

Open Domain Question Answering

Gautier Izacard

FACEBOOK AI



Inria

Open-Domain QA

- Build systems that automatically answer questions posed in a natural language.
- Open-Domain:
 - Questions about anything, without pre-defined domain, pre-selected evidence.
 - Goal: use world knowledge to answer questions.
 - In research:
 - Knowledge source = Wikipedia.

Q: What element did Marie Curie name after her native land?

A: Polonium


Q: Technically a shoal of fish becomes a school of fish when it is?

A: Swimming in the same direction




Q: Which river flows from northern Moscow to the Caspian Sea?







A: Volga

Google Search: from search to answer



What element did Marie Curie name after her native land?





 All  News  Images  Videos  Shopping  More Settings Tools

polonium

And Marie was proven right: in 1898 the Curies discovered two new radioactive elements: **radium** (named after the **Latin word** for ray) and **polonium** (named after Marie's home country, Poland).
Jan 22, 2008

www.nobelprize.org › prizes › physics › questions-and-an...
[Marie Curie - Questions and answers - NobelPrize.org](#)

 About Featured Snippets  Feedback

People also ask

| | |
|---|---|
| What elements did Marie Curie discover? | ▼ |
| What is Marie Curie famous for? | ▼ |
| Why is Marie Curie's body radioactive? | ▼ |
| What are 3 interesting facts about Marie Curie? | ▼ |

Feedback

Google Search



how many employees does credit agricole have?



All



News



Images



Maps



Shopping

More

Settings

Tools

About 3,470,000 results (0.90 seconds)

140,000 employees

Crédit Agricole Group's Corporate & Investment Banking arm

Founded in 1885, it now **has** a presence in all segments of the banking, finance and insurance sectors and is one of Europe's foremost banks. Every day, its 140,000 **employees** actively strive to meet the needs of individual customers and multinational companies.

[www.ca-cib.com](#) › [pressroom](#) › [news](#) › [5-things-you-may...](#)

[5 things you may not know about our Bank | Crédit Agricole CIB](#)



About Featured Snippets



Feedback

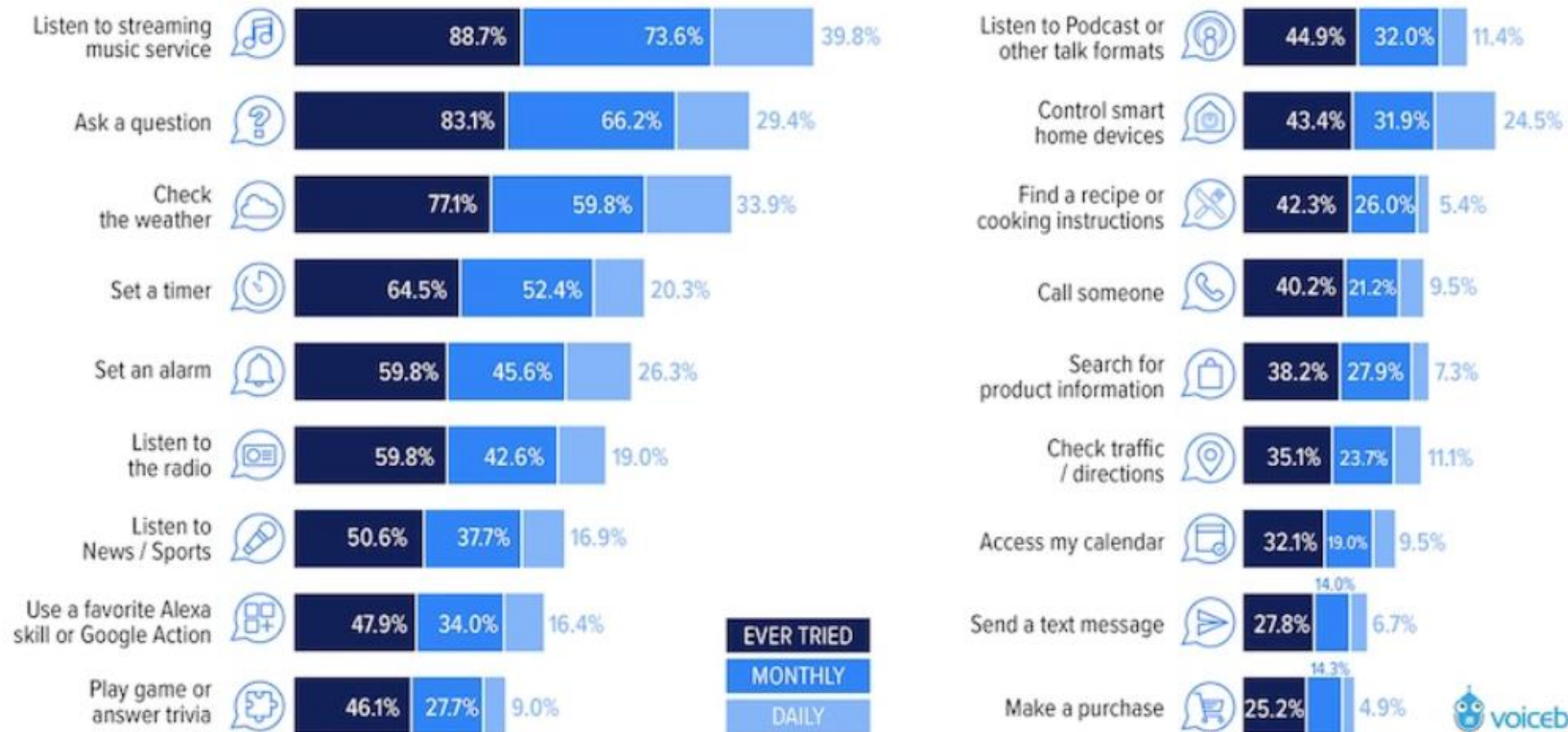
[www.credit-agricole.com](#) › [finance](#) › [finance](#) › [key-fig...](#) ▼

[Key figures Crédit Agricole S.A | Crédit Agricole](#)

Crédit Agricole S.A.'s ownership structure enables it to approach development with a view to long-term value creation. Together, the Regional Banks **have** ...

