Question Answering

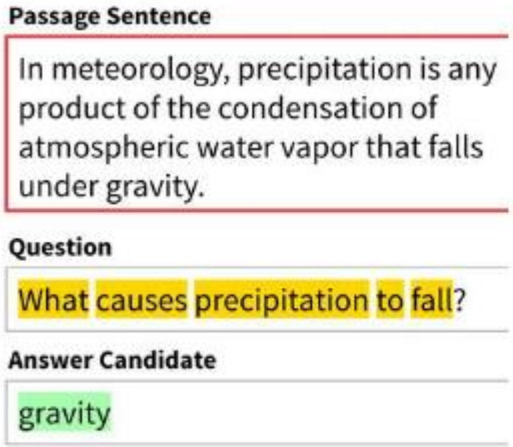
Prof. Luca Cagliero
Dipartimento di Automatica e Informatica
Politecnico di Torino

Outline

- Problem statement
- Common datasets
- Evaluation metrics
- State-of-the-art models

Question Answering

- Question Answering (QA) is the task of answering questions (typically reading comprehension questions) while abstaining when presented with a question that cannot be answered based on the provided context.



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

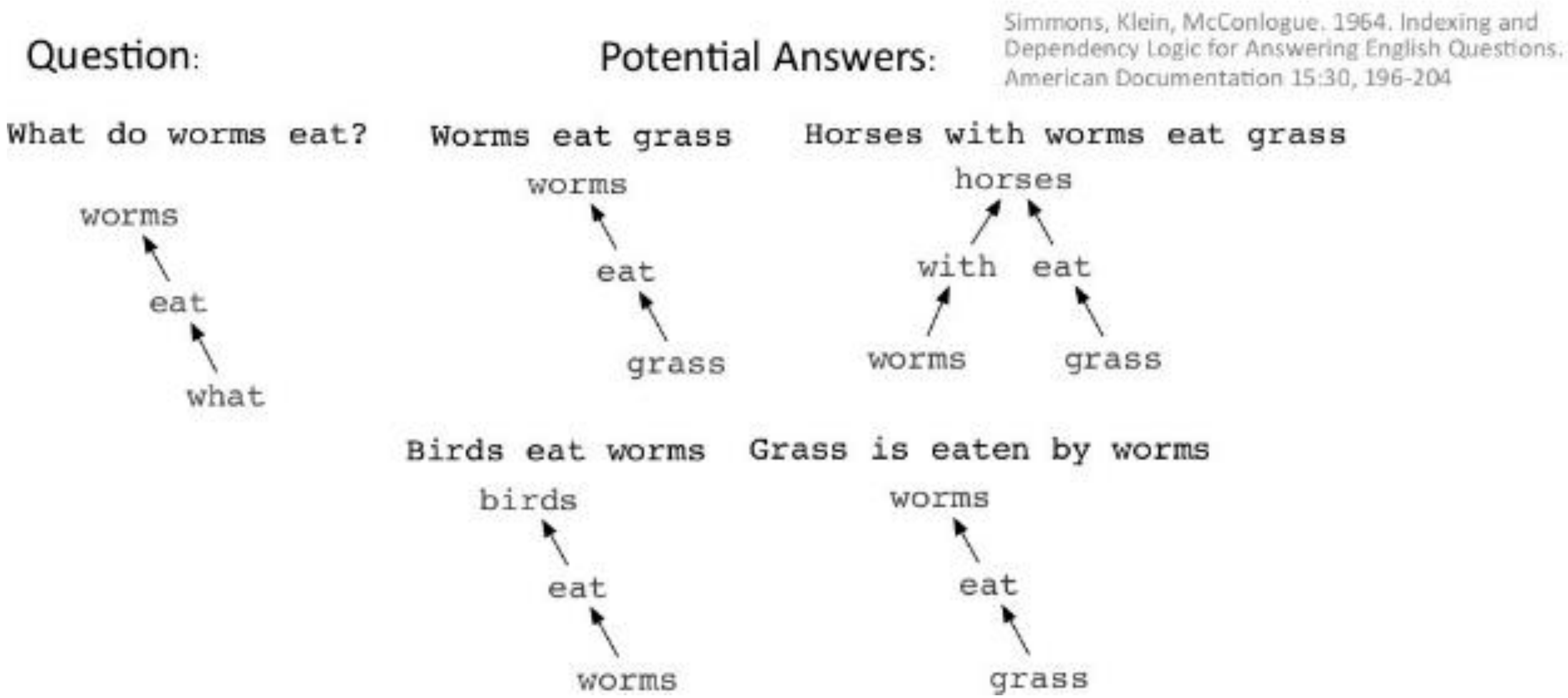
Sentiment analysis

- Wide range of question types
  - Facts
  - Lists
  - Definitions
  - How
  - Why
  - ...

