Question Answering

Slides adapted from Danqi Chen, Rui Zhang

The New York Times

Computer Wins on 'Jeopardy!': Trivial, It's Not

Two "Jeopardy!" champions, Ken Jennings, left, and Brad Rutter, competed against a computer named Watson, which proved adept at buzzing in quickly. Carol Kaelson/Jeopardy Productions Inc., via Associated Press

By John Markoff

Feb. 16, 2011

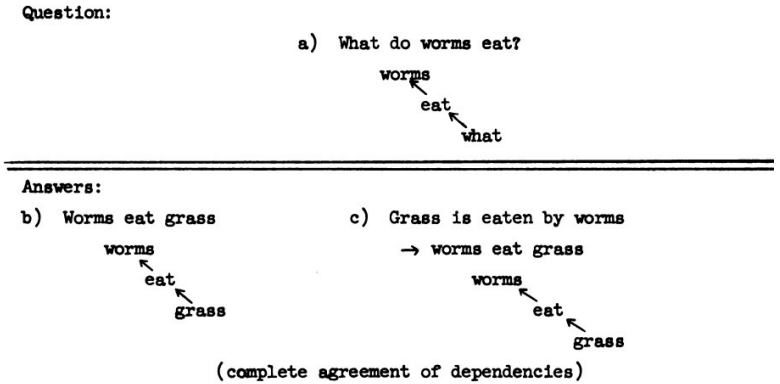
What is question answering?

Question (Q) →  → Answer (A)

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

The earliest QA systems dated back to 1960s!

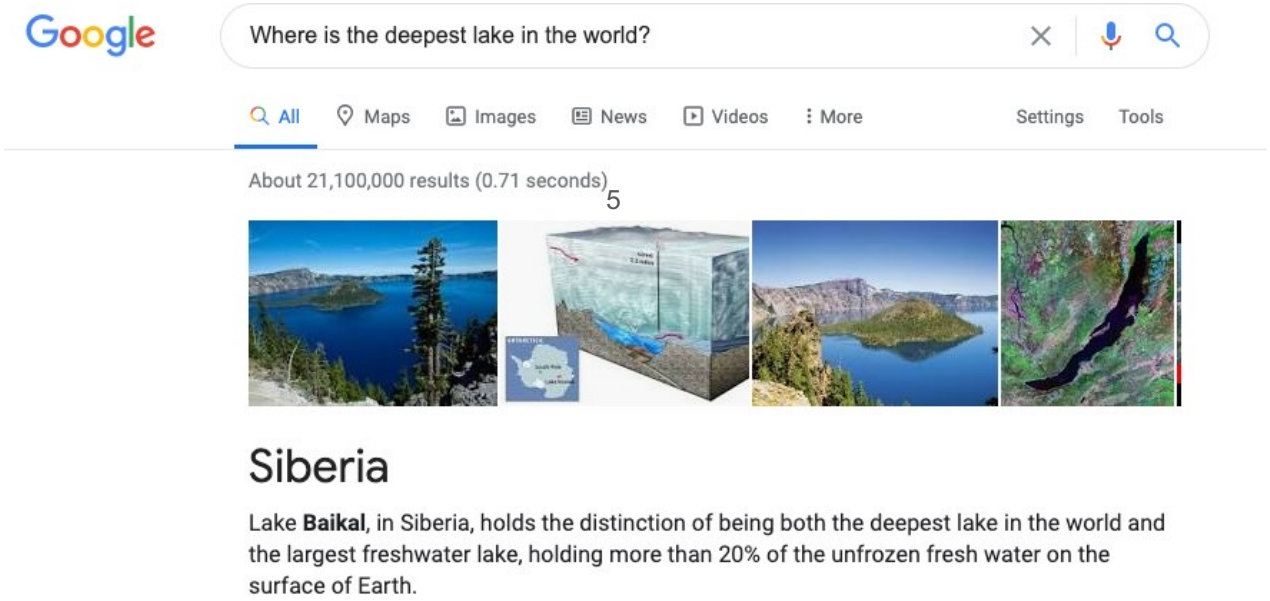
(Simmons et al., 1964)



Question answering: a taxonomy

- What information source does a system build on?
  - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
  - Factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional, ..
- Answer type
  - A short segment of text, a paragraph, a list, yes/no, ...

Lots of practical applications



The World of Question Answering

Question Answering produces **Answers** to **Natural Language Questions** based on **Knowledge Resources**.

- Knowledge Resources
  - Unstructured texts (SQuAD)
  - Knowledge Graphs (Freebase)
  - Tables (WikiTableQuestions)
  - Relational Databases (WikiSQL, Spider)
  - Multimodal (VQA, MultiModalQA)
  - Common Sense (CommonsenseQA)

The World of Question Answering

Question Answering produces **Answers** to **Natural Language Questions** based on **Knowledge Resources**.

- Knowledge Resources
  - Unstructured texts (SQuAD)
  - Knowledge Graphs (Freebase)
  - Tables (WikiTableQuestions)
  - Relational Databases (WikiSQL, Spider)
  - Multimodal (VQA, MultiModalQA)
  - Common Sense (CommonsenseQA)
- Natural Language Questions
  - Answerable (SQuAD 1.0) vs Unanswerable (SQuAD 2.0) vs Ambiguous (AmbigQA)
  - Factoid vs Explanatory (ELI5)
  - Single-Turn vs Multi-Turn (SQA) vs Conversational (QuAC, CoQA, CoSQL)

The World of Question Answering

Question Answering produces **Answers** to **Natural Language Questions** based on **Knowledge Resources**.

- Knowledge Resources
  - Unstructured texts (SQuAD)
  - Knowledge Graphs (Freebase)
  - Tables (WikiTableQuestions)
  - Relational Databases (WikiSQL, Spider)
  - Multimodal (VQA, MultiModalQA)
  - Common Sense (CommonsenseQA)
- Natural Language Questions
  - Answerable (SQuAD 1.0) vs Unanswerable (SQuAD 2.0) vs Ambiguous (AmbigQA)
  - Factoid vs Explanatory (ELI5)
  - Single-Turn vs Multi-Turn (SQA) vs Conversational (QuAC, CoQA, CoSQL)
- Answers
  - Short-form: Extractive, Multiple Choice, Yes/No
  - Long-form: Abstractive (NarrativeQA, Natural Question), Open-domain Long-form QA (ELI5)

The World of Question Answering

Question Answering produces **Answers** to **Natural Language Questions** based on **Knowledge Resources**.

- Knowledge Resources
  - **Unstructured texts (SQuAD, Natural Questions)**
  - Knowledge Graphs (Freebase)
  - Tables (WikiTableQuestions)
  - Relational Databases (WikiSQL, Spider)
  - Multimodal (VQA, MultiModalQA)
  - **Common Sense (CommonsenseQA)**

QA based on unstructured Text

Reading comprehension (MRC) ⬅️

How to answer questions over **a single passage of text**

Open-domain (textual) question answering

How to answer questions over **a large collection of documents**

CNN/Daily Mail (Hermann et al. 2015)

- Cloze Style (fill-in-the-blank) Questions
- Use the summarized bullet points of the news article to create questions
- Anonymize the entities to prevent model from using co-occurrence of entities to answer the question.

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	

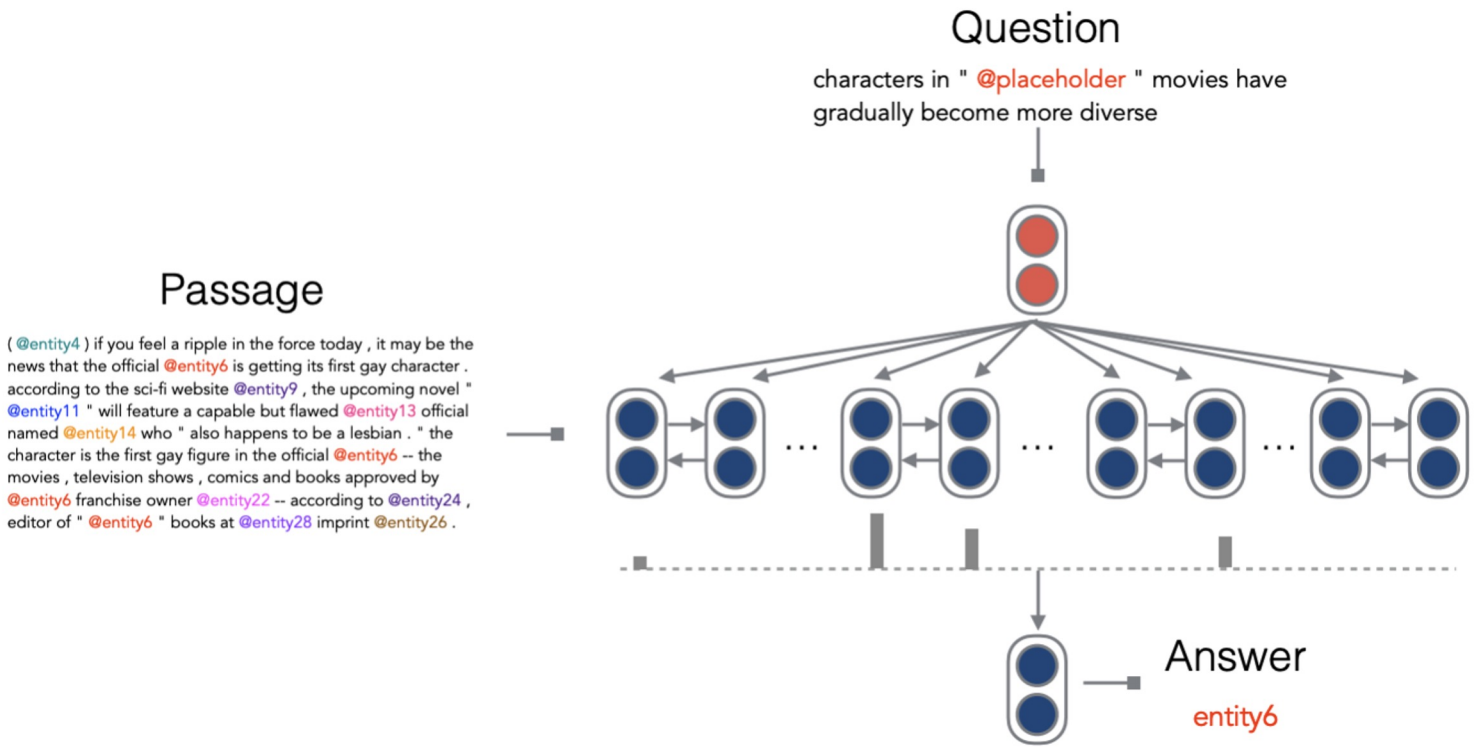
CNN/Daily Mail (Hermann et al. 2015)

- Cloze Style (fill-in-the-blank) Questions
- Use the summarized bullet points of the news article to create questions
- Anonymize the entities to prevent model from using co-occurrence of entities to answer the question.