Top use case of smart speakers

Smart Speaker Use Case Frequency January 2020

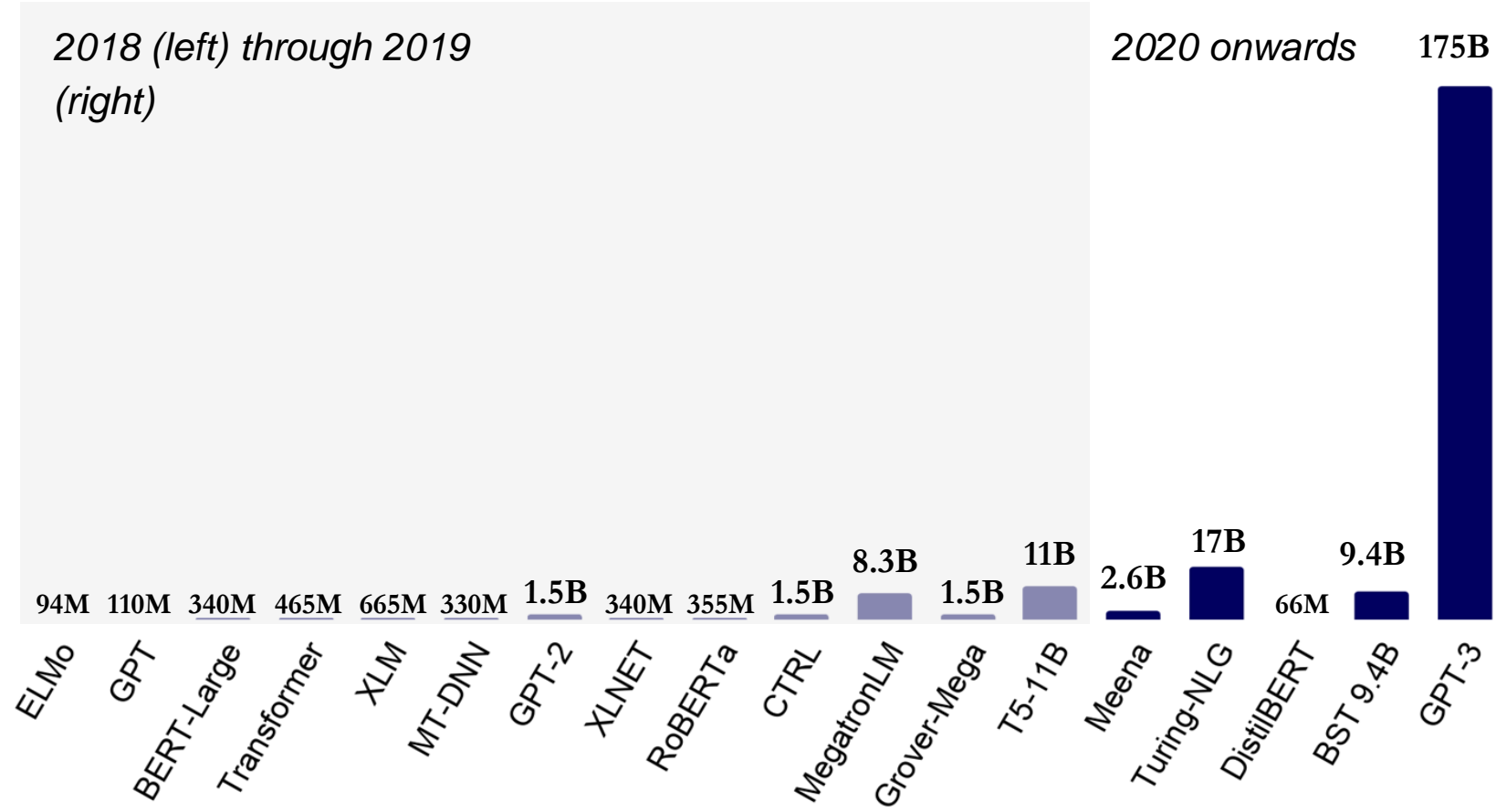


Natural Language Processing

- Key themes
 - Natural language understanding
 - Machine translation
 - Natural language generation
- Dominated by Deep Learning methods

Language models: Welcome to the Billion Parameter club

► Huge models, large companies and massive training costs dominate the hottest area of AI today, NLP.

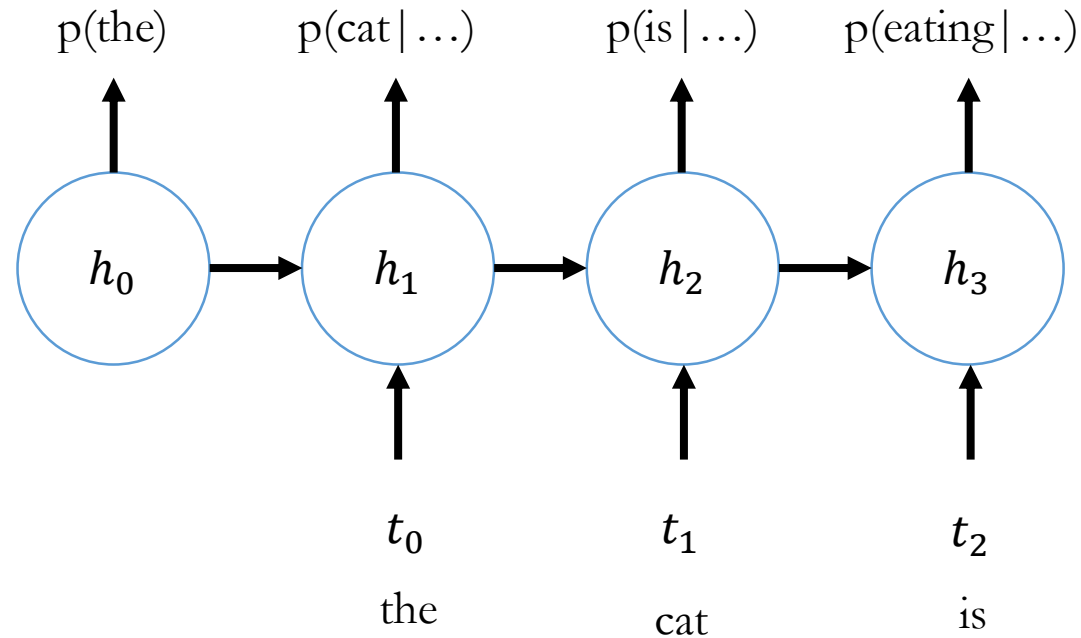


Note: The number of parameters indicates how many different coefficients the algorithm optimizes during the training process.

How to represent text?

