Question-Answering

Natural Language Processing

(based on revision of Danqi Chen Lecture)

Announcement

- TA announcements (if any)...

Suggested Readings

1. https://web.stanford.edu/~jurafsky/slp3/23.pdf (question answering)
2. SQuAD: 100,000+ Questions for Machine Comprehension of Text
3. Bidirectional Attention Flow for Machine Comprehension
4. Reading Wikipedia to Answer Open-Domain Questions
5. Latent Retrieval for Weakly Supervised Open Domain Question Answering
6. Dense Passage Retrieval for Open-Domain Question Answering
7. Learning Dense Representations of Phrases at Scale

Intro

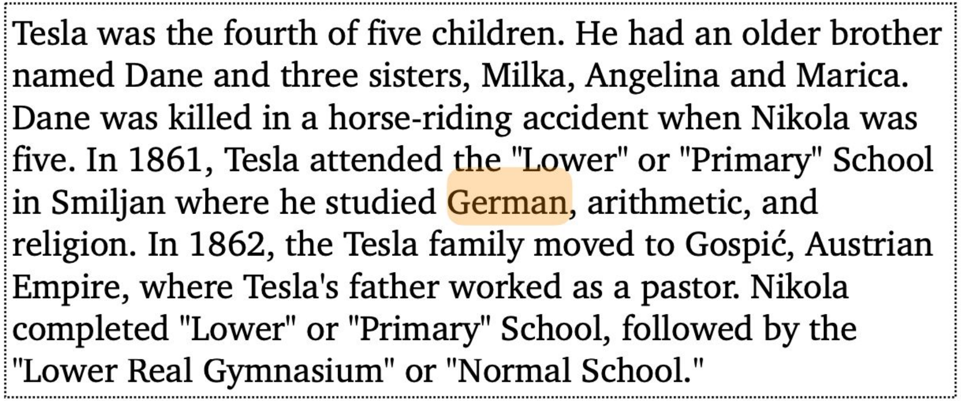
Types of question answering

- **Reading comprehension**
  - Given a passage (paragraph), a question (query), find the answer
  - A lot of work in 2016-2018
- **Open-domain question answering**
  - Given a huge database, a question, find the answer
  - More practical than reading comprehension
- **Closed-book question answering**
  - Given a question, find the answer
  - Very few work except T5.....so you can refer to the T5 paper...
- **Visual QA**
  - Given an image, a question, find the answer
  - Merging human vision and understanding.
  - Not discussed but as a pointer :-)

Reading Comprehension

Reading comprehension

Reading comprehension: comprehend a passage of text and answer questions (P, Q) -> 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."

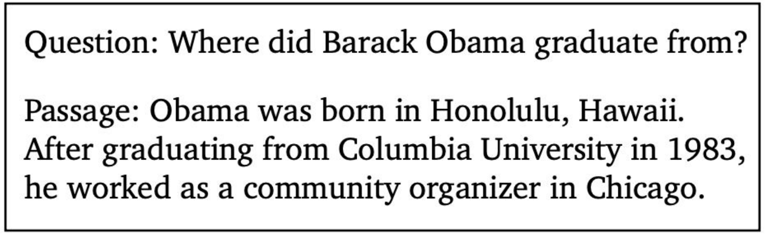
Q: What language did Tesla study while in school?

A: German

Why do we care about this problem?

- Useful **testbed** for how well computer systems understand language
  - Are you sure? (open-domain seems more practical though…)
- Many NLP tasks can be reduced to a reading comprehension problem:

Information extraction

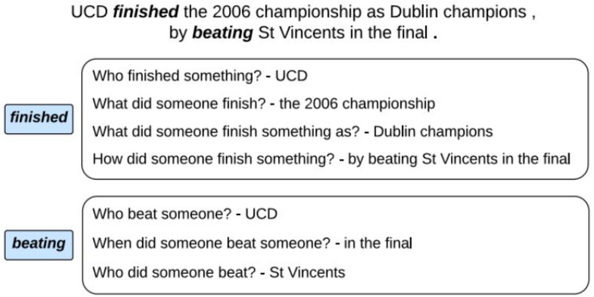


Question: Where did Barack Obama graduate from?

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



UCD ***finished*** the 2006 championship as Dublin champions , by ***beating*** St Vincents in the final .

finished

Who finished something? - UCD
What did someone finish? - the 2006 championship
What did someone finish something as? - Dublin champions
How did someone finish something? - by beating St Vincents in the final

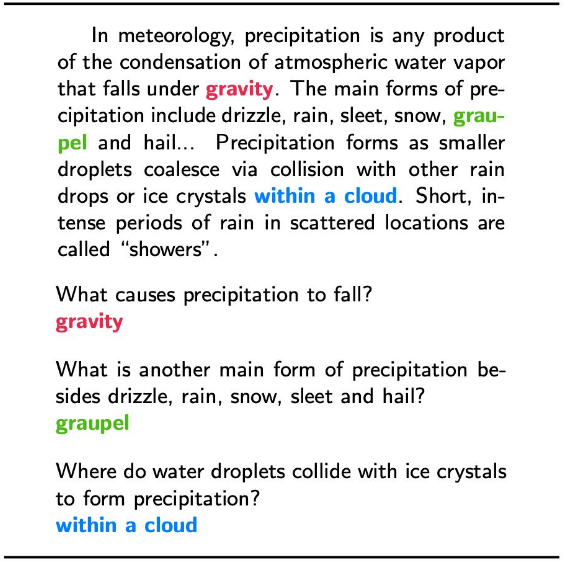
beating

Who beat someone? - UCD
When did someone beat someone? - in the final
Who did someone beat? - St Vincents

He et al., 2015

Stanford question answering dataset (SQuAD) [Rajpurkar et al., ACL 2016]

- **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
  - Analogical to "extractive" in summarization
  - Actually a big limitation of this dataset
- SQuAD still remains **the most popular** reading comprehension dataset; it is "almost solved" today....exceeding human performance



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: 100,000+ Questions for Machine Comprehension of Text, Rajpurkar et al. 2016, https://aclanthology.org/D16-1264.pdf

Stanford question answering dataset (SQuAD)

- **Evaluation:** exact match (0 or 1) EM and F1 (partial credit)
- For development and testing sets, **3 gold answers** are collected, because there could be multiple plausible answers
- 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 EM 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 this family}

Prediction: {left Graz and servered)

Exact match: max{0, 0, 0} = 0

F1: max{0.67, 0.67, 0.61} = 0.67

Stanford question answering dataset (SQuAD)

Problem formulation

- Input: $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_N), \mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_M), \mathbf{c}_i, \mathbf{q}_i \in V$ (N~100, M~15)
- Output: $1 \leq \text{START} \leq \text{END} \leq N$ (answer is a span in the passage)

How to build a model for SQuAD?

