Question Answering

Deep Natural Language Processing Class, 2022
Paweł Budzianowski

Practical 5 is out

Part 1: Task-oriented dialogue system

Part 2: Combining generative models with task-specific problems

Due date: 17th of May

Polish NLProc Group

Polish Natural Language Processing Meetup Group

Warsaw, Poland

527 members · Public group

Organized by **Darek Kleczek** and **1 other**

Edit group info

Share:

Plan today

1. Question-Answering
2. Closed-domain QA
3. Open-domain QA

4. Subwords model
5. Byte-level models

What is question-answering?

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

What is question-answering?

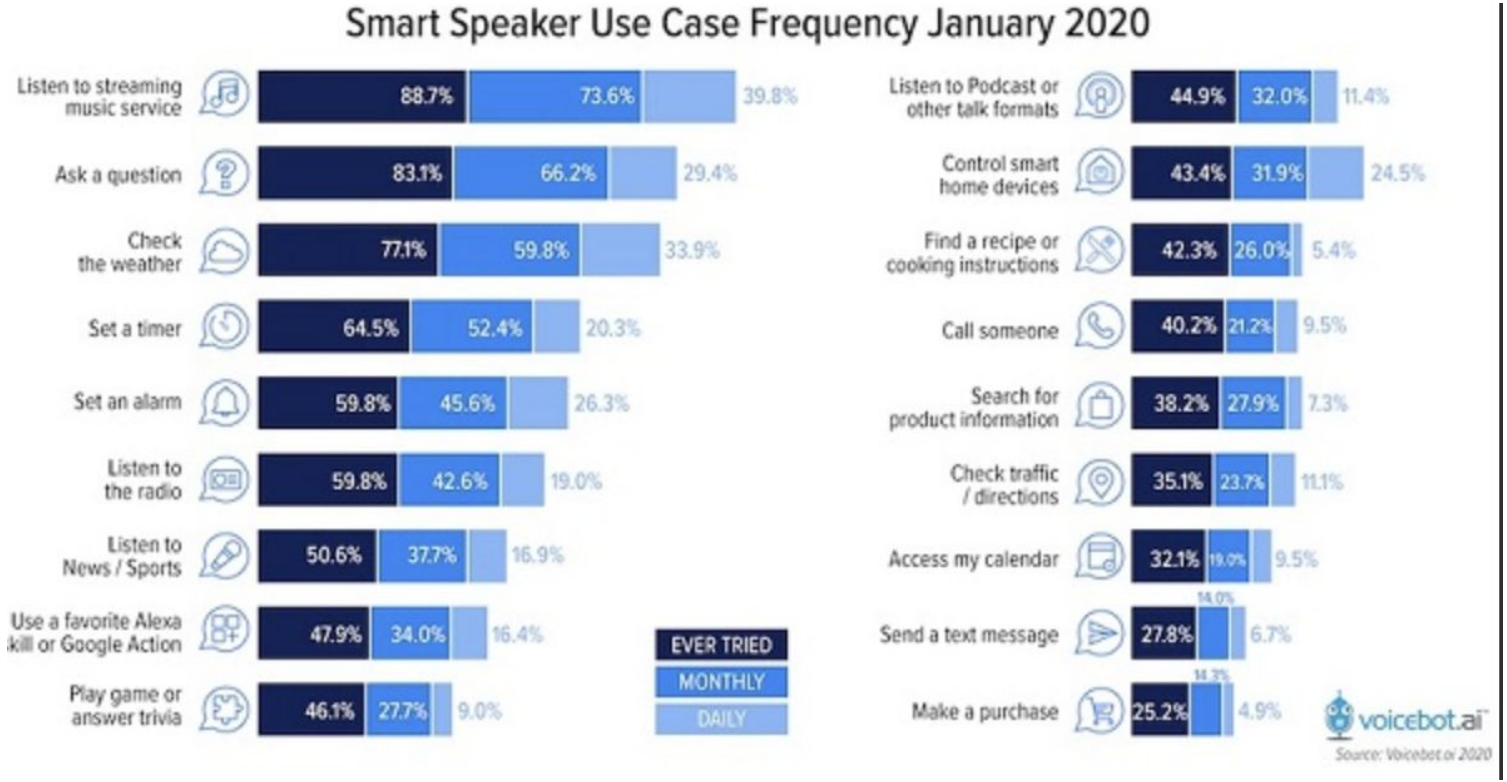
The goal of question-answering is to build systems that automatically answer questions posed by humans in a natural language.

The earliest QA systems dated back to 1960s [Simmons et al., 1964]

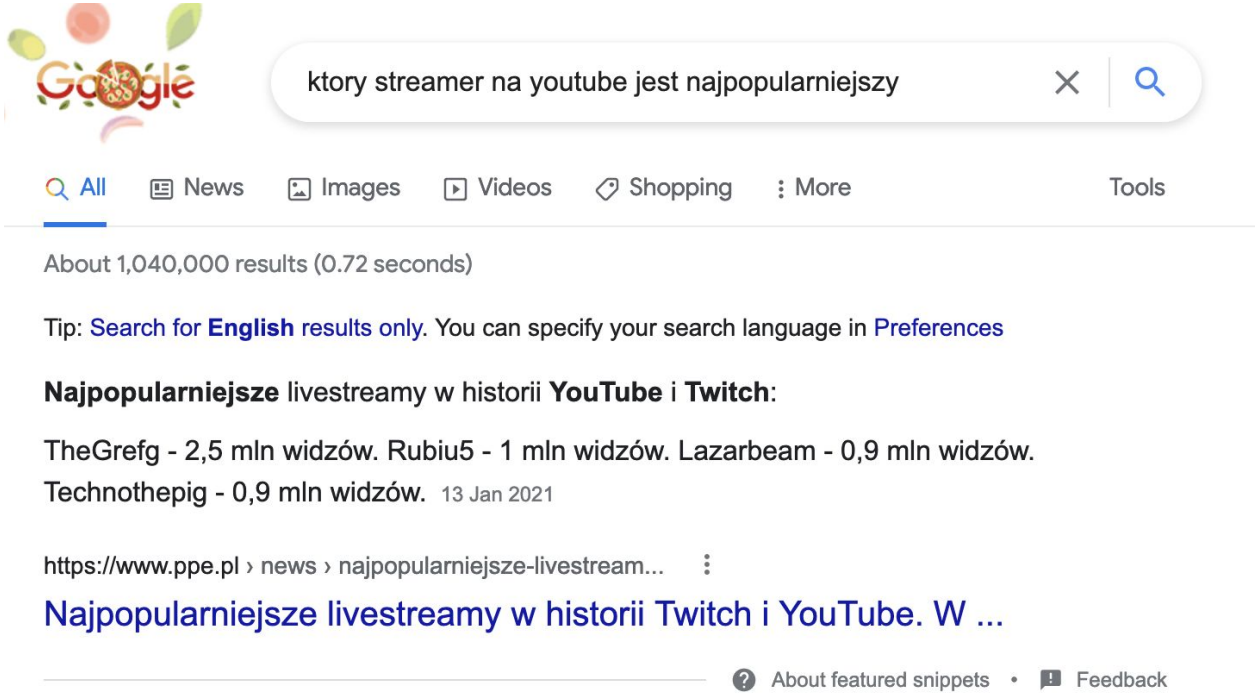
Taxonomy

- 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
- Answer type
  - A short segment of text, a list, yes/no