Sentiment analysis

- Wide range of answer types
    - A span
        - an extracted portion of the input text
    - Generated answer
        - the answer is generated by a seq2seq model
    - A choice
        - the model selects an option among a set of candidate answers

QA: old-fashion example



Question:

Potential Answers:

Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

What do worms eat?

Worms eat grass

Horses with worms eat grass

Birds eat worms

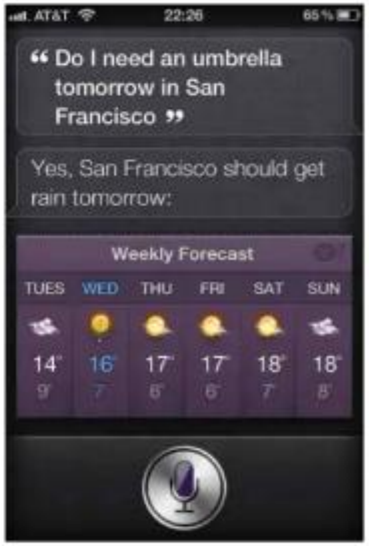
Grass is eaten by worms

Deep NLP for QA

- Usage of linguistic intuitions and machine learning methods to extract answers from retrieved snippet.

QA example: Virtual Assistant

Apple’s Siri



“Do I need an umbrella tomorrow in San Francisco”

Yes, San Francisco should get rain tomorrow:

Weekly Forecast

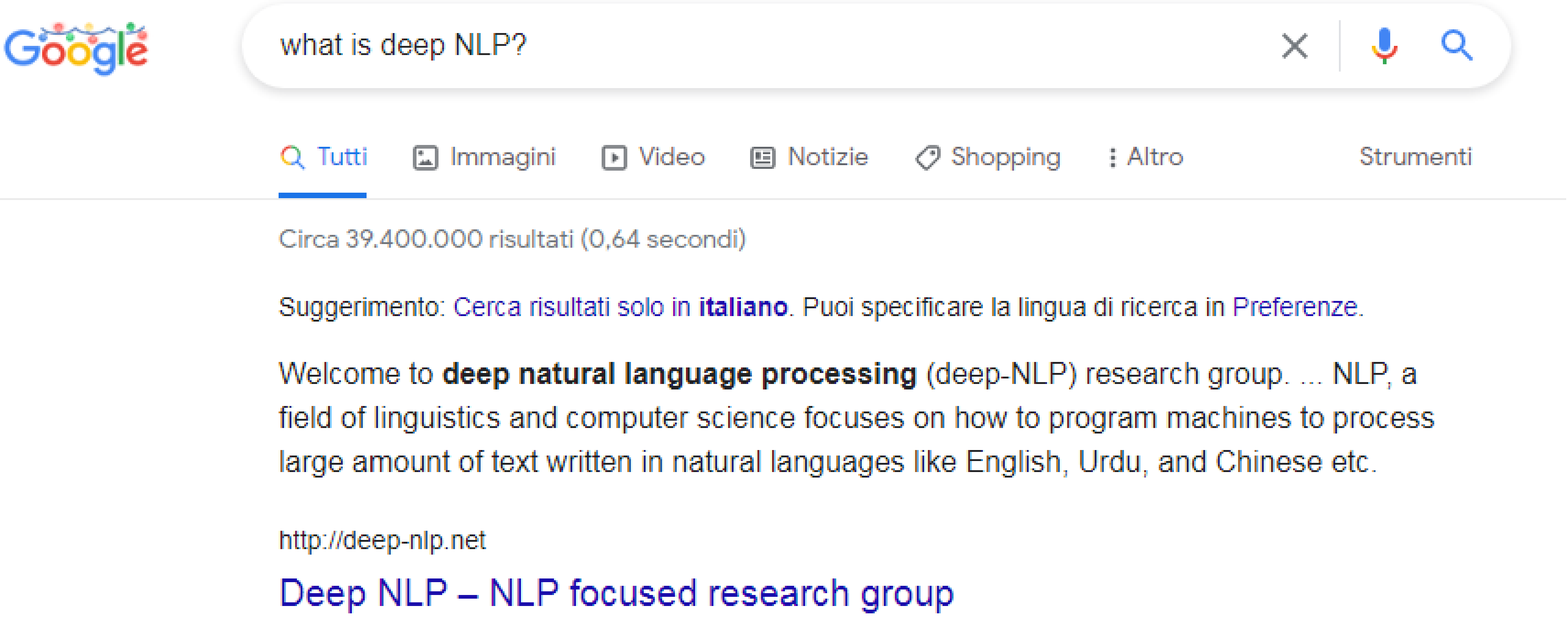
TUES	WED	THU	FRI	SAT	SUN
14°	16°	17°	17°	18°	18°
9°	7°	6°	6°	7°	8°

ChatGPT for QA

- Applications
  - Draft documents given specifications
  - Write computer code given requirements
  - Answer questions about a knowledge base
  - Give software a natural language interface
  - Tutor in a range of subjects

https://platform.openai.com/docs/guides/text-generation

QA example: Search Engine



what is deep NLP?

