Question Answering part 2

Module 9

THE UNIVERSITY of ADELAIDE

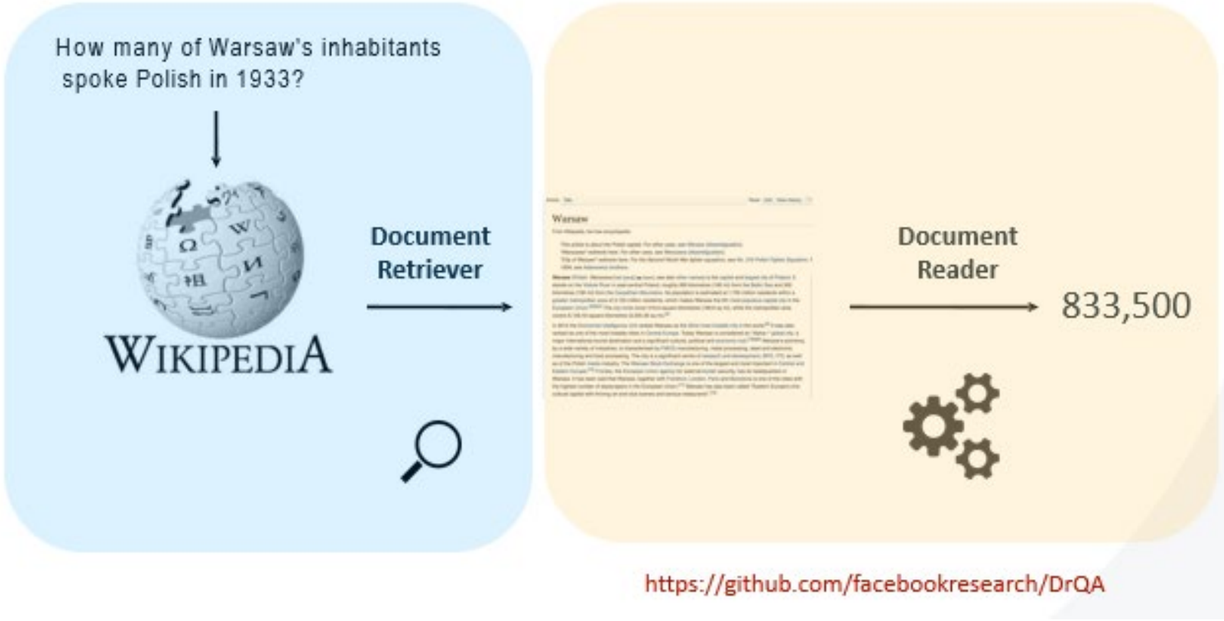
Topics for this session

- In this session we will look at some of QA systems, and also will try thinking what might be useful for your assignment 2.
- IR based QA (related to your assignment). Open vs closed domain QA?
- Evaluation of Factoid based QA: MRR vs MAP
- Using SQUaD for training and evaluation of QA systems
- Neural based QA systems
  - BIDAF, T5, SpanBERT

THE UNIVERSITY of ADELAIDE

Information Retrieval QA

Retriever-reader framework



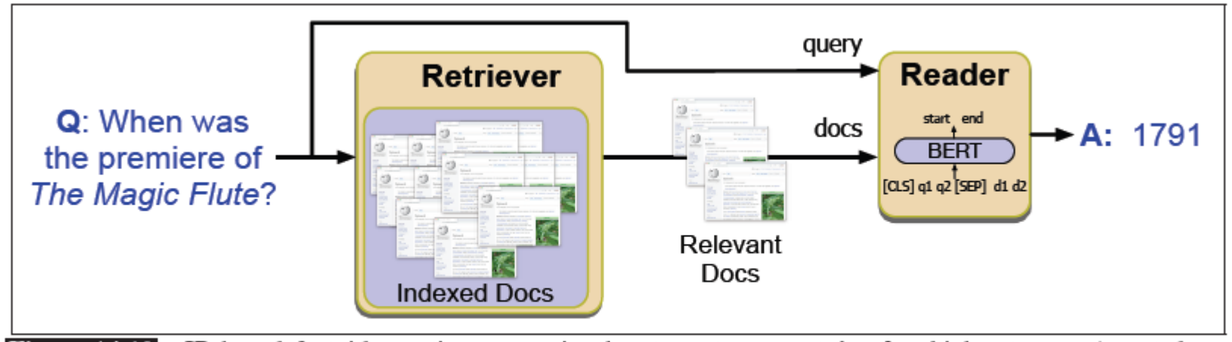


Figure 14.10 IR-based factoid question answering has two stages: **retrieval**, which returns relevant documents from the collection, and **reading**, in which a neural reading comprehension system extracts answer spans.

- Input: a large collection of documents $\mathcal{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A
- Retriever: $f(\mathcal{D}, Q) \longrightarrow P_1, \ldots, P_K$
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$

K is pre-defined (e.g., 100)

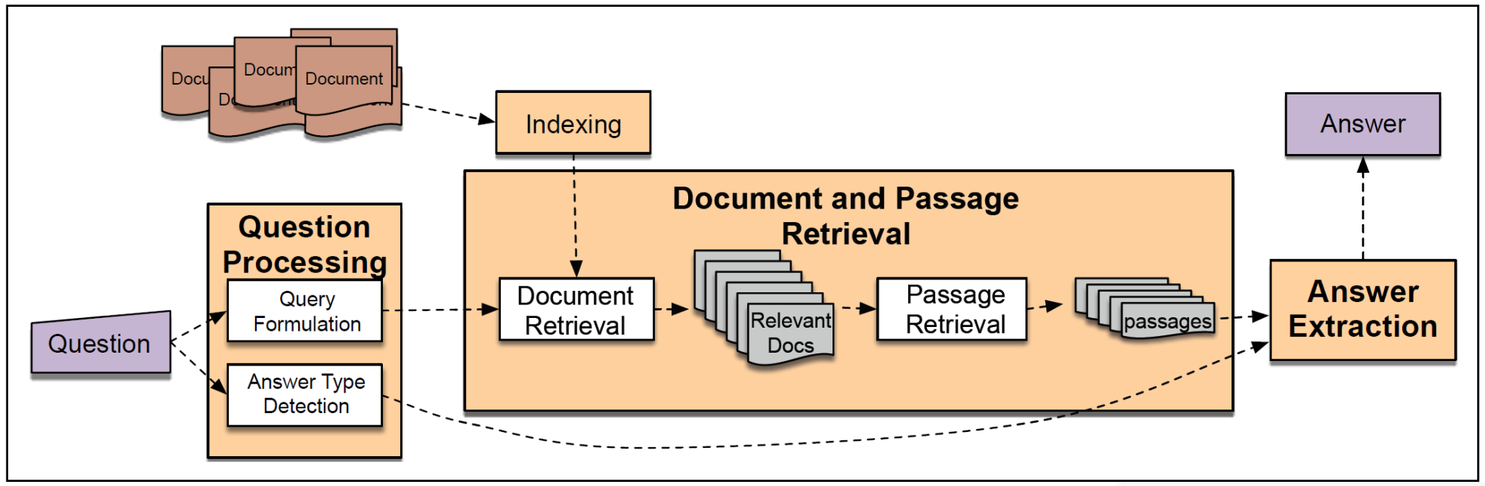
A reading comprehension problem!