Lots of practical applications



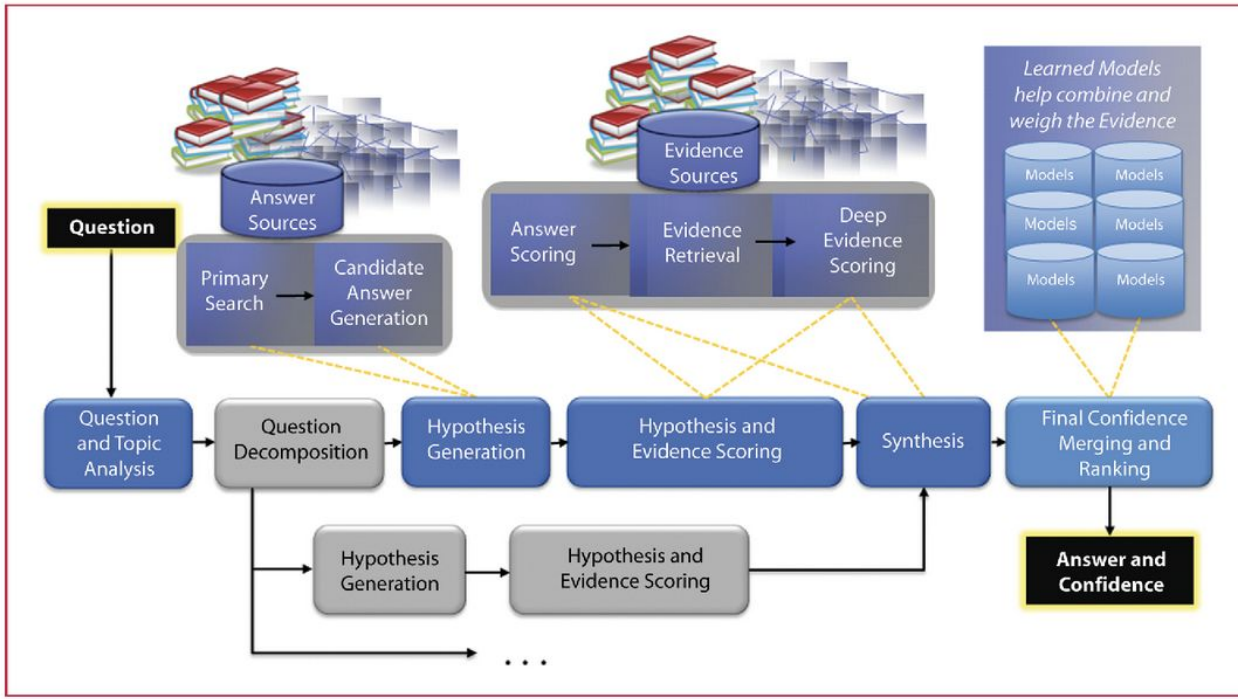
Lots of practical applications



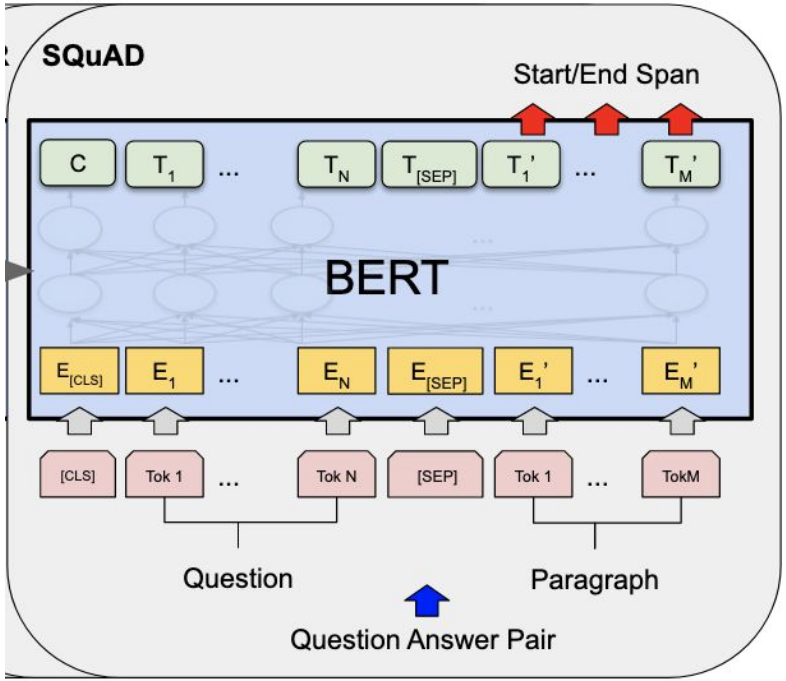
IBM Watson


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
$24,000
Who is Stoker?
(FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

IBM Watson


Learned Models help combine and weigh the Evidence
Models
Models
Models
Models
Models
Models
Evidence Sources
Answer Sources
Question
Primary Search
Candidate Answer Generation
Answer Scoring
Evidence Retrieval
Deep Evidence Scoring
Question and Topic Analysis
Question Decomposition
Hypothesis Generation
Hypothesis and Evidence Scoring
Synthesis
Final Confidence Merging and Ranking
Hypothesis Generation
Hypothesis and Evidence Scoring
Answer and Confidence
. . .

QA in 2022

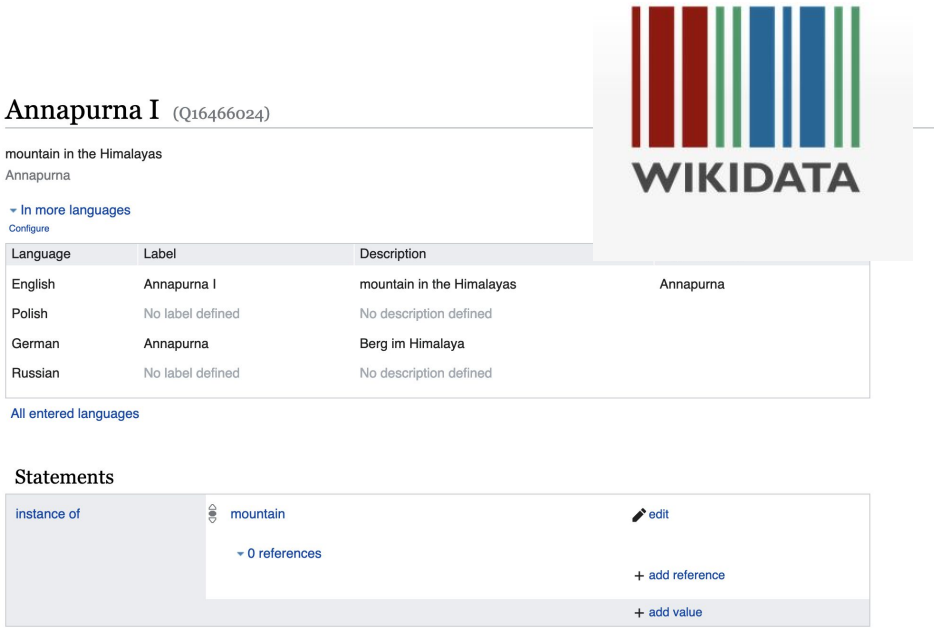

SQuAD
Start/End Span
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E1
...
EN
E[SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Question
Paragraph
Question Answer Pair

Beyond textual QA problems

We will mostly focus on how to answer questions based on unstructured text.

Knowledge based QA

Wikidata



Beyond textual QA problems

We will mostly focus on how to answer questions based on unstructured text.

Visual QA

Plenty of material in next lectures!

Reading comprehension

Reading comprehension

Reading comprehension - comprehend a passage of text and answer questions about its content: (P,Q) -> A

Reading comprehension

Reading comprehension - comprehend a passage of text and answer questions about its content: (P,Q) -> A

2.8. **ZABRANIA SIĘ** używania pralki bez prawidłowo zamontowanych wtyczek.

2.9. Przy podłączeniu pralki do sieci wodociągowej należy używać wyłącznie węża doprowadzającego wodę znajdującego się w zestawie.

2.10. Zabrania się używać do prania środków zawierających rozpuszczalniki, gdyż może to powodować emisję szkodliwych gazów, uszkodzenie pralki, a nawet wybuch.

2.11. W trakcie prania nie dotykać szklanych drzwiczek pralki, ponieważ nagrzewają się i mogą stać się przyczyną oparzenia.

2.12. Po zakończeniu prania należy wyłączyć pralkę, wyjąć wtyczkę z gniazda sieciowego i zakręcić zawór doprowadzający wodę (rys. 8).

2.13. W celu zapewnienia bezpieczeństwa przeciwpożarowego i elektrycznego zabrania się:

Question: Can you touch washing machine when it’s running?

Reading comprehension

Reading comprehension - comprehend a passage of text and answer questions about its content: (P,Q) -> A

2.8. **ZABRANIA SIĘ** używania pralki bez prawidłowo zamontowanych wtyczek.

2.9. Przy podłączeniu pralki do sieci wodociągowej należy używać wyłącznie węża doprowadzającego wodę znajdującego się w zestawie.

2.10. Zabrania się używać do prania środków zawierających rozpuszczalniki, gdyż może to powodować emisję szkodliwych gazów, uszkodzenie pralki, a nawet wybuch.

2.11. W trakcie prania nie dotykać szklanych drzwiczek pralki, ponieważ nagrzewają się i mogą stać się przyczyną oparzenia.

2.12. Po zakończeniu prania należy wyłączyć pralkę, wyjąć wtyczkę z gniazda sieciowego i zakręcić zawór doprowadzający wodę (rys. 8).

2.13. W celu zapewnienia bezpieczeństwa przeciwpożarowego i elektrycznego zabrania się:

Question: Can you touch washing machine when it’s running?

Answer: Maybe?

Why do we care about this problem?

- Useful for many practical applications

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.

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
  - Slot-labeling

Stanford Question Answering Dataset (SQuAD)

- 100k annotated (passage, question, answer) triples
- Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension

Passage

S_1 : Pharmacists are healthcare professionals with specialized education and training who perform various roles to ensure optimal health outcomes for their patients through the quality use of medicines.

S_2 : Pharmacists may also be **small-business proprietors**, owning the pharmacy in which they practice.

S_3 : Since pharmacists know about the mode of action of a particular drug, and its metabolism and physiological effects on the human body in great detail, they play an important role in optimization of a drug treatment for an individual.

Question: What other role do many pharmacists play?

Answer: **small-business proprietors**

Stanford Question Answering Dataset (SQuAD)

- 100k annotated (passage, question, answer) triples
- Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension
- Passages are selected from English Wikipedia, typically 100-150 words
- Questions are crowd-sourced
- Each answer is a short segment of text (or span) in the passage.
    - **This is a limitation** - not all answers are given in that way
- It still remains the most popular reading comprehension dataset: it is almost solved today and the SOTA exceeds the estimated human performance.