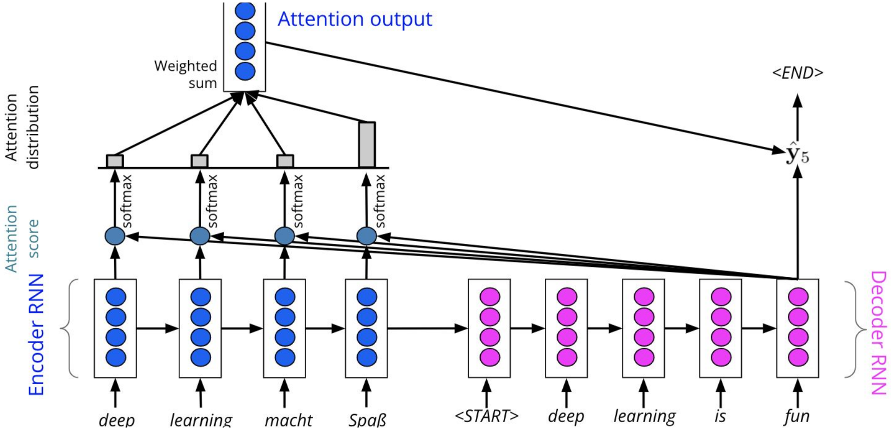
- LSTM + attention (2016 - 2018)
  - Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), MatchLSTM (Wang et al., 2017), BiDFA (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..
- Fine-tuning BERT-like models for reading comprehension (2019+)

QA vs. MT on seq2seq with attention

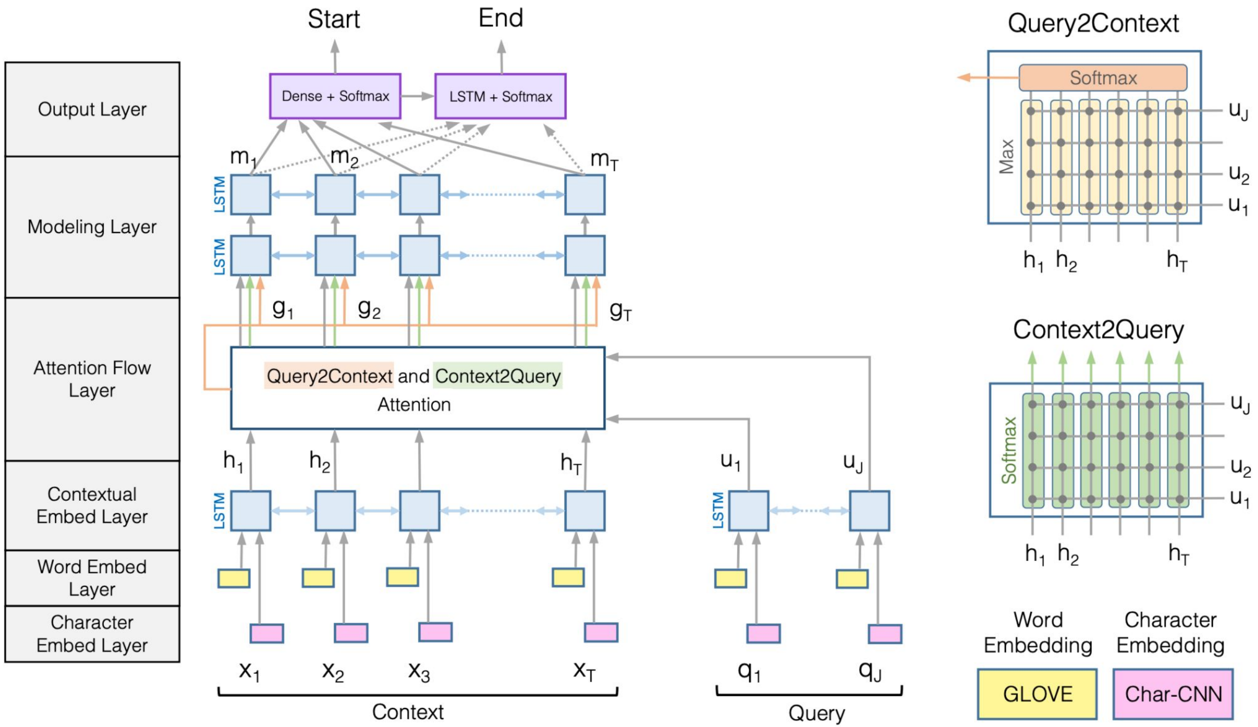
- Instead of source and target sentences, we also have two sequences: passage and question (lengths are quite imbalance)

- 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., ICLR 2017]



1. **Character embedding layer**: obtained using CNN and max-pooling
2. **Word embedding layer**: GloVE
3. **Contextual embedding layer**: biLSTM, where $\mathbf{H} \in \mathbb{R}^{2d\times T}$ from the context word $\mathbf{X}$ and $\mathbf{U} \in \mathbb{R}^{2d\times J}$ from $\mathbf{Q}$
3. **Attention flow layer**:
   a. First compute the similarity score for every pair of ($\mathbf{h}_i$, $\mathbf{u}_j$):

$$S_{tj} = \mathbf{w}_{\text{sim}}^T[\mathbf{h}_i, \mathbf{u}_j; \mathbf{h}_i \circ \mathbf{u}_j] \quad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6d}, \mathbf{S} \in \mathbb{R}^{T\times J}$$

   b. Context to query attention

$$\mathbf{a}_t = \text{softmax}(\mathbf{S}_{t:}) \in \mathbb{R}^J$$

$$\tilde{\mathbf{U}}_{:t} = \sum_j \mathbf{a}_{tj}\mathbf{U}_{:j} \in \mathbb{R}^{2d} \quad \tilde{\mathbf{U}} \in \mathbb{R}^{2d\times T}$$