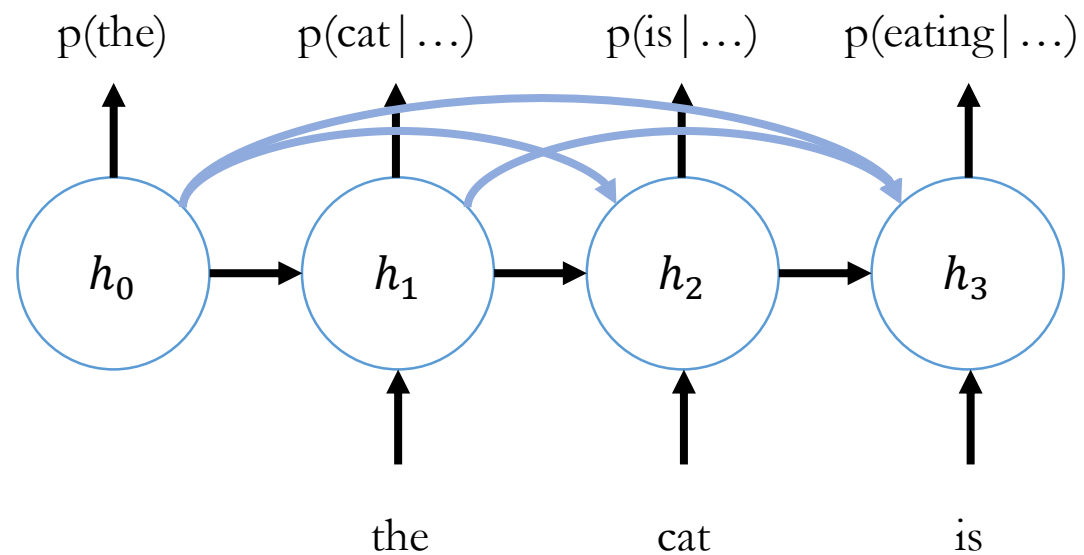
- Convert a piece of text into tokens, then 1 token \rightarrow 1 embedding vector
- The cat is eating an apple.
 - Word-level: The cat is eating an apple .
 - Character-level: I h e _ c a t _ i s _ e a t i n g _ a n _ a p p l e _ .
 - Byte Pair Encoding: The cat is eat ing an app le .
 - Start with a unigram vocabulary of all characters in data
 - Most frequent ngram pairs \rightarrow a new ngram
 - [Sennrich et al. Neural Machine Translation of Rare Words with Subword Units.]
 - Alternative: Wordpiece/Sentencepiece model
 - [Kudo, Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates]

Recurrent neural network

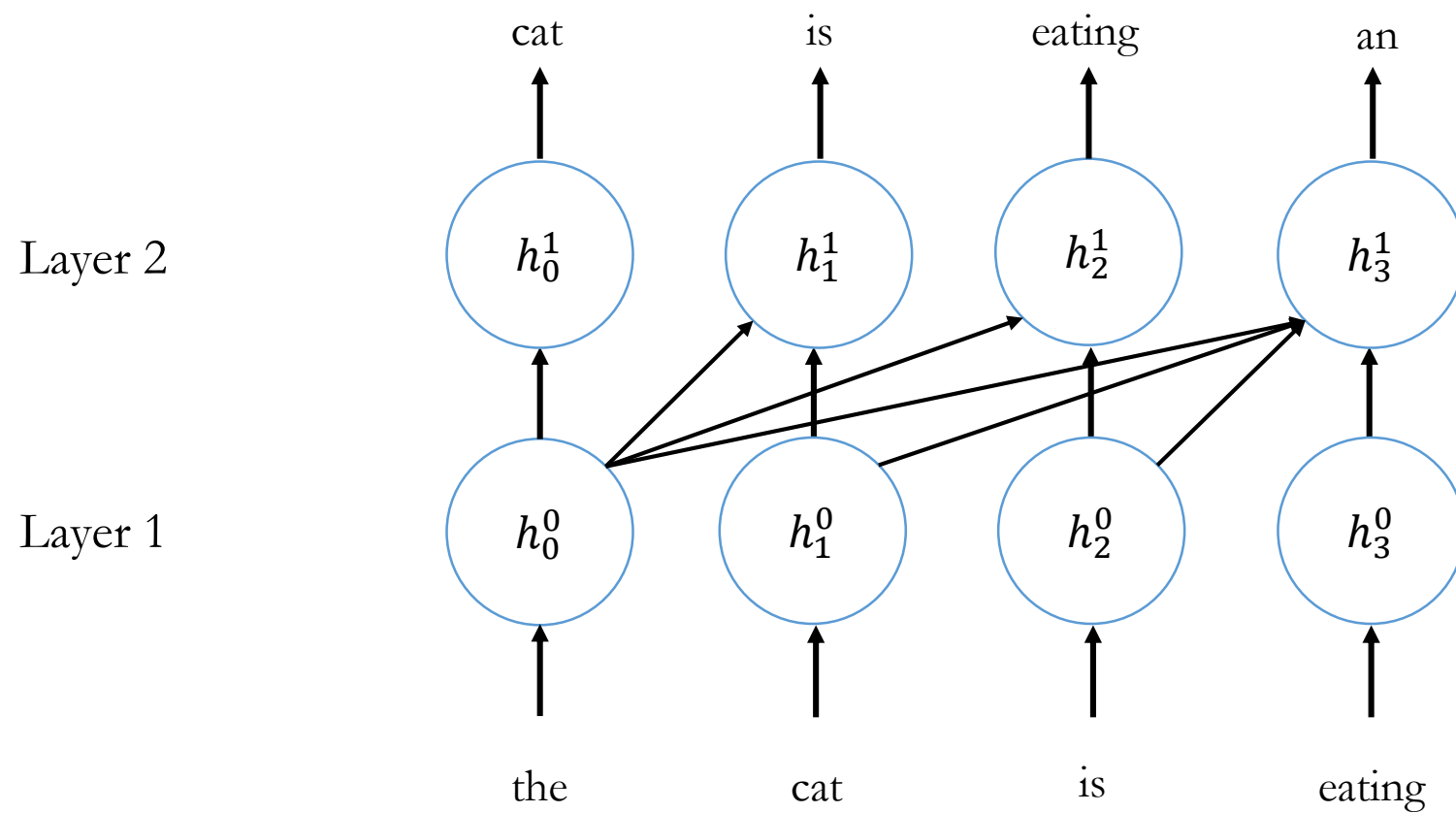


- $h_{i+1} = f_{\theta}(h_i, t_i)$
- LSTM (long short-term memory), GRU (gated recurrent unit)

Long term dependencies



Transformers



BERT

[Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding]