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

A Thorough Examination of CNN/Daily Mail (Chen et al., 2016)

- AttentiveReader

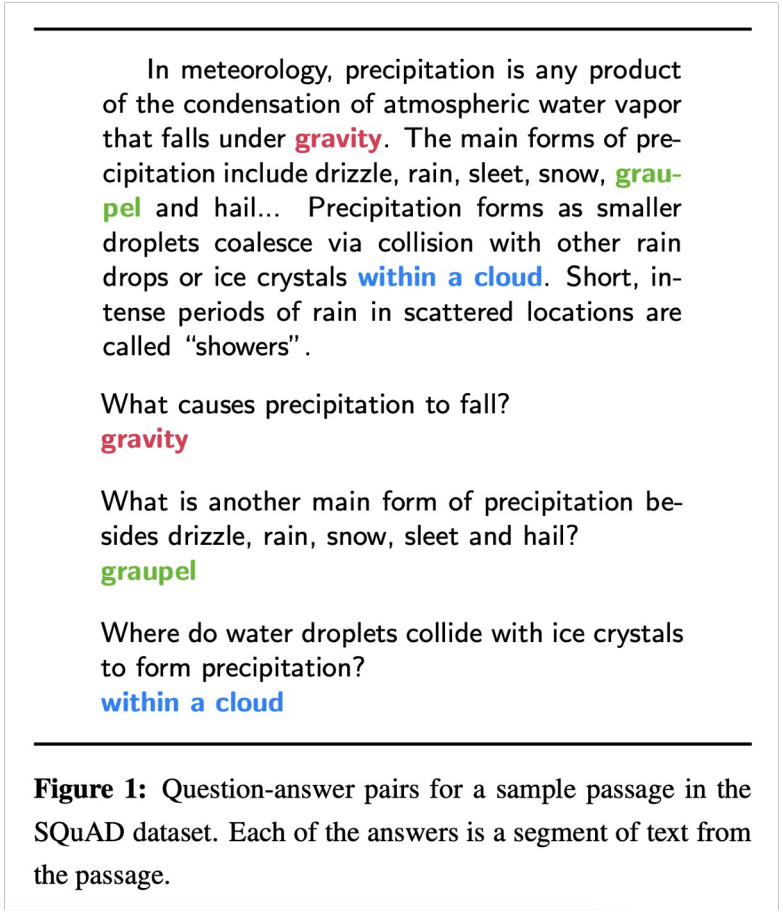


SQuAD (Rajpurkar et al., 2016)

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD)

- 100,000 Questions over 500 Wikipedia articles.
- Answer to every question is a segment of text, or ***span***, from the corresponding reading passage.

https://rajpurkar.github.io/SQuAD-explorer/



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

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

SQuAD 2.0 (Rajpurkar et al., 2018)

Know What You Don't Know: Unanswerable Questions for SQuAD.

50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

To do well on SQuAD 2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

https://rajpurkar.github.io/SQuAD-explorer/

Reading comprehension (MRC)

Reading comprehension: building systems to comprehend a passage of text and answer questions about its content $(P, Q) \rightarrow A$

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

Reading comprehension (MRC)

Reading comprehension: building systems to comprehend a passage of text and answer questions about its content (P, Q)→A

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

Why do we care about this problem?

- Useful for many practical applications
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
  - Wendy Lehnert 1977: "Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."
- Many other NLP tasks can be reduced to a reading comprehension problem:

Information extraction

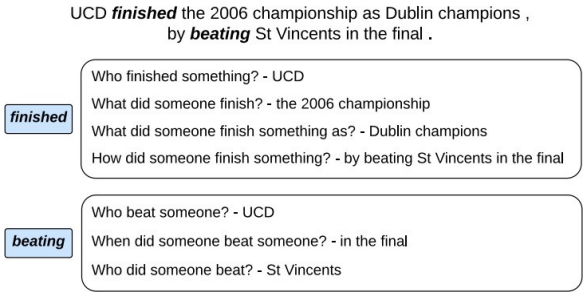
(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

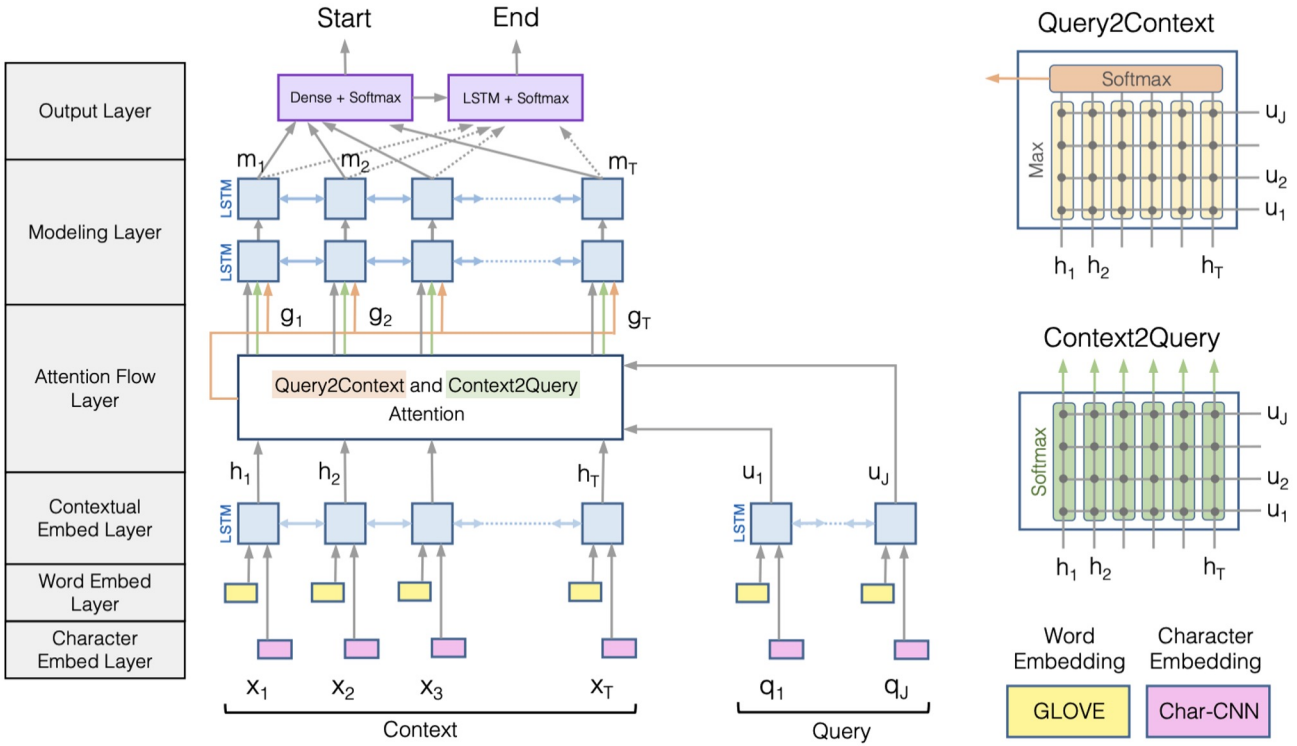
(Levy et al., 2017)

Semantic role labeling



(He et al., 2015)

BiDAF: Bidirectional Attention Flow (Seo et al., 2017)



LSTM-based vs BERT models

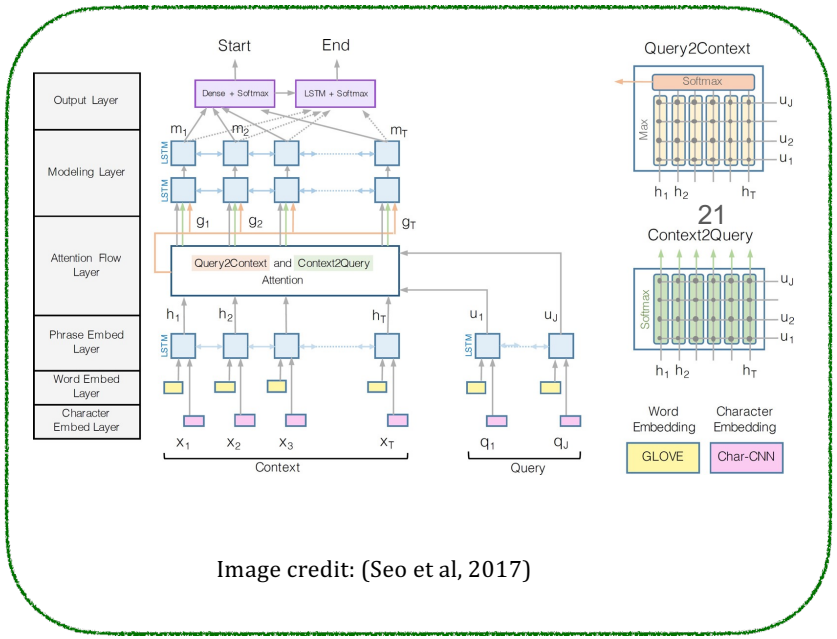


Image credit: (Seo et al, 2017)

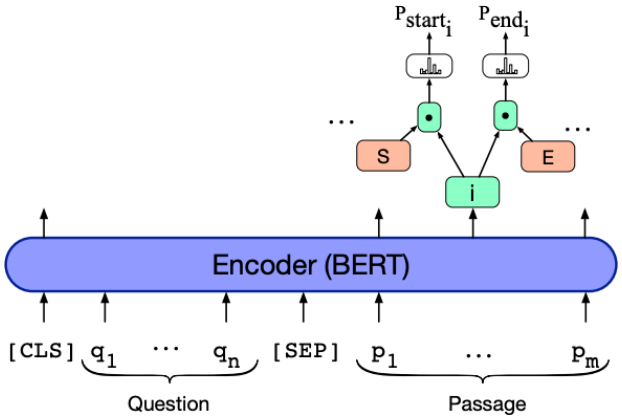


Image credit: J & M, edition 3

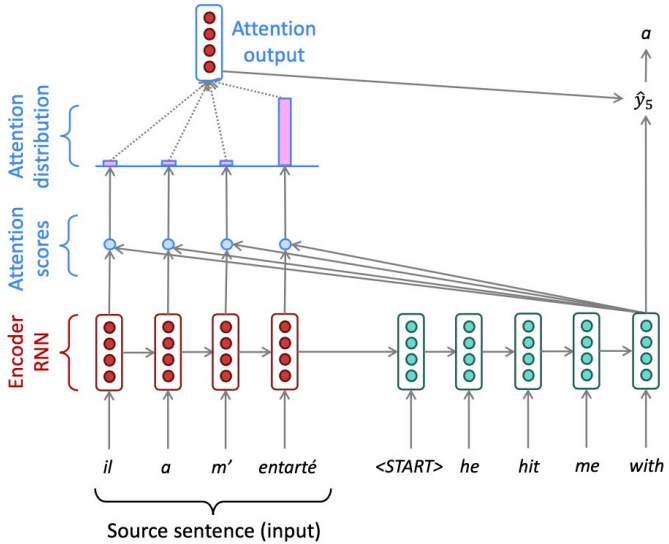
Recap: seq2seq model with attention

- Instead of source and target sentences, we also have two sequences: passage and question (lengths are unbalanced)

