Course Outline

- Question Answering systems (Q.A.) +lab (Q.A. fine-tune BERT model)
- Multilingual and Multimodal Q.A. +lab (Q.A. test BERT models)
- Information Retrieval (I.R.) +lab (I.R. train and test BM25 model)
- Open Domain Q.A. +assingment (Open Domain Q.A)

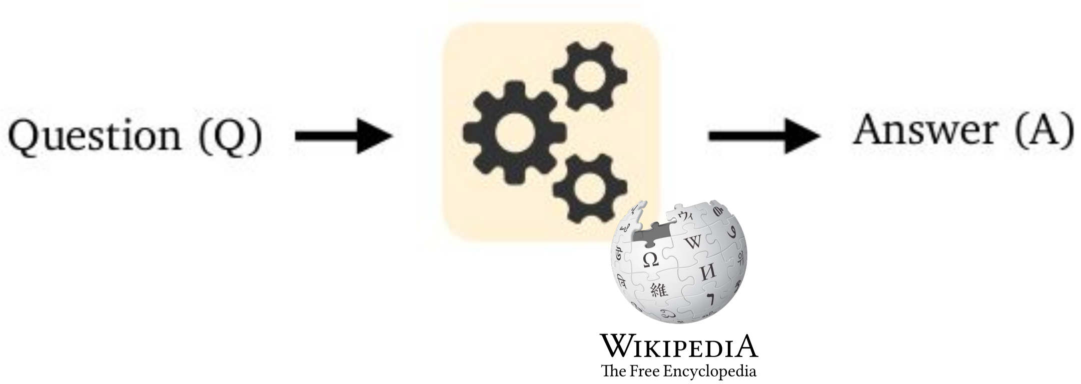
Course Outline

- Question Answering systems (Q.A.) +lab
- Multilingual and Multimodal Q.A. +lab
- Information Retrieval (I.R.) +lab
- **Open Domain Q.A.** +assignment(copyPaste…) + final proyect
  - Based on Danqi Chen (Princeton University) & Christopher Manning (Stanford University) slides. Special thanks to Jon Ander Campos (UPV/EHU)

Course Outline

- **Open Domain Question Answering systems**
    - **Introduction**
    - Open Domain Q.A. practical applications
    - SOA models

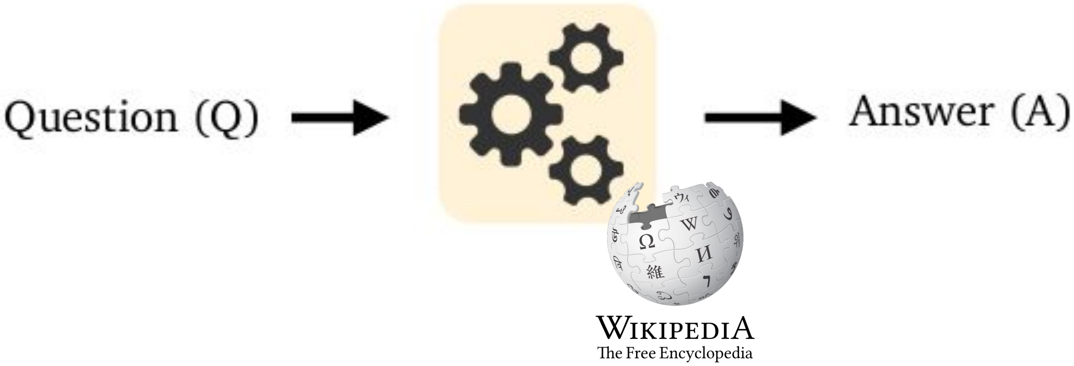
Open Domain Q.A.



- 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 but **a more practical problem!**

Open Domain Q.A.

In contrast to “closed-domain” systems that deal with questions under a specific domain...



- 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 but **a more practical problem!**

Open Domain Q.A.

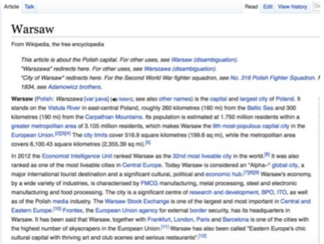
Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
The Free Encyclopedia

Document Retriever



Article Talk Read Edit View history Se

Warsaw

From Wikipedia, the free encyclopedia

This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

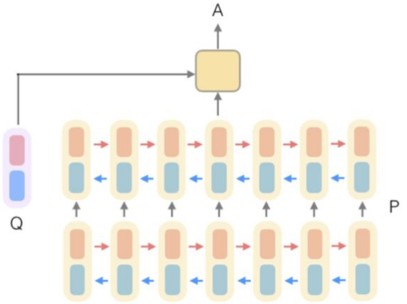
Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500

Information Retrieval Model (I.R.)

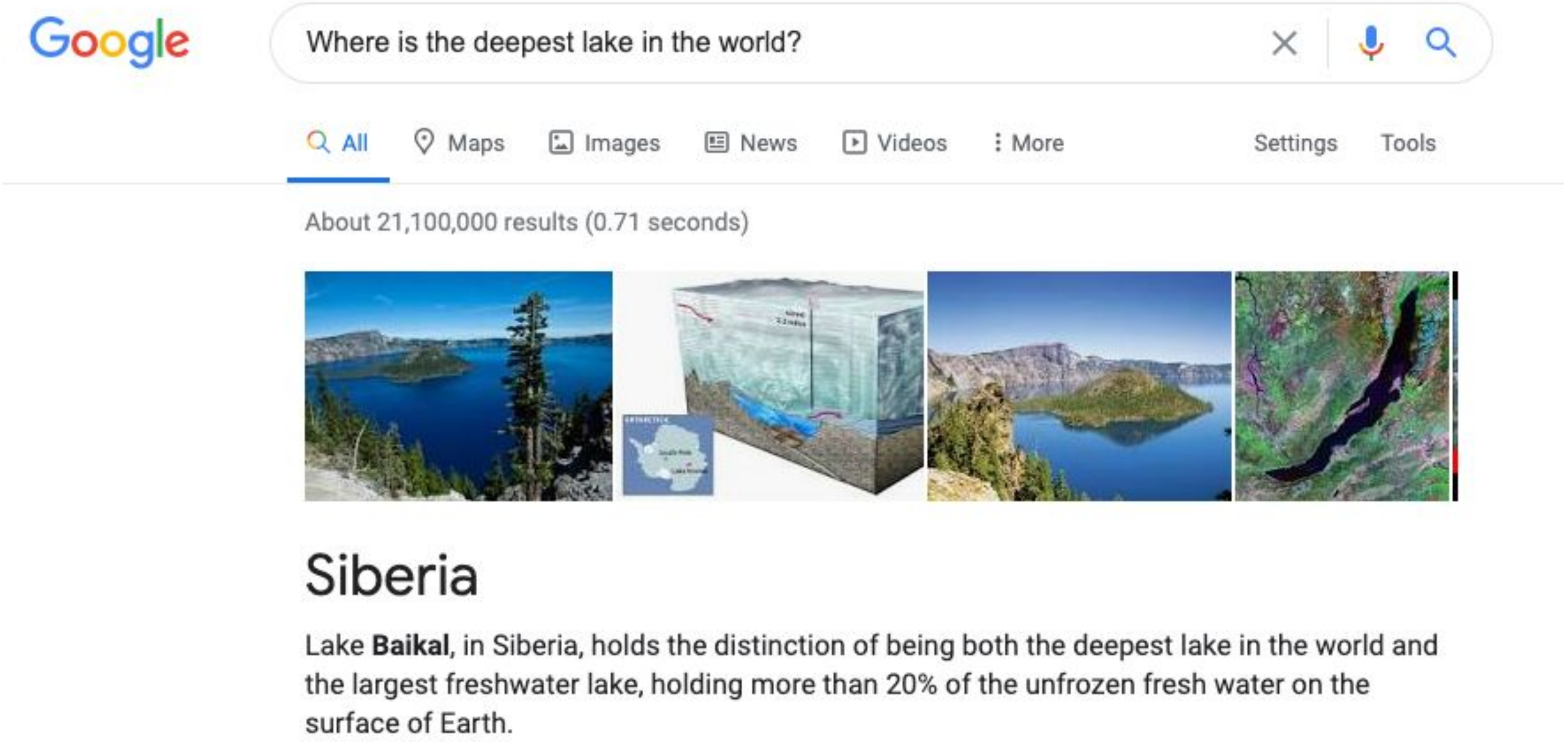


BERT
fine-tuned for Q.A.