SQuAD Evaluation pipeline

- Evaluation: exact match (0 or 1) and F1 (partial credit)
- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers
- We compare the predicted answer to *each* gold answer (a, an, the punctuations are removed) and take max scores. Finally, we take the average of all examples for both exact match and F1
- Estimated human performance: **EM=82.3, F1=91.2**

SQuAD 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

SQuAD 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
8 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
May 05, 2020	*QIANXIN*		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

93.100

93.011

Neural models for reading comprehension

- How can we build a model to solve SQuAD
  - passage, paragraph and context
  - question == query
- Problem formulation:
  - Input: $C = (c_1, \ldots, c_N), Q = (q_1, \ldots, q_M)$
  - Output: $1 \leq \text{start} \leq \text{end} \leq N$
- A family of LSTM-based models with attention (2016-2018)
- Fine-tuning BERT-like models for reading comprehension (2019+)

Seq2Seq model with attention

- Instead of source and target sentences, we have now **passage** and **question**

Seq2Seq model with attention

- Instead of source and target sentences, we have now **passage** and **question**
- We need to model which words in the passage are most relevant to the question

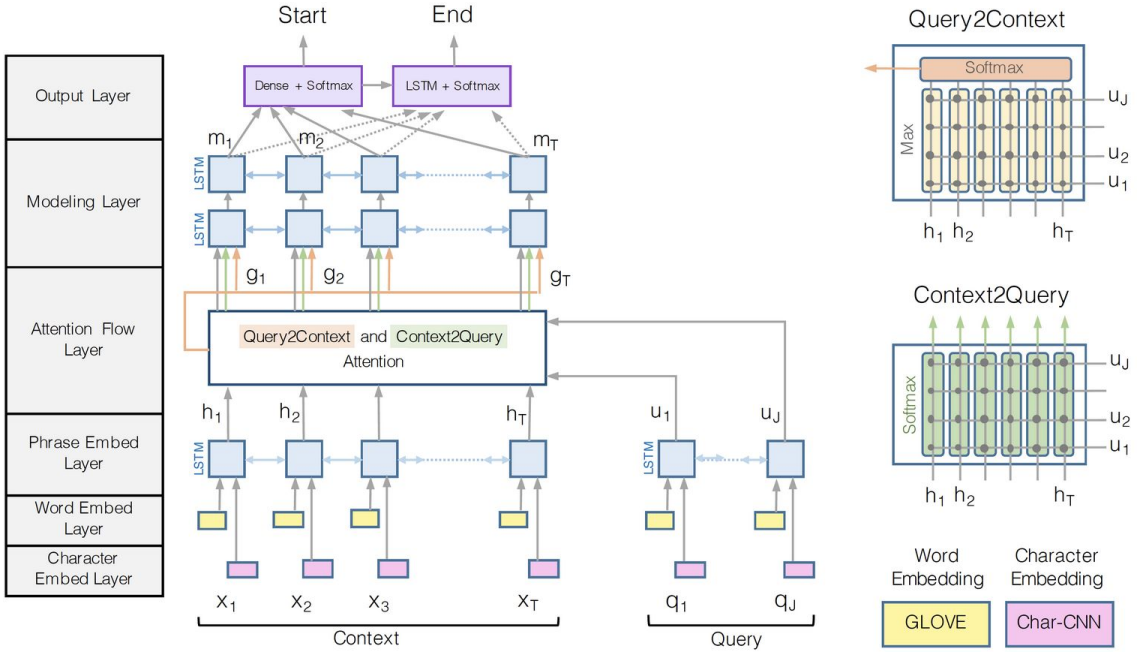
Seq2Seq model with attention

- Instead of source and target sentences, we have now **passage** and **question**
- We need to model which words in the passage are most relevant to the question
- Attention is the key ingredient here, similar to which words in the source sentence are most relevant to the current target word…

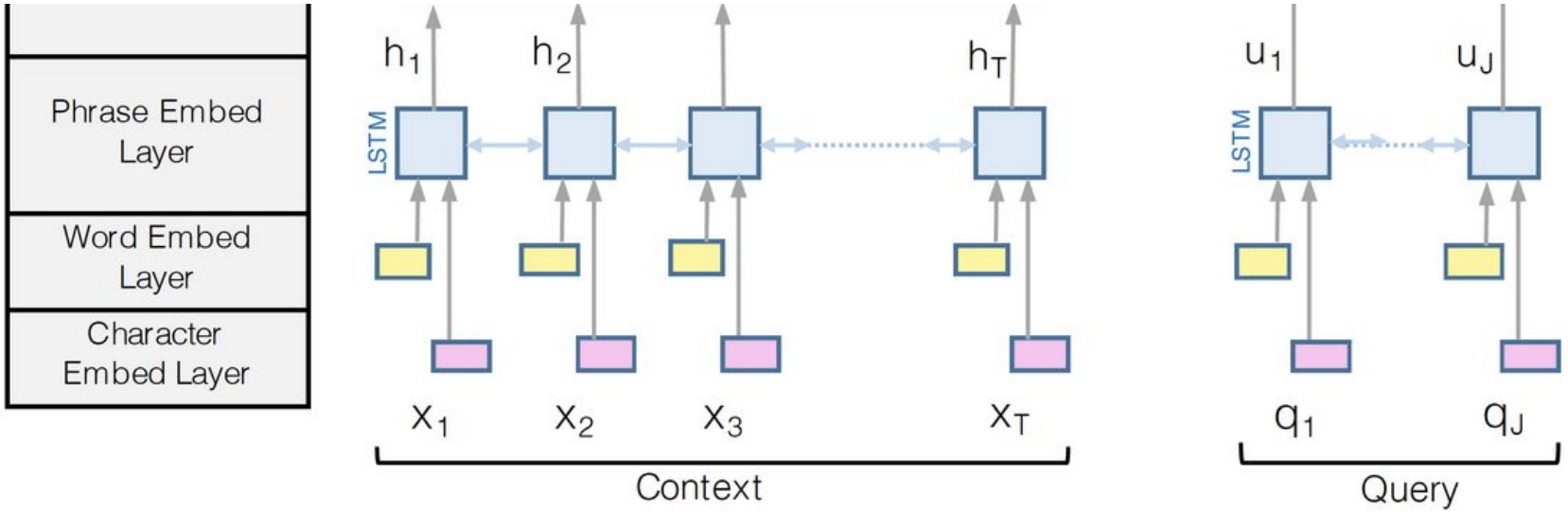
Seq2Seq model with attention

- Instead of source and target sentences, we have now **passage** and **question**
- We need to model which words in the passage are most relevant to the question
- 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. Instead, we just need to train two classifier to predict the start and the end positions of the answer

BiDAF: Bidirectional Attention Flow Model [Seo et al., 2017]

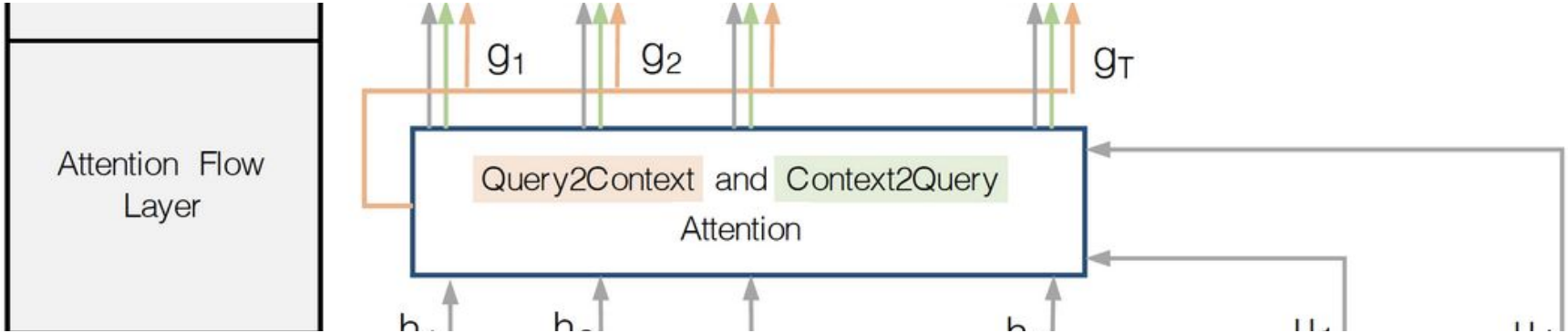


BiDAF: Encoding



- Concatenation of word embeddings and character embeddings for each word
- Two bidirectional LSTMs separately for contextual embeddings

BiDAF: Attention

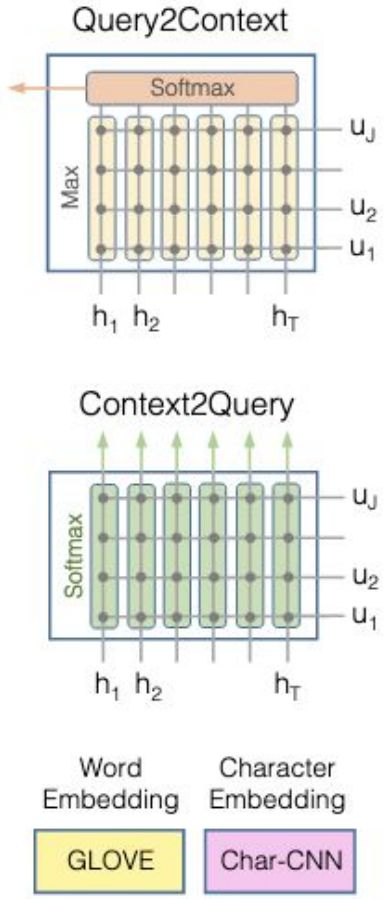


BiDAF: Attention

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

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

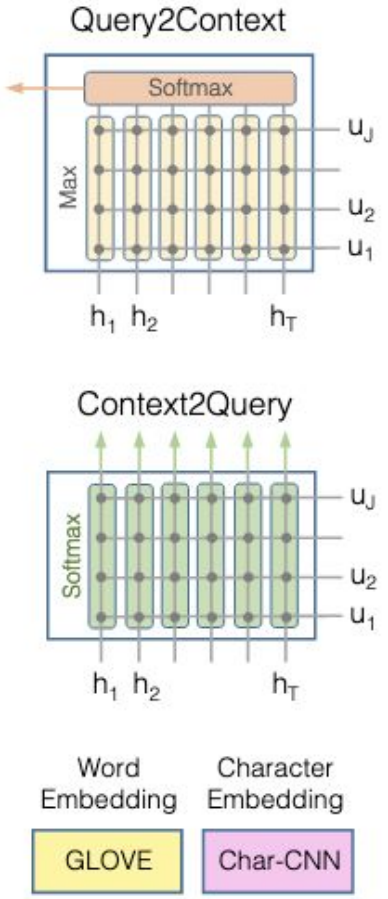
This follows standard attention with context words as queries and query words as keys.