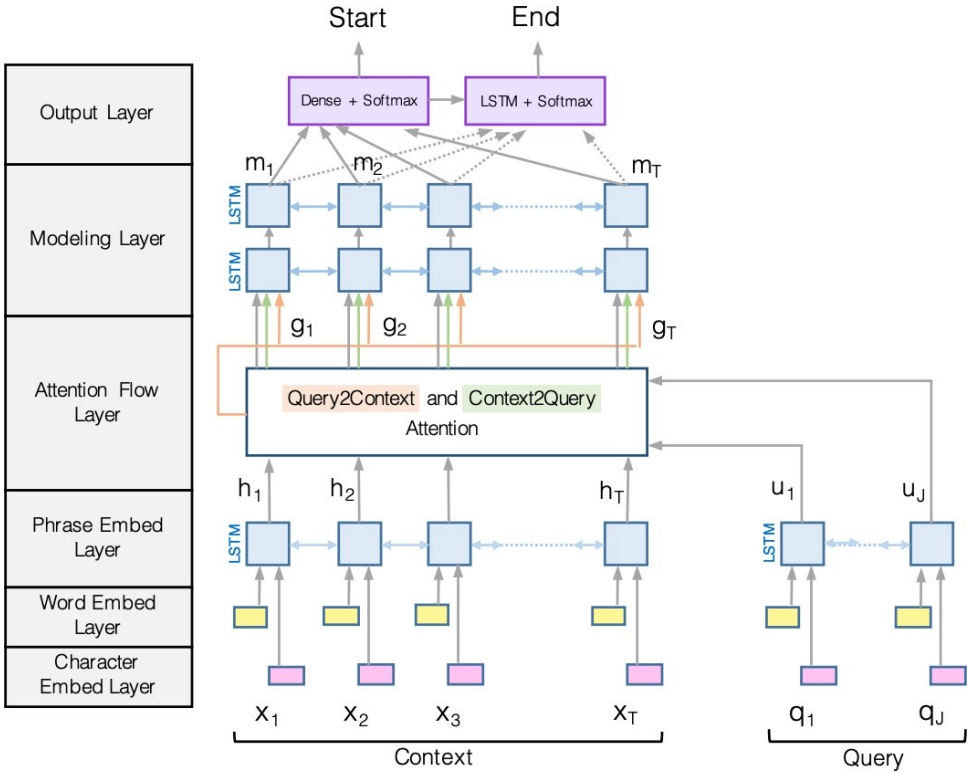
- We need to model which words in the passage are most relevant to the question (and which question words)

Attention is the key ingredient here, similar to which words in the source sentence are most relevant to the current target word…

- We don't need an autoregressive decoder to generate the target sentence word-by-word. Instead, we just need to train two classifiers to predict the start and end positions of the answer!

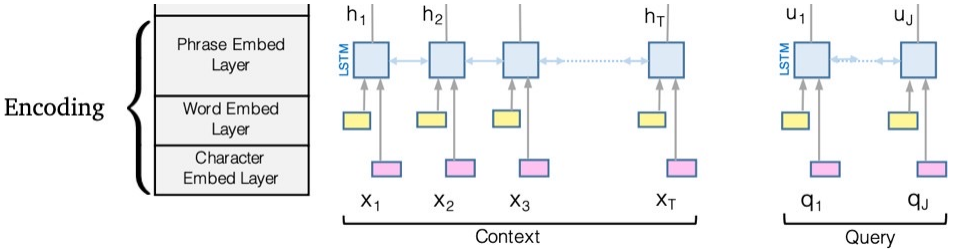


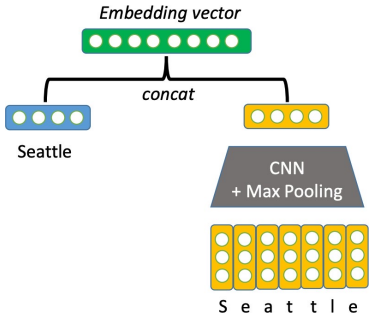
BiDAF: the Bidirectional Attention Flow model



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

BiDAF: Encoding





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

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

f: high-way networks omitted here

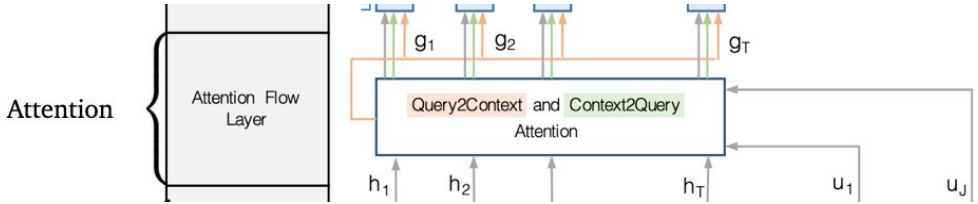
- Then, use two **bidirectional** LSTMs separately to produce contextual embeddings for both context and query.

$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H} \qquad \mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Attention



- Context-to-query attention: For each context word, choose the most relevant words from the query words.

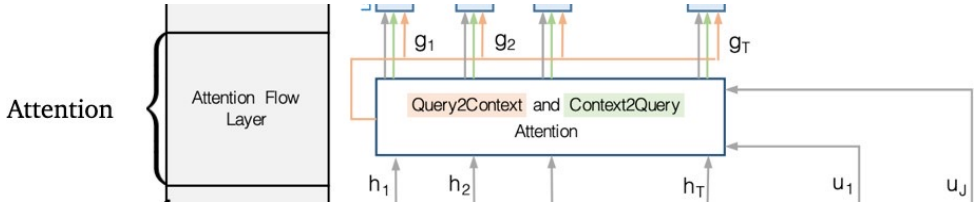
Q: *Who leads the United States?*

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

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

(Slides adapted from Minjoon Seo)

BiDAF: Attention



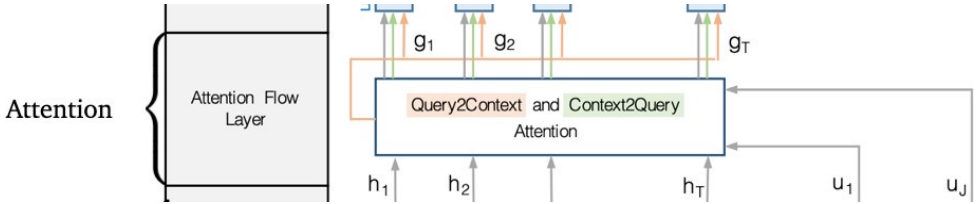
- Query-to-context attention: choose the context words that are most relevant to one of query words.

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?

(Slides adapted from Minjoon Seo)

BiDAF: Attention



The final output is
$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$

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

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

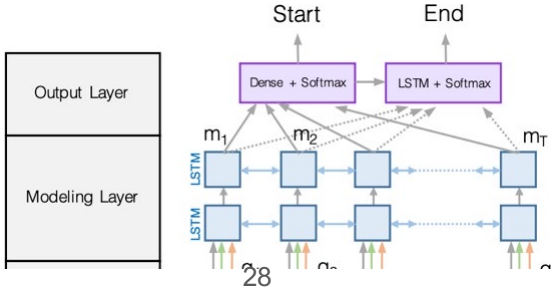
- Context-to-query attention (which question words are more relevant to c_i):

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

- 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: Modeling and output layers



The final training loss is

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

Modeling layer: pass $\mathbf{g}_i$ 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

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_i) \in \mathbb{R}^{2H}$$

Output layer: two classifiers predicting the start and end positions:

$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^\intercal [\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^\intercal [\mathbf{g}_i; \mathbf{m}_i'])$$

$$\mathbf{m}_i' = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

BiDAF: Performance on SQuAD

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

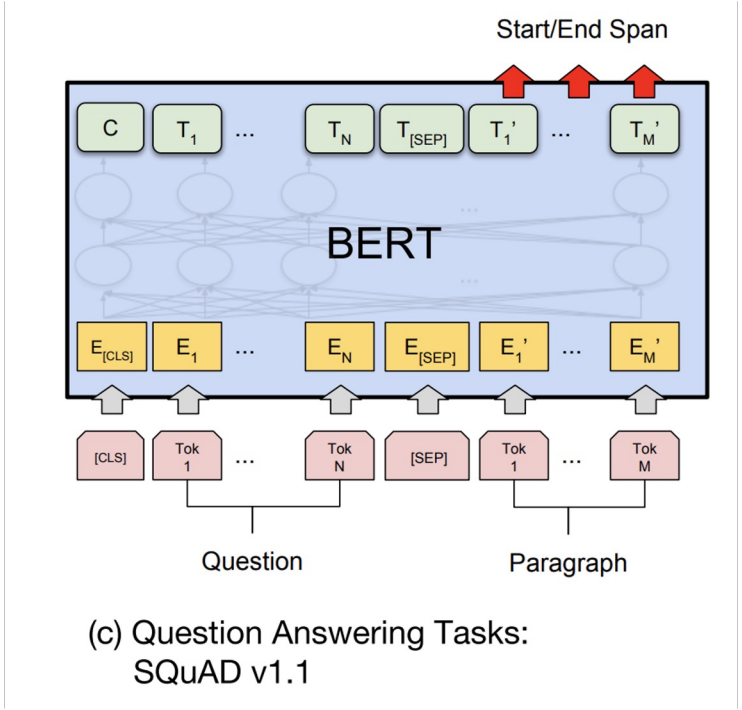
	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]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

64

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

BERT on Different Tasks

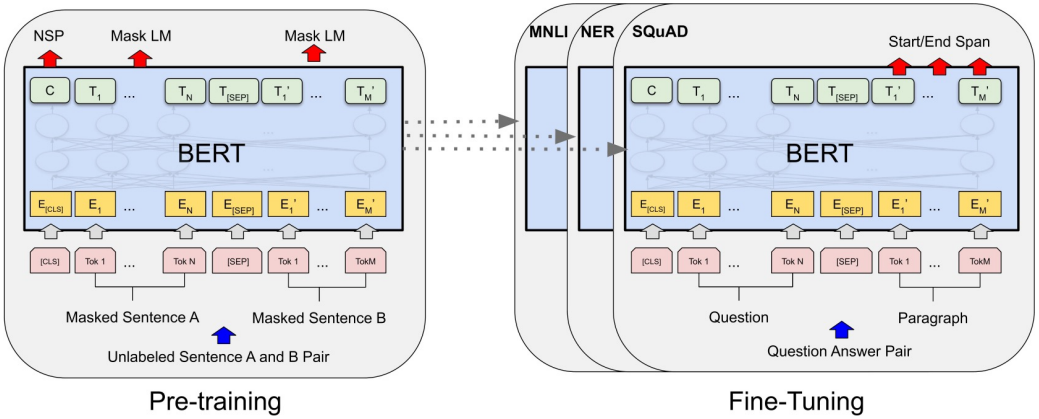
Question Answering



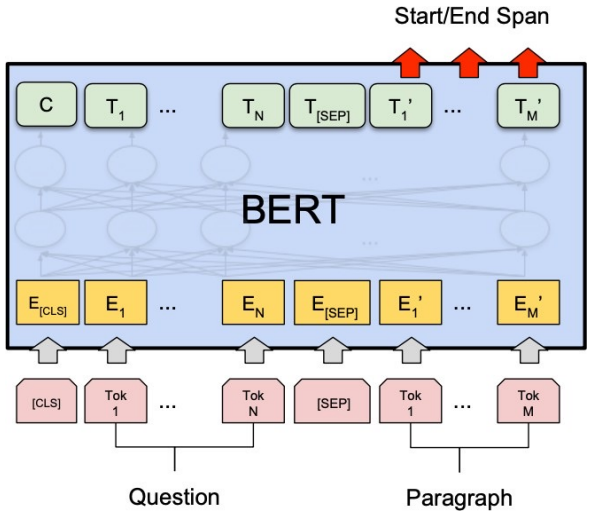
(c) Question Answering Tasks:
SQuAD v1.1

