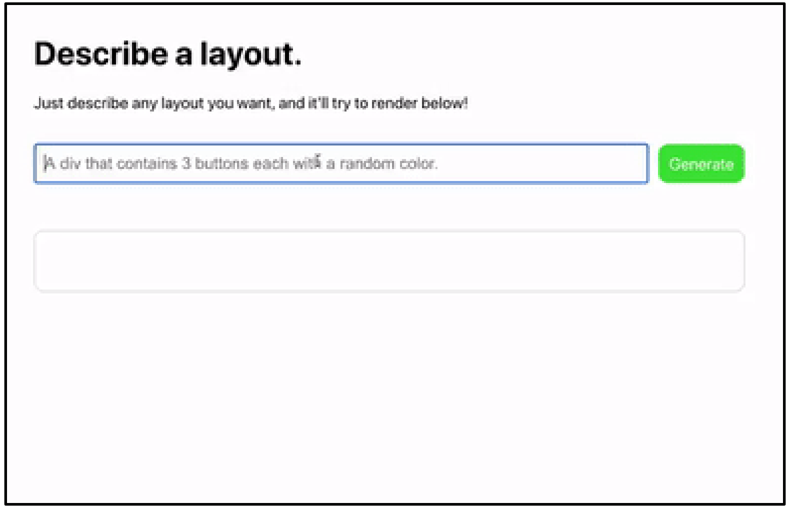
*: Pre-trained on a multitask mixture including an **unsupervised "span corruption" task** on unlabeled text as well as supervised translation, summarization, classification, and reading comprehension tasks

ChatGPT


AI generated faces

x GPT-3

Enter description to generate


Describe a layout.

Just describe any layout you want, and it'll try to render below!

A div that contains 3 buttons each with a random color.

Q&A

Summary

- Retrieval and Question Answering Introduction
- Retrieval-based Question Answering
  - Traditional Method
  - Deep Learning Method
  - Hybrid
- Question Answering
  - Answer Sentence Selection
  - Reading Comprehension
  - Generative Model

Thank you

Email me
yenvth@vnu.edu.vn

Course: Large Language Models and Its Applications
sponsored by **KEPCO KDN Co., Ltd.**
Eco-friendly & Digital Centered Energy ICT Platform Leader