THE UNIVERSITY of ADELAIDE

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

IR-based Factoid QA

- **Question Processing**
  - Detect question type, answer type, focus, relations
  - Formulate a query to send to **Retriever**
- **Retrieve a ranked list of documents (Retriever)**
- **Passage Retrieval (Reader)**
  - Retrieve ranked documents
  - Break into suitable passages and rerank.
- **Answer Processing (Reader)**
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources



Question Processing

Answer Type Detection

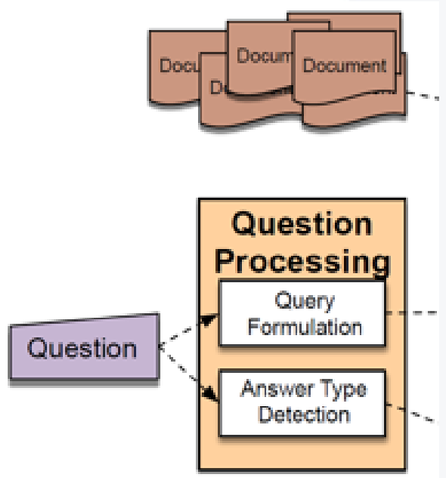
- Decide the named entity type (person, place) of the answer

Query Formulation

- Create query for the IR system, and make it a declarative sentence

Question Type classification

- Is this a definition question, a math question, a list question?



Focus Detection

- Find the question words that are replaced by the answer

  e.g. *When hit by electrons, a phosphor gives off electromagnetic energy in this form (which form…).*

Relation Extraction

- Find relations between entities in the question

  e.g., *They're the two states you could be reentering if you're crossing Florida's northern border.*

  borders(Florida, ?x, north)

THE UNIVERSITY of ADELAIDE

Document and Passage Retrieval

Step 1: IR engine retrieves documents using query terms

Step 2: Segment the documents into shorter units, i.e., passages

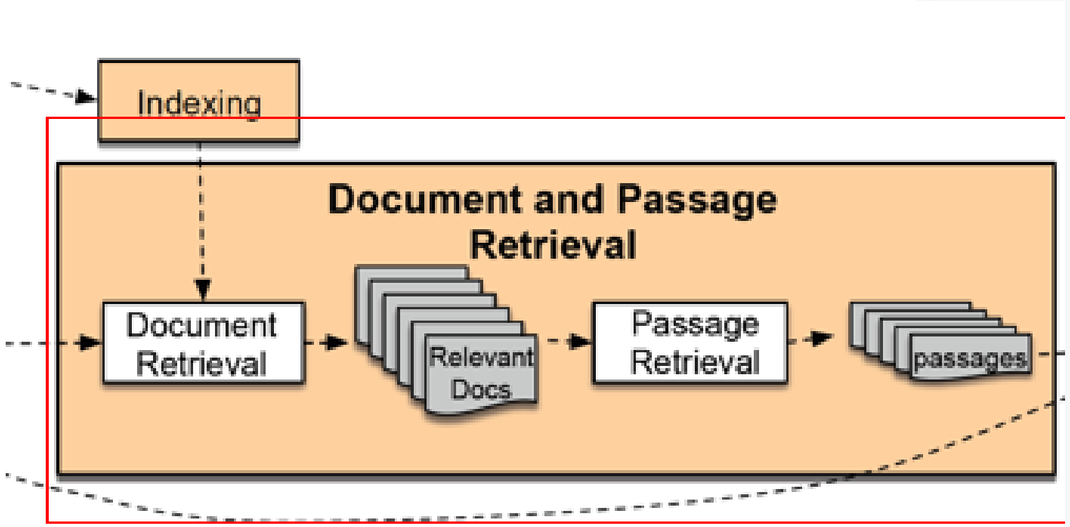
Could be sections, paragraphs, or sentences

Step 3: Passage ranking

Use answer type to help re-rank passages.

Use features for ranking, e.g:

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N--grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words in the passage
- Rank of the document containing passage
- …