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)
- $\text{BERT}_{\text{base}}$ has 12 layers and 110M parameters, $\text{BERT}_{\text{large}}$ has 24 layers and 330M parameters



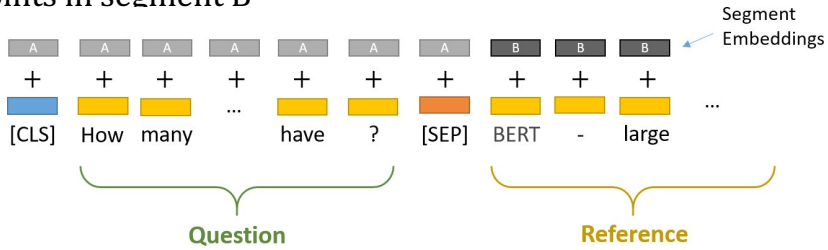
BERT for reading comprehension



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B

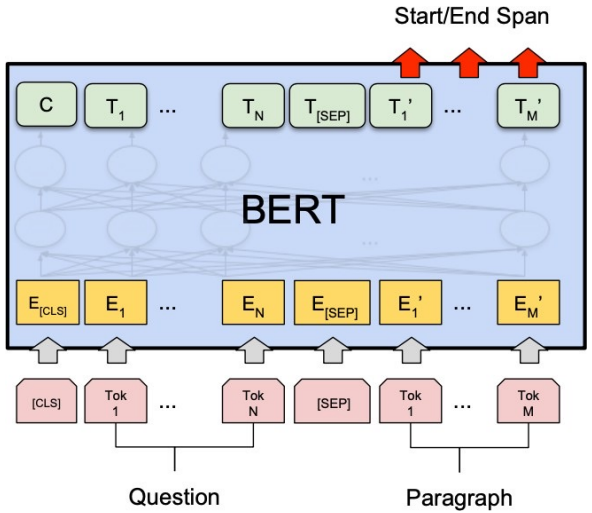


Question: How many parameters does BERT-large have?

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/

BERT for reading comprehension

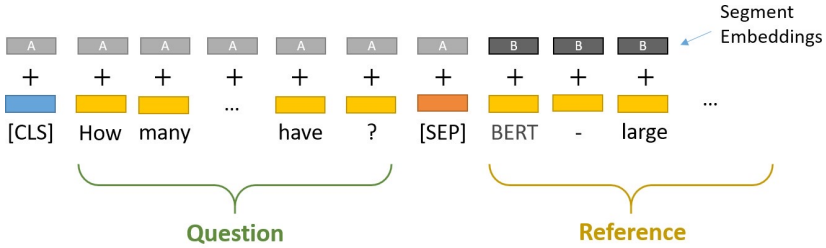


$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B

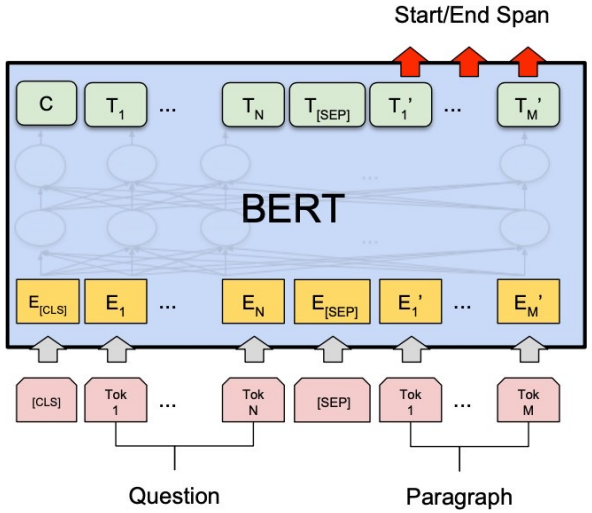


Question: How many parameters does BERT-large have?

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/

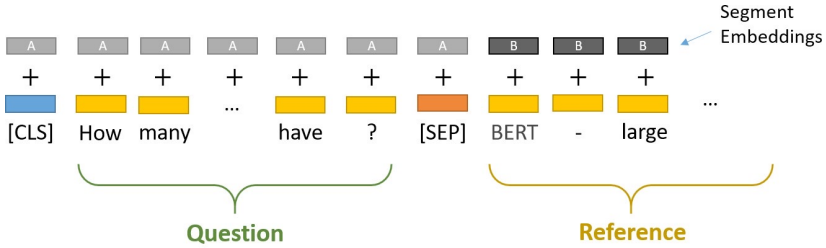
BERT for reading comprehension



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/

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

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

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

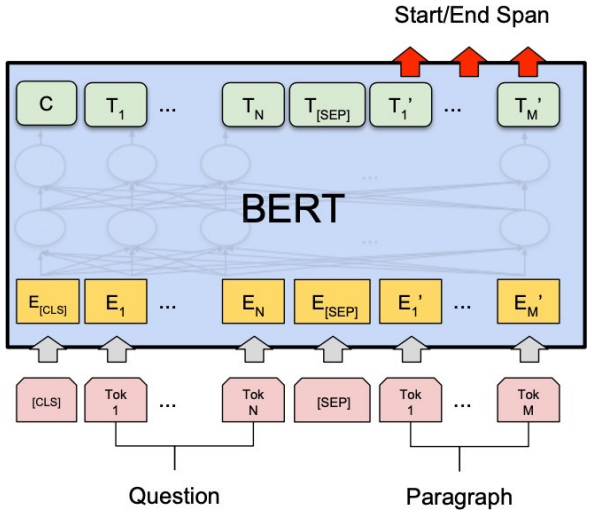
where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

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 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.



	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)

Comparisons between BiDAF and BERT models

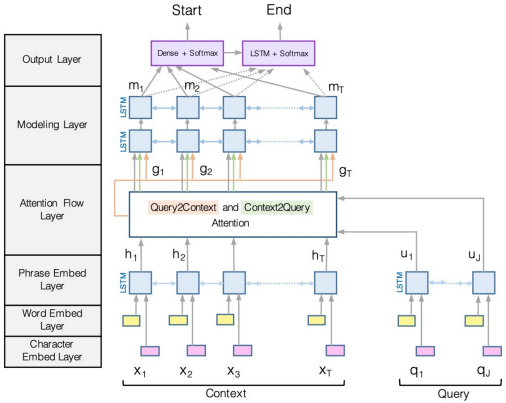
- BERT model has many many more parameters (110M or 330M)
  BiDAF has ~2.5M parameters.
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers (no recurrence architecture and easier to parallelize).
- BERT is **pre-trained** while BiDAF is only built on top 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..

Comparisons between BiDAF and BERT models

Are they fundamentally different?

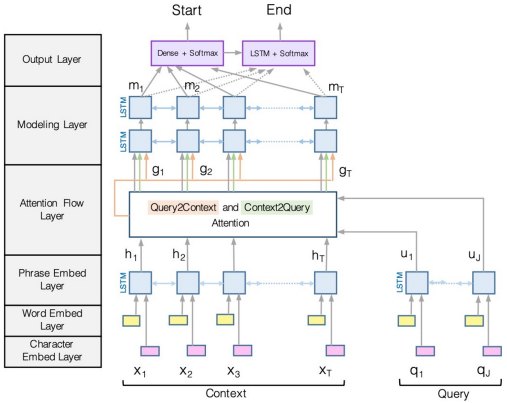
- BiDAF and other models aim to model the interactions between question and passage.
- BERT uses self-attention between the **concatenation** of question and passage = attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
- Can we improve BiDAF drawing inspirations from BERT ?

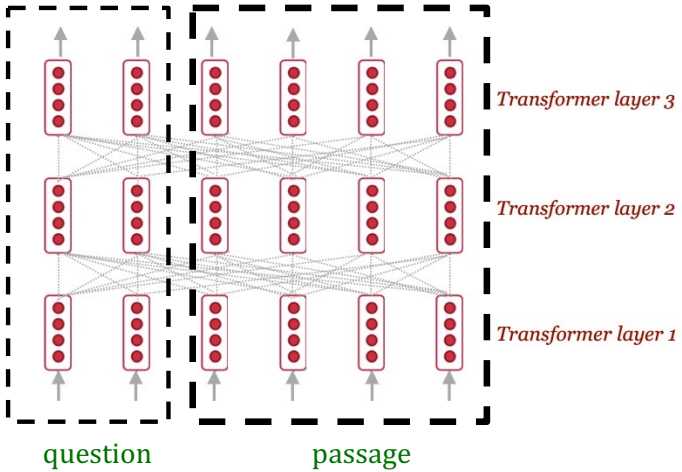


Comparisons between BiDAF and BERT models

Are they fundamentally different?

- BiDAF and other models aim to model the interactions between question and passage.
- BERT uses self-attention between the **concatenation** of question and passage = attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
- (Clark and Gardner, 2018) shows that adding a self-attention layer for the passage attention(P, P) to BiDAF also improves performance.





SQuAD 2.0 Leaderboard

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899

Is reading comprehension solved?

Adversarial Examples for Evaluating Reading Comprehension Systems (Jia and Liang, 2017)

Create examples by inserting sentences to distract the computer systems.

In this adversarial setting, the accuracy of sixteen published models drops from an average of 75% F1 score to 36%.