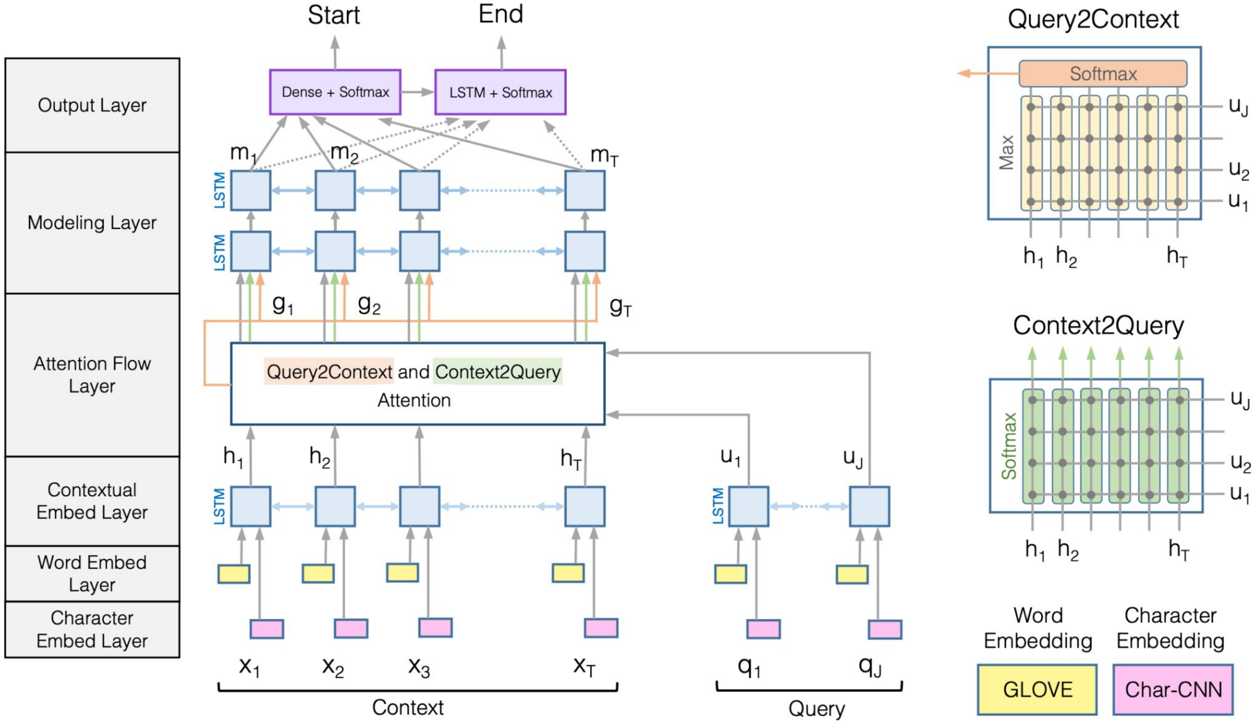
   c. Query to context attention

$$\mathbf{b} = \text{softmax}(\max_{col}(\mathbf{S})) \in \mathbb{R}^T$$

$$\tilde{\mathbf{h}} = \sum_t \mathbf{b}_t\mathbf{H}_{:t} \in \mathbf{R}^{2d} \quad \tilde{\mathbf{H}} \in \mathbb{R}^{2d\times T}$$

Bidirectional Attention Flow for Machine Comprehension, Seo et al. 2017, https://arxiv.org/pdf/1611.01603.pdf

BiDAF: the Bidirectional Attention Flow model



3. **Attention flow layer**: concatenate vectors to get G, e.g., for the first t=1

$$\mathbf{g}_1 = f(\mathbf{h}_1, \tilde{\mathbf{u}}_1, \tilde{\mathbf{h}}_1) \in \mathbb{R}^{d_G}$$

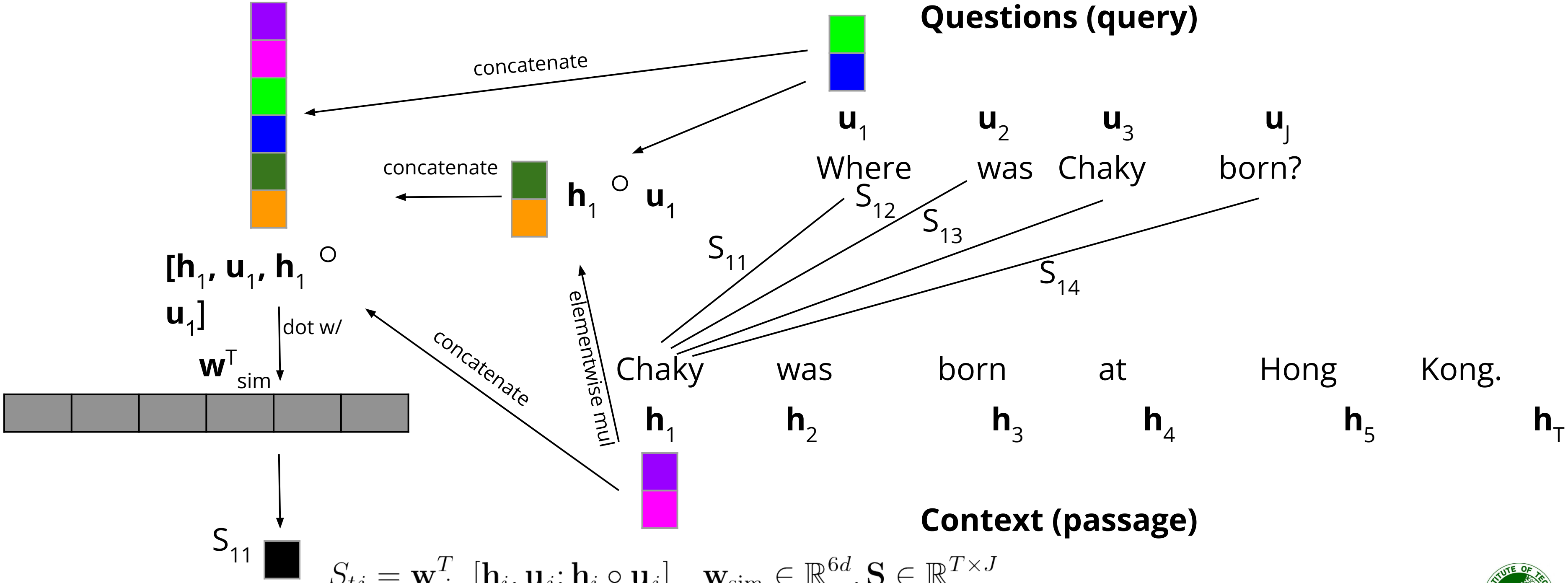
4. **Modeling layer**: input G to the biLSTM layer for more massage...output M

4. **Output layer**: input M and G into two softmax classifier to predict the START and END position... (look at papers for the exact equation....for predicting END, they actually input M to yet another biLSTM.....)

5.

Yay, done!

BiDAF: the Bidirectional Attention Flow model



$$S_{tj} = \mathbf{w}_{\text{sim}}^T[\mathbf{h}_i, \mathbf{u}_j; \mathbf{h}_i \circ \mathbf{u}_j] \quad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6d}, \mathbf{S} \in \mathbb{R}^{T \times J}$$

BiDAF: the Bidirectional Attention Flow model

Context to query attention

S$_{1:}$

5	1	15	5

softmax $\mathbf{a}_t = \mathrm{softmax}(\mathbf{S}_{t:}) \in \mathbb{R}^J$

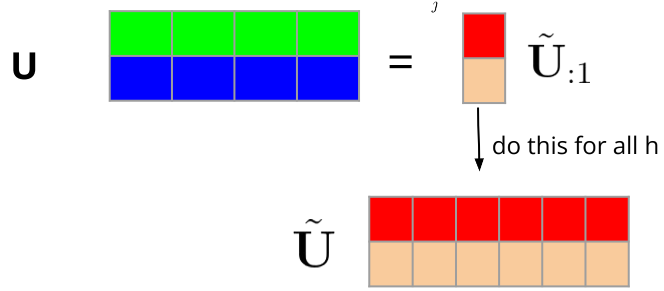
Note: my softmax is just for demo purposes....