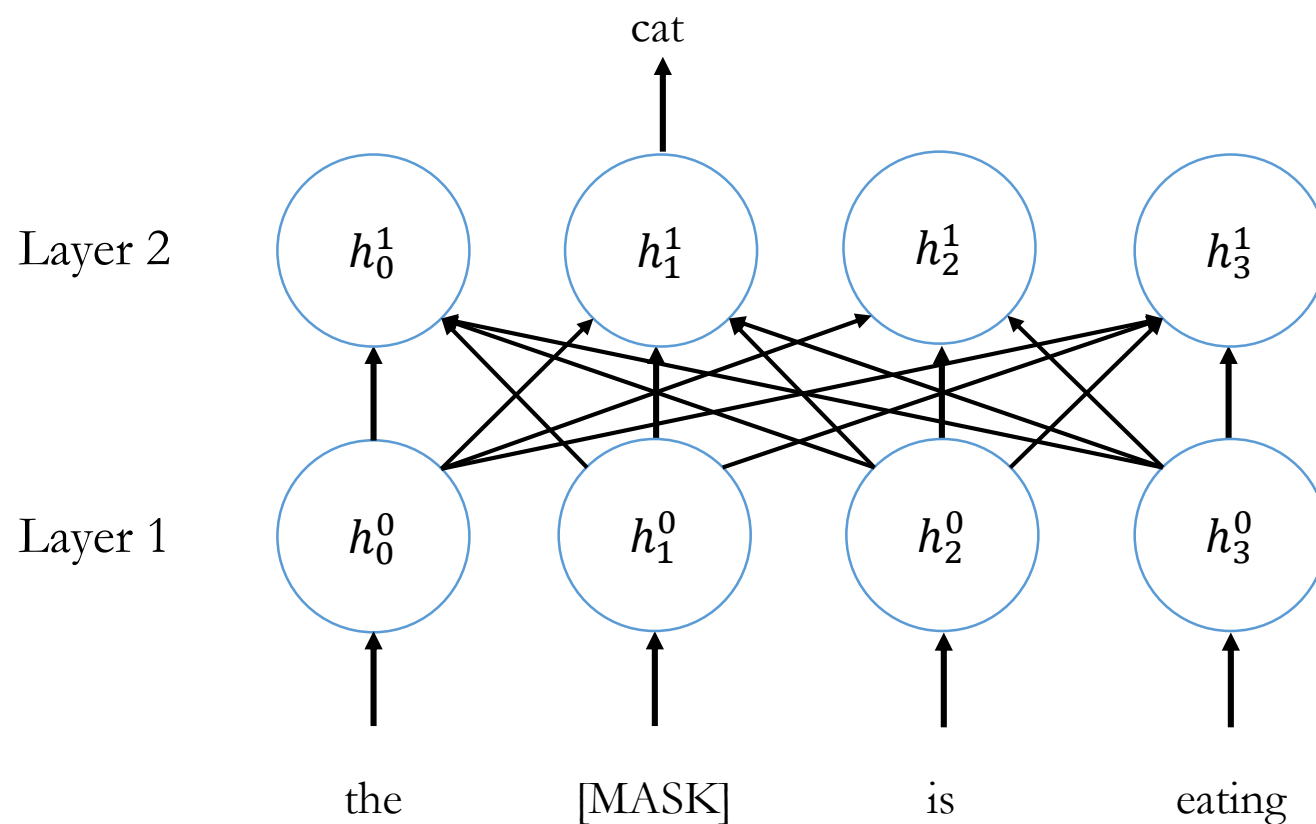
Pretraining task: Masked Language Modelling (MLM), mask out 15% of the input words

The man went to the [MASK] to buy a [MASK] of milk

↓ ↓

store gallon

BERT



In general:

- 12 or 24 layers
- 100M to 400M parameters
- Trained on text crawled on the web + Wikipedia + some book corpus
- Expensive to pre-train

BERT [Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding]

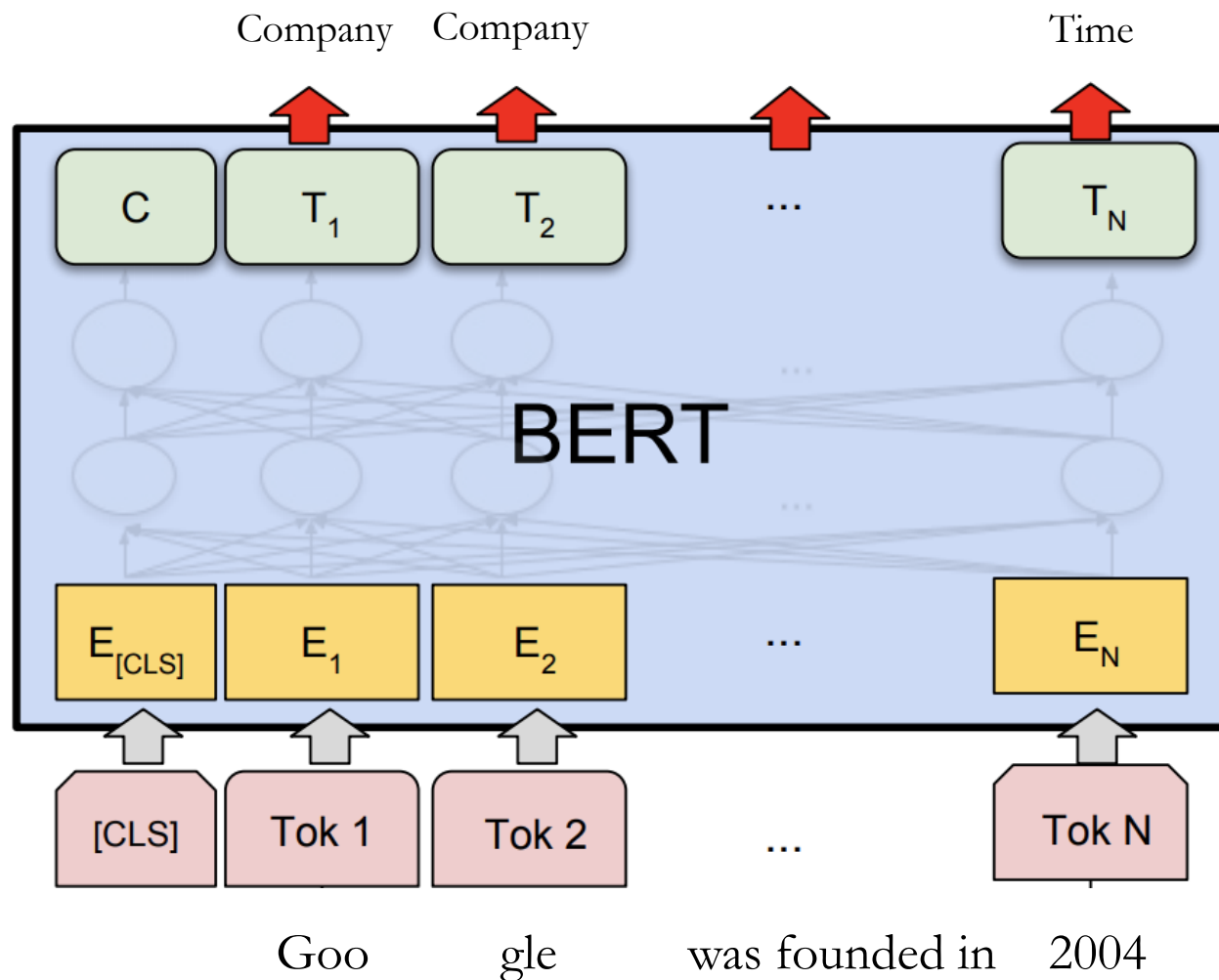
- After pretraining Finetuning on downstream tasks of interest:
 - Text classification: sentiment classification
 - Question answering
 - Named entity recognition
 - Entity linking, coreference resolution
 - Semantic parsing
- But also Roberta, Electra, XLNet
- Pretrained Seq2Seq models: T5, BART
- En français: Camembert, Flaubert.