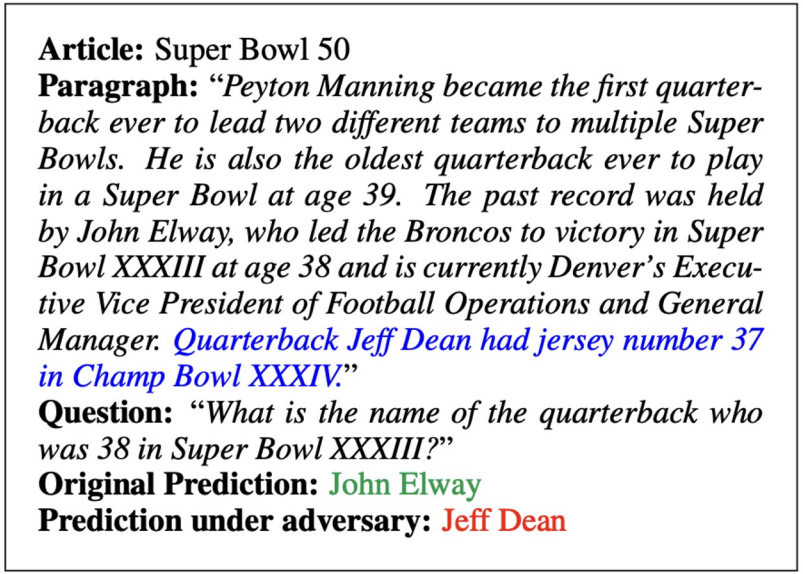
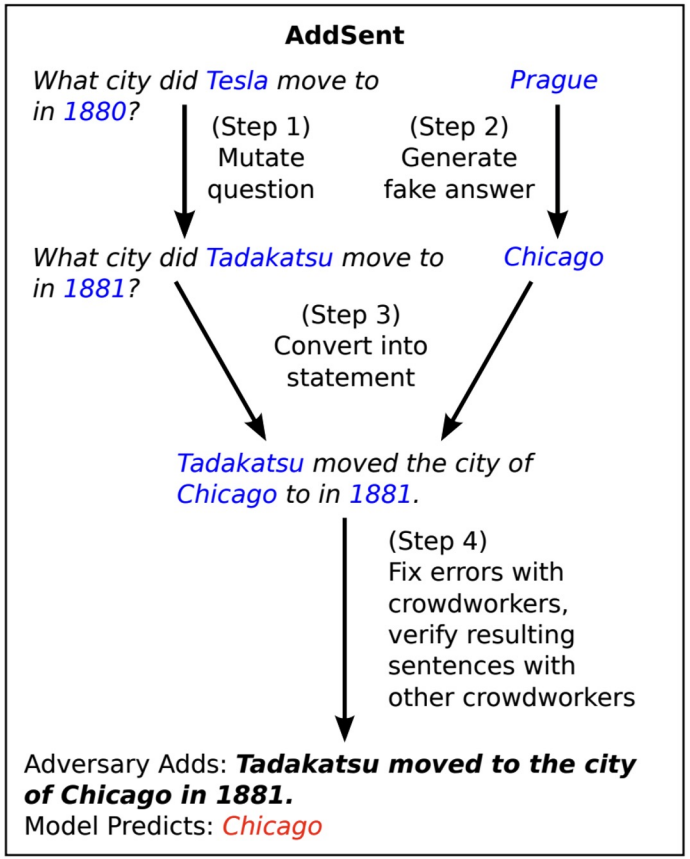


Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Adversarial Examples for Evaluating Reading Comprehension Systems (Jia and Liang, 2017)

Create examples by inserting sentences to distract the computer systems.

In this adversarial setting, the accuracy of sixteen published models drops from an average of 75% F1 score to 36%.



Trick Me If You Can: **Human-in-the-loop** Generation of Adversarial Examples for Question Answering (Wallace et al., 2018)

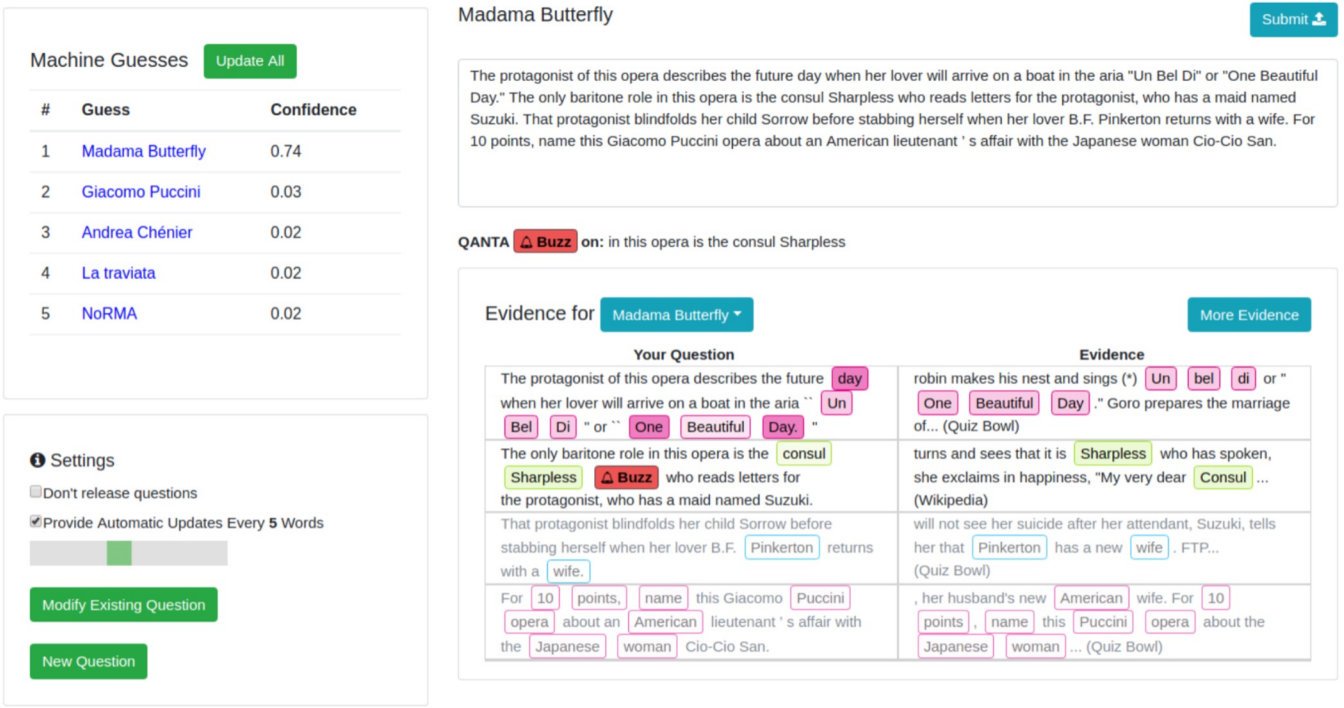


Figure 3: The author writes a question (top right), the QA system provides guesses (left), and explains why it makes those guesses (bottom right). The author can then adapt their question to "trick" the model.

What do Models Learn from Question Answering Datasets? (Sen et al., 2020)

Models trained on one dataset do not generalize well on the others.

Fine-tuned on	Evaluated on SQuAD	TriviaQA	NQ	QuAC	NewsQA
SQuAD	**75.6**	46.7	48.7	20.2	41.1
TriviaQA	49.8	**58.7**	42.1	20.4	10.5
NQ	53.5	46.3	**73.5**	21.6	24.7
QuAC	39.4	33.1	33.8	**33.3**	13.8
NewsQA	52.1	38.4	41.7	20.4	**60.1**

Table 3: F1 scores of each fine-tuned model evaluated on each test set

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Ribeiro et al., 2020)

	Test *TYPE* and Description	Failure Rate (🤖)	Example Test cases (with expected behavior and 🤖 prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan 🤖: Victoria
	MFT: intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna 🤖: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny 🤖: purple
	MFT: nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant 🤖: Indian accountant
	MFT: animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella 🤖: Jonathan
	MFT: comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly 🤖: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor 🤖: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV 🤖: 7 million → 5 million
	INV: add irrelevant sentence to **C**	9.8	(no example)
Temporal	***MFT:*** change in one person only	41.5	**C:** Both Jason and Abigail were journalists, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail 🤖: Abigail were journalists, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle 🤖: Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca 🤖: Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron 🤖: Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio 🤖: Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria 🤖: Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly 🤖: Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth 🤖: Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa 🤖: Jose

Table 3: A selection of tests for Machine Comprehension.

NarrativeQA (Kočiský et al. 2018)

- Existing datasets have questions that can be solved by selecting answers using superficial information; can be solved by shallow pattern matching.
- Generative Question Answering: **Generate** an output (similar to Natural Language Generation).
- Answer questions on stories about books and movie scripts.

Title: Ghostbusters II
Question: How is Oscar related to Dana?
Answer: her son
Summary snippet: . . . Peter's former girlfriend Dana Barrett has had a son, Oscar. . .
Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

ELI5: Long Form Question Answering (Fan et al., 2019)

Models must write multi-sentence answers given questions and supporting web documents.

Question: How do Jellyfish function without brains or nervous systems? [...] (60 words)

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. Sometimes just a couple of hours. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water. (327 words)

Documents: [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they dont possess brains, the animals still have neurons that send all sorts of signals throughout their body. [...] They may accomplish this through the assistance of their nerve rings. Jellyfish don't have brains, and that's just where things begin. They don't have many of the body parts that are typical in other animals. [...] (1070 words)

ELI5: Long Form Question Answering (Fan et al., 2019)

Models must write multi-sentence answers given questions and supporting web documents.

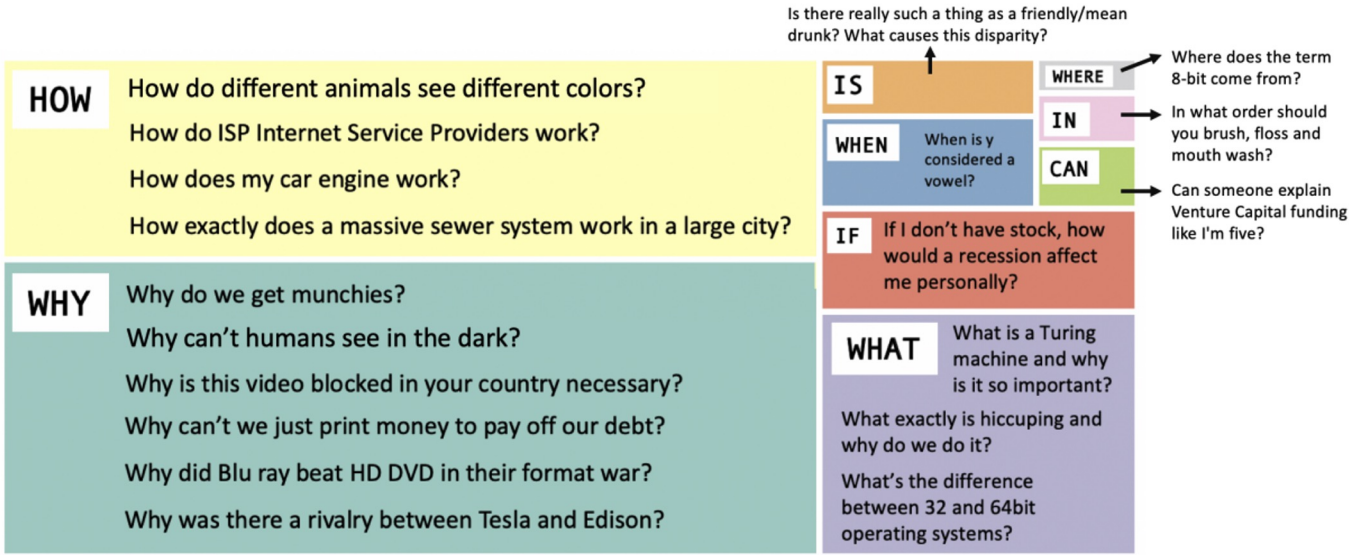


Figure 2: ELI5 questions by starting word, where box size represents frequency. Questions are open ended and diverse.

Natural Questions (Kwiatkowski et al., 2019)

Visualization

Natural Questions (Kwiatkowski et al., 2019)

Questions consist of real anonymized, aggregated queries issued to the Google search engine.