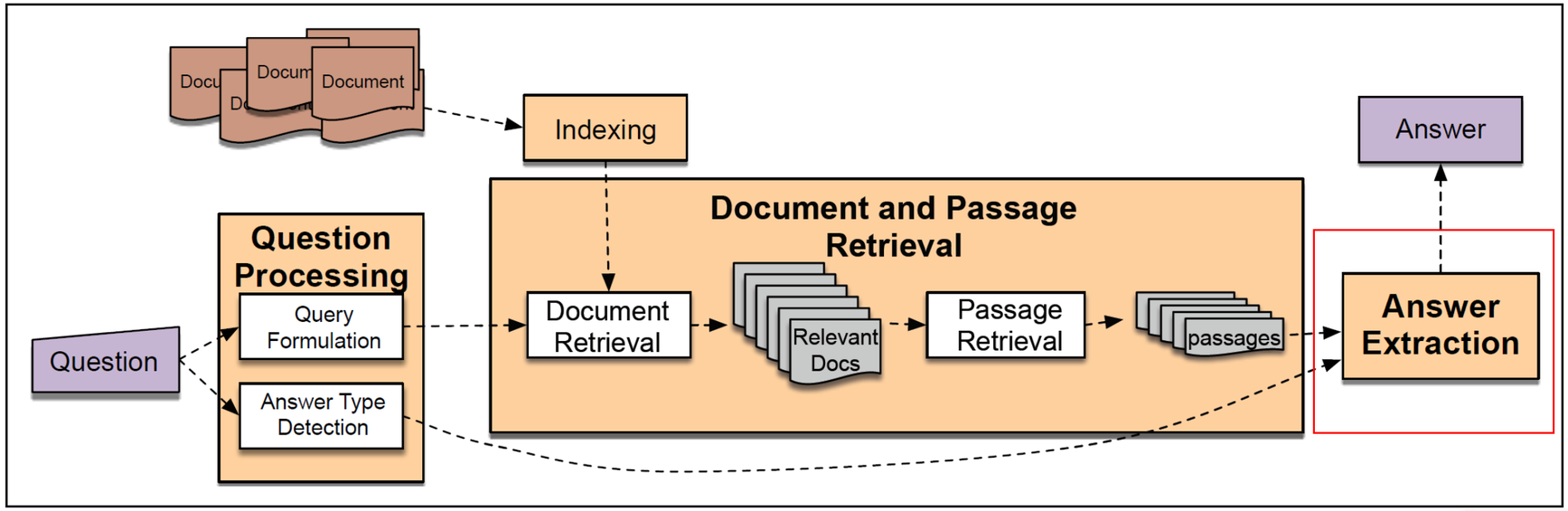
THE UNIVERSITY of ADELAIDE

Answer Extraction

Type-based Answer Extraction

Feature-based Answer Extraction

Neural Answer Extraction



Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

THE UNIVERSITY of ADELAIDE

Answer Extraction: Type-based

Run an answer--type named--entity tagger on the passages

Each answer type requires a named--entity tagger that detects it.

If answer type is CITY, tagger has to tag CITY

Return the string with the right type:

Who is the prime minister of India (PERSON)

Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.

How tall is Mt. Everest? (LENGTH)

The official height of Mount Everest is 29035 feet!

THE UNIVERSITY of ADELAIDE

Answer Extraction: Feature-based

- Answer type match: Candidate contains a phrase with the correct answer type.
- Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.
- Keyword distance: Distance in words between the candidate and query keywords
- Pattern match: Regular expression pattern matches the candidate.
- Question keywords: # of question keywords in the candidate.
- Apposition features: The candidate is an appositive to question terms
  - e.g., *John, the dentist*
- Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

THE UNIVERSITY of ADELAIDE

Machine Comprehension (Burges 2013)

- "A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

THE UNIVERSITY of ADELAIDE

Reading comprehension

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

- Useful for many practical applications
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
- Many other NLP tasks can be reduced to a reading comprehension problem:

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

Answer Extraction: Neural Approach

Neural network approaches to answer extraction draw on the intuition that a question and its answer are semantically similar in some appropriate way. This intuition can be fleshed out by

- computing an embedding for the question and an embedding for each token of the passage, and
- then selecting passage spans whose embeddings are closest to the question embedding.
  - parsing questions into phrases,
  - then embedding phrases of Q and A,
  - and comparing embeddings