Course Outline

- **Open Domain Question Answering systems**
  - Introduction
  - **Open Domain Q.A. practical applications**
  - SOA models

Open Domain Q.A.: Lots of practical application



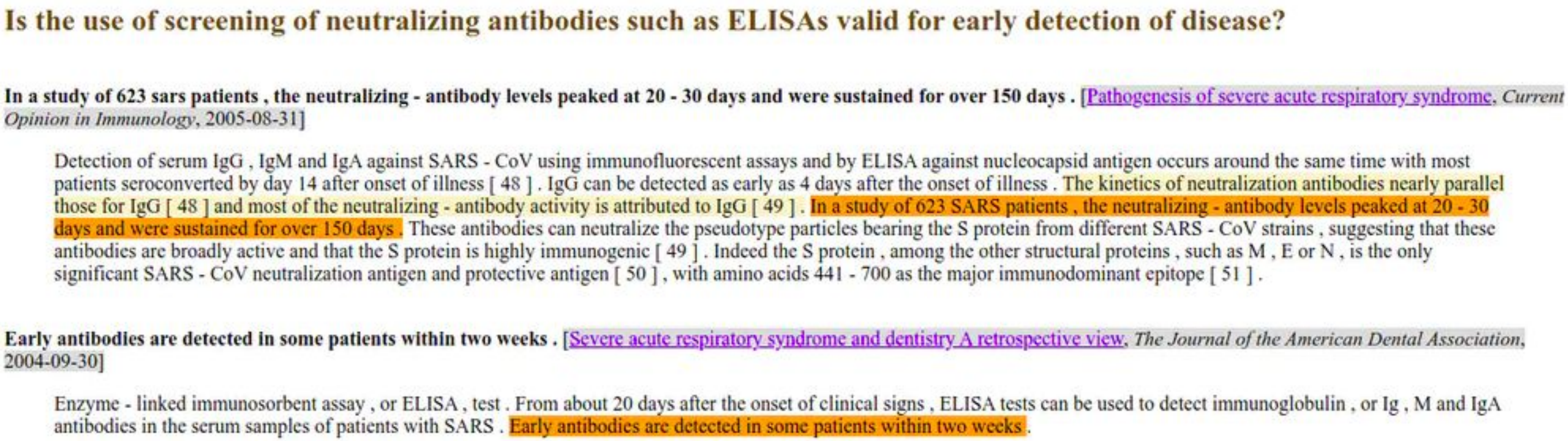
Where is the deepest lake in the world?

About 21,100,000 results (0.71 seconds)

Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

Open Domain Q.A.: Lots of practical application



Is the use of screening of neutralizing antibodies such as ELISAs valid for early detection of disease?

In a study of 623 sars patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . [Pathogenesis of severe acute respiratory syndrome, *Current Opinion in Immunology*, 2005-08-31]

Detection of serum IgG , IgM and IgA against SARS - CoV using immunofluorescent assays and by ELISA against nucleocapsid antigen occurs around the same time with most patients seroconverted by day 14 after onset of illness [48] . IgG can be detected as early as 4 days after the onset of illness . The kinetics of neutralization antibodies nearly parallel those for IgG [48] and most of the neutralizing - antibody activity is attributed to IgG [49] . In a study of 623 SARS patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . These antibodies can neutralize the pseudotype particles bearing the S protein from different SARS - CoV strains , suggesting that these antibodies are broadly active and that the S protein is highly immunogenic [49] . Indeed the S protein , among the other structural proteins , such as M , E or N , is the only significant SARS - CoV neutralization antigen and protective antigen [50] , with amino acids 441 - 700 as the major immunodominant epitope [51] .

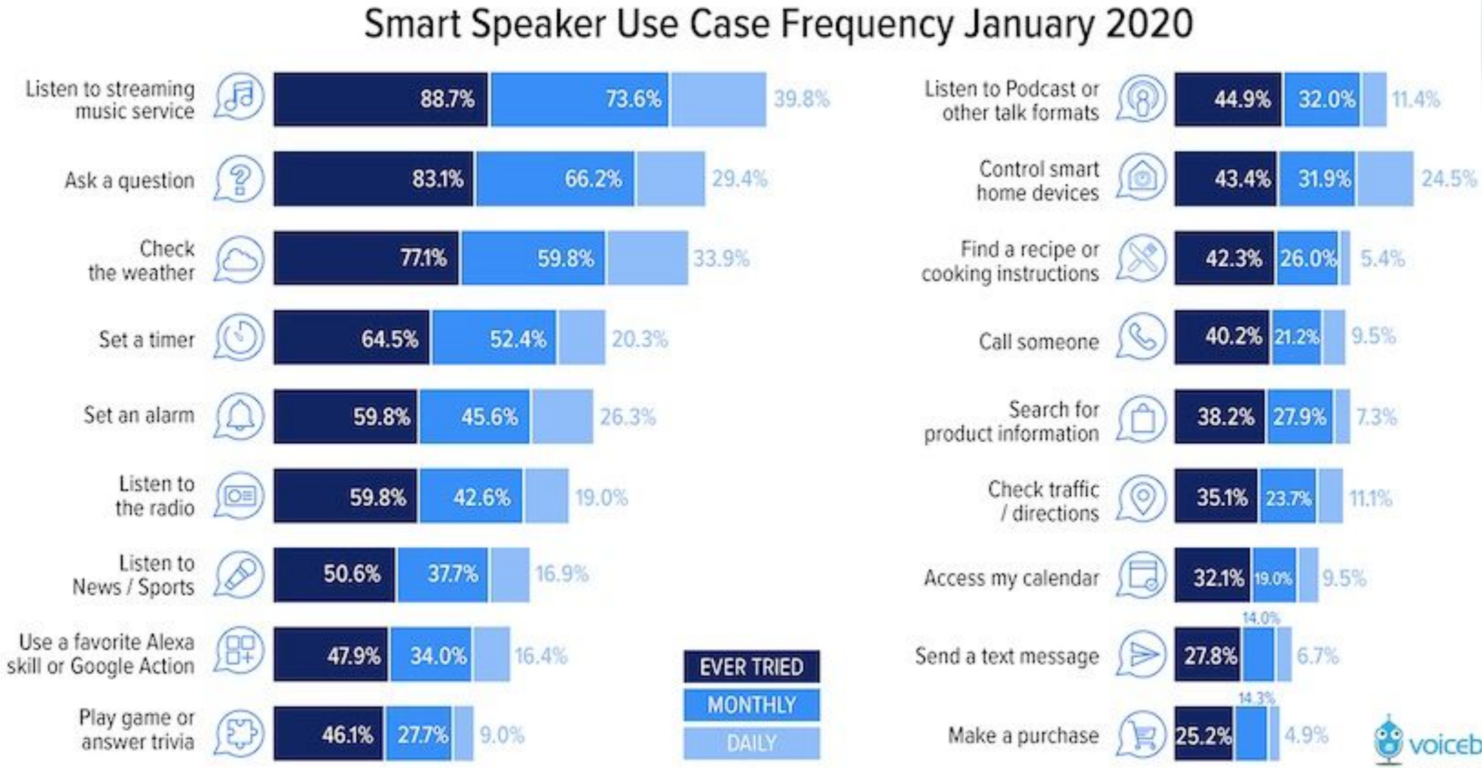
Early antibodies are detected in some patients within two weeks . [Severe acute respiratory syndrome and dentistry A retrospective view, *The Journal of the American Dental Association*, 2004-09-30]

Enzyme - linked immunosorbent assay , or ELISA , test . From about 20 days after the onset of clinical signs , ELISA tests can be used to detect immunoglobulin , or Ig , M and IgA antibodies in the serum samples of patients with SARS . Early antibodies are detected in some patients within two weeks .

System to collect answers of Covid-related questions in scientific publications
Winner in two competitions (White House, NIH)

Open Domain Q.A.: Lots of practical applicatio

Smart Speaker Use Case Frequency January 2020



voicebot.ai

Source: Voicebot.ai 2020

Course Outline

- **Open Domain Question Answering systems**
  - Introduction
  - Open Domain Q.A. practical applications
  - **SOA models**

Open Domain Q.A. baseline:

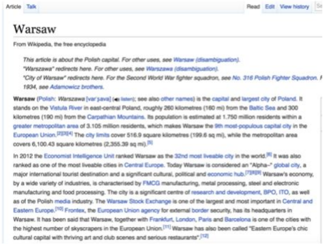
Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Wikipedia
The Free Encyclopedia

Document Retriever



Article Talk Read Edit View history Se

Warsaw

From Wikipedia, the free encyclopedia

This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

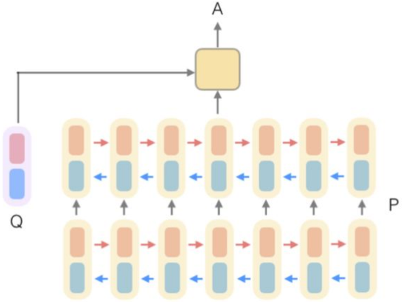
Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500

Information Retrieval Model (I.R.)



BERT

fine-tuned for Q.A.

Open Domain Q.A. baseline:

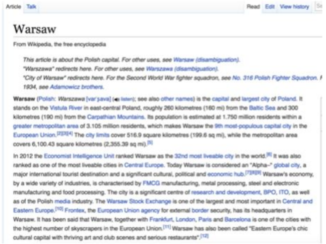
Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

SubOptimal!
we can not fix the Retriever errors

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
The Free Encyclopedia

Document Retriever



Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

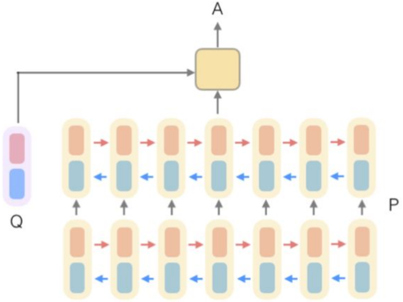
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500



Information Retrieval Model (I.R.)

BERT
fine-tuned for Q.A.