Named-entity recognition

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index (BMI) of 33.5 kg/m2 , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin (HbA1c) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The β -hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again . The patient was treated with an insulin drip for euDKA and HTG with a reduction in the anion gap to 13 and triglycerides to 1400 mg/dL , within 24 hours . Her euDKA was thought to be precipitated by her respiratory tract infection in the setting of SGLT2 inhibitor use . The patient was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely . She had close follow-up with endocrinology post discharge .

Color codes: Patient problem, Test, Treatment

Named-entity recognition



Machine Reading

- Input is a question Q and a passage P
 - Passage: paragraph, document, arbitrary-length text
- Expected output is an answer A
- SQuAD dataset: Stanford Question Answering Dataset [Rajpurkar et al., 2016]
- Restricted setting: A is a segment of P
- → extractive question answering

Passage

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.

Question: Which NFL team won Super Bowl 50?

Answer: Denver Broncos

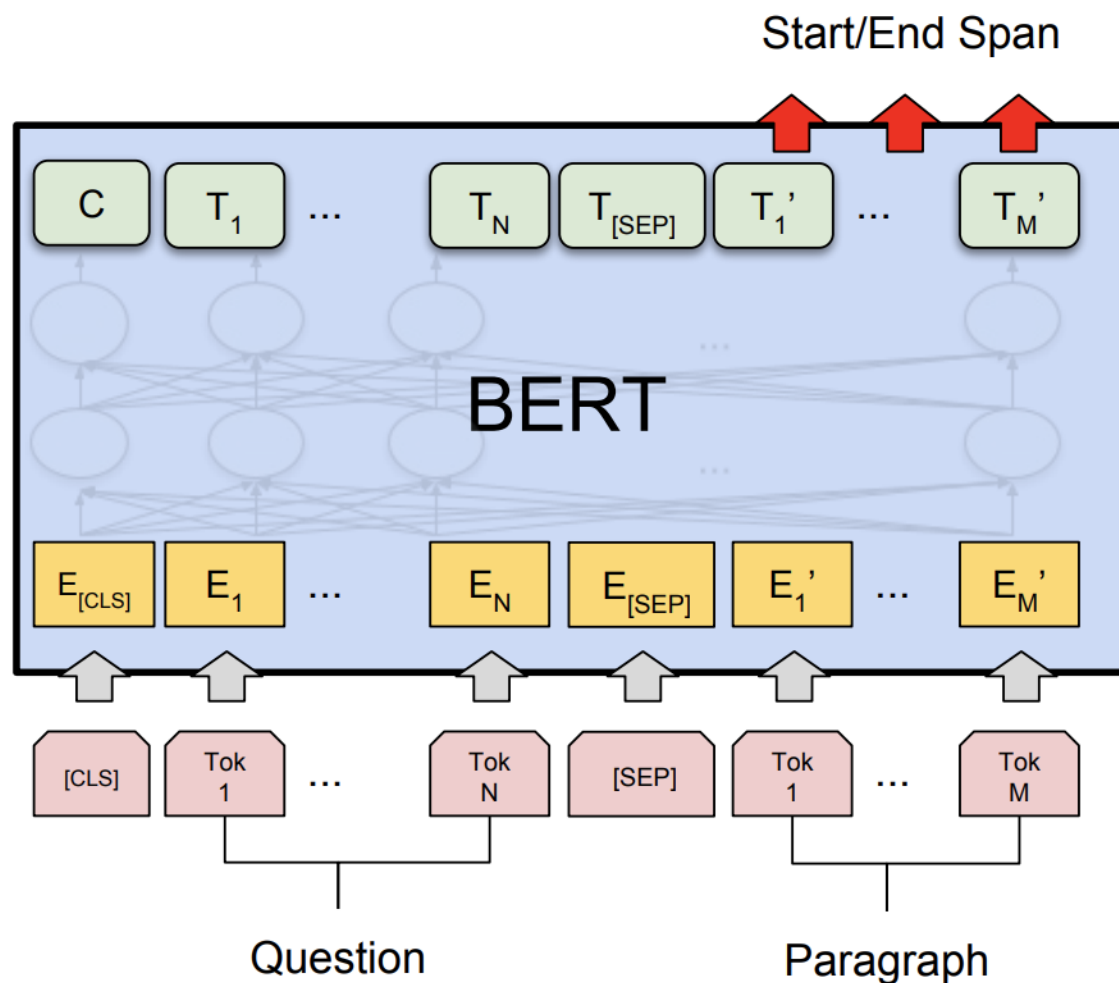
Question: What does AFC stand for?

Answer: American Football Conference

Question: What year was Super Bowl 50?

Answer: 2016

Machine reading



- Question and support passage side by side
- For each token of the support passage:
 1. Is it the start of the answer span?
 2. Is it the end of the answer span?