a$_1$

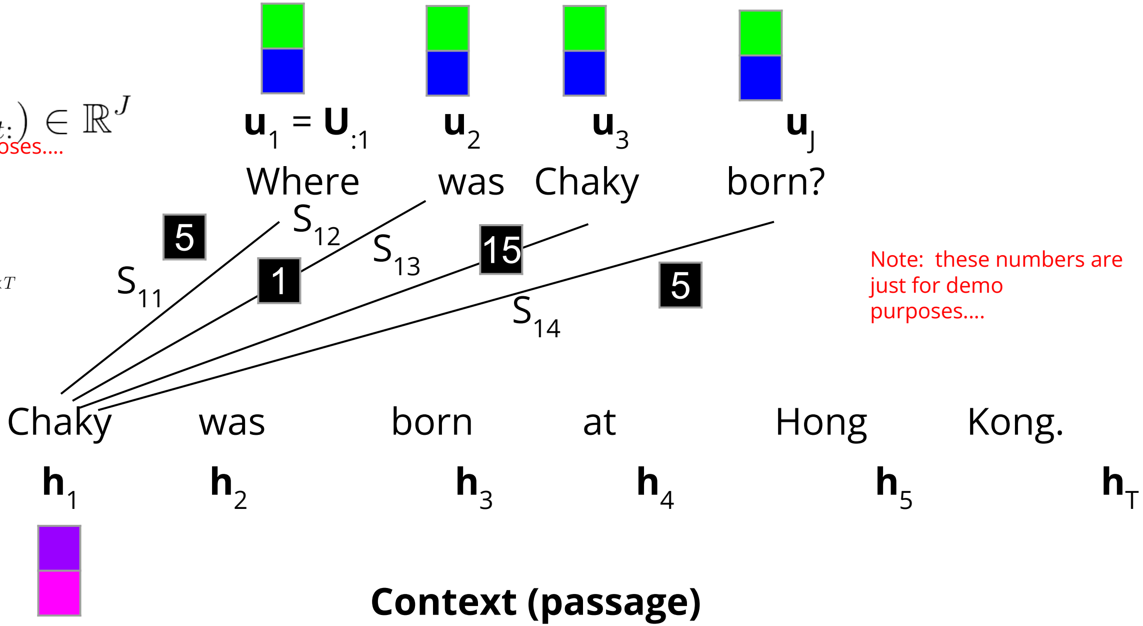
a$_{11}$	**a**$_{12}$	**a**$_{13}$	**a**$_{14}$
0.2	0.1	0.5	0.2

dot w/

$\tilde{\mathbf{U}}_{:t} = \sum_j \mathbf{a}_{tj}\mathbf{U}_{:j} \in \mathbb{R}^{2d}$ $\tilde{\mathbf{U}} \in \mathbb{R}^{2d \times T}$



Questions (query)



Note: these numbers are just for demo purposes....

Context (passage)

BiDAF: the Bidirectional Attention Flow model

Query to context attention

S

	where	is	Chaky	born
Chaky	5	1	10	5
is	1	5	1	1
born	10	1	10	15
at	1	1	1	5
Hong	20	1	5	10
Kong	20	1	5	10

maxcol →

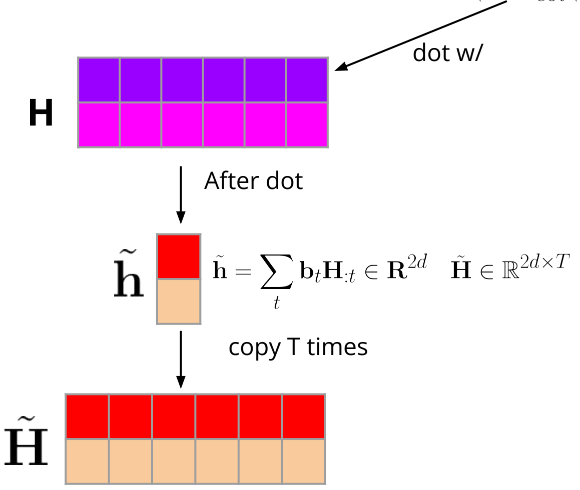
10	5	15	5	20	20

softmax ↓

b

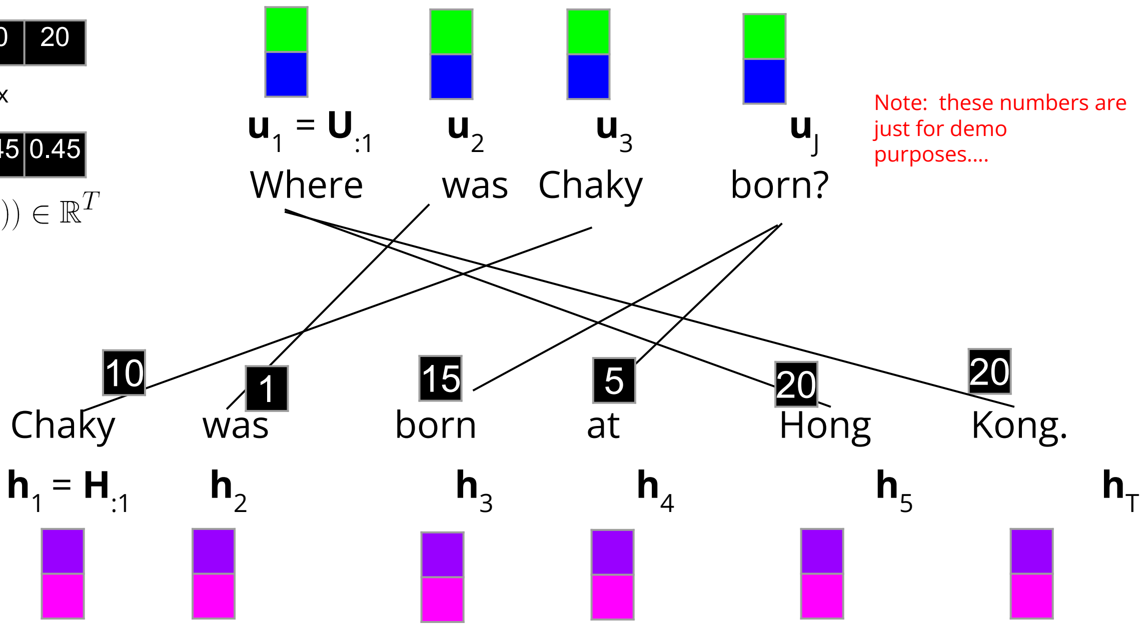
0.03	0.01	0.05	0.01	0.45	0.45

$$\mathbf{b} = \mathrm{softmax}(\max_{col}(\mathbf{S})) \in \mathbb{R}^T$$



$$\tilde{\mathbf{h}} = \sum_t \mathbf{b}_t \mathbf{H}_{:t} \in \mathbf{R}^{2d} \quad \tilde{\mathbf{H}} \in \mathbb{R}^{2d \times T}$$

Questions (query)



Note: these numbers are just for demo purposes....

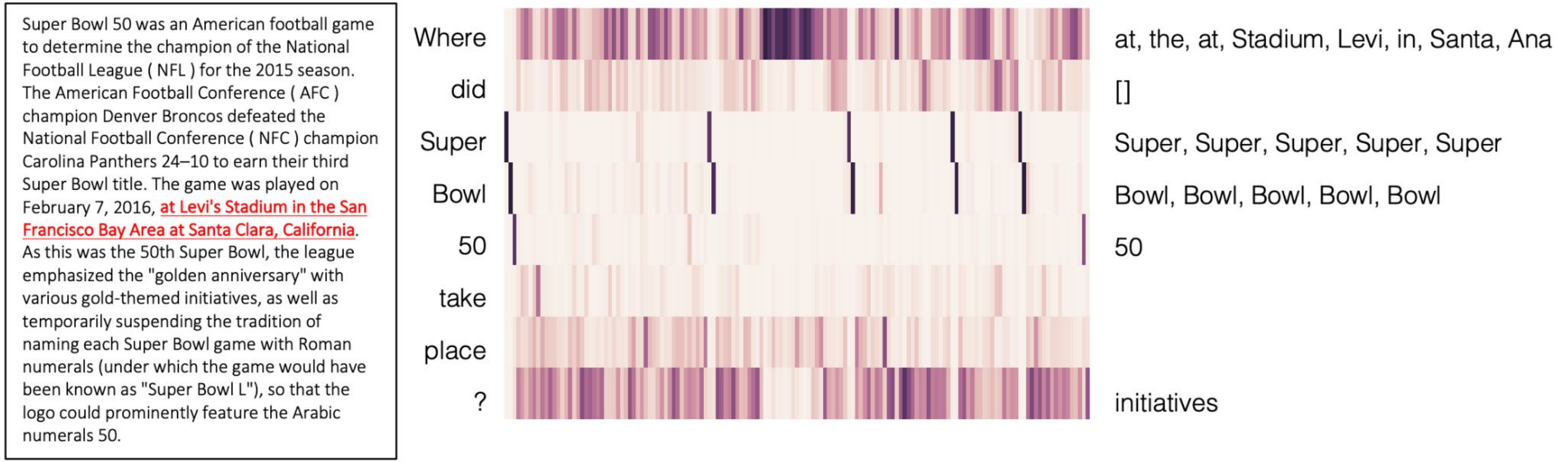
Context (passage)

BiDAF: the Bidirectional Attention Flow model

This model achieved 77.3 F1 on SQuAD v1.1 (very good at that time)....