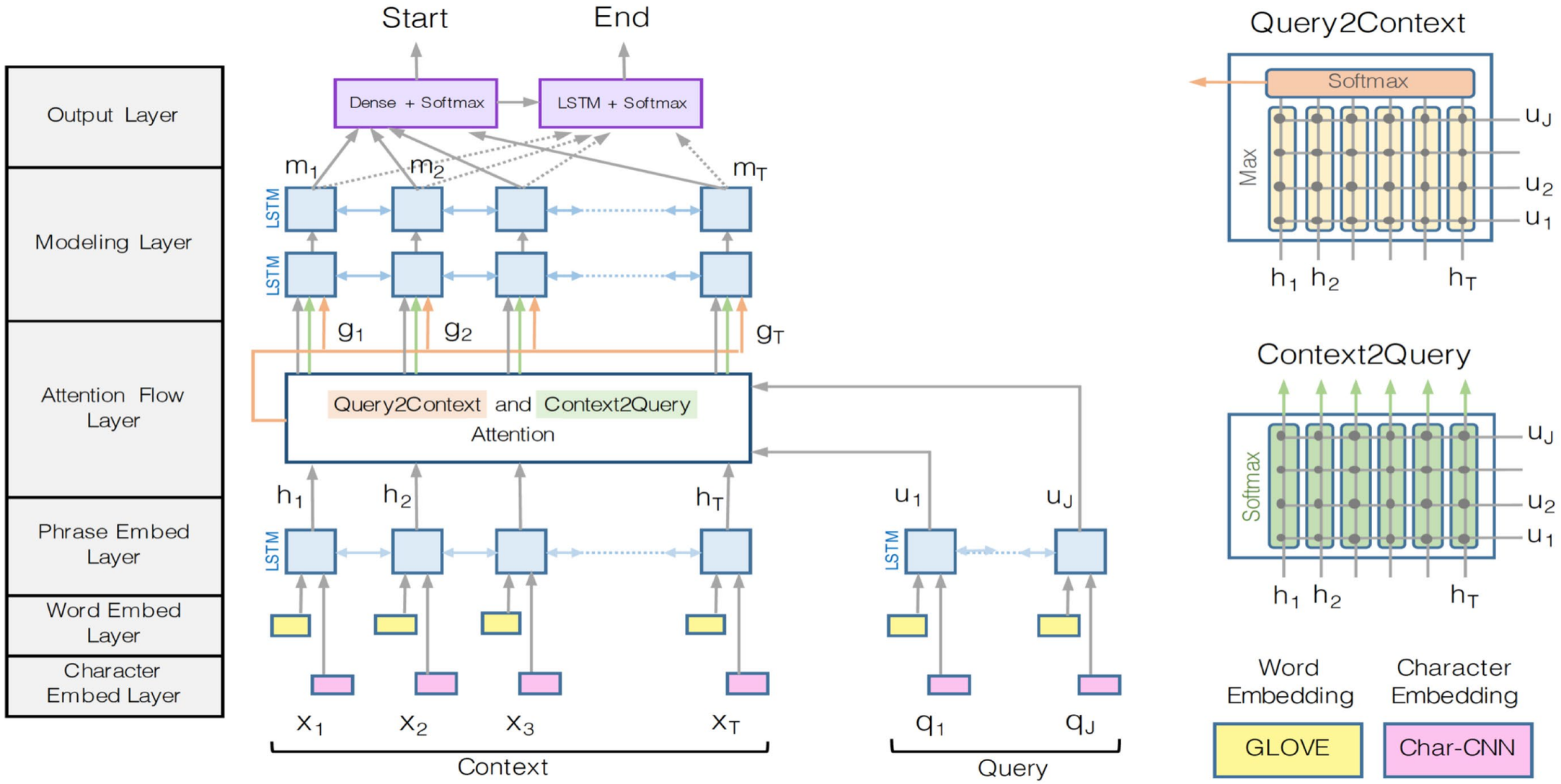
THE UNIVERSITY of ADELAIDE

Modules in BiDAF

1. **Character Embedding Layer** maps each word to a vector space using character-level CNNs.
2. **Word Embedding Layer** maps each word to a vector space using a pre-trained word embedding model.
3. **Contextual Embedding Layer** utilizes contextual cues from surrounding words to refine the embedding of the words. These first three layers are applied to both the query and context.
4. **Attention Flow Layer** couples the query and context vectors and produces a set of query-aware feature vectors for each word in the context.
5. **Modeling Layer** employs a Recurrent Neural Network to scan the context.
6. **Output Layer** provides an answer to the query.

- First three layers separately applied to query and context

BiDAF: the Bidirectional Attention Flow model



Stanford cs224n-2019-lecture10-QA

BiDAF

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is **the Attention Flow layer**
- **Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with **w** of dimension 6*d*):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T [\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention:
(which query words are most relevant to each context word)

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF

- **Attention Flow Idea:** attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  (the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max)

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$

Stanford cs224n-2019-lecture10-QA

BiDAF

- There is then a “modelling” layer:
  - Another deep (2-layer) BiLSTM over the passage
- And answer span selection is more complex:
  - Start: Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax
  - End: Put output of modelling layer M through another BiLSTM to give M_2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax

BiDAF: Performance on SQuAD

Ablation test

This model achieved 77.3 F1 on SQuAD v1.1.

- Without context-to-query attention ⟹ 67.7 F1
- Without query-to-context attention ⟹ 73.7 F1
- Without character embeddings ⟹ 75.4 F1

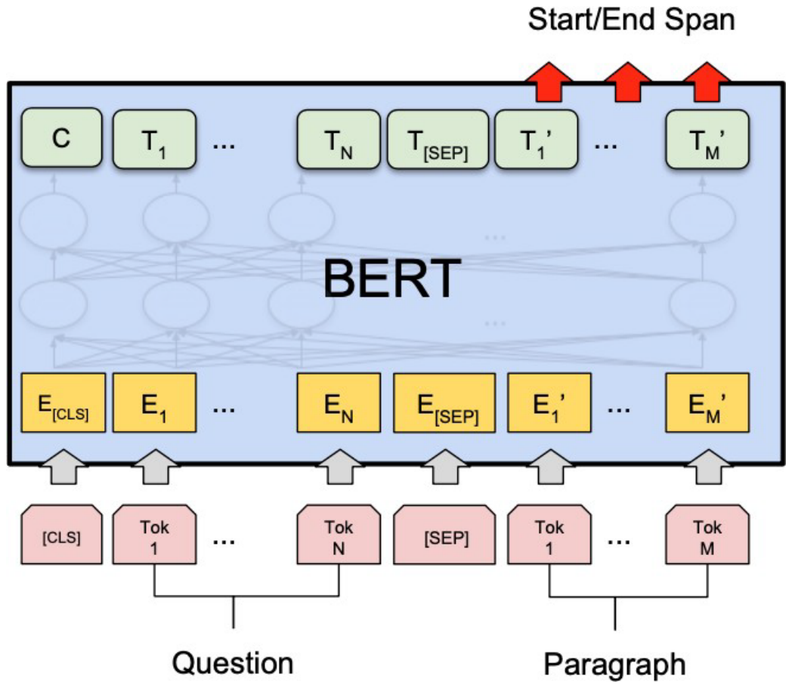
	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]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

64

THE UNIVERSITY of ADELAIDE

(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

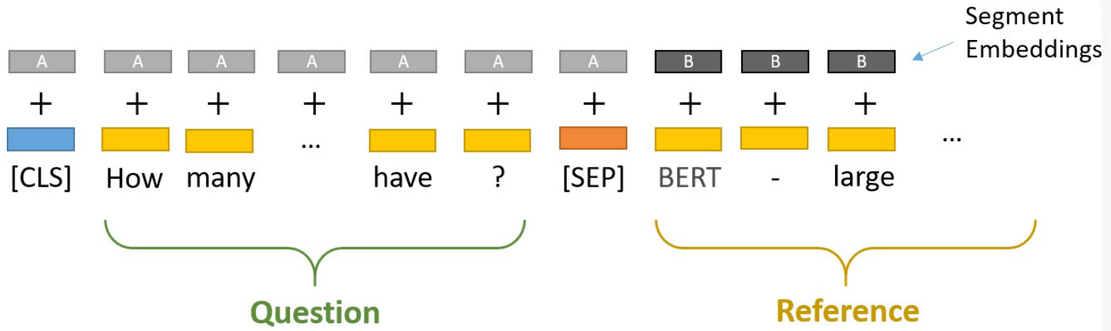
BERT for reading comprehension



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B



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.`

Image credit: https://mccormickml.com/

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^\intercal \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^\intercal \mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