An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present.

Visualization

Reading Wikipedia to Answer Open-Domain Questions (Chen et al., 2017)

Open-domain QA = Retriever + Reader

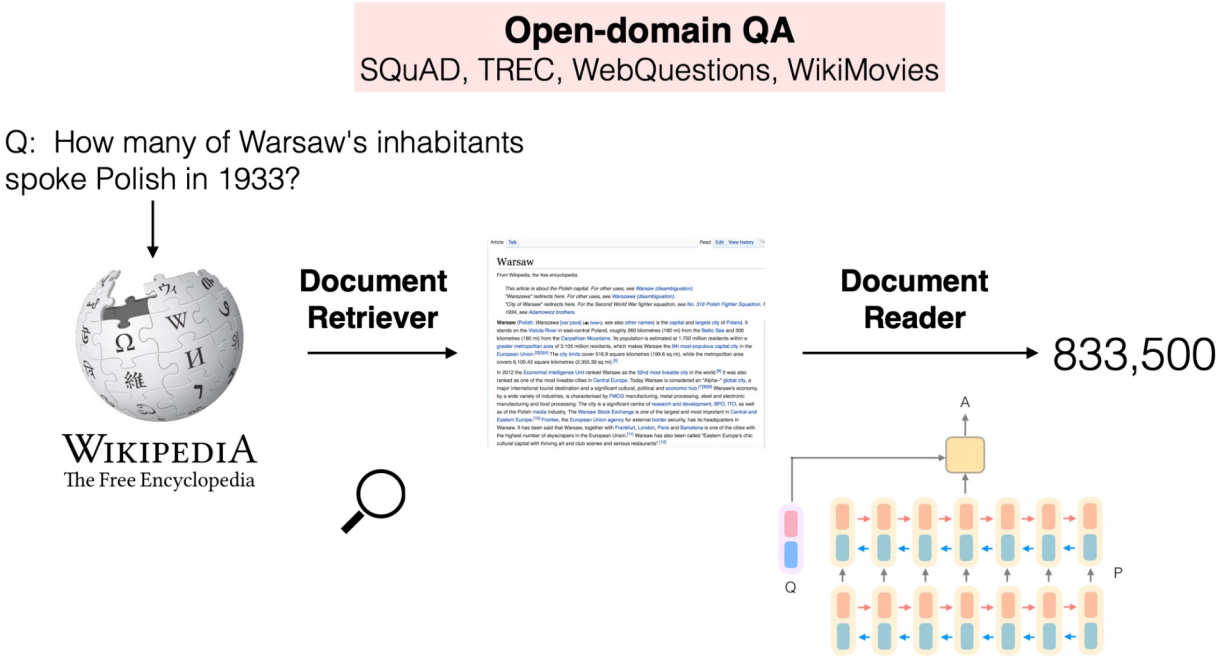


Figure 1: An overview of our question answering system DrQA.

QA based on unstructured Text

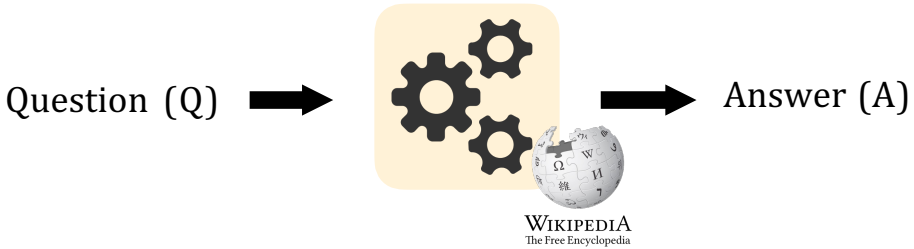
Reading comprehension (MRC)

How to answer questions over **a single passage of text**

Open-domain (textual) question answering ⬅️

How to answer questions over **a large collection of documents**

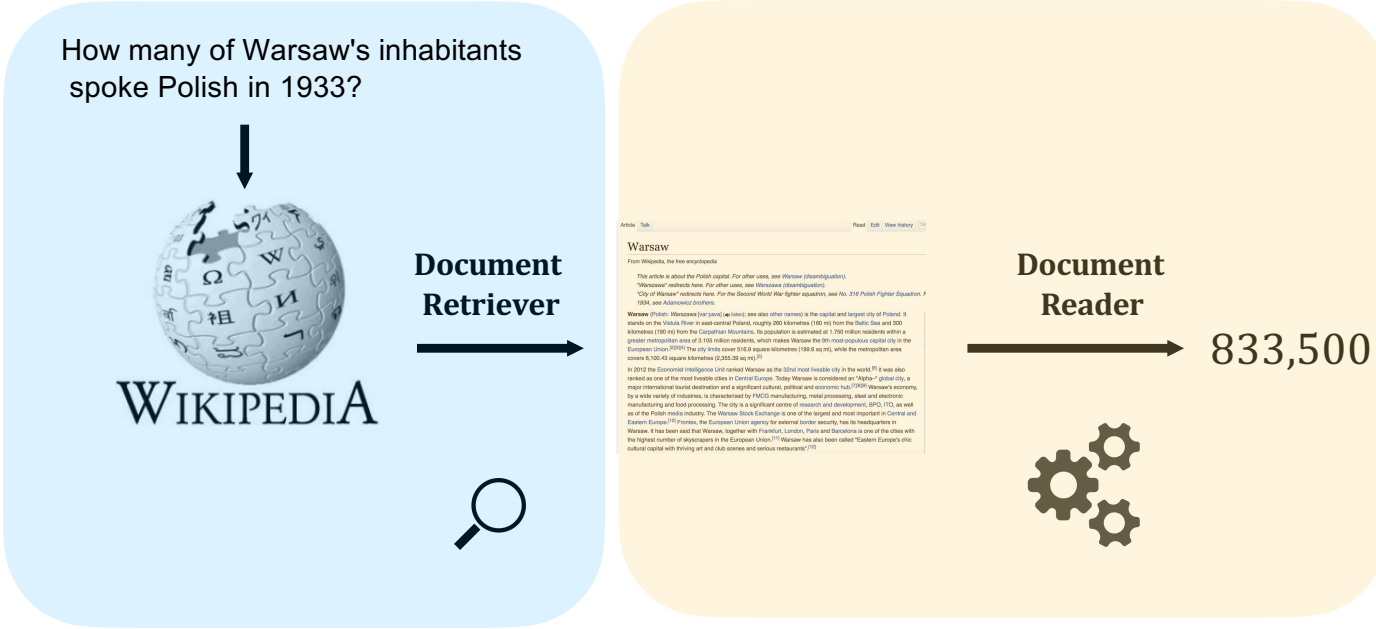
Open-domain question answering



- Different from reading comprehension, we don't assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging and a more practical problem!

- *In contrast to **closed-domain** systems that deal with questions under a specific domain (medicine, technical support).*

Retriever-reader framework



https://github.com/facebookresearch/DrQA

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

Retriever-reader framework

- Input: a large collection of documents $\mathcal{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A

- Retriever: $f(\mathcal{D}, Q) \rightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \rightarrow A$ A reading comprehension problem!

In DrQA,

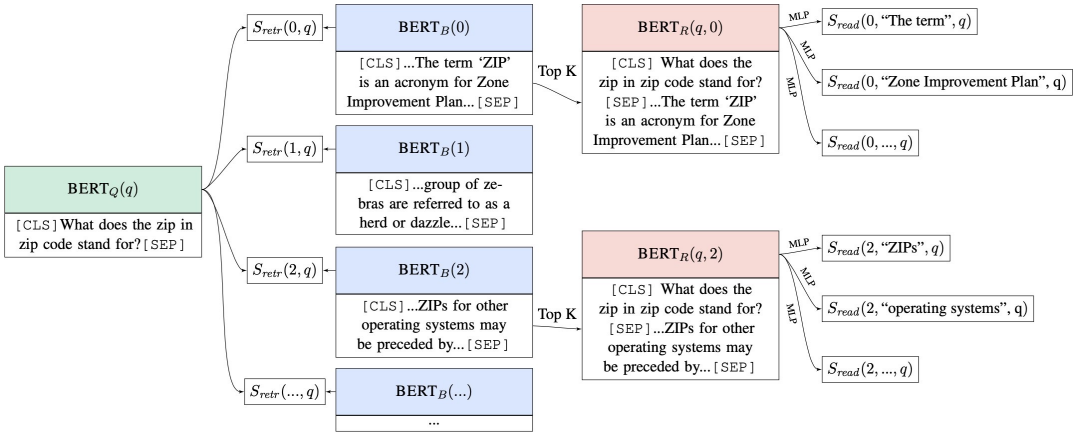
- Retriever = A standard TF-IDF information-retrieval sparse model (a fixed module)
- Reader = a neural reading comprehension model that we just learned
  - Trained on SQuAD and other distantly-supervised QA datasets

Distantly-supervised examples: (Q, A) (P, Q, A)

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

We can train the retriever too

- **Joint training** of retriever and reader

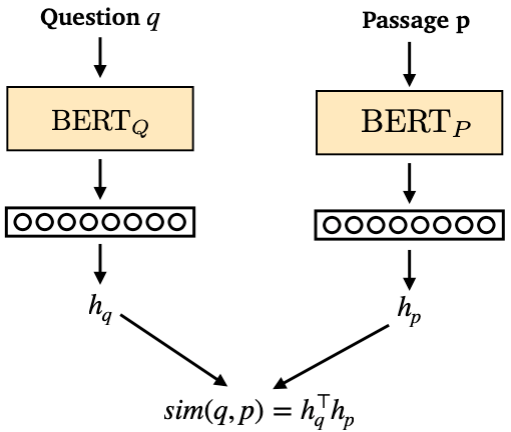


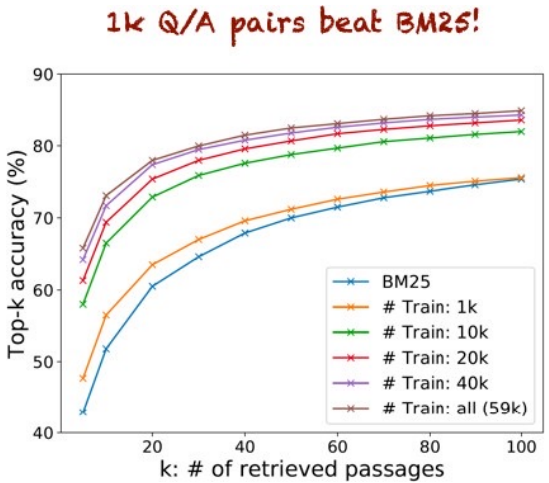
- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