Course Outline

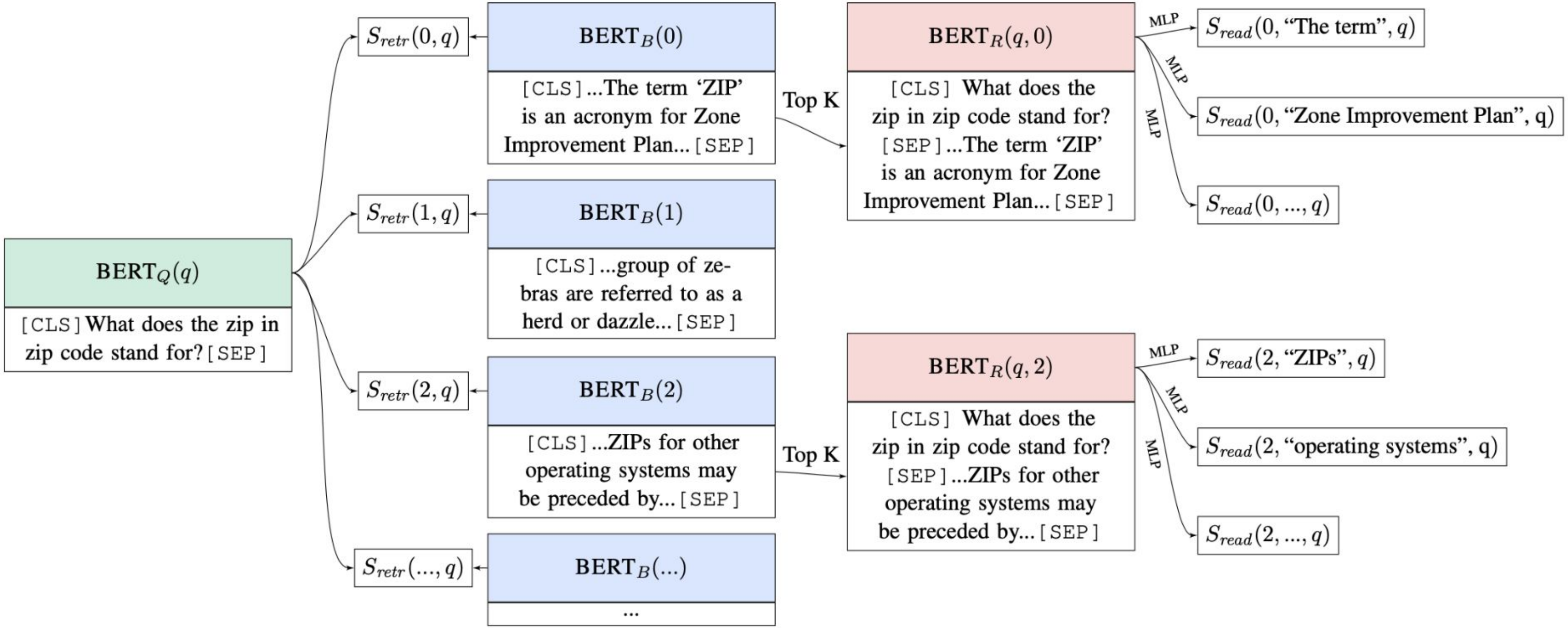
- **Open Domain Question Answering systems**
  - Introduction
  - Open Domain Q.A. practical applications
  - **SOA models**
    - Lee et al., 2019. **Latent Retrieval for Weakly Supervised Open Domain Question Answering**

Open Domain Q.A. we can train the retriever too! forget about I.R. system

- 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

Open Domain Q.A. we can train the retriever too



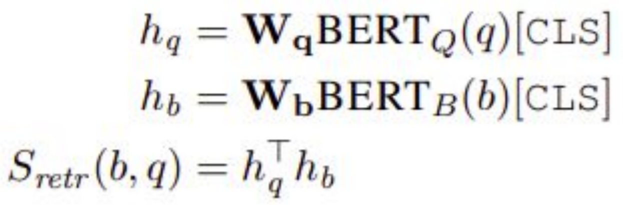
$\text{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0,q)$
$S_{retr}(1,q)$
$S_{retr}(2,q)$
$S_{retr}(\ldots,q)$
$\text{BERT}_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_B(1)$
[CLS] ...group of zebras are referred to as a herd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

- This model outpeforms BM25 by up to 19 points in exact match.

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

Open Domain Q.A. we can train the retriever too

- **Retriever**
  - In order for the retriever to be learnable, we define the retrieval score as the inner product of dense vector representations of the question q and the **evidence block b**.



$$h_q = \mathbf{W_q}\mathrm{BERT}_Q(q)[\mathtt{CLS}]$$

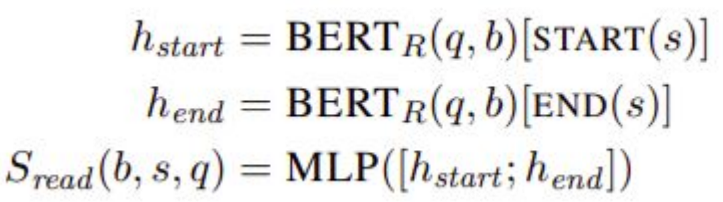
$$h_b = \mathbf{W_b}\mathrm{BERT}_B(b)[\mathtt{CLS}]$$

$$S_{retr}(b, q) = h_q^\top h_b$$

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

Open Domain Q.A. we can train the retriever too

- **Reader**
  - ○ The reader is a span-based variant of the reading comprehension model (finnetuned BERT):



$$\begin{aligned} h_{start} &= \mathrm{BERT}_R(q, b)[\mathrm{START}(s)] \\ h_{end} &= \mathrm{BERT}_R(q, b)[\mathrm{END}(s)] \\ S_{read}(b, s, q) &= \mathrm{MLP}([h_{start}; h_{end}]) \end{aligned}$$

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

Open Domain Q.A. we can train the retriever too

- **Joint learn Reader & Retriever**

$$S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$$

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

Open Domain Q.A. we can train the retriever too

- We still have millions of evidence blocks (passages)...
- Reduce the search space by:
  - Pretraining of retrieval (Inverse Cloze Task)
  - Pre-compiled index for inference (Locality Sensitive Hashing)
  - Beam-search over top 5 candidates
  - ...

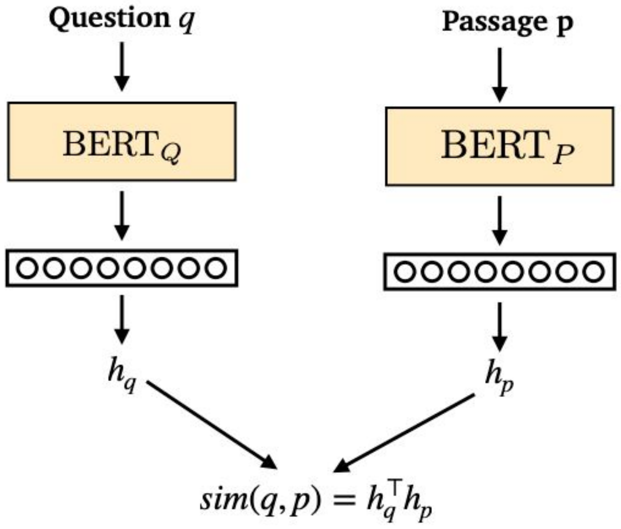
Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Course Outline

- **Open Domain Question Answering systems**
  - Introduction
  - Open Domain Q.A. practical applications
  - **SOA models**
    - Lee et al., 2019. **Latent Retrieval for Weakly Supervised Open Domain Question Answering**
    - Karpukhin et al., 2020. **Dense Passage Retrieval for Open-Domain Question Answering**

Open Domain Q.A. 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



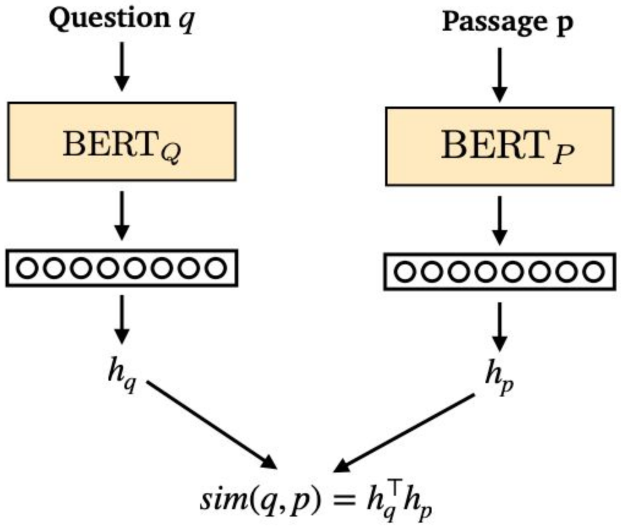
How to train the retriever?

The goal is to create a vector space such that relevant pairs of questions and passages will have smaller distance

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

Open Domain Q.A. 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



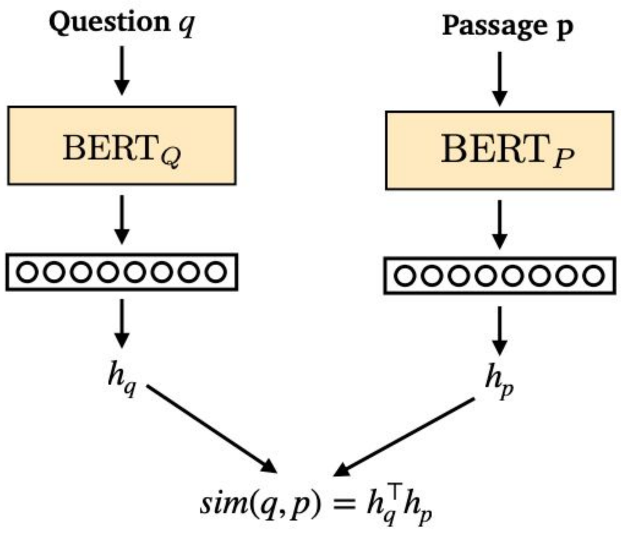
How to train the retriever?