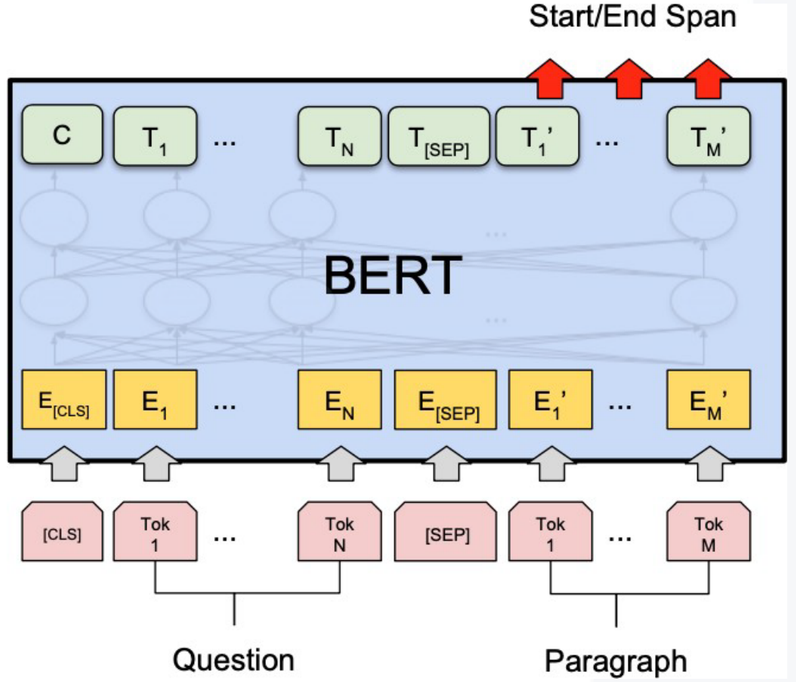
Stanford cs224n-2023-lecture12-QA

THE UNIVERSITY of ADELAIDE

BERT for reading comprehension

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

- All the BERT parameters (e.g., 110M) as well as the newly introduced parameters $\mathbf{h}_{\text{start}}, \mathbf{h}_{\text{end}}$ (e.g., 768 x 2 = 1536) are optimized together for $\mathcal{L}$.
- It works amazing well. Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pre-trained models (EM is Exact Match).



	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

(dev set, except for human performance)

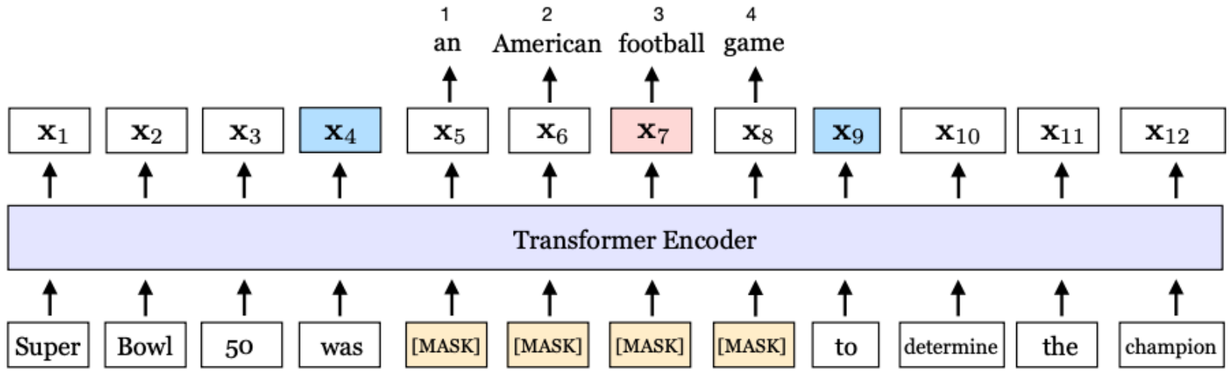
BiDAF vs BERT models

- 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).

Pre-training is clearly a game changer but it is expensive..

spanBERT

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



Two ideas:

1) masking contiguous spans of words instead of 15% random words

2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

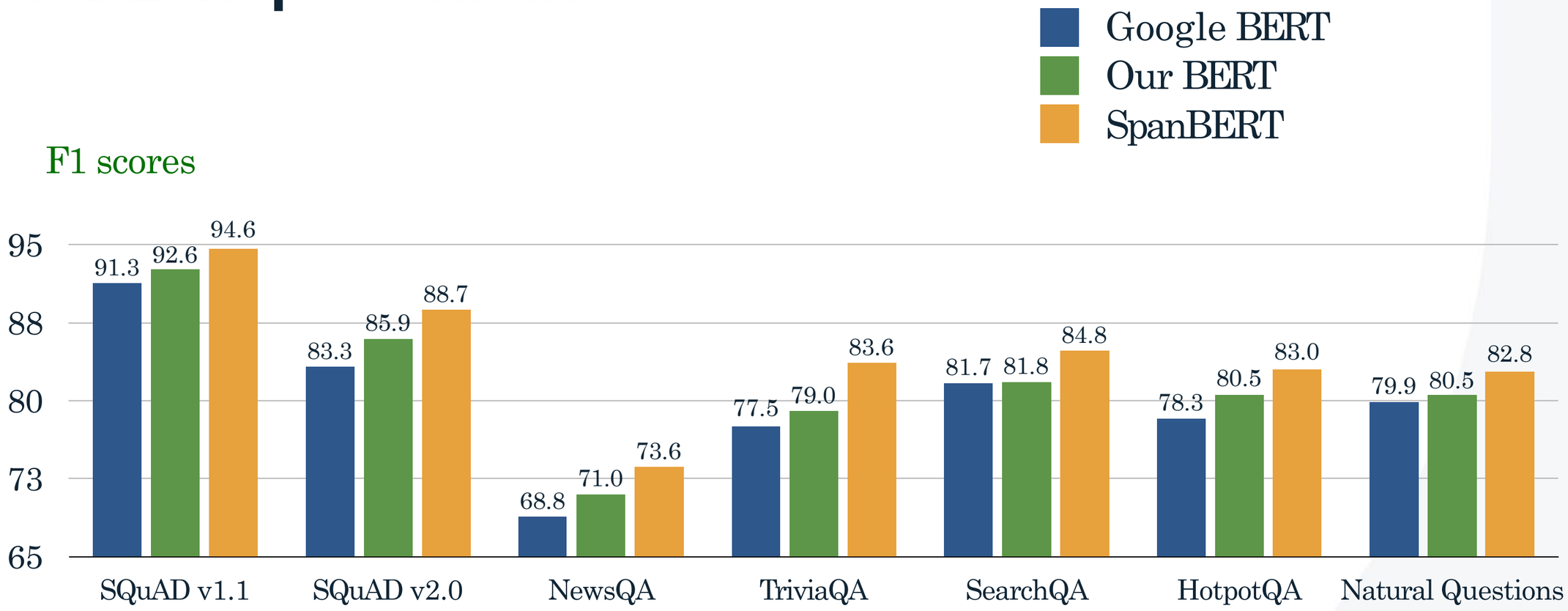
$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