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

We can train the retriever too

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

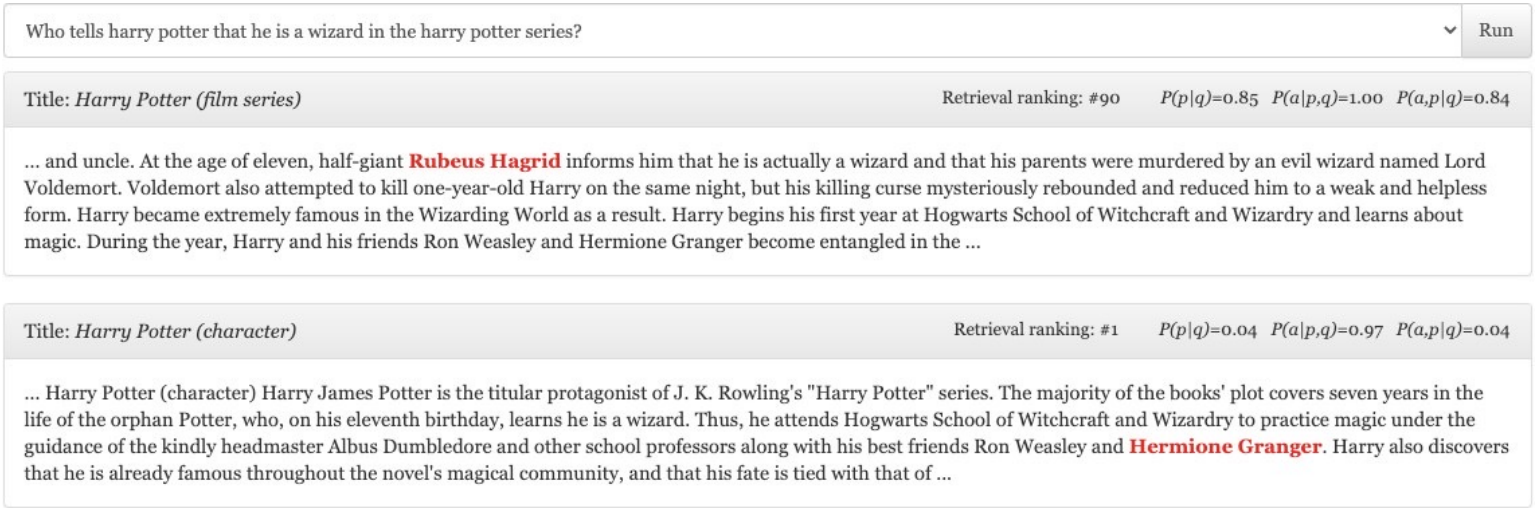




- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

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

We can train the retriever too



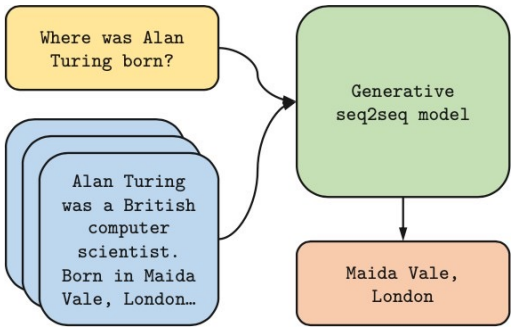
http://qa.cs.washington.edu:2020/

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

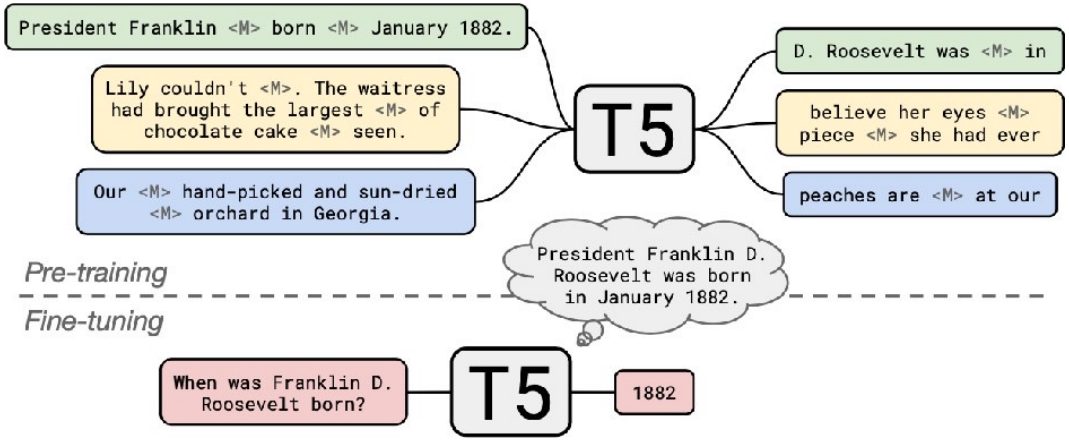


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9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

Large language models can do open-domain QA well

- ? ? no explicit IR



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

Conversational Question Answering

QuAC, CoQA, SParC, CoSQL

Multi-turn Questions

Answers depend on the context in dialogs

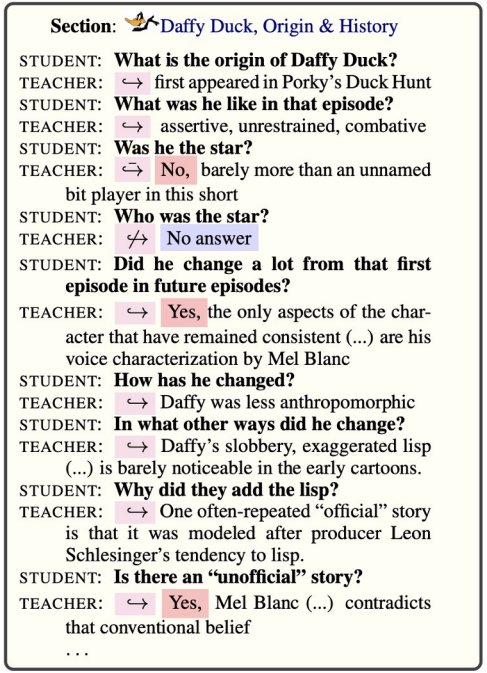


Figure 1: An example dialog about a Wikipedia section. The student, who does not see the section text, asks questions. The teacher provides a response in the form of a text span (or No answer), optionally yes or no (Yes / No), and encouragement about continuing a line of questioning (should, ↪, could ↩, or should not ↛ ask a follow-up question).

HotpotQA (Yang et al., 2018)

Reasoning Type	%	Example(s)
	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP*. **Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? (**yes**)
	15	**Paragraph A:** Several *current and former members of the Pittsburgh Pirates* ... John Milner, **Dave Parker**, and Rod Scurry... **Paragraph B:** **David Gene Parker**, *nicknamed "The Cobra"*, is an American former player in Major League Baseball... **Q:** Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"?
	6	**Paragraph A:** *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at *Marine Corps Air Station Cherry Point*... **Paragraph B:** *Marine Corps Air Station Cherry Point* ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ... **Q:** What city is the Marine Air Control Group 28 located in?
	2	**Paragraph A:** ... the towns of Yodobashi, **Okubo, Totsuka, and Ochiai town** *were merged into Yodobashi ward*. ... *Yodobashi Camera is a store with its name taken from the town and ward*. **Paragraph B**: *Yodobashi Camera* Co., Ltd. is *a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment*. **Q:** Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it's name?

HotpotQA (Yang et al., 2018)

Reasoning Type	%	Example(s)
Inferring the *bridge entity* to complete the 2nd-hop question (Type I)	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP*. **Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
Comparing two entities (Comparison)	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? (**yes**)
Locating the **answer entity** by checking multiple properties (Type II)	15	**Paragraph A:** Several *current and former members of the Pittsburgh Pirates* ... John Milner, **Dave Parker**, and Rod Scurry... **Paragraph B:** **David Gene Parker**, *nicknamed "The Cobra"*, is an American former player in Major League Baseball... **Q:** Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"?
Inferring about the property of an entity in question through a *bridge entity* (Type III)	6	**Paragraph A:** *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at *Marine Corps Air Station Cherry Point*... **Paragraph B:** *Marine Corps Air Station Cherry Point* ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ... **Q:** What city is the Marine Air Control Group 28 located in?
Other types of reasoning that require more than two supporting facts (Other)	2	**Paragraph A:** ... the towns of Yodobashi, **Okubo, Totsuka, and Ochiai town** *were merged into Yodobashi ward*. ... *Yodobashi Camera is a store with its name taken from the town and ward*. **Paragraph B**: *Yodobashi Camera* Co., Ltd. is *a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment*. **Q:** Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it's name?

Table 3: Types of multi-hop reasoning required to answer questions in the HOTPOTQA dev and test sets. We show in *orange bold italics* bridge entities if applicable, *blue italics* supporting facts from the paragraphs that connect directly to the question, and **green bold** the answer in the paragraph or following the question. The remaining 8% are single-hop (6%) or unanswerable questions (2%) by our judgement.

HotpotQA (Yang et al., 2018)

Multi-hop Reasoning: Explicitly multisteps through the text.

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Question Decomposition

Multi-hop Reading Comprehension through Question Decomposition and Rescoring (Min et al., 2019)

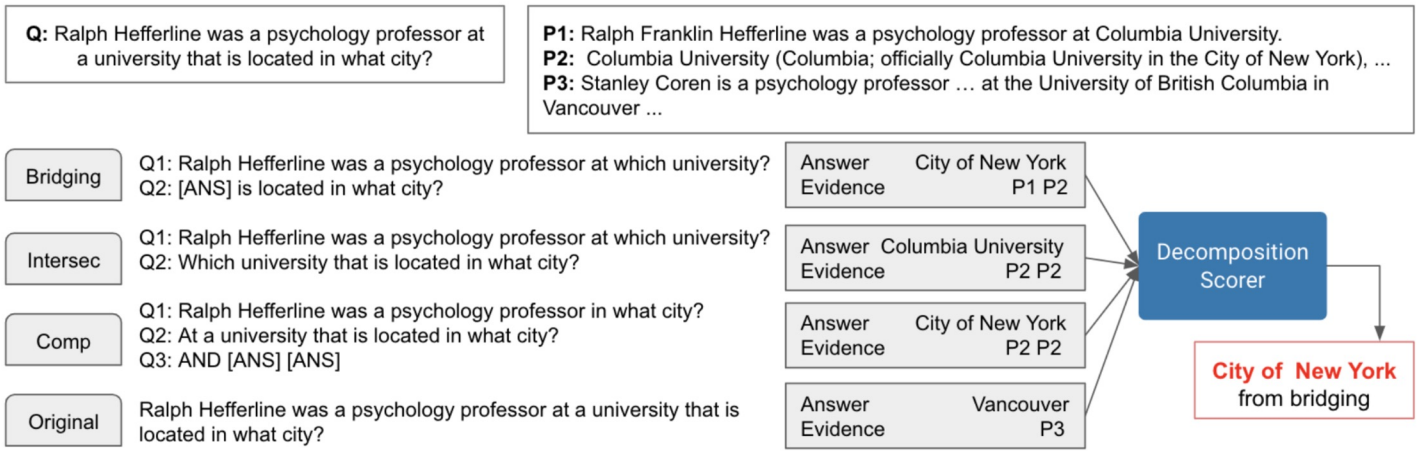


Figure 1: The overall diagram of how our system works. Given the question, DECOMPRC decomposes the question via all possible reasoning types (Section 3.2). Then, each sub-question interacts with the off-the-shelf RC model and produces the answer (Section 3.3). Lastly, the decomposition scorer decides which answer will be the final answer (Section 3.4). Here, "City of New York", obtained by bridging, is determined as a final answer.