Circa 39.400.000 risultati (0,64 secondi)

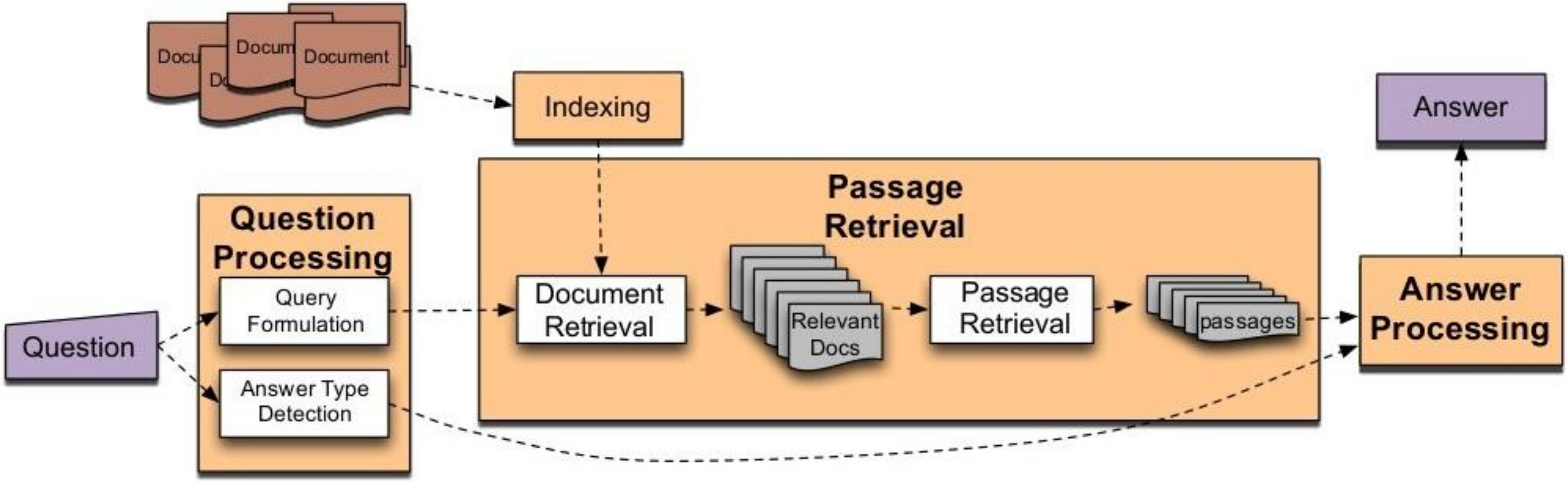
Suggerimento: Cerca risultati solo in **italiano**. Puoi specificare la lingua di ricerca in Preferenze.

Welcome to **deep natural language processing** (deep-NLP) research group. ... NLP, a field of linguistics and computer science focuses on how to program machines to process large amount of text written in natural languages like English, Urdu, and Chinese etc.