Stanford cs224n-2023-lecture12-QA

(Joshi & Chen et al., 2020): SpanBERT: Improving Pre-training by Representing and Predicting Spans

THE UNIVERSITY of ADELAIDE

SpanBERT performance



Stanford cs224n-2023-lecture12-QA

THE UNIVERSITY of ADELAIDE

Evaluation of IR based and Extractive QA

Common Evaluation Metrics for Extractive QA

Accuracy (F1 score)

- Does answer match gold--labelled answer

Mean Reciprocal Rank (MRR)

MAP (Mean Average Precision (MAP)

Number of Exact Matches (EM)

THE UNIVERSITY of ADELAIDE

Stanford question answering dataset (SQuAD)

- 100k annotated (passage, question, answer) triples
  - Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.
  - This is a limitation— not all the questions can be answered in this way!

- SQuAD still remains the most popular reading comprehension dataset; it is "almost solved" today and the state-of-the-art exceeds the estimated human performance. (demo https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

(Rajpurkar et al., 2016): SQuAD: 100,000+ Questions for Machine Comprehension

THE UNIVERSITY of ADELAIDE

Stanford question answering dataset (SQuAD)

- **Evaluation on two metrics**: 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 BOW 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: exact match 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

THE UNIVERSITY of ADELAIDE

SQuAD 2.0, limitations

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one.
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
- For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Limitations of SQuAD:
  - No yes/no questions, only span based
  - Anotators make questions given text so uses just co-reference, so easier than in real life
  - In real like, people type questions, then look for answers, more semantic inference.
  - No multi-sentence

Other QA datasets

- Hotspot QA dataset (https://hotpotqa.github.io/)
  - created by showing crowd workers multiple context documents and asked to come up with questions that require reasoning about all of the documents

- Natural Questions dataset (https://ai.google.com/research/NaturalQuestions/download)
  - Bases on Google queries
  - Annotators are given Wikipedia page and annotate a paragraph AND a short span, OR null.