On SQuAD 1.1:

- Humans: F1 = 90.9
- RoBERTa F1 = 94.6
- Best Model (LUKE) F1 = 95.4

Open-Domain QA

- Open-domain: no pre-selected support passage
- Long form question answering

Q: Why Is the Sky Blue?

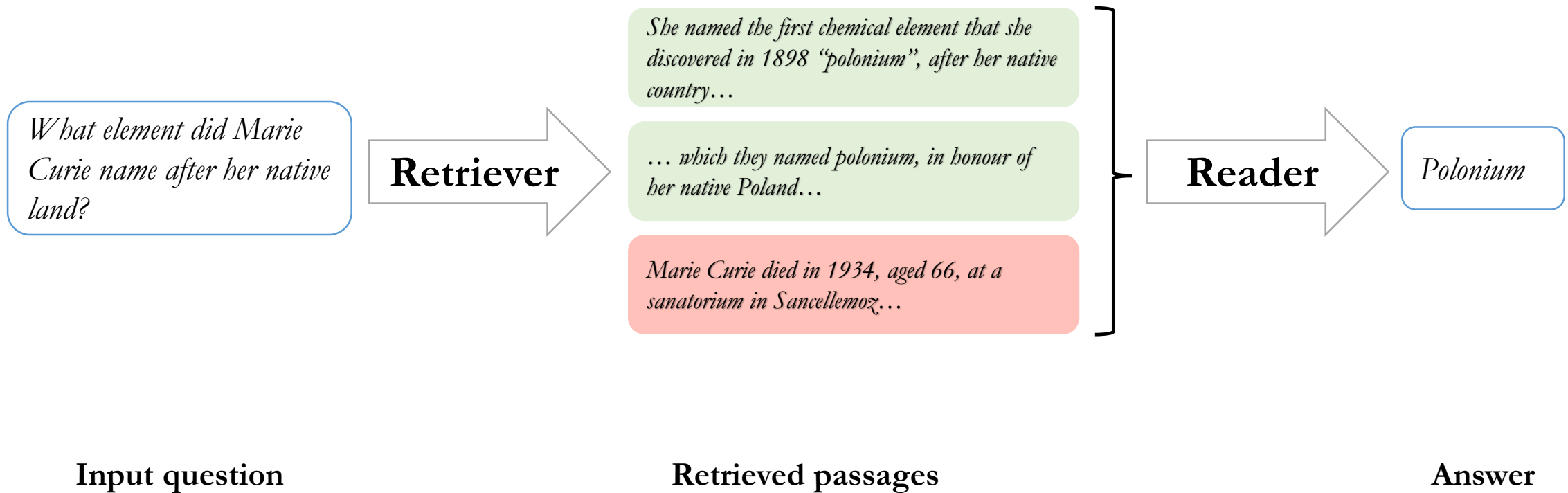
A: Gases and particles in Earth's atmosphere scatter sunlight in all directions. Blue light is scattered more than other colors because it travels as shorter, smaller waves. This is why we see a blue sky most of the time.

- Extractive models don't work for this

Datasets

- NaturalQuestions
 - Google Search queries, with human annotated answers.
- TriviaQA
 - Human generated question-answer pairs, collected on quiz league websites.
- SQuAD
 - Open-version without support passages.
 - Each answer is a span extracted from a Wikipedia paragraph.
- NarrativeQA
 - Human generated question-answer pairs based on summaries of stories (books or movie scripts).
 - *A: He dismantles it and attaches it to his mother's jeep*
 - *Q: What does Mark do with his radio station?*

Retriever-reader approach



Knowledge sources

- Which knowledge sources?
- Unstructured: Wikipedia, Web documents.
- Semi-structure: Tables, Wikipedia infoboxes
- Structured: Wikidata graph
- Here: Wikipedia text
- Why Wikipedia ?
 - Contains a lot of human knowledge.
 - Manageable size for research lab.
 - Already offers an interesting scale problem.

Retriever

- Select relevant passages in the knowledge source
- Need to operate at scale
 - Applying Bert on Wikipedia = ~ 50 GPU-hours
- Preprocessing, for each passage p a representation $E(p)$ is precomputed
- Then for an input question q with representation $E(q)$
- Similarity score between passage p and question q : $S_{\theta}(q, p) = E(q)^T E(p)$
- Retrieve passages with the highest similarity score with the question
 - MIPS with efficient libraries [FAISS, Johnson et al., 2017]
- Knowledge source
 - Wikipedia ~ 5 millions documents
 - Divided into 22 Millions passages of 100.

Dense vs Sparse representation

- Sparse vs dense representation $d_1 \gg d_2$
 - Sparse repr. $[0 \dots 1 \ 1 \ 0 \dots 1] \in \mathbb{R}^{d_1}$
 - Dense repr. $[0.51 \dots 0.25 \ 0.41] \in \mathbb{R}^{d_2}$
- Sparse: What part of the atom did Chadwick discover?
- Dense: Who is the bad guy in lord of the rings?
 - Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the villain Sauron in the Lord of the Rings trilogy by Peter Jackson.
- Sparse representation works well for rare entities
- Easier to capture context/general meaning with a dense representation

Sparse Retriever

- Classic methods based on TF-IDF
 - tf = term frequency, idf = inverse document frequency
 - Improved version BM25
 - t : term, p : passage (= 100 words of a Wiki. article), D : corpus (= Wikipedia)
 - $tf - idf(t, p, D) = tf(t, p) idf(t, D)$
 - $tf(t, p) = \log(1 + freq(t, p))$
 - $idf(t, D) = \log(\frac{|D|}{|d \in D: t \in d|})$
- No training
- Sparse representation