Using positive and negative question passage pairs!

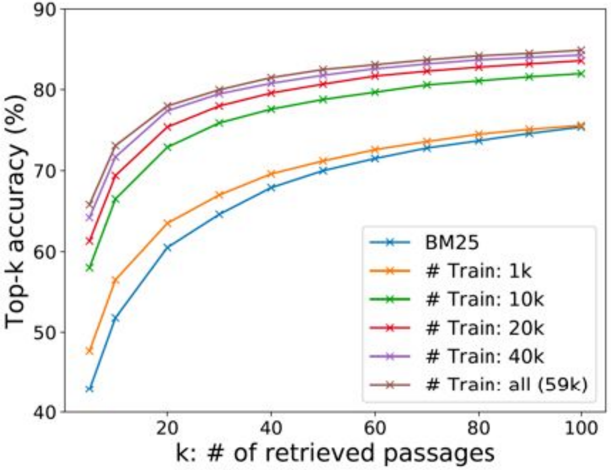
Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Open Domain Q.A. 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



1k Q/A pairs beat BM25!



Code here: https://github.com/facebookresearch/DPR

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

Open Domain Q.A. we can train the retriever too

- We still have millions of evidence blocks (passages)...
- Reduce the search space by:
  - Precompute all vector representation for passages
  - **FAISS** for indexing representations
    - open-source library for similarity search and clustering of dense vectors, which can easily be applied to billions of vectors

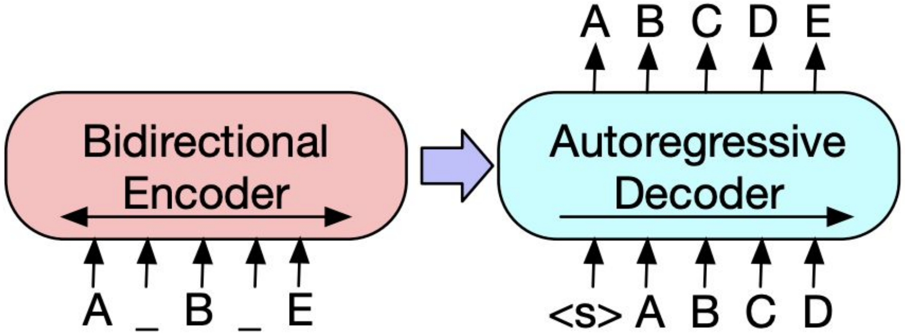
Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Course Outline

- **Open Domain Question Answering systems**
  - Introduction
  - Open Domain Q.A. practical applications
  - **SOA models**
    - Lee et al., 2019. **Latent Retrieval for Weakly Supervised Open Domain Question Answering**
    - Karpukhin et al., 2020. **Dense Passage Retrieval for Open-Domain Question Answering**
    - Izacard and Grave 2020. **Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering**

Open Domain Q.A. Dense retrieval + Generative Models

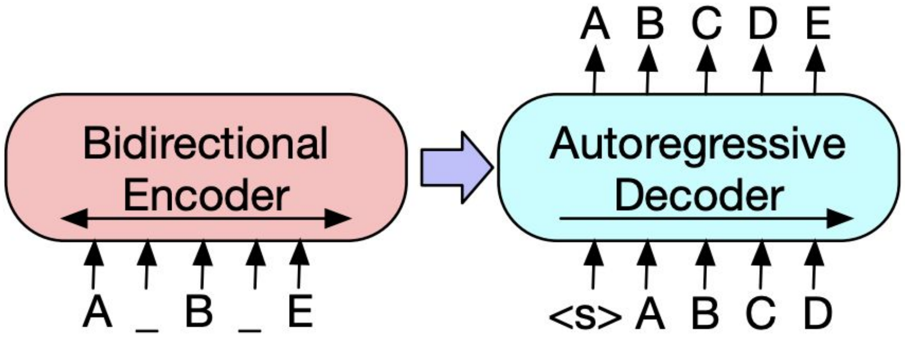
- Recent work shows that it is beneficial to **generate answers** instead of to **extract answers**. Seq2Seq archityecture to **generate** text:



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

Open Domain Q.A. Dense retrieval + Generative Models

- Recent work shows that it is beneficial to **generate answers** instead of to **extract answers**. Seq2Seq archityecture to **generate** text:

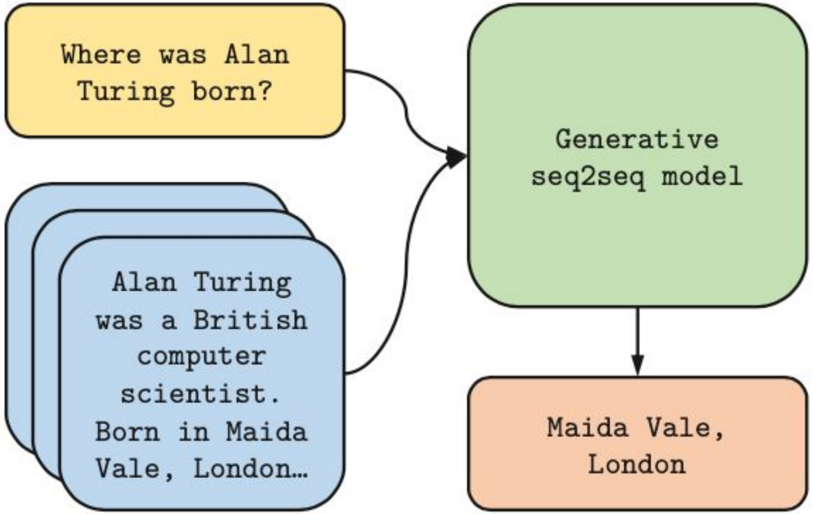


BERT is old fashioned!

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

Open Domain Q.A. Dense retrieval + Generative Models

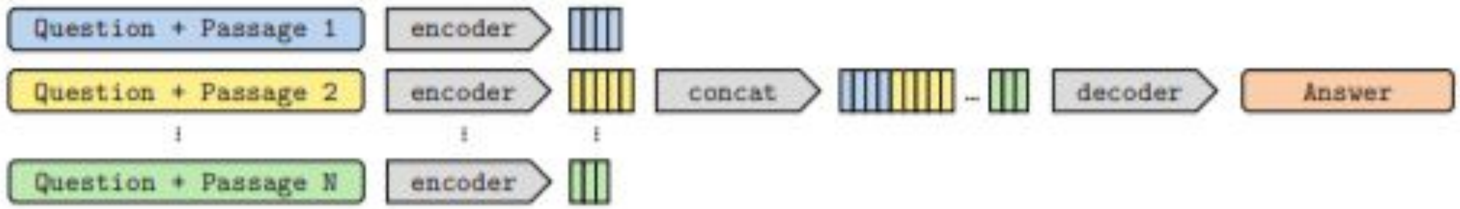
- Recent work shows that it is beneficial to **generate answers** instead of to **extract answers**.


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

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

Open Domain Q.A. Dense retrieval + Generative Models

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



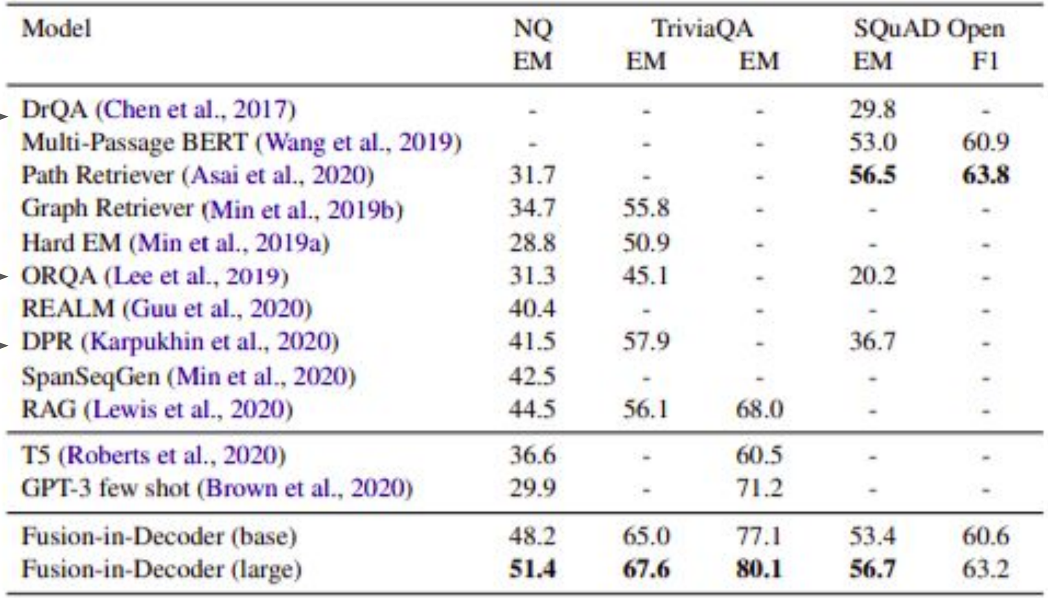
5, 10, 25, 50, 100…
the more the better!!