BiDAF: Attention

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

For the given query word, the attended vector indicates the weighted sum of the most important words in the context with respect to the query.

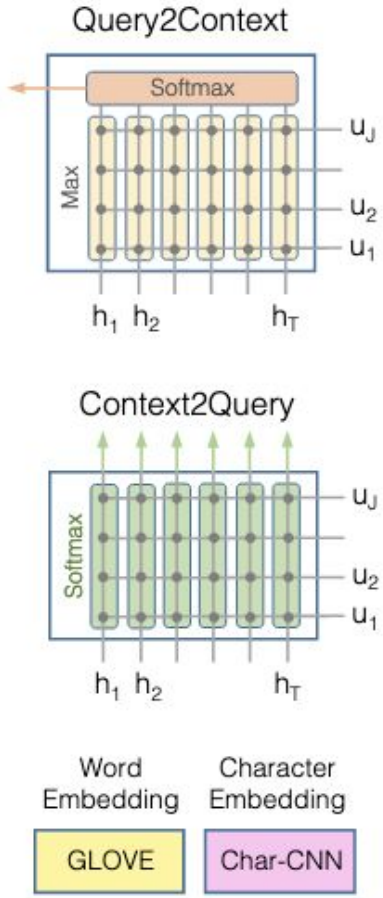


BiDAF: Attention

Similarity matrix is obtained by:

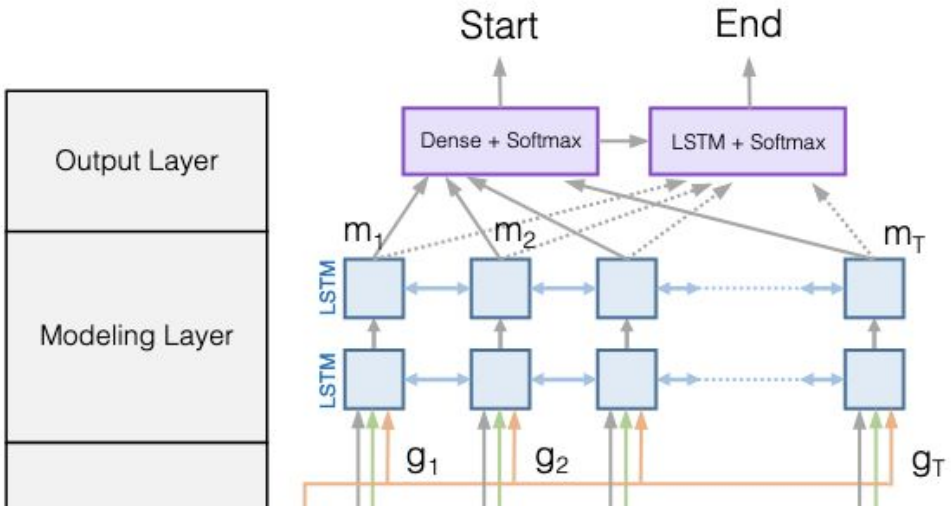
$$S = \alpha(C, Q) \in \mathbb{R}$$

$$\alpha(C, Q) = w_S[C, Q, C \circ Q]$$



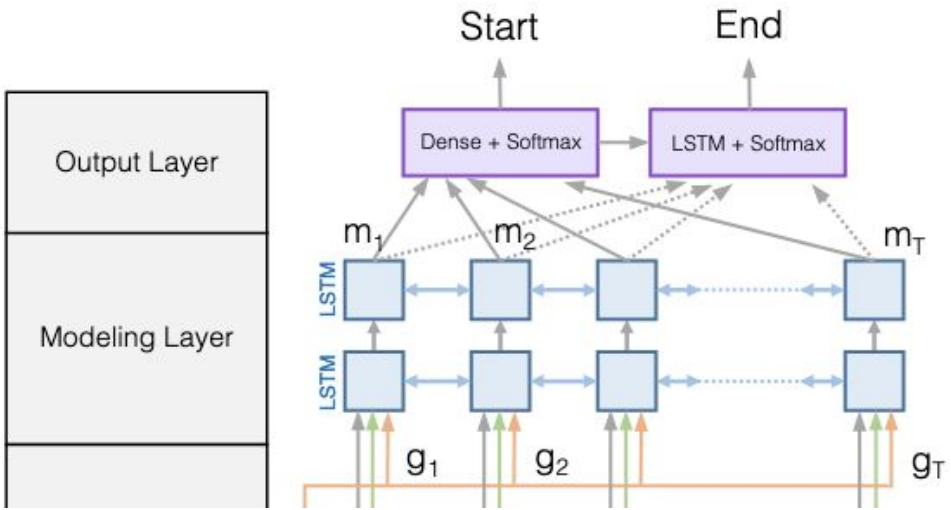
BiDAF: Modelling and output layers

- Modelling layer: pass the output of the attention layer 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.



BiDAF: Modelling and output layers

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



BiDAF: Modeling and output layers

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

$$\mathbf{p}^1 = \text{softmax}(\mathbf{w}_{(\mathbf{p}^1)}^\top [\mathbf{G}; \mathbf{M}])$$

$$\mathbf{p}^2 = \text{softmax}(\mathbf{w}_{(\mathbf{p}^2)}^\top [\mathbf{G}; \mathbf{M}^2])$$

$$L(\theta) = -\frac{1}{N} \sum_{i}^{N} \log(\mathbf{p}_{y_i^1}^1) + \log(\mathbf{p}_{y_i^2}^2)$$

Ablation studies

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

SQuAD results

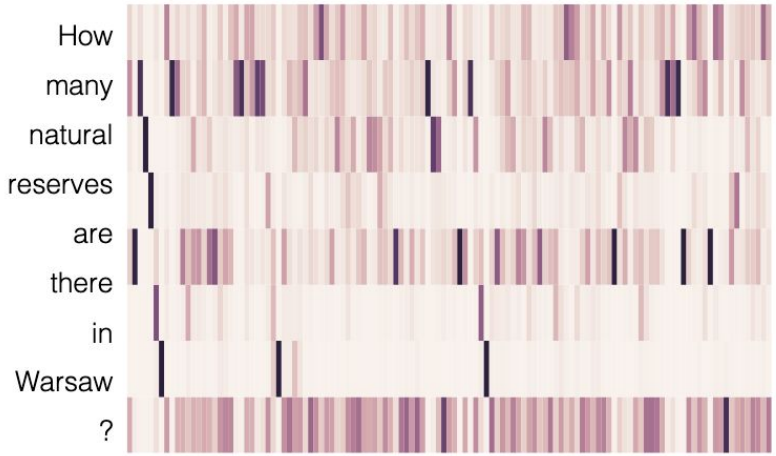
	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

Performance across different tasks

	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	-	-
EpiReader (Trischler et al., 2016)	73.4	74.0	-	-
Stanford AR (Chen et al., 2016)	73.8	73.6	77.6	76.6
GAReader (Dhingra et al., 2016)	73.0	73.8	76.7	75.7
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6
BIDAF (Ours)	**76.3**	**76.9**	**80.3**	**79.6**
MemNN* (Hill et al., 2016)	66.2	69.4	-	-
ASReader* (Kadlec et al., 2016)	73.9	75.4	78.7	77.7
Iterative Attention* (Sordoni et al., 2016)	74.5	75.7	-	-
GA Reader* (Dhingra et al., 2016)	76.4	77.4	79.1	78.1
Stanford AR* (Chen et al., 2016)	77.2	77.6	80.2	79.2

Attention visualization

There are 13 natural reserves in Warsaw–among others, Bielany Forest, Kabaty Woods, Czerniaków Lake . About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw – mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent–the majority are emptied before winter to clean them of plants and sediments.



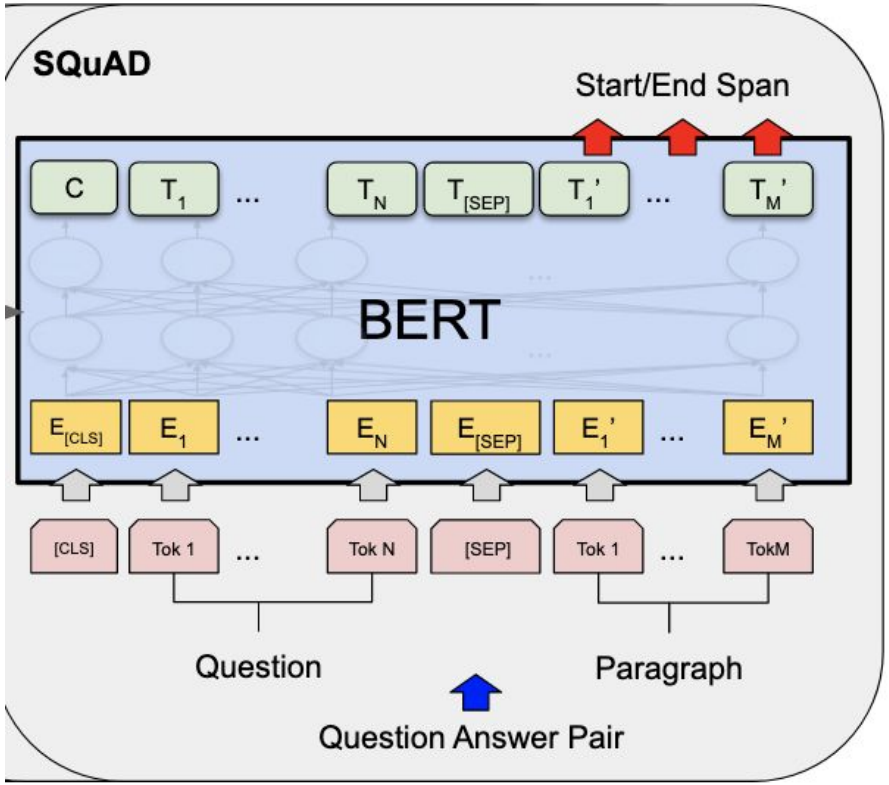
Transformers for reading comprehension

BERT for reading comprehension

BERT is a deep bidirectional Transformer encoder pre-trained on large amount of text data (Wikipedia + BookCorpus)

MLM and NSP losses.