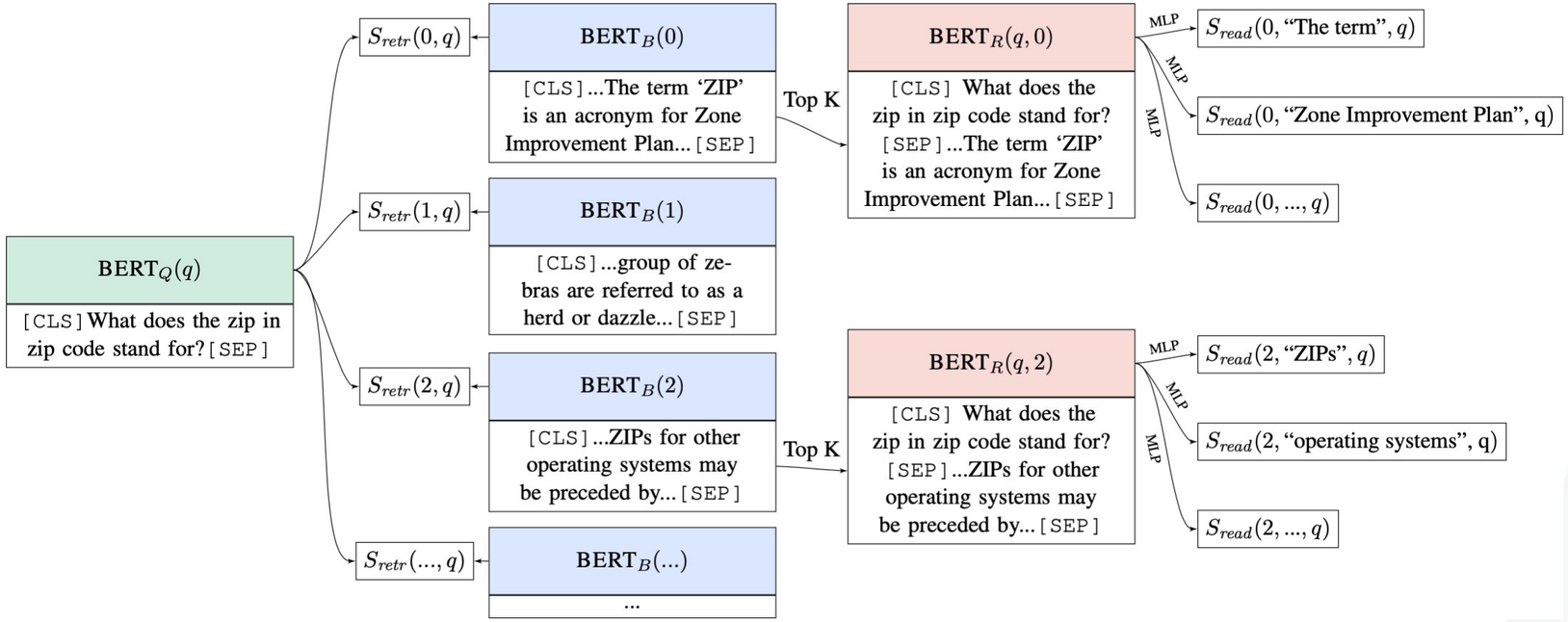
THE UNIVERSITY of ADELAIDE

Other neural based QA

THE UNIVERSITY of ADELAIDE

Training the retriever

- **Joint training** of retriever and reader

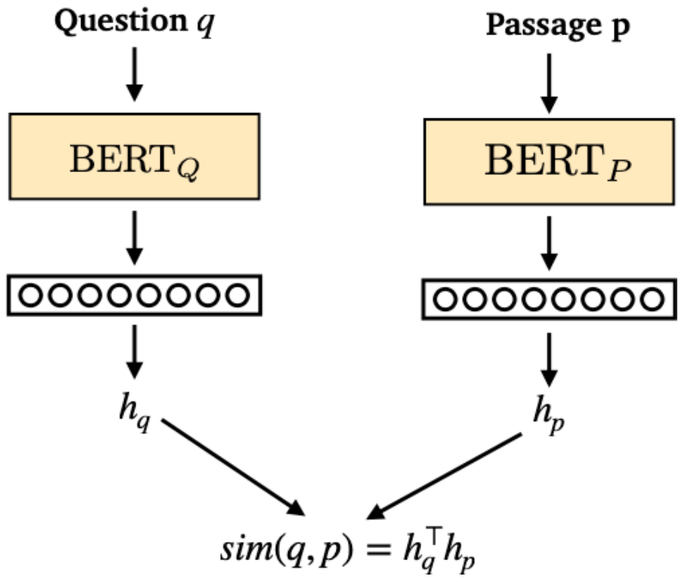


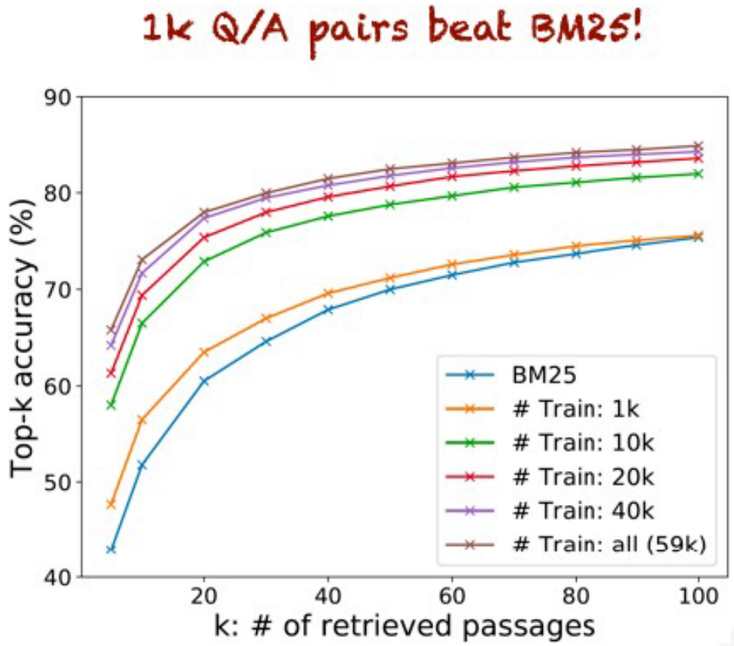
- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Training the retriever

- Dense passage retrieval (DPR) - We can also just train the retriever using question-answer pairs!





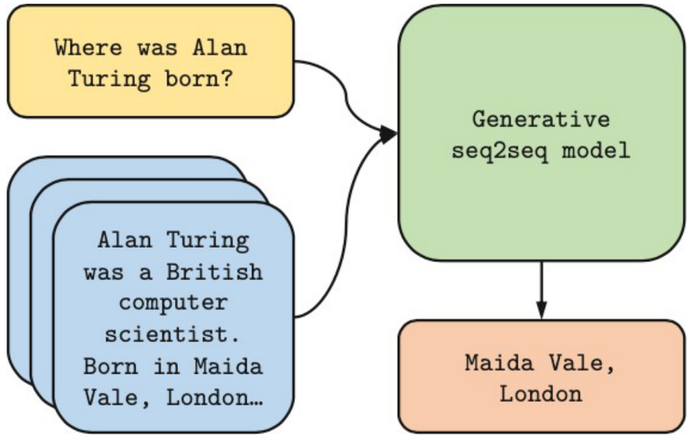
- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

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

Dense retrieval + generative models

Recent work shows that it is beneficial to generate answers instead of to extract answers.

Fusion-in-decoder (FID) = DPR + T5

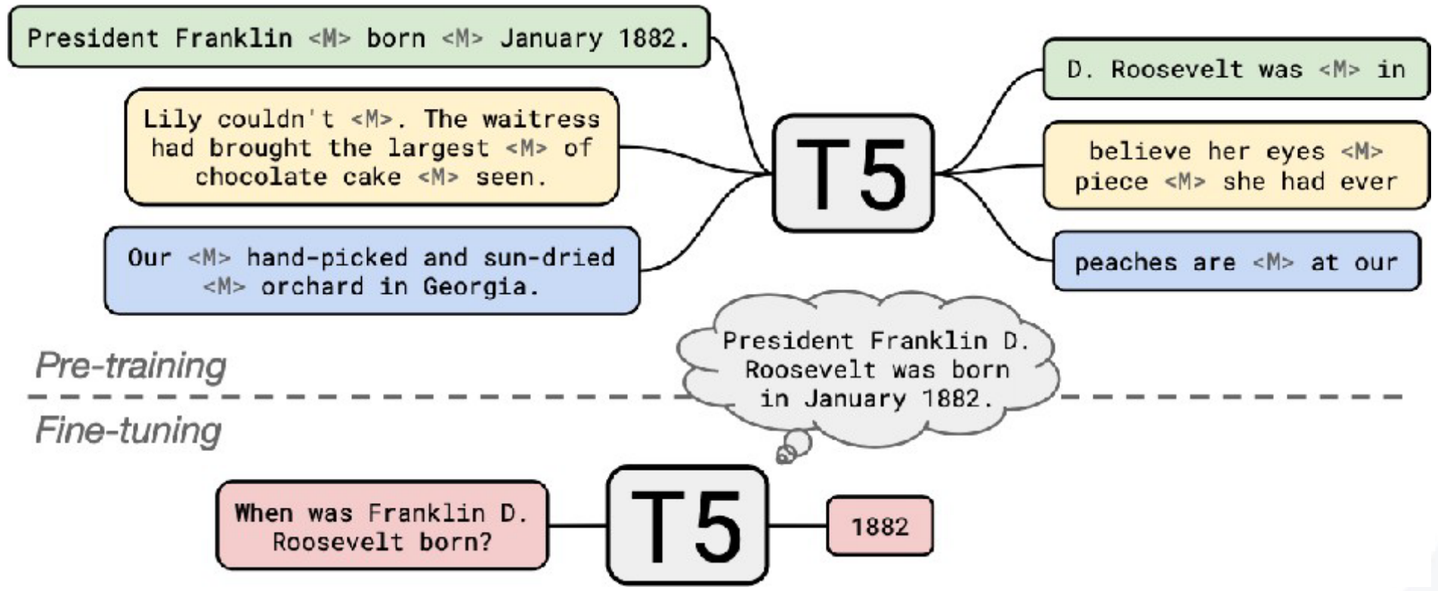


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	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