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

Open Domain Q.A. Dense retrieval + Generative Models

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

BM25 + BERT



Model	NQ	TriviaQA		SQuAD Open	
	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	29.8	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-	53.0	60.9
Path Retriever (Asai et al., 2020)	31.7	-	-	**56.5**	**63.8**
Graph Retriever (Min et al., 2019b)	34.7	55.8	-	-	-
Hard EM (Min et al., 2019a)	28.8	50.9	-	-	-
ORQA (Lee et al., 2019)	31.3	45.1	-	20.2	-
REALM (Guu et al., 2020)	40.4	-	-	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-	36.7	-
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	53.4	60.6
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**	**56.7**	63.2

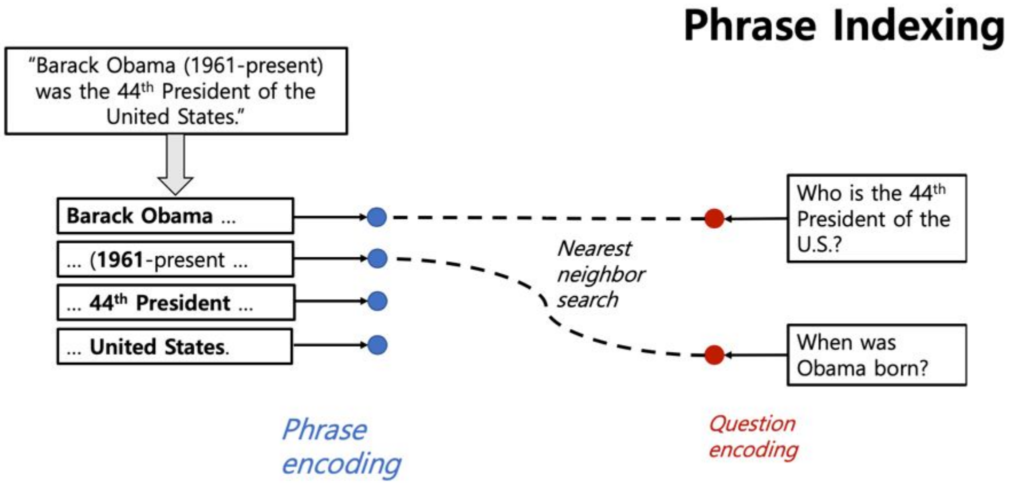
Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Course Outline

- **Open Domain Question Answering systems**
    - Introduction
    - Open Domain Q.A. practical applications
    - **SOA models**
        - Lee et al., 2019. **Latent Retrieval for Weakly Supervised Open Domain Question Answering**
        - Karpukhin et al., 2020. **Dense Passage Retrieval for Open-Domain Question Answering**
        - Izacard and Grave 2020. **Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering**
        - Seo et al., 2019. **Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index**
        - Lee et al., 2020. **Learning Dense Representations of Phrases at Scale**

Open Domain Q.A. without reader...

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

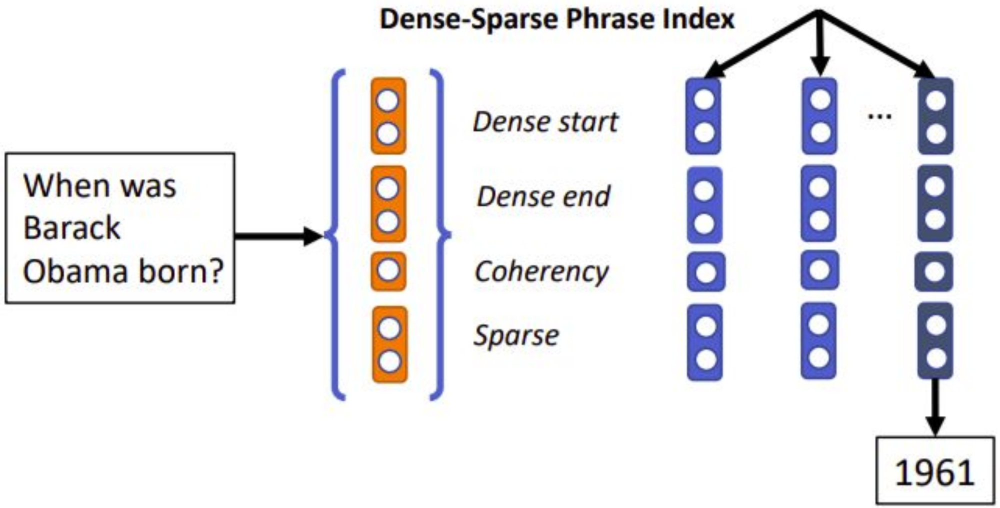


https://github.com/princeton-nlp/DensePhrases

Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index
Lee et al., 2020. Learning Dense Representations of Phrases at Scale

Open Domain Q.A. without reader...

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

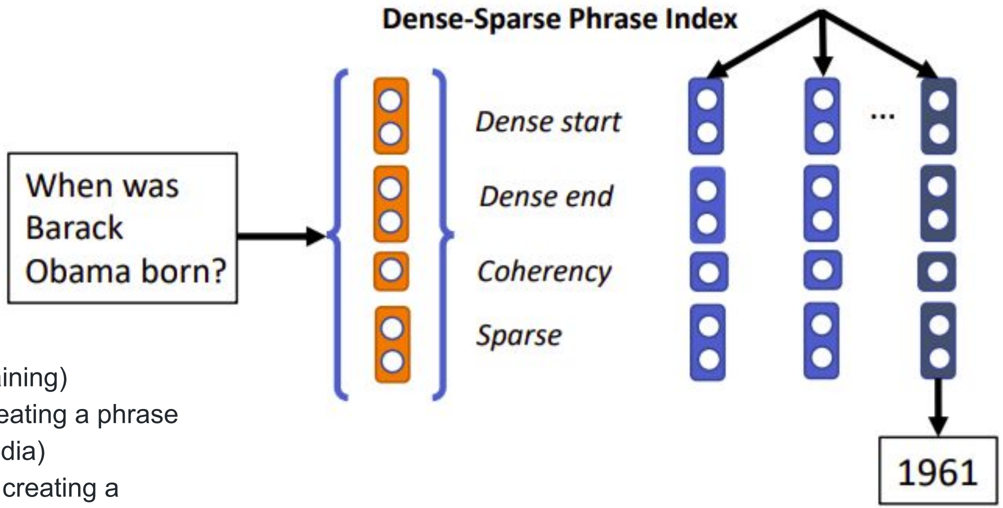


https://github.com/princeton-nlp/DensePhrases

Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index
Lee et al., 2020. Learning Dense Representations of Phrases at Scale

Open Domain Q.A. without reader...

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



https://github.com/princeton-nlp/DensePhrases

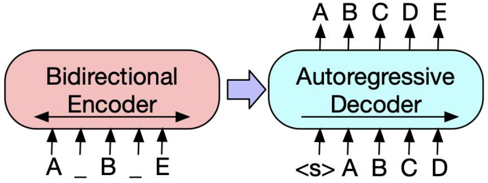
- Single 24GB GPU (for training)
- up to 150GB RAM (for creating a phrase index of the entire Wikipedia)
- up to 500GB storage (for creating a phrase dump of the entire Wikipedia)

Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index
Lee et al., 2020. Learning Dense Representations of Phrases at Scale

Course Outline

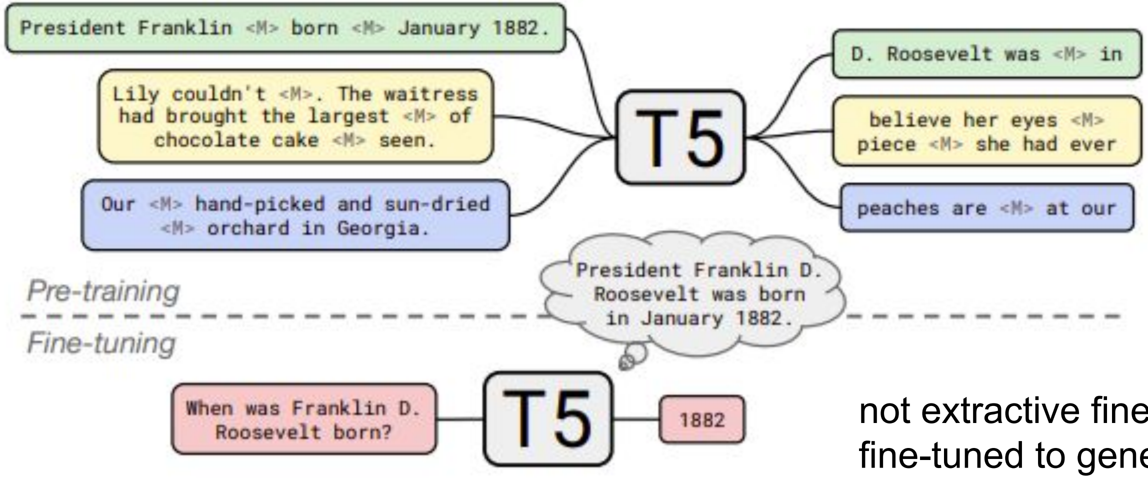
- **Open Domain Question Answering systems**
  - Introduction
  - Open Domain Q.A. practical applications
  - **SOA models**
    - Lee et al., 2019. **Latent Retrieval for Weakly Supervised Open Domain Question Answering**
    - Karpukhin et al., 2020. **Dense Passage Retrieval for Open-Domain Question Answering**
    - Izacard and Grave 2020. **Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering**
    - Seo et al., 2019. **Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index**
    - Lee et al., 2020. **Learning Dense Representations of Phrases at Scale**
    - Roberts et al., 2020. **How Much Knowledge Can You Pack Into the Parameters of a Language Model?**

Open Domain Q.A. without retriever...