BERT for reading comprehension



BERT for reading comprehension

Softmax over the spikes between start and word tokens:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}.$$

Training takes both probabilities of the start and end spikes:

$$L(\theta) = -\frac{1}{N} \sum_{i}^{N} \log(\mathbf{p}^1_{y^1_i}) + \log(\mathbf{p}^2_{y^2_i})$$

BERT for reading comprehension

All the BERT parameters as well as the newly introduced parameters are optimized together

To put simply, it works amazingly well!

BERT for reading comprehension

All the BERT parameters as well as the newly introduced parameters are optimized together

To put simply, it works amazingly well!

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

BiDAF vs BERT

- BERT model has much more parameters than BiDAF (110M, 330M vs 2.5M)
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers
- BERT is pre-trained while BiDAF is built only on top of GloVe (all remaining parameters need to be learned from the supervision datasets).

Question: Main takeaway?

BiDAF vs BERT

- BERT model has much more parameters than BiDAF (110M, 330M vs 2.5M)
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers
- BERT is pre-trained while BiDAF is built only on top of GloVe (all remaining parameters need to be learned from the supervision datasets).

Question: Main takeaway?

Answer: Pre-training

BiDAF vs BERT

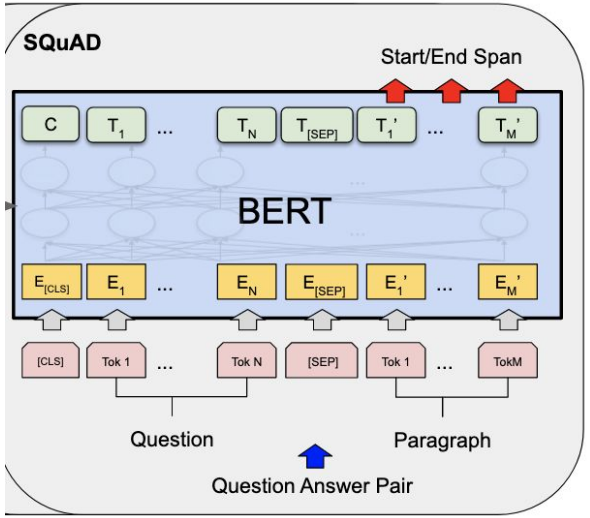
- Are they really fundamentally different?
  - BiDAF and other models aim to model the interactions between question and passage.
  - BERT uses self-attention between the concatenation of question and passage.

BiDAF vs BERT

- Are they really 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.

- Essentially computing:

  att(P, P) + att(P, Q) + att(Q, P) + att(Q, Q)

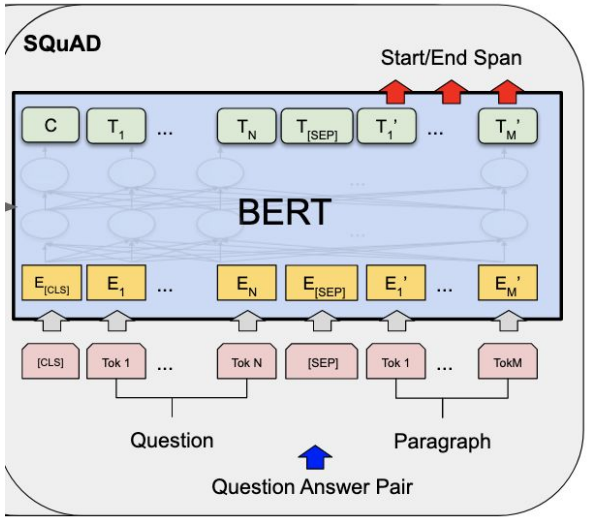


BiDAF vs BERT

- Are they really 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.

- Essentially computing:

  att(P, P) + att(P, Q) + att(Q, P) + att(Q, Q)



Can we design better pre-training objectives?

- The answer is yes.

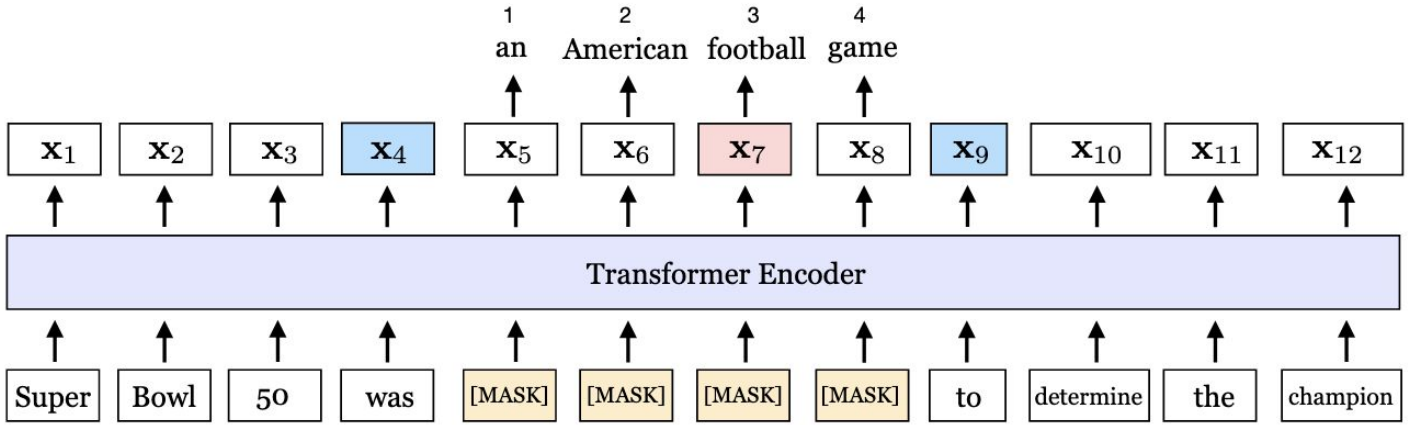
Can we design better pre-training objectives?

- The answer is yes.

- Two ideas:
  - Masking contiguous spans of words instead of 15% random words.
  - Using the two end-points of span to predict all the masked words in between == compressing the information of a span into its two endpoints.

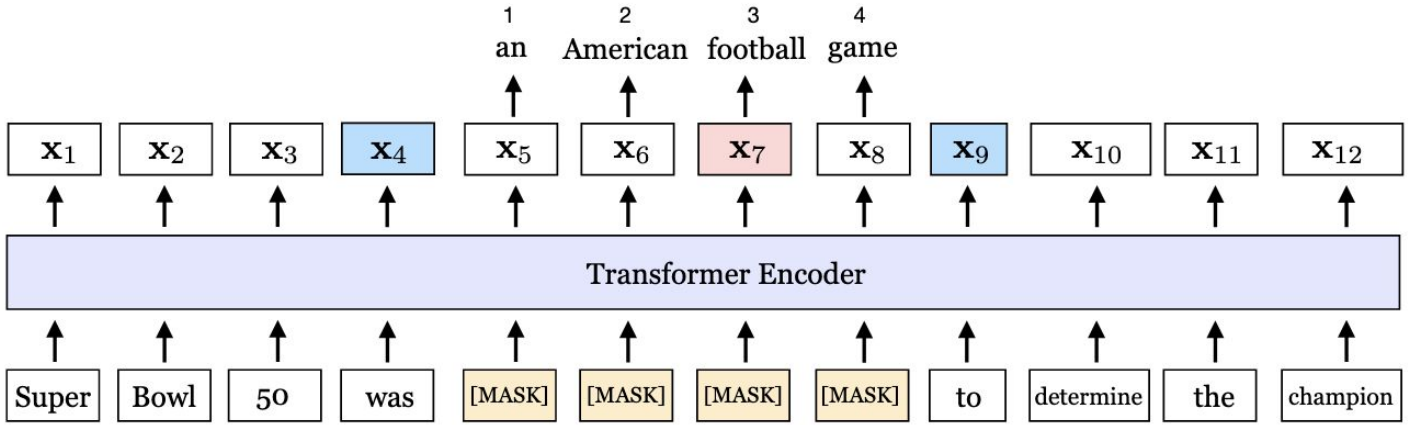
SpanBERT [Joshi et al., 2020]

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



SpanBERT [Joshi et al., 2020]

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



Performance comparison

	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Human Perf.	82.3	91.2	86.8	89.4
Google BERT	84.3	91.3	80.0	83.3
Our BERT	86.5	92.6	82.8	85.9
Our BERT-1seq	87.5	93.3	83.8	86.6
SpanBERT	**88.8**	**94.6**	**85.7**	**88.7**

Table 1: Test results on SQuAD 1.1 and SQuAD 2.0.

Performance

	NewsQA	TriviaQA	SearchQA	HotpotQA	Natural Questions	Avg.
Google BERT	68.8	77.5	81.7	78.3	79.9	77.3
Our BERT	71.0	79.0	81.8	80.5	80.5	78.6
Our BERT-1seq	71.9	80.4	84.0	80.3	81.8	79.7
SpanBERT	**73.6**	**83.6**	**84.8**	**83.0**	**82.5**	**81.5**

Table 2: Performance (F1) on the five MRQA extractive question answering tasks.

Where are we in 2022?

- We have already surpassed human performance on SQuAD. Does it mean that reading comprehension is already solved?
- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions.

Robustness issues

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

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).