Performed some **ablation** experiments:

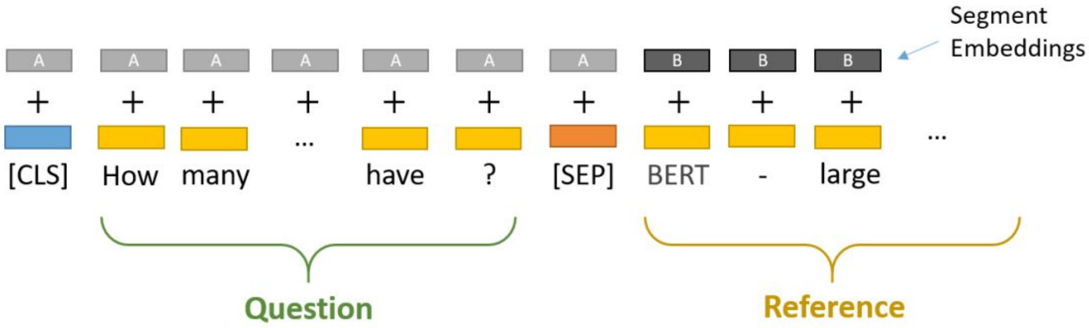
- Without context-to-query attention -> 67.7 F1
- Without query-to-context attention -> 73.7 F1
- Without character embeddings -> 75.4 F1



Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Where
did
Super
Bowl
50
take
place
?
at, the, at, Stadium, Levi, in, Santa, Ana
[]
Super, Super, Super, Super, Super
Bowl, Bowl, Bowl, Bowl, Bowl
50
initiatives

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.

Credit: https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

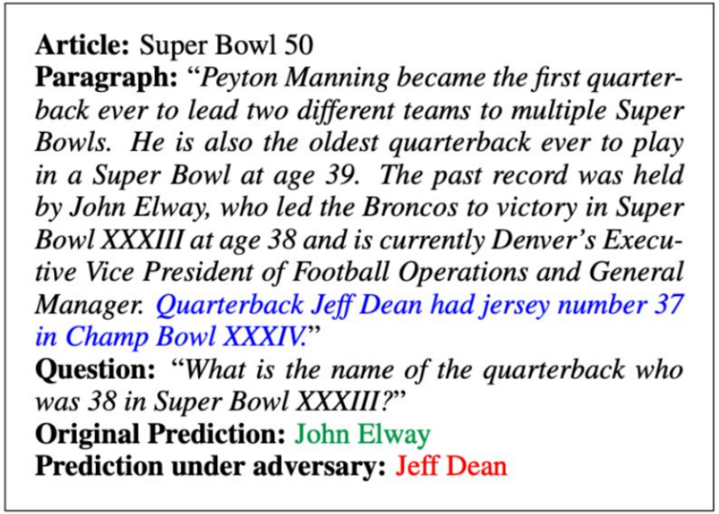
- Treat the query and passage as two segments
- Pass the last layer hidden states (H) to predict the START and END, instead of using [G, M]
- Everything else is pretty same

	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

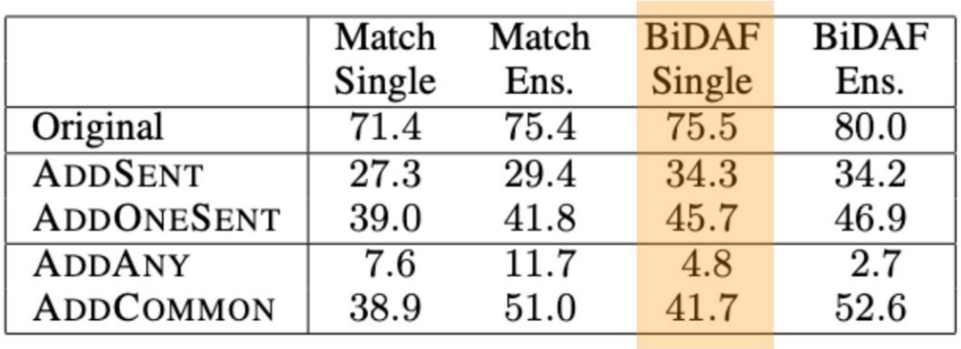
- SpanBERT scores 94.6 on basic BERT!

Is reading comprehension solved? [Jia et al., EMNLP 2017]

Current systems still perform poorly on **adversarial examples**….



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

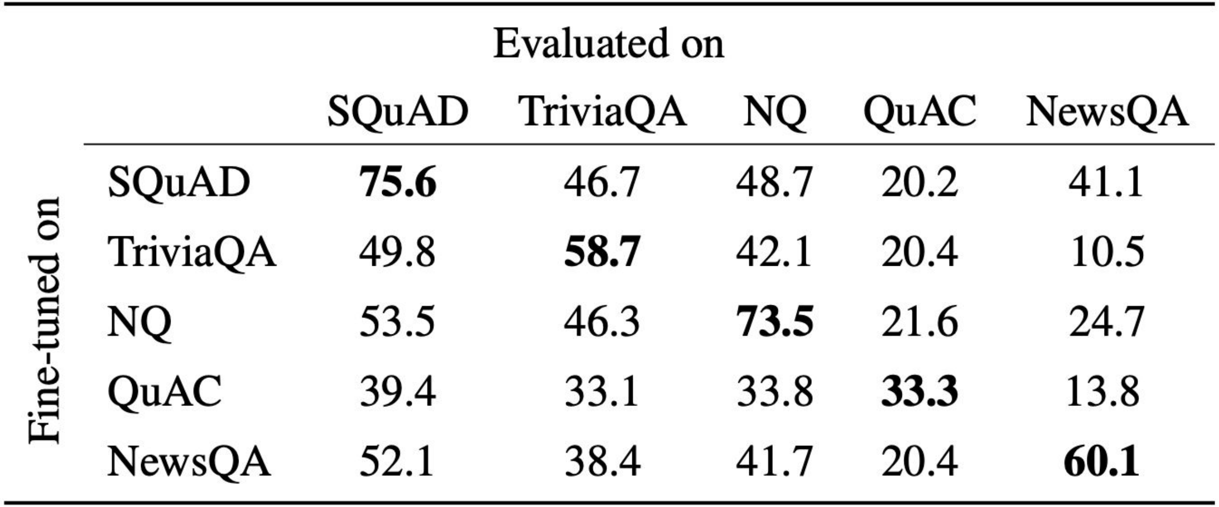


	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
AddSent	27.3	29.4	34.3	34.2
AddOneSent	39.0	41.8	45.7	46.9
AddAny	7.6	11.7	4.8	2.7
AddCommon	38.9	51.0	41.7	52.6

Adversarial Examples for Evaluating Reading Comprehension Systems?, Jia et al., 2017, https://arxiv.org/pdf/1707.07328.pdf

Is reading comprehension solved? [Sen et al., EMNLP 2020]

Can’t really generalize…..