Dense Retriever

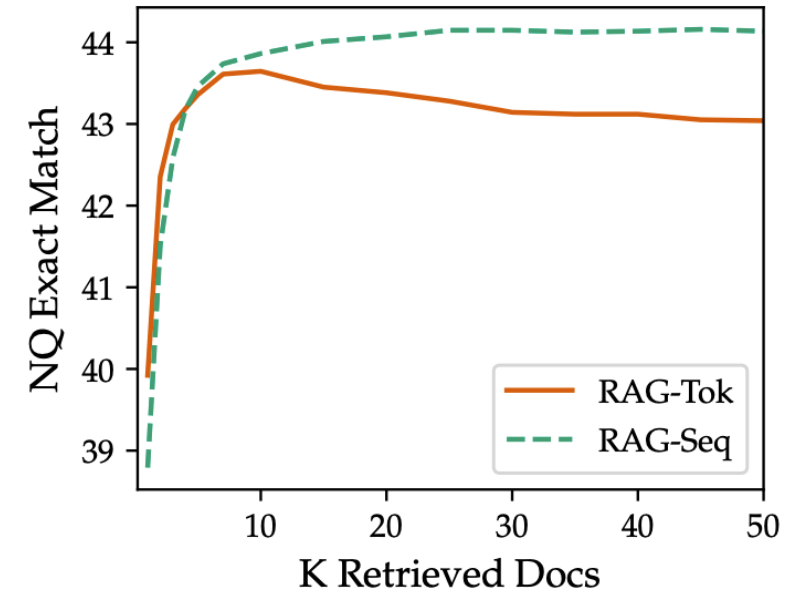
- Dense retriever
 - Learn a dense vector per passage (Karpukhin et al., 2019)
- Sparse methods have outperformed dense retriever until 2019!
 - Sparse representation works well for rare entities
 - Learned dense retriever can adapt to the question distribution
- Classification problem
 - Out of n passages, the retriever is trained to select a single positive passage

Retriever-Reader

1. Given an input question select n relevant documents in Wikipedia
2. For each passage, generate an answer
3. Given n answers, select the answer with the highest confidence

Reader for Open-Domain QA

- How to leverage a knowledge source?
 - How to combine multiple passages?
 - Is it useful to scale to many passages?
 - How to read a lot of passages simultaneously?
- Standard approach: Extractive Reader
 - Better than generative models for machine reading.
 - Various schemes to transfer this approach to multiple passages
 - No evidence that more than 20 passages improve performance

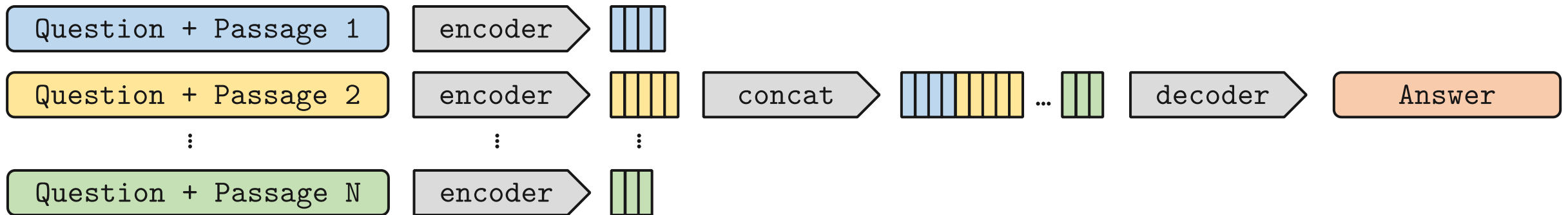


Generative reader

- Use pretrained seq2seq models
- Recently, generative models:
 - Without support knowledge source
 - Closed-book T5, Roberts et al., 2020
 - With support knowledge source
 - RAG, Lewis et al., 2020; SpanSeqGen, Min et al., 2020
 - RAG, no improvement after 25 passages

Fusion-in-Decoder

- Initialize seq2seq model with pretrained T5
 1. Apply the encoder on each passage concatenated with the question.
 2. Concatenate the representations of all passages.
 3. Apply the decoder on the concatenation.

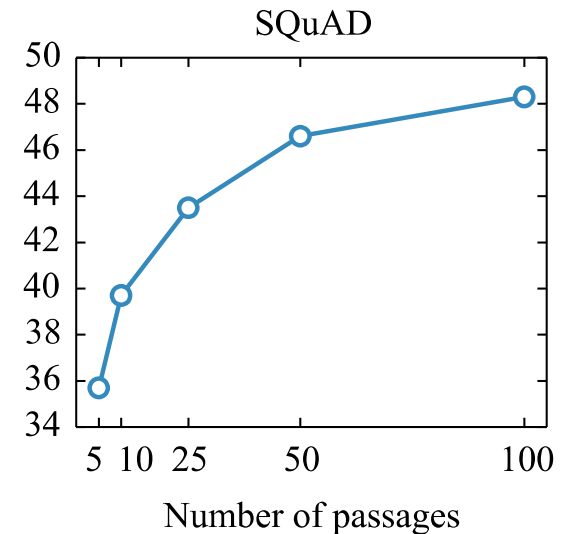
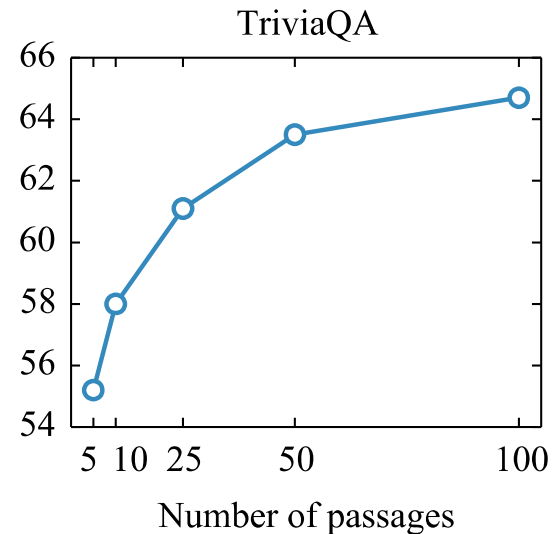
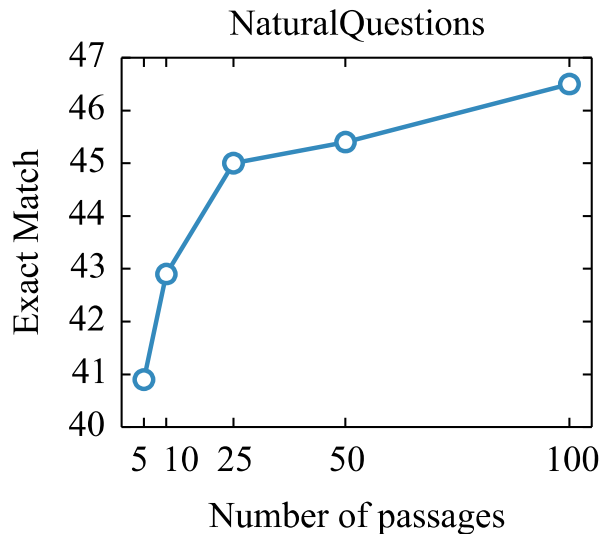


Results

| Model | NaturalQuestions | TriviaQA | | SQuAD Open |
|--|------------------|-------------|-------------|-------------|
| DrQA (Chen et al., 2017) | - | - | - | 29.8 |
| Multi-Passage BERT (Wang et al., 2019) | - | - | - | 53.0 |
| Path Retriever (Asai et al., 2020) | 31.7 | - | - | 56.5 |
| 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 |
| Fusion-in-Decoder (large) | 51.4 | 67.6 | 80.1 | 56.7 |

Scaling with the number of passages

- Scale well with the number of retrieved passages
- Improvement up to 100 passages
 - Other methods peak around 10 to 20 passages



How to train a retriever without annotations or heuristics?