	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

Table 2: Adversarial evaluation on the Match-LSTM and BiDAF systems. All four systems can be fooled by adversarial examples.

Generalization issues

		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

CheckList [Ribeiro et al., 2020]

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

	Test *TYPE* and Description	Failure Rate (%)					Example test cases & expected behavior
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba→ Canada... INV @VirginAmerica I miss the #nerdbird in San Jose→ Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon→ Erin was your saviour INV @united 8602947, Jon→ Sean at http://t.co/58tuTgli0D, thanks. INV

CheckList [Ribeiro et al., 2020]

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

	Test *TYPE* and Description	Failure Rate (%)					Example test cases & expected behavior
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba→ Canada... INV @VirginAmerica I miss the #nerdbird in San Jose→ Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon→ Erin was your saviour INV @united 8602947, Jon→ Sean at http://t.co/58tuTgli0D, thanks. INV
Temporal	***MFT:*** Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this airline, although now I like it. pos In the past I thought this airline was perfect, now I think it is creepy. neg
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	***MFT:*** Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

Table 1: A selection of tests for sentiment analysis. All examples (right) are failures of at least one model.

Open-domain question-answering

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. We do not know where the answer is located, and the goal is to return the answer for any open-domain questions
- Much more challenging but more practical as well

Retriever-reader framework (DrQA) [Chen et al. 2017]

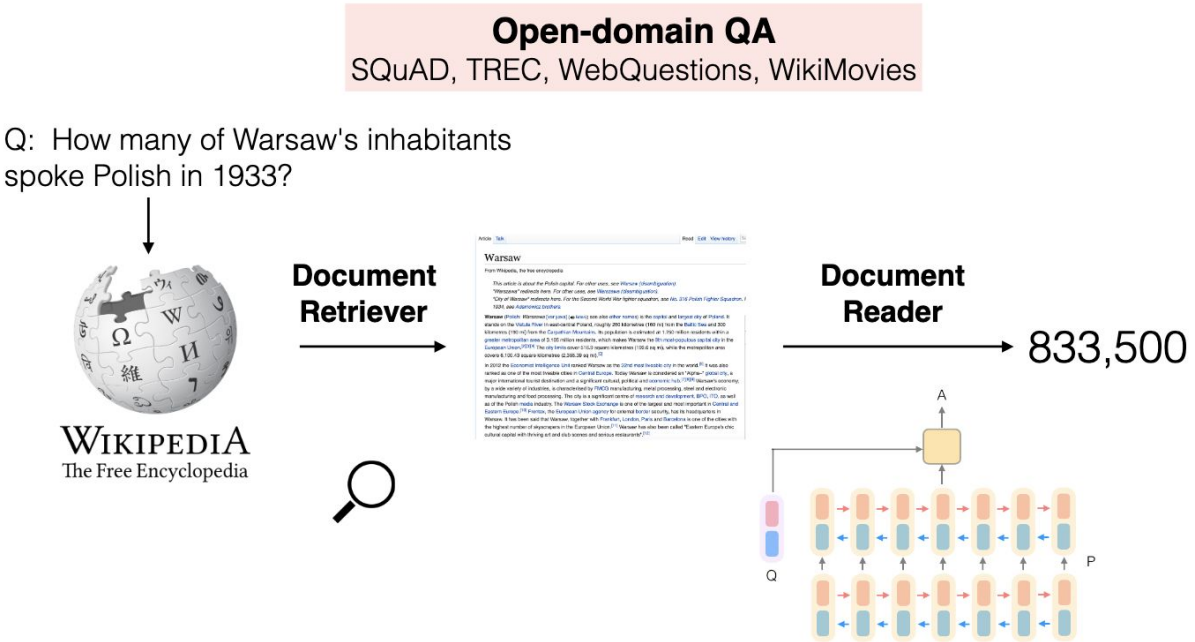


Figure 1: An overview of our question answering system DrQA.

Retriever-reader framework (DrQA) [Chen et al. 2017]

- Input: a large collection of documents $D_1, D_2, \ldots, D_N$ and Q .
- Output: an answer string *A*.

- Retriever: $f(D, Q) \rightarrow P_1, \ldots, P_k$
- Reader: $g(Q, \{P_1, \ldots., P_k\}) \rightarrow A$

Retriever-reader framework (DrQA) [Chen et al. 2017]

- Input: a large collection of documents $D_1, D_2, \ldots, D_N$ and Q .
- Output: an answer string *A*.

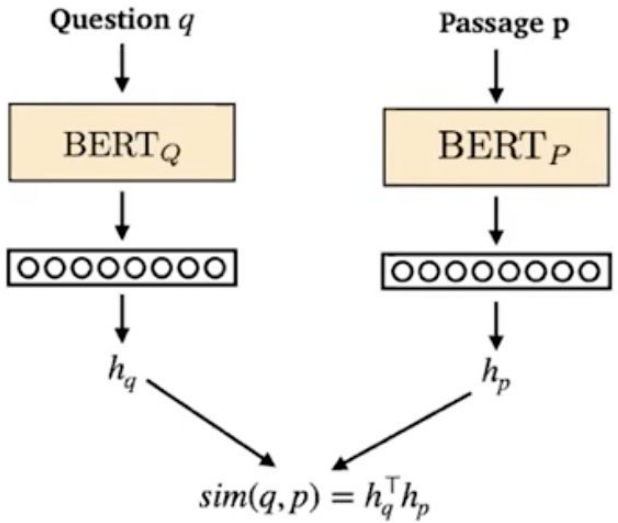
- Retriever: $f(D, Q) \rightarrow P_1, \ldots, P_k$
- Reader: $g(Q, \{P_1, \ldots, P_k\}) \rightarrow A$

In DrQA:

- Retriever - a standard TF-IDF information-retrieval sparse model
- Reader - a neural reading comprehension model that we just learned
  - Trained on SQuAD an other distantly-supervised QA domains.

Training the retriever [Karpukhin et al., 2020]

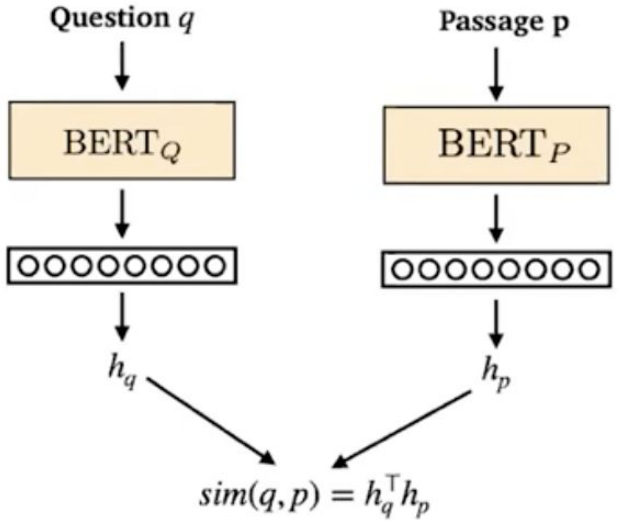
- Dense passage retrieval (DPR) - we can also just train the retriever using question-answer pairs!



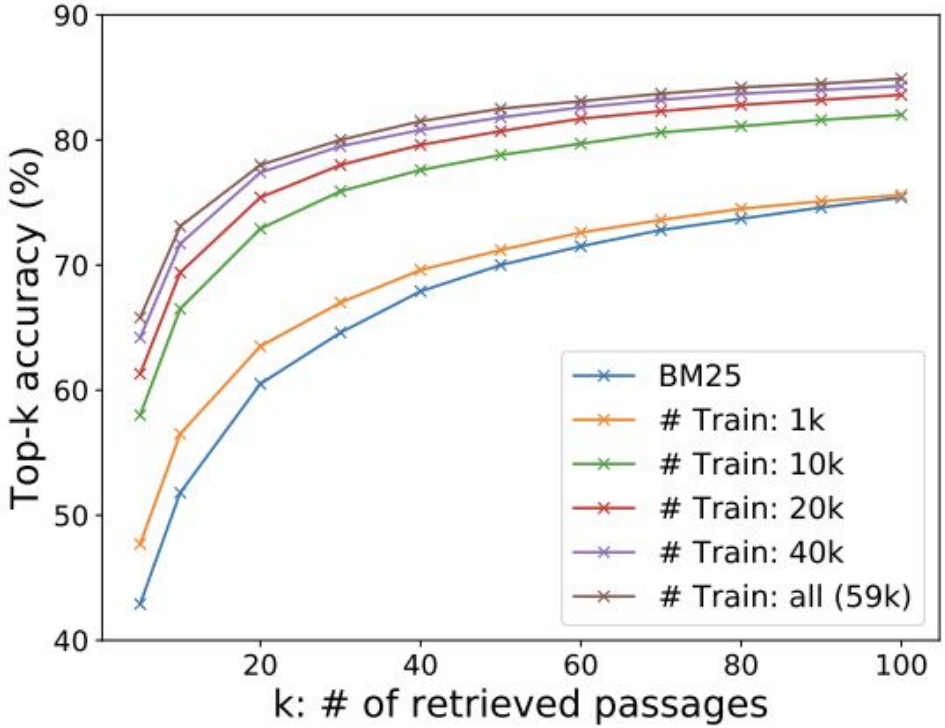
Training the retriever [Karpukhin et al., 2020]

- 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.