Pretraining encoder-decoders: T5

Large language model can do open-domain QA well without an explicit retriever stage



Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Pretraining encoder-decoders: T5

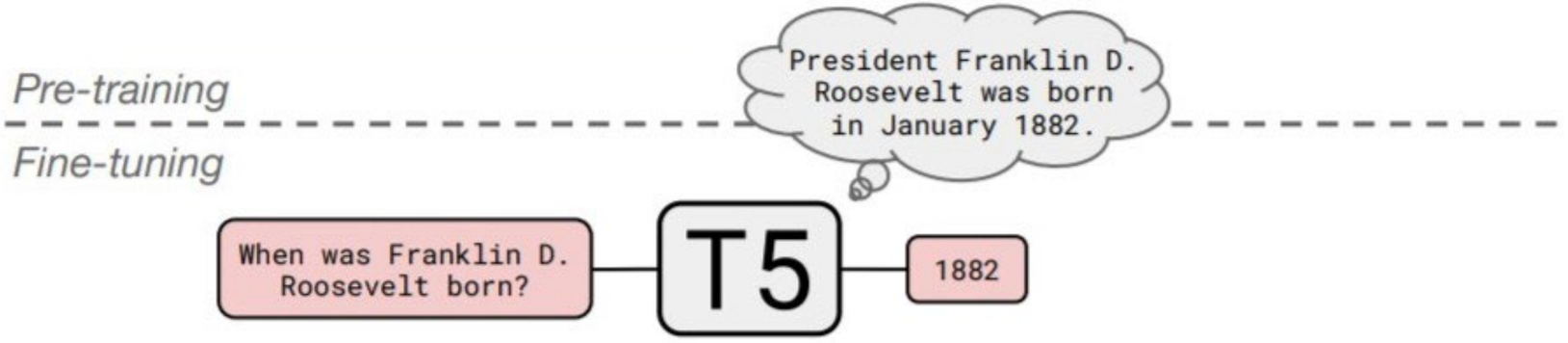
A fascinating property of T5: it can be finetuned to answer a wide range of questions, retrieving knowledge from its parameters.

NQ: Natural Questions

WQ: WebQuestions

TQA: Trivia QA

All “open-domain” versions

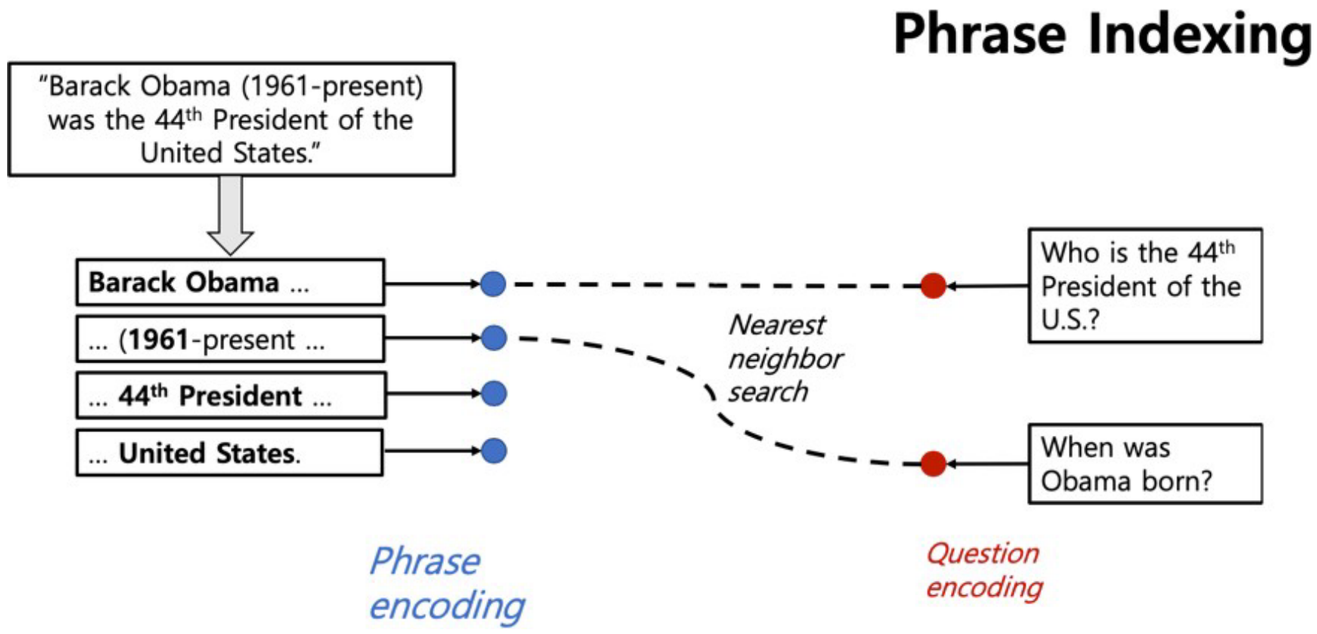


	NQ	WQ	TQA dev	TQA test	
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–	
T5.1.1-Base	25.7	28.2	24.2	30.6	**220 million params**
T5.1.1-Large	27.3	29.5	28.5	37.2	**770 million params**
T5.1.1-XL	29.5	32.4	36.0	45.1	**3 billion params**
T5.1.1-XXL	32.8	35.6	42.9	52.5	**11 billion params**
T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**	

[Raffel et al., 2018]

Retriever without reader

It is possible to encode all the phrases (60 billion phrases in Wikipedia) using **dense** vectors and only do nearest neighbor search without a BERT model at inference time!



https://github.com/princeton-nlp/DensePhrases?tab=readme-ov-file

Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index

Lee et al., 2020. Learning Dense Representations of Phrases at Scale

References

- Publicly available lecture slides were used from Stanford courses CS224u, and CS224n by Christopher Manning and Christopher Potts (https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/slides/cs224n-2021-lecture05-rnnlm.pdf)

THE UNIVERSITY of ADELAIDE

THE UNIVERSITY
of ADELAIDE