- Annotated pairs of question-support document
 - Expensive to obtain
- Alternative use an heuristic to identify relevant passage
 - Positive passage = passage containing the answer
 - Not applicable to long form question answering
- Use signal extracted by the reader from the question-answer pairs
- Knowledge distillation between the reader attention scores and the retriever

| Model | NQ | | TriviaQA | |
|--|-------------|-------------|-------------|-------------|
| | dev. | test | dev. | test |
| DPR (Karpukhin et al., 2020) | - | 41.5 | - | 57.9 |
| RAG (Lewis et al., 2020b) | - | 44.5 | - | 56.1 |
| ColBERT-QA (Khattab et al., 2020) | - | 48.2 | - | 63.2 |
| Fusion-in-Decoder (T5 base) (Izacard & Grave, 2020) | - | 48.2 | - | 65.0 |
| Fusion-in-Decoder (T5 large) (Izacard & Grave, 2020) | - | 51.4 | - | 67.6 |
| Ours (starting from BERT, T5 base) | 39.3 | 40.0 | 62.5 | 62.7 |
| Ours (starting from BM25, T5 base) | 47.9 | 48.9 | 67.7 | 67.7 |
| Ours (starting from DPR, T5 base) | 48.0 | 49.6 | 68.6 | 68.8 |
| Ours (starting from DPR, T5 large) | 51.3 | 52.5 | 71.6 | 71.5 |

| Method | Iter. | Rouge-L | | Bleu-1 | | Bleu-4 | | Meteor | |
|----------------------------|-------|-------------|-------------|-------------|-------------|------------|------------|-------------|-------------|
| | | dev. | test | dev. | test | dev. | test | dev. | test |
| Best result original paper | - | 14.5 | 14.0 | 20.0 | 19.1 | 2.23 | 2.1 | 4.6 | 4.4 |
| DPR + FiD | - | 29.7 | 30.8 | 33.0 | 34.0 | 6.7 | 6.9 | 10.3 | 10.8 |
| Ours starting from BM25 | 0 | 29.9 | 30.3 | 34.6 | 33.7 | 7.1 | 6.5 | 10.5 | 10.4 |
| Ours starting from BM25 | 1 | 31.6 | 32.0 | 34.9 | 35.3 | 7.6 | 7.5 | 11.0 | 11.1 |

Izacard & Grave, 2020: Distilling Knowledge from Reader to Retriever for Question Answering

Conclusion

- Seq2seq models offer a flexible framework to aggregate passages
- Reading a lot of passages can significantly improve performance
- Reader cross-attention scores is a good proxy to train a retriever without strong supervision

Thank you!

References

Roberts, A., Raffel, C., & Shazeer, N. (2020). How Much Knowledge Can You Pack Into the Parameters of a Language Model? *ArXiv*, *abs/2002.08910*.

Min, S., Michael, J., Hajishirzi, H., & Zettlemoyer, L. (2020). AmbigQA: Answering Ambiguous Open-domain Questions. *ArXiv*, *abs/2004.10645*.

Kwiatkowski, Tom et al. “Natural Questions: A Benchmark for Question Answering Research.” *Transactions of the Association for Computational Linguistics* 7 (2019): 453-466.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv*, *abs/2005.11401*.

Joshi, M., Choi, E., Weld, D.S., & Zettlemoyer, L. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *ACL*.

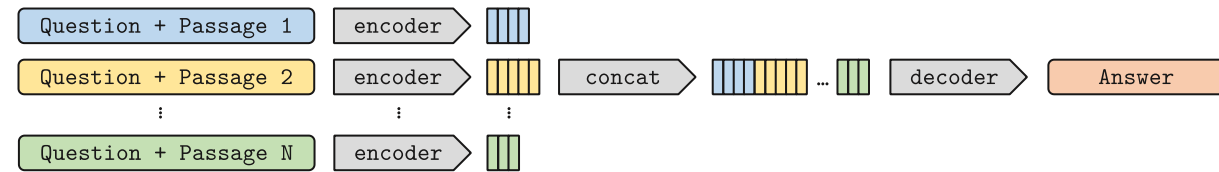
Brown, Tom B. et al. “Language Models are Few-Shot Learners.” *ArXiv* *abs/2005.14165* (2020)

Karpukhin, V., Ouguz, B., Min, S., Lewis, P., Wu, L.Y., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv*, *abs/2004.04906*.

Dense Bi-encoder for passage retrieval

- We would like to rank passages according to the cross-attention scores
 - Impossible \rightarrow the passages and the question need to be processed simultaneously
- Train a retriever to estimate a gold score $(G_{q,p})_{q \in Q, p \in D_q}$ obtained with the cross-attention scores
 - $S_\theta(q, p) = E(q)^T E(p)$
 - Target: $(G_{q,p})_{q \in Q, p \in D_q}$
 - for each passage: mean cross-attention over tokens/layers/heads
 - Loss: KL-divergence
- Knowledge distillation
 - Student = retriever
 - Teacher = Reader cross attentions

Cross-attention mechanism



Input: \mathbf{H} representation of the previous decoder layer,
 \mathbf{X} concatenated output of the encoder

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{H}, \quad \mathbf{K} = \mathbf{W}_K \mathbf{X}, \quad \mathbf{V} = \mathbf{W}_V \mathbf{X}.$$

Output of the cross-attention: $\mathbf{o}_i = \mathbf{w}_o \sum_j \tilde{\alpha}_{i,j} \mathbf{v}_{i,j}$

with $\alpha_{i,j} = \mathbf{Q}_i^T \mathbf{K}_j, \quad \tilde{\alpha}_{i,j} = \frac{\exp(\alpha_{i,j})}{\sum_m \exp(\alpha_{i,m})}$

$\alpha_{:,j}$ Measures the importance of j -th key and value to compute the next representation

Hypothesis: the more the tokens in a text segment are attended to, the more relevant the text segment is to answer the question.