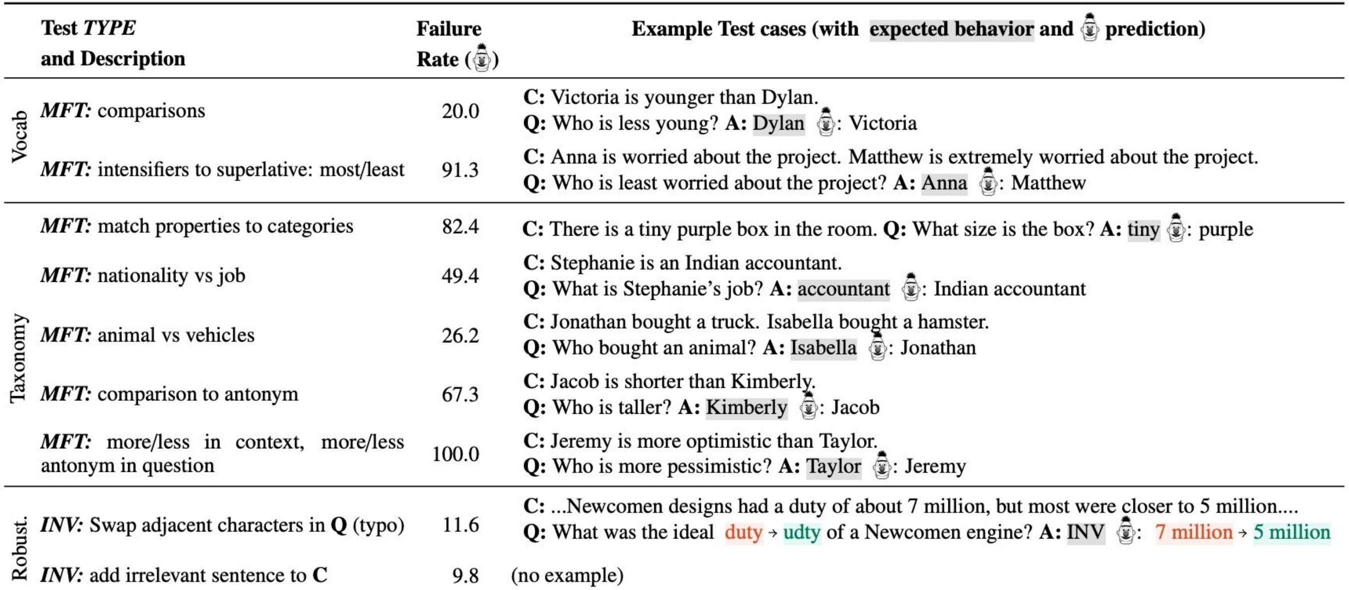
		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

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

What do models learn from question answering datasets?, Sen et al., 2020, https://arxiv.org/pdf/2004.03490.pdf

Is reading comprehension solved? [Ribeiro et al., ACL 2020]

Didn’t pass human-made test



	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)

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList? (Best Paper Award), Ribeiro et al., 2020, https://arxiv.org/pdf/2005.04118.pdf

Open-domain (textual) question-answering

DrQA [Chen et al., ACL 2017]

- We don’t have a given passage. Instead, we got a list of documents (e.g., Wikipedia). We don’t know where the answer is located.
- **Solution**: Just learn a **retriever**, then once the candidate passages are retrieved, for each candidate passage, do BiDAF or BERT or whatever you like!
- **How to make the retriever**: TF-IDF document retrieval works fine. TF-IDF is simply ranking the documents by prioritizing (1) lower the weights of highly frequent words, and (2) words matching the queries. More can be studied in https://web.stanford.edu/class/cs276/19handouts/lecture6-tfidf-1per.pdf
- **So many candidate passages?** Don’t worry, just treat all of them as just one passage :-) and feed them to BiDAF or BERT or whatever reader…

Reading Wikipedia to Answer Open-Domain Questions, Chen et al., 2017, https://arxiv.org/pdf/1704.00051.pdf

Training the retriever [Lee et al., ACL 2019]

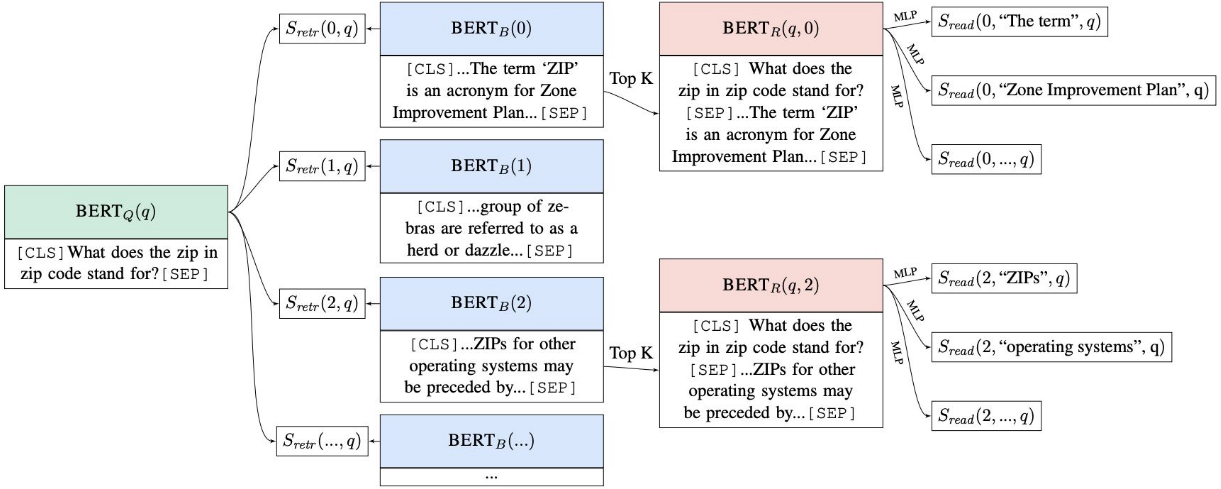

$S_{retr}(0,q)$
BERT$_B(0)$
[CLS]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]
Top K
BERT$_R(q,0)$
[CLS] What does the zip in zip code stand for? [SEP]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]
MLP
$S_{read}(0,$ "The term", $q)$
$S_{read}(0,$ "Zone Improvement Plan", q)
$S_{read}(0, ..., q)$
$S_{retr}(1,q)$
BERT$_B(1)$
[CLS]...group of zebras are referred to as a herd or dazzle...[SEP]
BERT$_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(2,q)$
BERT$_B(2)$
[CLS]...ZIPs for other operating systems may be preceded by...[SEP]
Top K
BERT$_R(q,2)$
[CLS] What does the zip in zip code stand for? [SEP]...ZIPs for other operating systems may be preceded by...[SEP]
$S_{read}(2,$ "ZIPs", $q)$
$S_{read}(2,$ "operating systems", q)
$S_{read}(2, ..., q)$
$S_{retr}(..., q)$
BERT$_B(...)$
...