- How Much Knowledge Can You Pack Into the Parameters of a Language Model?

T5 is pre-trained to fill in dropped-out spans of text



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.

D. Roosevelt was <M> in

believe her eyes <M> piece <M> she had ever

peaches are <M> at our

President Franklin D. Roosevelt was born in January 1882.

When was Franklin D. Roosevelt born?

1882

not extractive fine-tuning
fine-tuned to generate answer

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

Open Domain Q.A. without retriever...

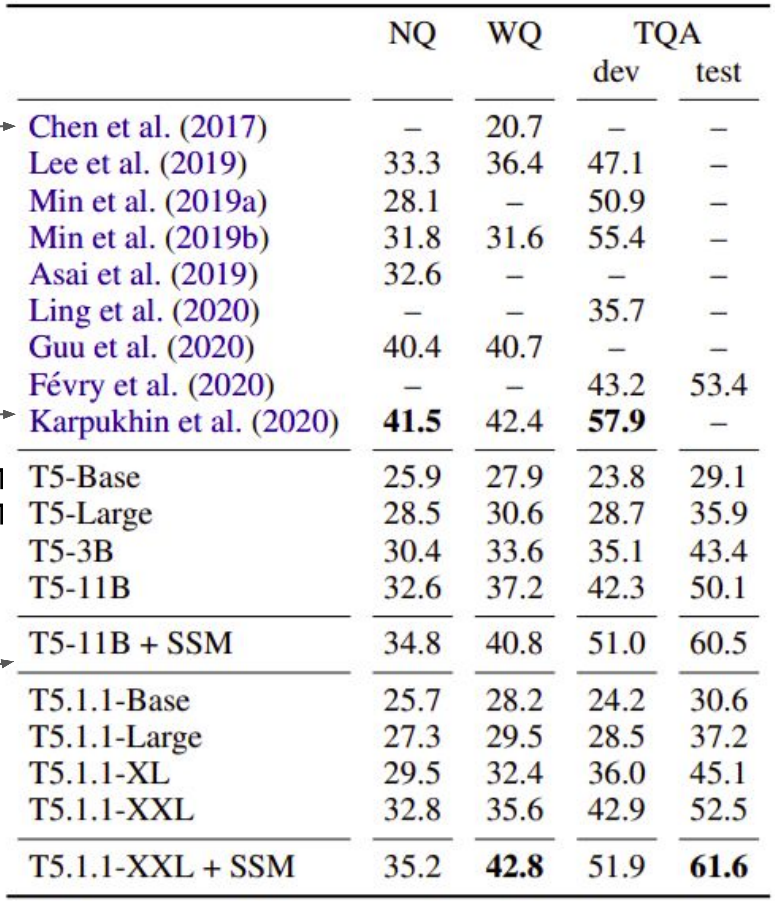
BM25 + BERT

DPR

How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Salient span masking

220M
770M
3B
11B



	NQ	WQ	TQA	
			dev	test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Févry et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–
T5-Base	25.9	27.9	23.8	29.1
T5-Large	28.5	30.6	28.7	35.9
T5-3B	30.4	33.6	35.1	43.4
T5-11B	32.6	37.2	42.3	50.1
T5-11B + SSM	34.8	40.8	51.0	60.5
T5.1.1-Base	25.7	28.2	24.2	30.6
T5.1.1-Large	27.3	29.5	28.5	37.2
T5.1.1-XL	29.5	32.4	36.0	45.1
T5.1.1-XXL	32.8	35.6	42.9	52.5
T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**

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

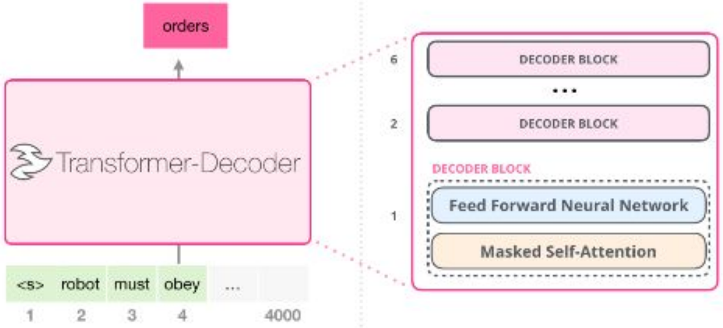
Open Domain Q.A. only L.M. **GPT2**

GPT2 is a very large transformer based language model trained on a massive dataset

48 layers, hidden size 1600, 1.5B parameters

WebText: 8 million documents, excluding Wikipedia (!)

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

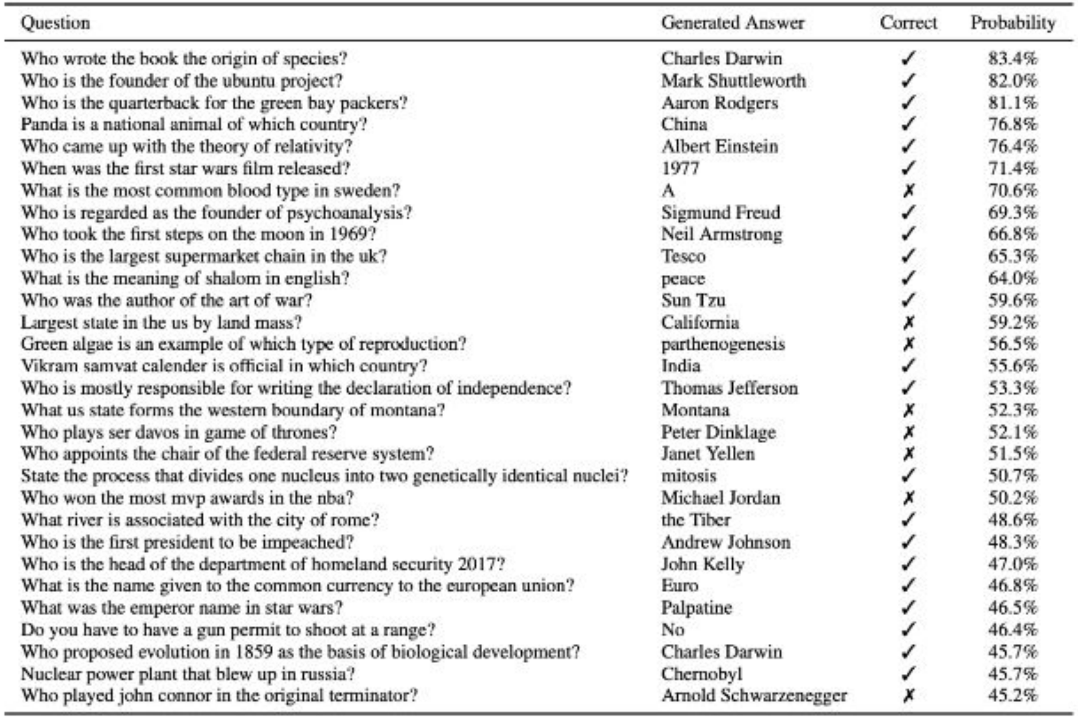


Radforf et al. 2019 Language Models are Unsupervised Multitask Learners

Open Domain Q.A. **GPT2**

Evaluated on Natural Questions and **no training at all**

63.1% on the **1%** of questions it is most confident in

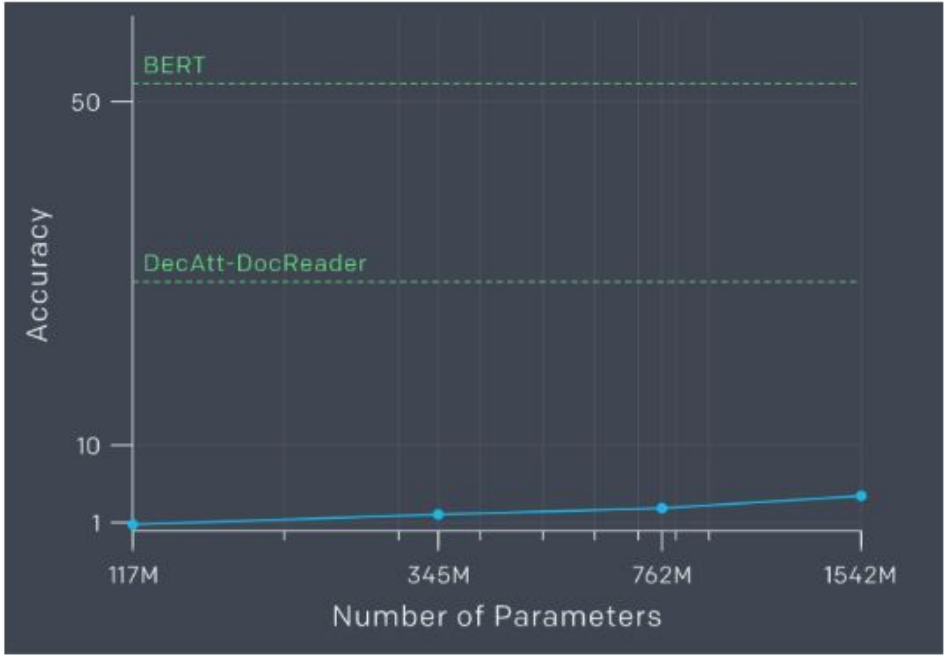


Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Radforf et al. 2019 Language Models are Unsupervised Multitask Learners