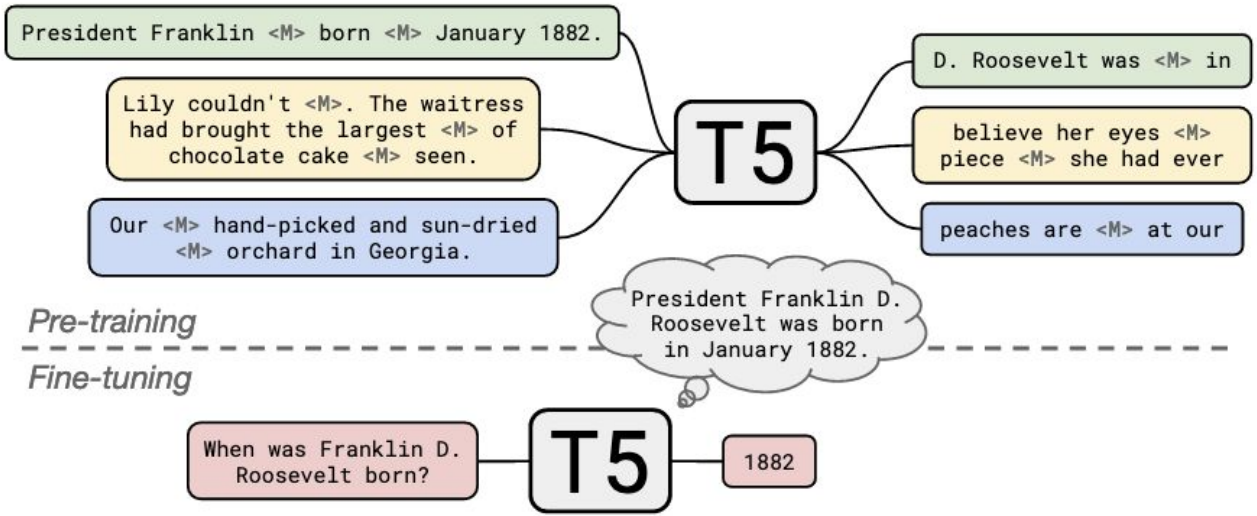


Training the retriever [Karpukhin et al., 2020]



Large language models can do open-domain QA well

- Without an explicit retriever stage:

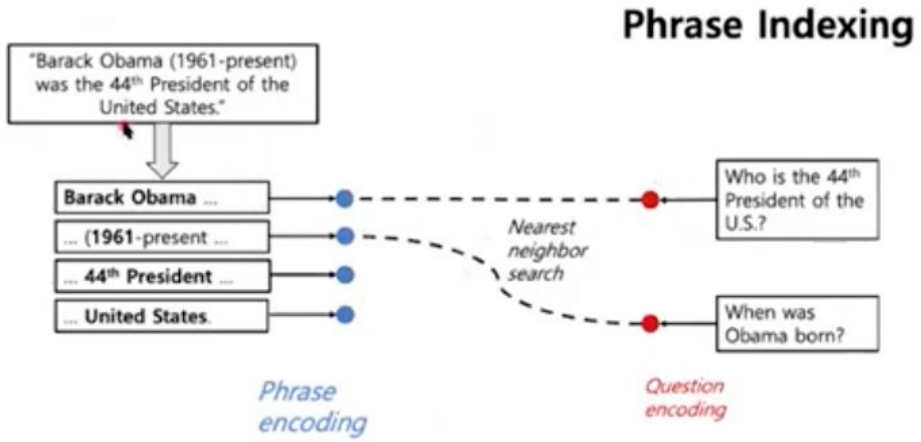


Maybe the reader model is not necessary too!

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

Maybe the reader model is not necessary too!

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



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Lewis et al., 2020]

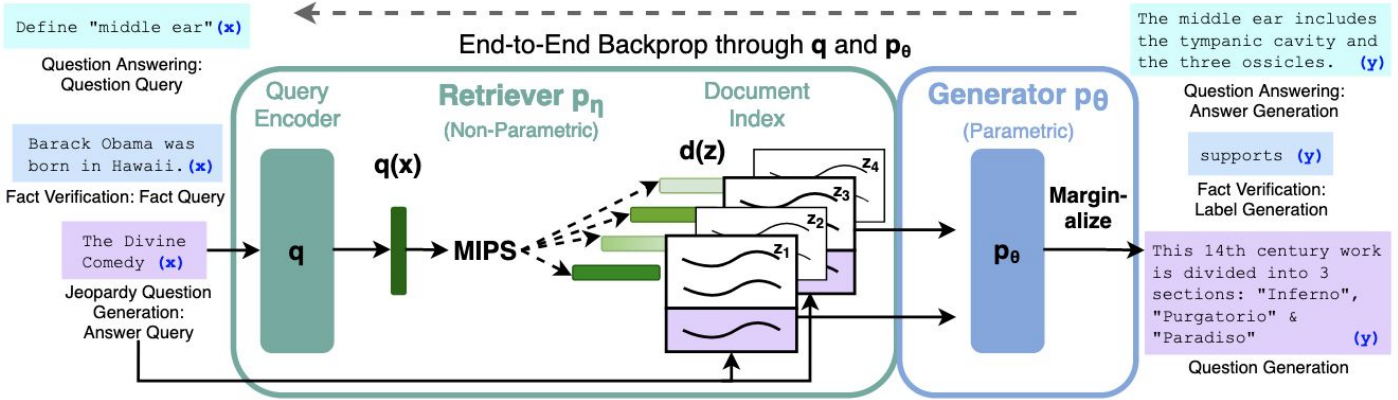


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

RAG [Lewis et al., 2020]

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

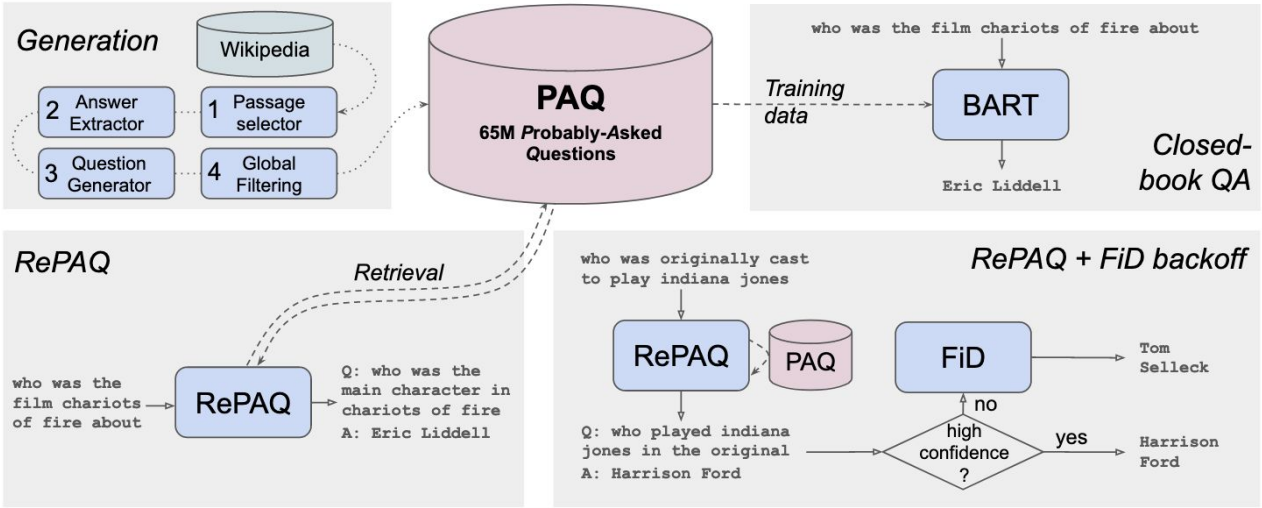
	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them



PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

Recipe:

1) A passage selection model *p(c)*, to identify passages which humans are likely to ask questions about

PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

Recipe:

1) A passage selection model *p(c)*, to identify passages which humans are likely to ask questions about
2) An answer extraction model *p(a | c)*, for identifying spans in a passage that are more likely to be answers to a question.

PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

Recipe:

1) A passage selection model *p(c)*, to identify passages which humans are likely to ask questions about
2) An answer extraction model *p(a | c)*, for identifying spans in a passage that are more likely to be answers to a question.
3) A question generator model *p(q | a, c)* that, given a passage and an answer, generates a question

PAQ [Lewis et al., 2021]

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

Recipe:

1) A passage selection model *p(c)*, to identify passages which humans are likely to ask questions about
2) An answer extraction model *p(a | c)*, for identifying spans in a passage that are more likely to be answers to a question.
3) A question generator model *p(q | a, c)* that, given a passage and an answer, generates a question
4) A filtering QA model *p(a | q, C)* that generates an answer for a given question.