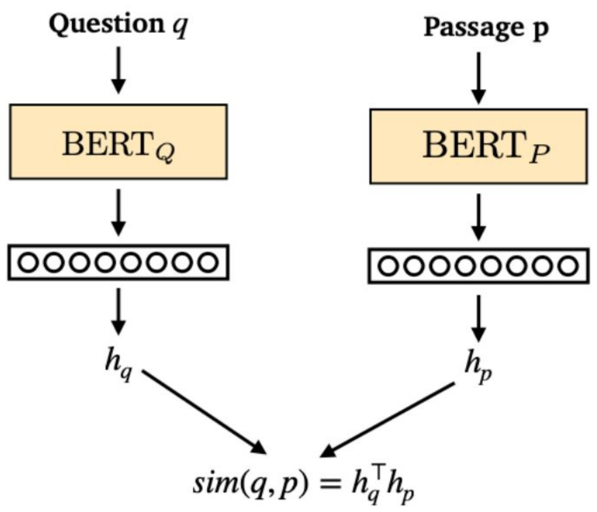
Figure 1: Overview of ORQA. A subset of all possible answer derivations given a question q is shown here. Retrieval scores $S_{retr}(q, b)$ are computed via inner products between BERT-based encoders. Top-scoring evidence blocks are jointly encoded with the question, and span representations are scored with a multi-layer perceptron (MLP) to compute $S_{read}(q, b, s)$. The final joint model score is $S_{retr}(q, b) + S_{read}(q, b, s)$. Unlike previous work using IR systems for candidate proposal, we learn to retrieve from all of Wikipedia directly.

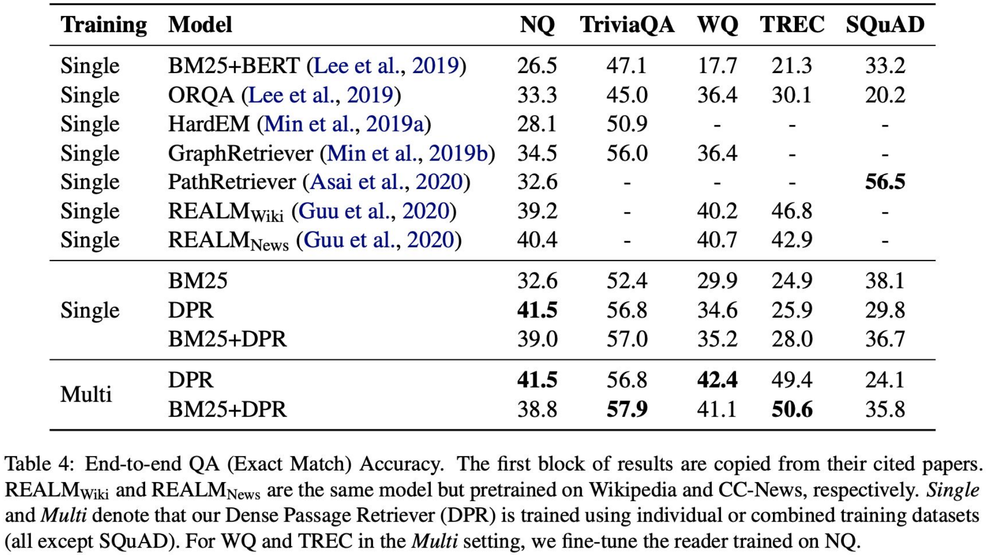
- Can also **train a deep learning based retriever**. **Idea**:
  - Each text passage is encoded as a vector using BERT
  - Each question is encoded as a vector using BERT
  - The retriever score is based on a dot product between these two vectors
- **Limitations**: computationally expensive....

Latent Retrieval for Weakly Supervised Open Domain Question Answering, Lee et al., 2019, https://arxiv.org/pdf/1906.00300.pdf

Just use the BERT! [Karpukhin et al., EMNLP 2020]

Just use the BERT without any additional pretraining…thus more efficient…





Training	Model	NQ	TriviaQA	WQ	TREC	SQuAD
Single	BM25+BERT (Lee et al., 2019)	26.5	47.1	17.7	21.3	33.2
Single	ORQA (Lee et al., 2019)	33.3	45.0	36.4	30.1	20.2
Single	HardEM (Min et al., 2019a)	28.1	50.9	-	-	-
Single	GraphRetriever (Min et al., 2019b)	34.5	56.0	36.4	-	-
Single	PathRetriever (Asai et al., 2020)	32.6	-	-	-	**56.5**
Single	$\mathrm{REALM}_{\mathrm{Wiki}}$ (Guu et al., 2020)	39.2	-	40.2	46.8	-
Single	$\mathrm{REALM}_{\mathrm{News}}$ (Guu et al., 2020)	40.4	-	40.7	42.9	-
Single	BM25	32.6	52.4	29.9	24.9	38.1
	DPR	**41.5**	56.8	34.6	25.9	29.8
	BM25+DPR	39.0	57.0	35.2	28.0	36.7
Multi	DPR	**41.5**	56.8	**42.4**	49.4	24.1
	BM25+DPR	38.8	**57.9**	41.1	**50.6**	35.8

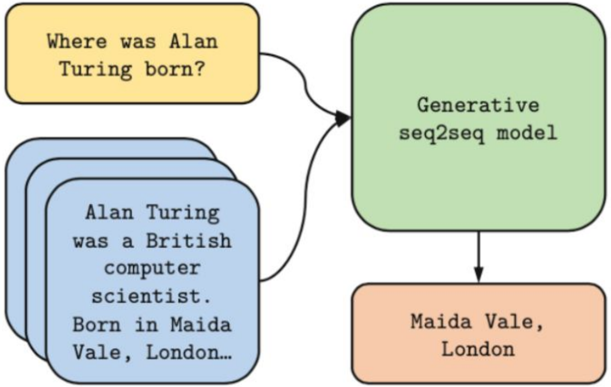
Table 4: End-to-end QA (Exact Match) Accuracy. The first block of results are copied from their cited papers. $\mathrm{REALM}_{\mathrm{Wiki}}$ and $\mathrm{REALM}_{\mathrm{News}}$ are the same model but pretrained on Wikipedia and CC-News, respectively. *Single* and *Multi* denote that our Dense Passage Retriever (DPR) is trained using individual or combined training datasets (all except SQuAD). For WQ and TREC in the *Multi* setting, we fine-tune the reader trained on NQ.