Open Domain Q.A. **GPT2**

Evaluated on Natural Questions and **no training at all**



4% accuracy:
Much much worse than supervised systems

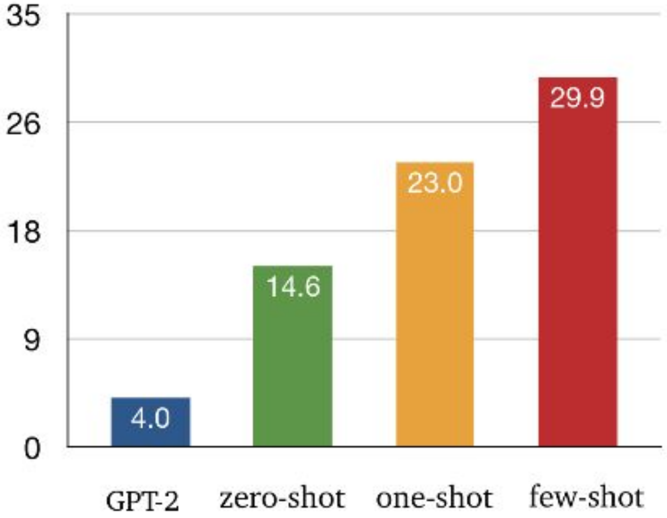
Radforf et al. 2019 Language Models are Unsupervised Multitask Learners

Open Domain Q.A. **GPT3** x100

96 layers, hidden size 12288, **175B** parameters

Larger corpora: Common Crawl + WebText + Books + **English Wikipedia**

Evaluated on Natural Questions:



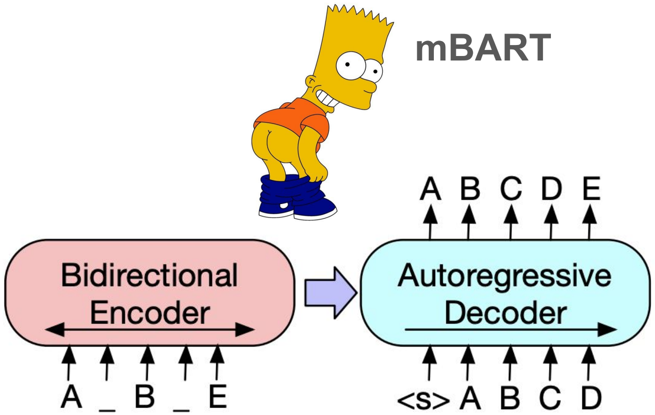
Few-shot learner

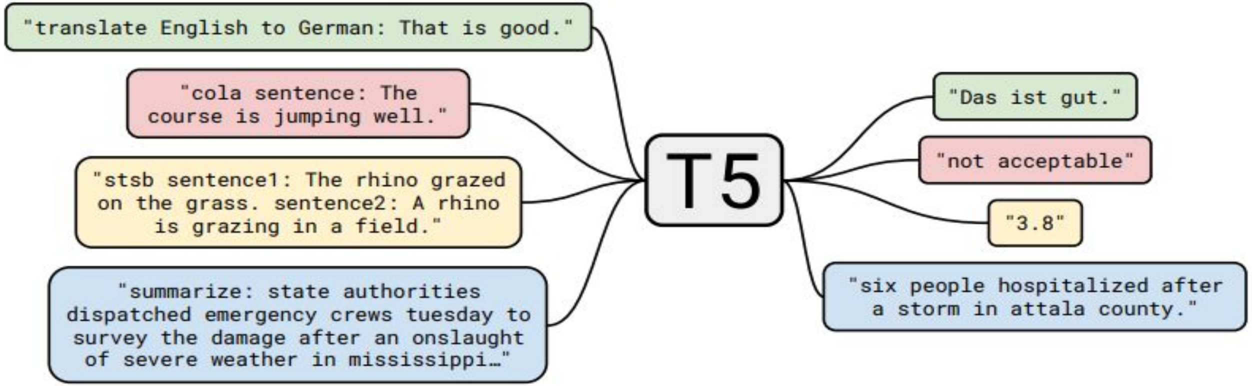
- No weight updates
- Q_1, A_1, Q_2, A_2, .., Q_K, A_K, Q ?
- One-shot setting is a special case when only **one** example is given.

Brown et al. 2020 Language Models are Few Shot Learners

Last words

- **N.L.P. -> language generation task?**
  - **Pretrained GPT, BART or T5 models for:**
    - Question Answering
    - Named Entity Disambiguation
    - Information extraction
    - Machine Translation*
    - Text Classification
    - ...





"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

"Das ist gut."

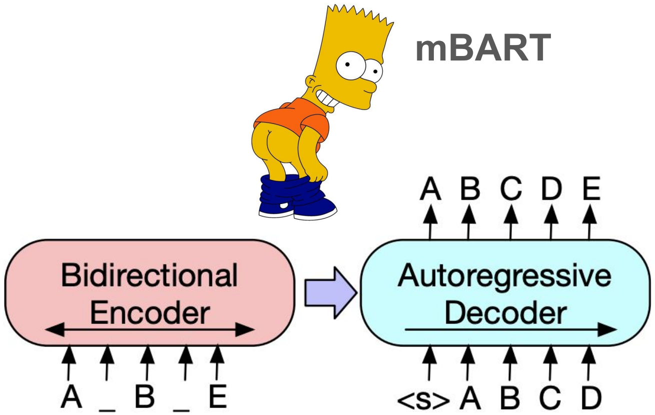
"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

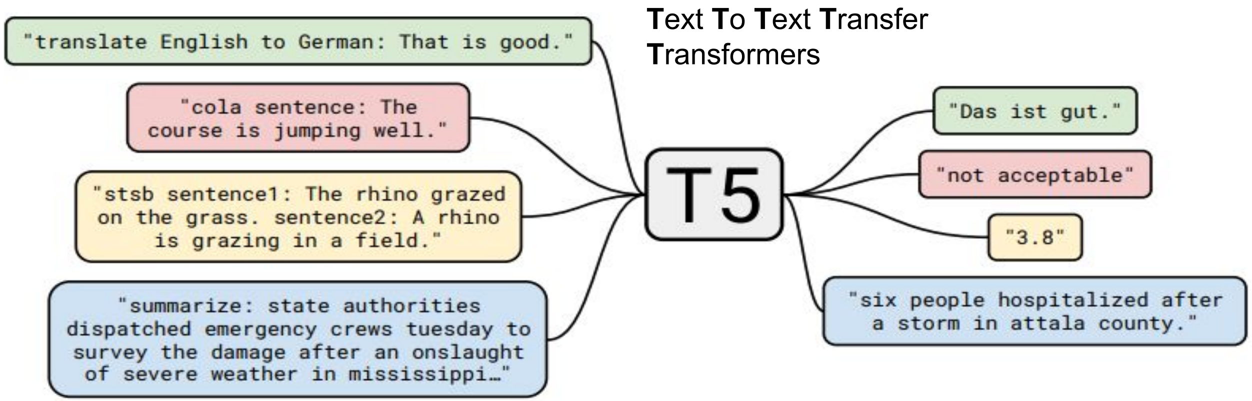
Last words

- **N.L.P. -> language generation task?**
  - **Pretrained GPT, BART or T5 models for:**
    - Question Answering
    - Named Entity Disambiguation
    - Information extraction
    - Machine Translation*
    - Text Classification
    - ...



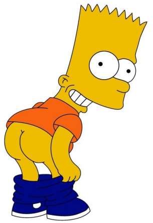
R.I.P.

??

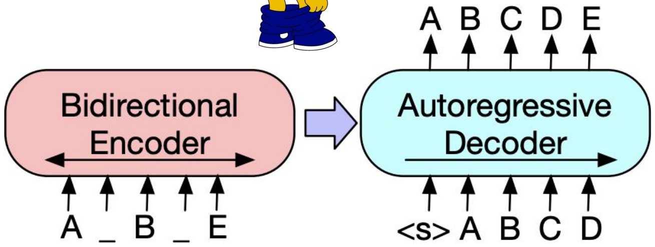

Text To Text Transfer Transformers
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Last words

- N.L.P. -> language generation task?
  - Pretrained GPT, BART or T5:
    - The performance is largely impacted by the model size.
    - A 11B T5 model = A 330M DPR BERT

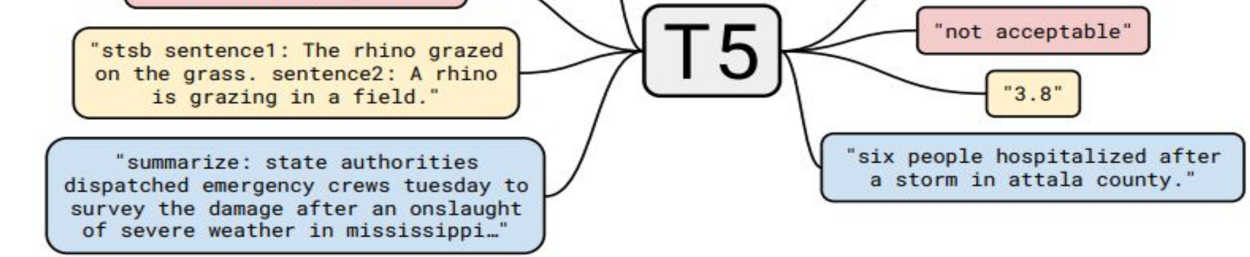


mBART



R.I.P.

??