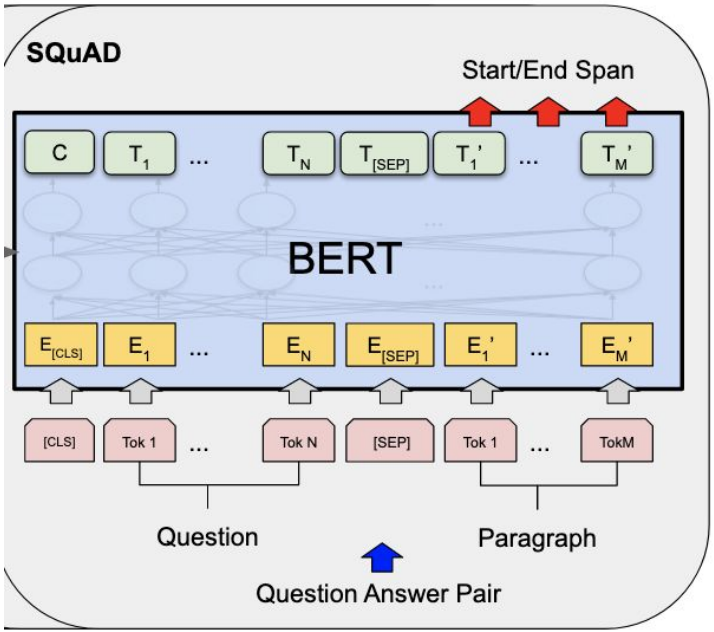
PAQ [Lewis et al., 2021]

#	Model Type	Model	NaturalQuestions	TriviaQA
1	Closed-book	T5-11B-SSM (Roberts et al., 2020)	35.2	51.8
2	Closed-book	BART-large (Lewis et al., 2020b)	26.5	26.7
3	QA-pair retriever	Dense retriever (Lewis et al., 2020b)	26.7	28.9
4	Open-book, retrieve-and-read	RAG-Sequence (Lewis et al., 2020a)	44.5	56.8
5	Open-book, retrieve-and-read	FiD-large, 100 docs (Izacard and Grave, 2020)	51.4	**67.6**
6	Open-book, phrase index	DensePhrases (Lee et al., 2021)	40.9	50.7
7	Closed-book	BART-large, pre-finetuned on PAQ	32.7	33.2
8	QA-pair retriever	RePAQ (retriever only)	41.2	38.8
9	QA-pair retriever	RePAQ (with reranker)	47.7	50.7
10	QA-pair retriever	RePAQ-multitask (retriever only)	41.7	41.3
11	QA-pair retriever	RePAQ-multitask (with reranker)	47.6	52.1
12	QA-pair retriever	RePAQ-multitask w/ FiD-Large Backoff	**52.3**	67.3

PAQ [Lewis et al., 2021]

#	KB	Filtering	Size	Exact Match	
				Retrieve	Rerank
1	NQ-Train	-	87.9K	27.9	31.8
2	$\text{PAQ}_{L,1}$	None	58.0M	21.6	30.6
3	$\text{PAQ}_{L,1}$	Local	31.7M	28.3	34.9
4	$\text{PAQ}_{L,1}$	Global	14.1M	38.6	44.3
5	$\text{PAQ}_{L,4}$	Global	53.8M	40.3	45.2
6	$\text{PAQ}_{NE,1}$	Global	12.0M	37.3	42.6
7	PAQ	Global	64.9M	**41.6**	**46.4**

Rethinking the Objectives of Extractive Question-Answering [Fajcik 2020]



Rethinking the Objectives of Extractive Question-Answering [Fajcik 2020]

Jointly model the probability of start and end tokens by modelling similarity of hidden vectors:

$$P_\theta(a_s, a_e) = \text{softmax}(\text{vec}(f_{sim}(\boldsymbol{H}_s, \boldsymbol{H}_e)))$$

A **multi-task compound objective**:

$$-\sum_{(q,D,a)\in\mathcal{D}} \log P_\theta(a_s, a_e) P_\theta(a_s) P_\theta(a_e)$$

- A dot product:

$$f_{dot}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{h_s}^\top \boldsymbol{h_e} \quad (5)$$

- A weighted dot product:

$$f_{wdot}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s} \circ \boldsymbol{h_e}] \quad (6)$$

- An additive similarity:

$$f_{add}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s}; \boldsymbol{h_e}] \quad (7)$$

- An additive similarity combined with weighted product:

$$f_{madd}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top [\boldsymbol{h_s}; \boldsymbol{h_e}; \boldsymbol{h_s} \circ \boldsymbol{h_e}] \quad (8)$$

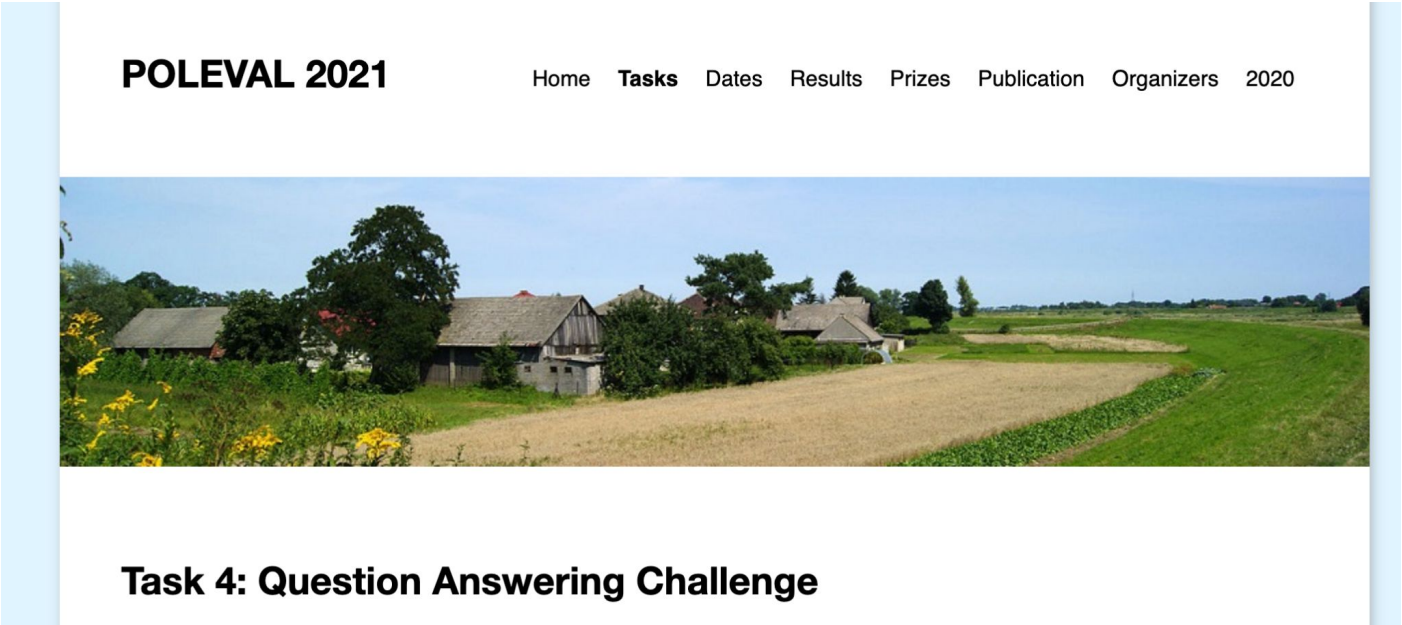
- A multi-layer perceptron (MLP) as proposed by Lee et al. (2019):

$$f_{MLP}(\boldsymbol{h_s}, \boldsymbol{h_e}) = \boldsymbol{w}^\top \sigma(\boldsymbol{W}[\boldsymbol{h_s}; \boldsymbol{h_e}] + \boldsymbol{b}) + \boldsymbol{b}_2$$

Rethinking the Objectives of Extractive Question-Answering [Fajcik 2020]

Model	Obj	SQ1	SQ2	AdvSQ	TriviaQA	NQ	NewsQA
BERT	I	81.31/88.65	**73.89**/76.74	47.04/52.62	62.88/69.85	65.66/78.20	52.39/67.17
	J	81.33/88.13	72.66/75.04	48.10/53.54	**63.93**/69.90	**67.75**/78.70	52.73/66.41
	JC	81.22/88.29	71.51/74.38	46.07/51.35	62.82/69.94	66.48/77.34	52.39/67.05
	I+J	***81.83***/88.52	73.53/76.14	***48.32***/53.47	*63.73*/69.75	***67.75***/78.81	***52.96***/66.83
ALBERT	I	88.55/94.62	87.07/90.02	68.12/73.54	74.7/80.33	70.78/83.42	59.95/75.0
	J	88.84/94.64	86.87/89.71	68.90/74.17	75.11/80.41	**73.36/84.01**	60.19/74.28
	JC	88.60/94.59	86.78/89.73	68.0/73.25	-	72.33/83.35	58.52/72.74
	I+J	***89.02***/94.77	**87.13**/89.98	***69.57***/74.76	***75.31***/80.43	*73.32*/84.08	***60.41***/74.46

Group project proposal #2



Group project proposal #2

POLEVAL 2021

Home **Tasks** Dates Results Prizes Publication Organizers 2020

Task 4: Question Answering C

Polish QA model

- Q: *Jak nazywa się bohaterka gier komputerowych z serii Tomb Raider?*
- A: *Lara Croft*

- Q: *Paź królowej to gatunek których owadów?*
- A: m*otyli*

Polish QA model

The goal of the task is to develop a solution capable of providing answers to general-knowledge questions typical for popular TV quiz shows, such as 1 z 10.

The evaluation will be carried out on the test-B dataset (https://github.com/poleval/2021-question-answering)

Asking Google on the fly is **not** a viable solution.

Evaluation

Checking if the two answers match will depend on the question type:

1. For non-numerical questions, assess textual similarity. A Levenshtein distance should be computed between the two (lowercased) strings and if it is less than ½ of the length of the gold standard answer, the candidate answer is accepted.

Evaluation

Checking if the two answers match will depend on the question type:

1. For non-numerical questions, assess textual similarity. A Levenshtein distance should be computed between the two (lowercased) strings and if it is less than ½ of the length of the gold standard answer, the candidate answer is accepted.

2. For numerical questions (e.g. *In which year…*), assess numerical similarity. Specifically, use a regular expression to extract a sequence of characters that could be interpreted as a number. If such sequences can be found in both answers and represent the same number, the prediction is accepted.

Last year result

Task 4: Question answering challenge

1. Mateusz Piotrowski – 71.68
2. Aleksander Smywiński–Pohl (AGH / eNeLPol) – 50.96
   Piotr Rybak (ML Research at Allegro.pl) – 50.96
3. Darek Kłeczek (skok.ai) – 46.44

Literature

1. Rajpurkar et al., 2016 - https://arxiv.org/abs/1606.05250
2. Seo et al., 2017 - https://arxiv.org/abs/1611.01603
3. Lewis et al., 2020 - https://arxiv.org/pdf/2005.11401.pdf