Commonsense QA

CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge(Talmor et al., 2018)

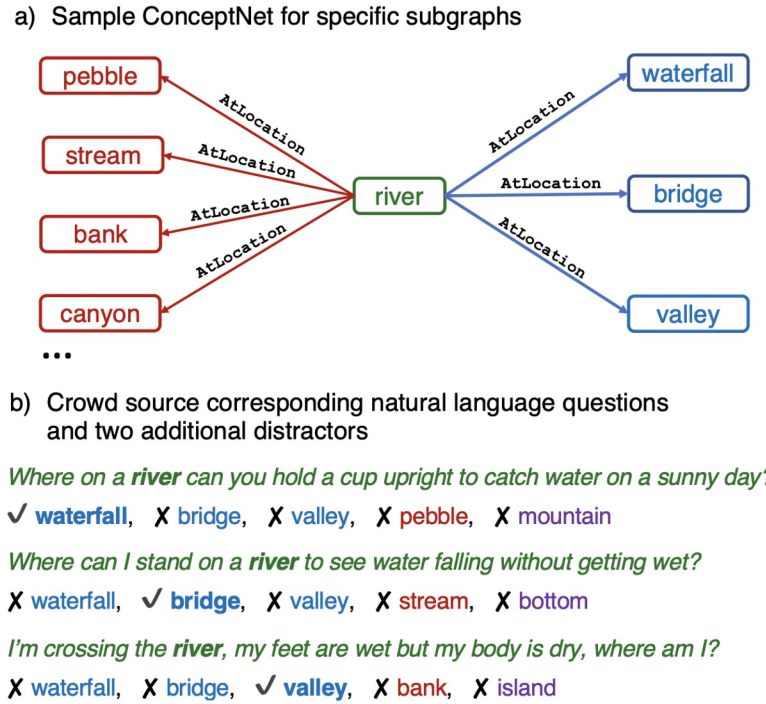


Figure 1: (a) A source concept (in green) and three target concepts (in blue) are sampled from CONCEPTNET (b) Crowd-workers generate three questions, each having one of the target concepts for its answer (✓), while the other two targets are not (✗). Then, for each question, workers choose an additional distractor from CONCEPTNET (in red), and author one themselves (in purple).

Numerical Reasoning QA

DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs (Dua et al., 2019)

A system must resolve multiple references in a question, map them onto a paragraph, and perform **discrete operations** over them (such as addition, counting, or sorting).

Reasoning	Passage (some parts shortened)	Question	Answer
	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile
	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller
	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992
	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay

UnifiedQA (Khashabi et al., 2020)

UnifiedQA system: Crossing Format Boundaries With a Single QA System

EX	**Dataset**	SQuAD 1.1
	Input	At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
	Output	16,000 rpm
AB	**Dataset**	NarrativeQA
	Input	What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...
	Output	fall in love with themselves
MC	**Dataset**	ARC-challenge
	Input	What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar
	Output	sugar
	Dataset	MCTest
	Input	Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...
	Output	The big kid
YN	**Dataset**	BoolQ
	Input	Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
	Output	no

New Directions in QA

Question Answering over Tables

Q: Which European countries have players who won the Australian Open at least 3 times?

Table 1: Matches

Id	Tourney	Year	Winner id	… …
1	Australian Open	2018	3	… …

Table 2: Ranking

Ranking	Points	Player id	Tours	… …
1	9,985	3	11	… …

Table 3: Players

Id	Name	Nation	Continent	… …
1	Djokovic	Serbia	Europe	… …
2	Osaka	Japan	Asia	… …

Semantic Parser

```
SELECT T1.nation
FROM players AS T1 JOIN matches AS T2
  ON T1.id = T2.winner_id
WHERE T2.Tourney = "Australian Open"
  AND T1.continent = "Europe"
GROUP BY T2.winner_id
HAVING COUNT(*) >= 3
```

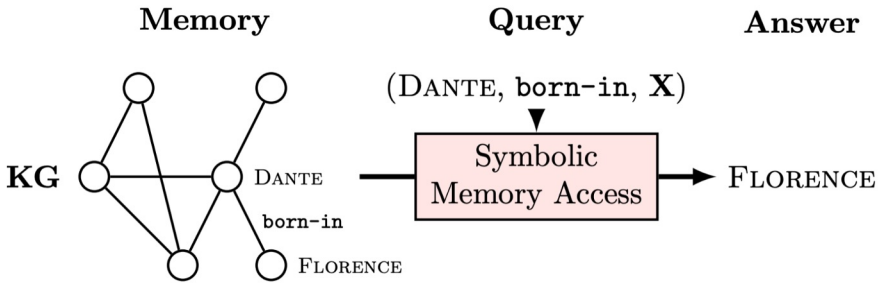
Switzerland
Serbia
...

Answers are
- extractive
- factoid
- short-form

WikiTableQuestion (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), WikiSQL (Zhong et al., 2017), Spider (Yu et al., 2018)

Knowledge Bases

Below is a symbolic knowledge base (Subject, Relation, Object)
Can we use LMs directly for QA? How?



Language Models as Knowledge Bases (Petroni et al., 2019)

Pretrained Language models are pretrained on large amounts of text; do they already capture knowledge in Wikipedia etc? Can we use LMs directly for QA?

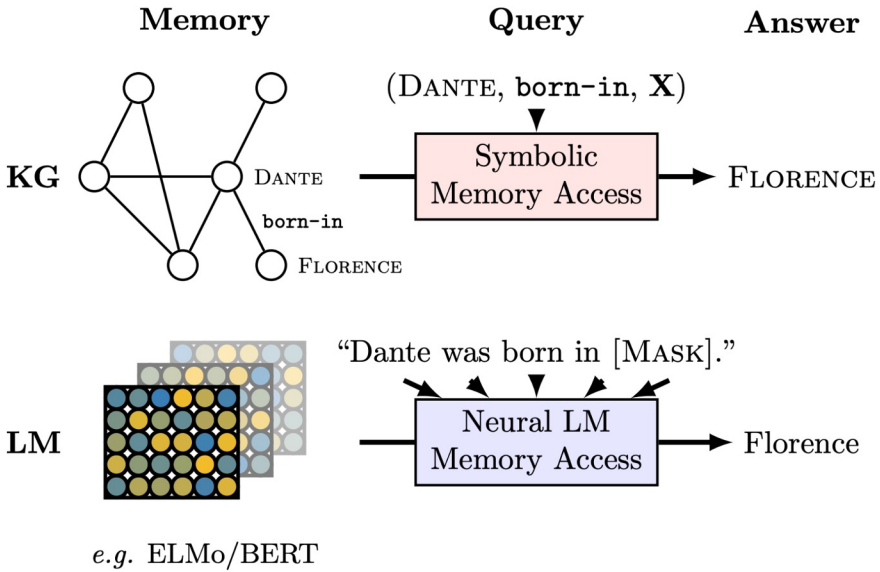


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Language Models as Knowledge Bases (Petroni et al., 2019)

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]**, Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]**, Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4], Pope [-2.5], **pope [-2.6]**, President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], **Capitol [-3.2]**, Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]**, Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], **ITV [-3.3]**, Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]**, Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], **Philippines [-3.6]**, February [-3.7], December [-3.8], Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

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

Close-book T5: Fine-tune Language models only with QA pairs without context such as documents.

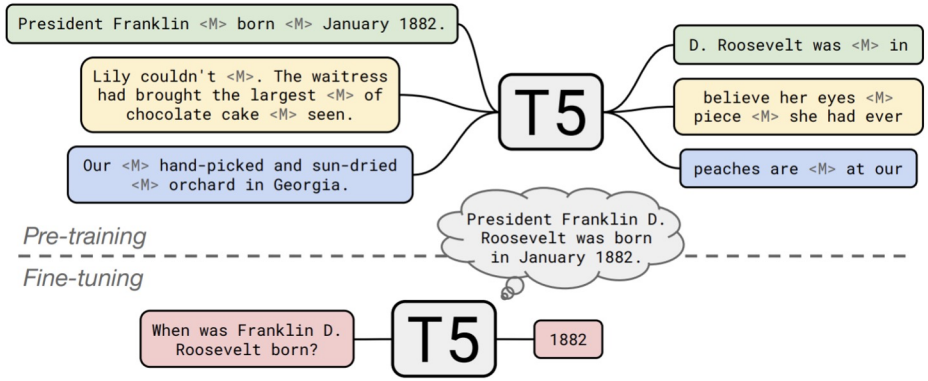


Figure 1: T5 is pre-trained to fill in dropped-out spans of text (denoted by <M>) from documents in a large, unstructured text corpus. We fine-tune T5 to answer questions without inputting any additional information or context. This forces T5 to answer questions based on "knowledge" that it internalized during pre-training.

Use Semiparametric Models to Retrieve Context

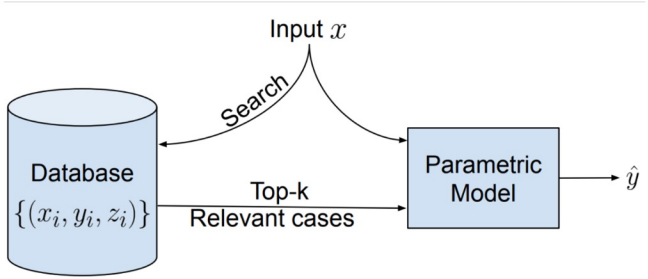
Semiparametric Methods in NLP: Decoupling Logic from Knowledge

Spa-NLP @ ACL 2022 Home Call for Papers Organization

Overview

Large parametric language models have achieved dramatic empirical success across many applications. However, these models lack several desirable properties such as explainability (providing provenance), privacy (ability to remove knowledge from the model), robust controllability, and debuggability. On the other hand, nonparametric models provide many of these features by design such as provenance, ability to incorporate/remove information. However, these models often suffer from weaker empirical performance as compared to deep parametric models.

Recently, many works have independently proposed a middle ground that combines a parametric model (that encodes logic) with a nonparametric model (that retrieves knowledge) in various areas from question answering over natural languages to complex reasoning over knowledge bases to even protein structure predictions. Given the increasingly promising results on various tasks of such semiparametric model, we believe this area is ripe for targeted investigation on understanding efficiency, generalization, limitations, widening its applicability, etc. As a result, we want to host a workshop on this topic.



www.semiparametric.ml

REALM (Guu et al., 2020)

REALM: **Retrieval-Augmented** Language Model Pre-Training

These pre-trained models, such as BERT and RoBERTa, have been shown to *memorize a surprising amount of world knowledge*, such as “the birthplace of Francesco Bartolomeo Conti”, “the developer of JDK” and “the owner of Border TV”. … these models memorize knowledge *implicitly* — i.e., world knowledge is captured in an abstract way in the model weights …

Instead, what if there was a method for pre-training that could access knowledge *explicitly*, e.g., by referencing an additional large external text corpus, in order to achieve accurate results without increasing the model size or complexity?

https://ai.googleblog.com/2020/08/realm-integrating-retrieval-into.html