Text To Text Transfer Transformers

"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

Last words

- N.L.P. -> language generation task?
    - Pretrained GPT, BART or T5:
        - The performance is largely impacted by the model size.
        - A 11B T5 model = A 330M DPR BERT
        - …
        - GPTneox, FLAN, instructGPT

Aligning Language Models to Follow Instructions:

Ouyang et al. 2022 *Training language models to follow instructions with human feedback*

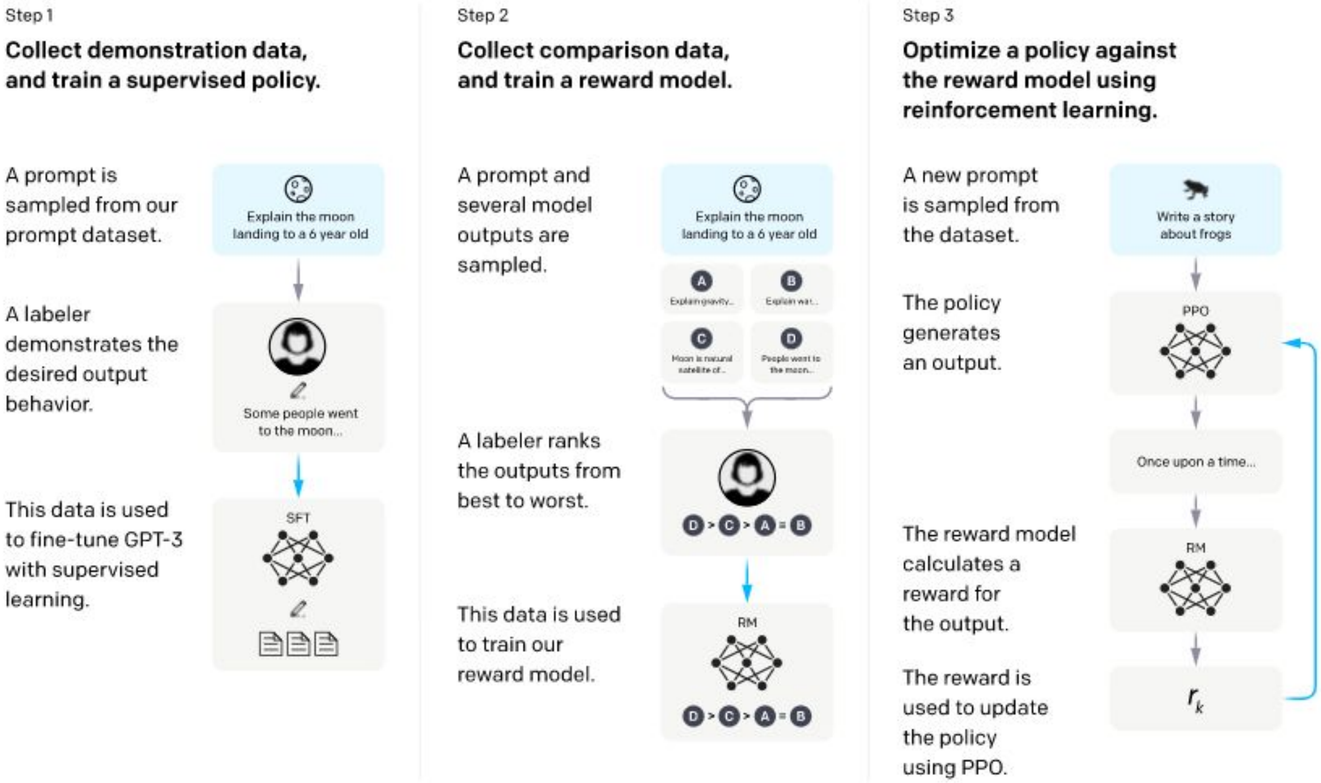
Our labelers prefer outputs from our 1.3B **InstructGPT** model over outputs from a 175B **GPT-3** model, despite having more than 100x fewer parameters.

R.I.P.

??

https://openai.com/blog/instruction-following/

Aligning Language Models to Follow Instructions



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

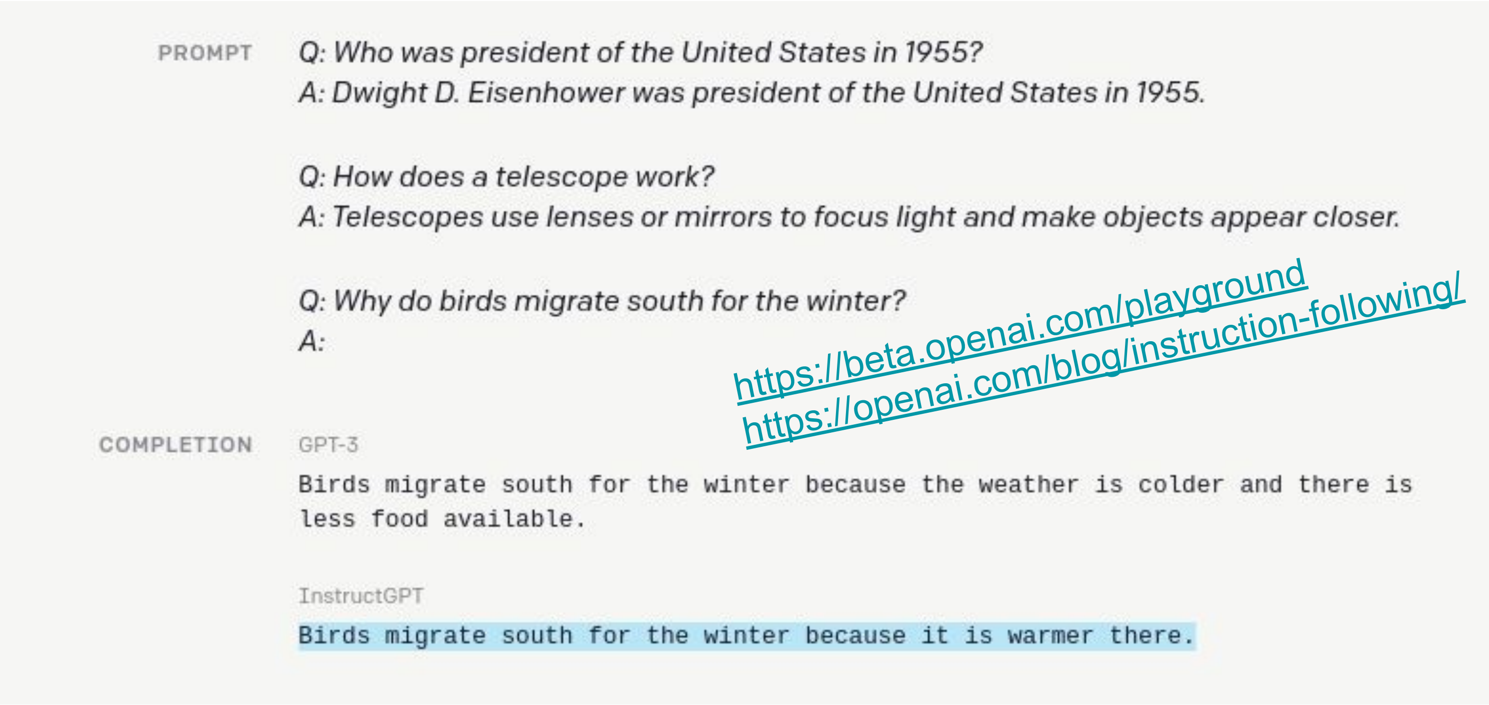
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Aligning Language Models to Follow Instructions



PROMPT

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Why do birds migrate south for the winter?
A:

COMPLETION

GPT-3

```
Birds migrate south for the winter because the weather is colder and there is
less food available.
```

InstructGPT

```
Birds migrate south for the winter because it is warmer there.
```

https://beta.openai.com/playground
https://openai.com/blog/instruction-following/

instructGPT -> text-davinci-002

Aligning Language Models to Follow Instructions

Q: Who was president of the United States in 1955? Promt
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: How does a telescope work? Promt
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

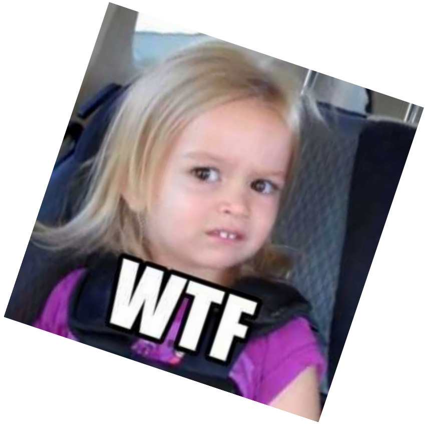
Q: Why is still BERT performing better than GPT3?
A:

Q: How many minutes does the Master of Puppets song last?
A:

Q: Who is better Goku or Vegeta?
A:

Q: 345+111?
A:

Q: Who are you? Skynet?
A:



Aligning Language Models to Follow Instructions

Q: Who was president of the United States in 1955? Promt
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: How does a telescope work? Promt
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Why is still BERT performing better than GPT3?
A:

Q: How many minutes does the Master of Puppets song last?
A:

Q: Who is better Goku or Vegeta?
A:

Q: 345+111?
A:

Q: Who are you? Skynet?
A:

Not only Q.A.
-> read a file line by line and store it in a list in python
-> read a file line by line and print it in python
-> Spaghetti boloñesa recipe

Thanks!

ander.barrena@ehu.eus
@4nderB