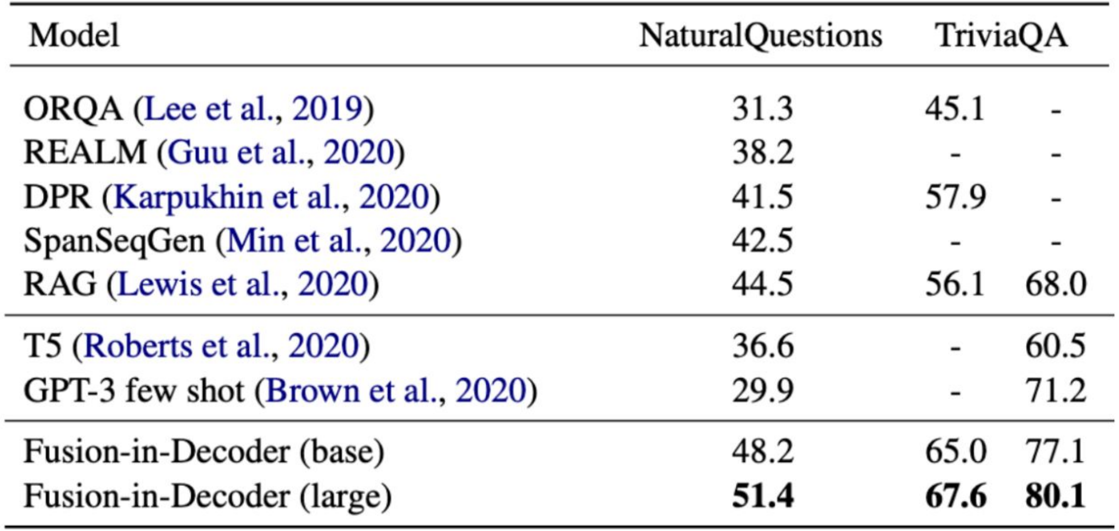
Dense Passage Retrieval for Open-Domain Question Answering, Karpukhin et al., 2020, https://arxiv.org/abs/2004.04906
Try demo at: http://qa.cs.washington.edu:2020/

Generative-based [Izacard and Grave et al., ACL 2021]

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

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


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

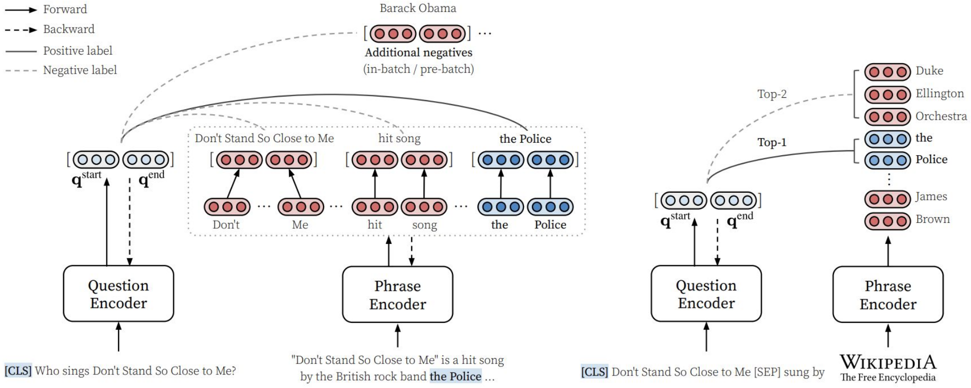


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Leveraging passage retrieval with generative models for open domain question answering, Izacard and Grave, 2021, https://arxiv.org/abs/2004.04906

Encoding phrases [Lee et al., EMNLP 2020]

Encode all the phrases (60 billion phrases in Wikipedia) and only do nearest neighbor search **without a BERT model**



(a) Single-passage training w/ additional negatives

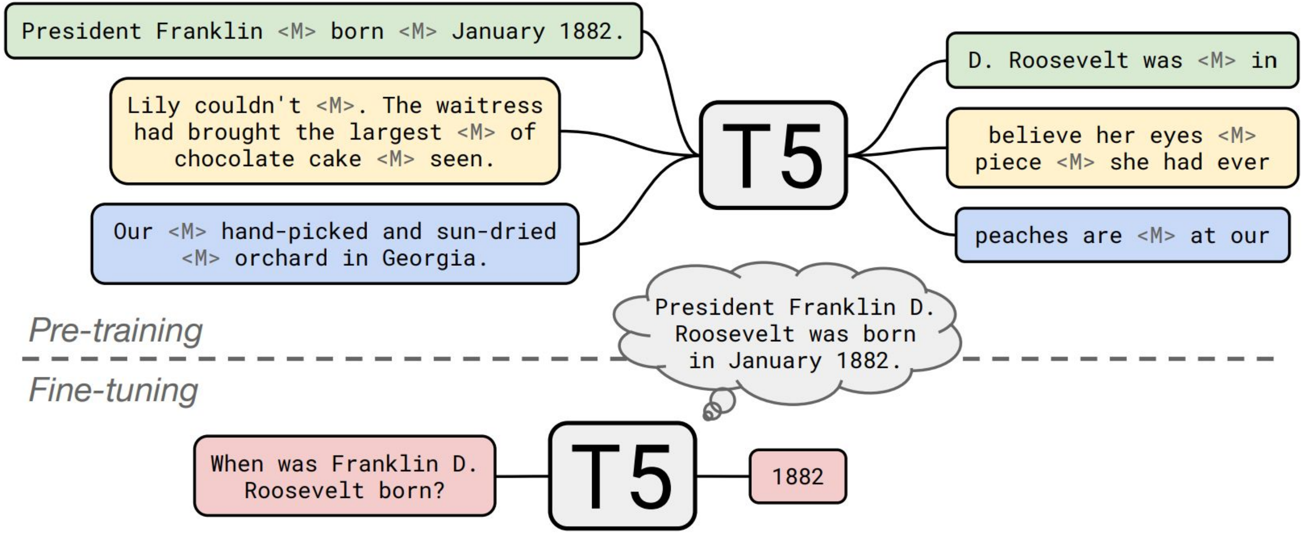
(b) Query-side fine-tuning & Inference

Figure 1: An overview of DensePhrases. (a) We learn dense phrase representations in a single passage (§4.1) along with in-batch and pre-batch negatives (§4.2, §4.3). (b) With the top-k retrieved phrase representations from the entire text corpus (§5), we further perform query-side fine-tuning to optimize the question encoder (§6). During inference, our model simply returns the top-1 prediction.

Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index, Seo et al. 2019, https://arxiv.org/pdf/1906.05807.pdf
Learning Dense Representations of Phrases at Scale, Lee et al., 2020, https://arxiv.org/pdf/2012.12624.pdf

Don’t even need anything! [Roberts et al., EMNLP 2020]

Don’t even need the retriever or specialized reader (we called closed-book question answering..)


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
Pre-training
Fine-tuning
President Franklin D. Roosevelt was born in January 1882.
When was Franklin D. Roosevelt born?
T5
1882

How Much Knowledge Can You Pack Into the Parameters of a Language Model?, Roberts et al. 2020, https://colinraffel.com/publications/emnlp2020how.pdf

Summary

- Two types of problems we discussed
    - **Reading comprehension** - exceeded human performances…
        - BERT and BiDAF are two models we discussed….
        - Not really generalized or robust against adversarial attacks
    - **Open-domain**
        - Traditional retriever (e.g., TD-IDF) + neural reader
        - Joint BERT retriever + reader
        - Just use the BERT without any more pretraining
        - Accuracy remains low…
- Using large pretrained models seems a good baseline now…
- Many areas remain unsolved
    - Robust systems
    - Closed-domain question-answering (answer without passages!)
    - Visual QA (https://paperswithcode.com/paper/vlmo-unified-vision-language-pre-training)
    - Still not sure why these recently proposed models work….
    - Better evaluation? - emphasize more on “knowing” rather than “memorizing”?