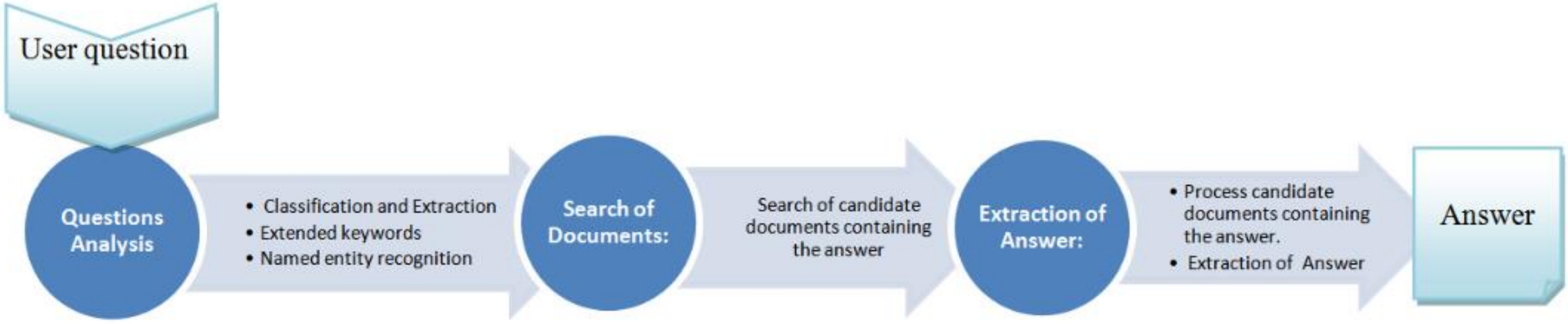
http://deep-nlp.net

Deep NLP – NLP focused research group

Classical QA pipeline

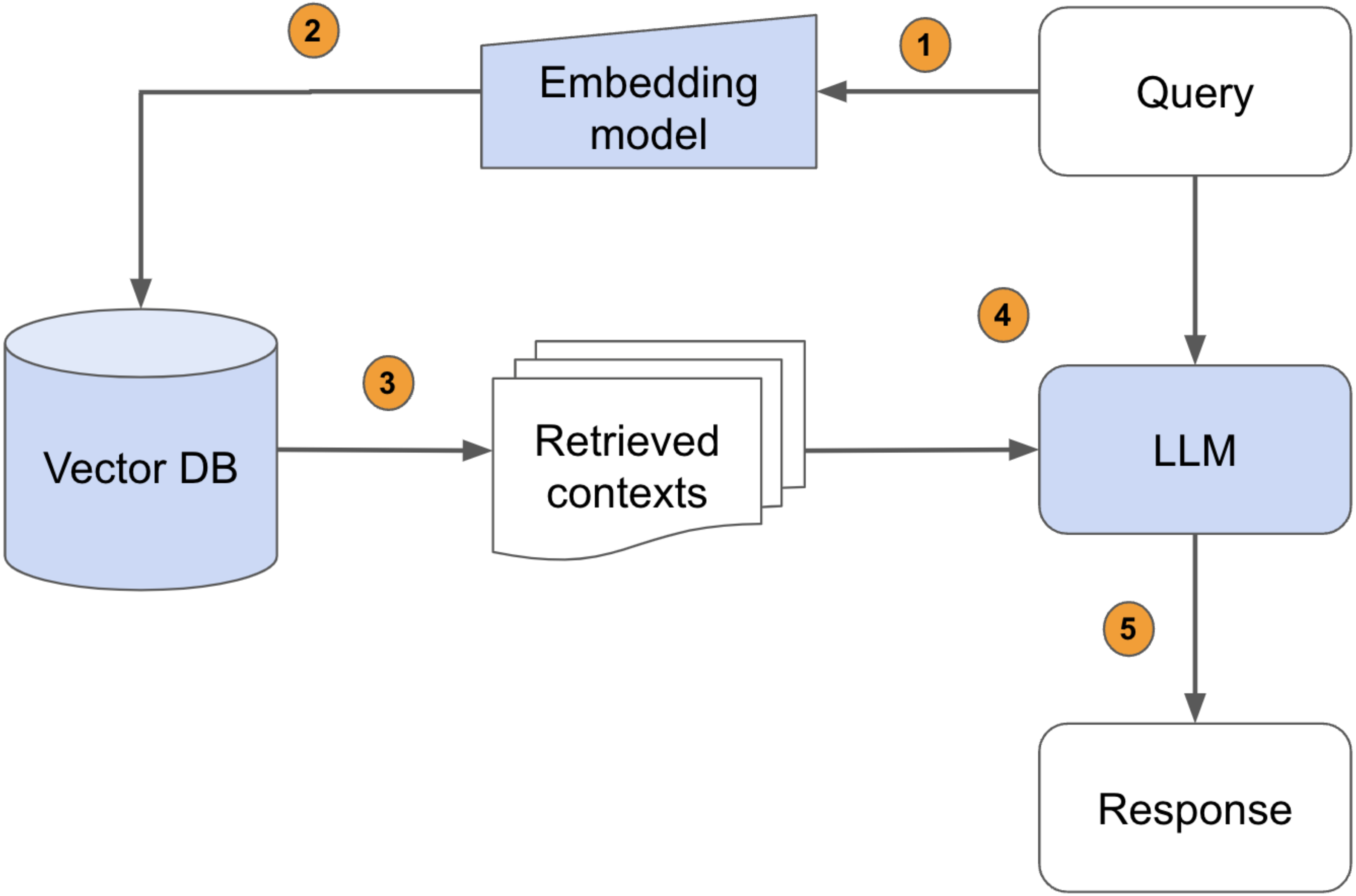


Classical QA pipeline



User question
Questions Analysis
• Classification and Extraction
• Extended keywords
• Named entity recognition
Search of Documents:
Search of candidate documents containing the answer
Extraction of Answer:
• Process candidate documents containing the answer.
• Extraction of Answer
Answer

RAG pipeline



https://www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1

RAG pipeline

1. Pass the query to the embedding model to semantically represent it as an embedded query vector.
2. Pass the embedded query vector to our vector DB.
3. Retrieve the top-k relevant contexts – measured by distance between the query embedding and all the embedded chunks in our knowledge base.
4. Pass the query text and retrieved context text to our LLM.
5. The LLM will generate a response using the provided content.

LlamaIndex: a RAG-based approach

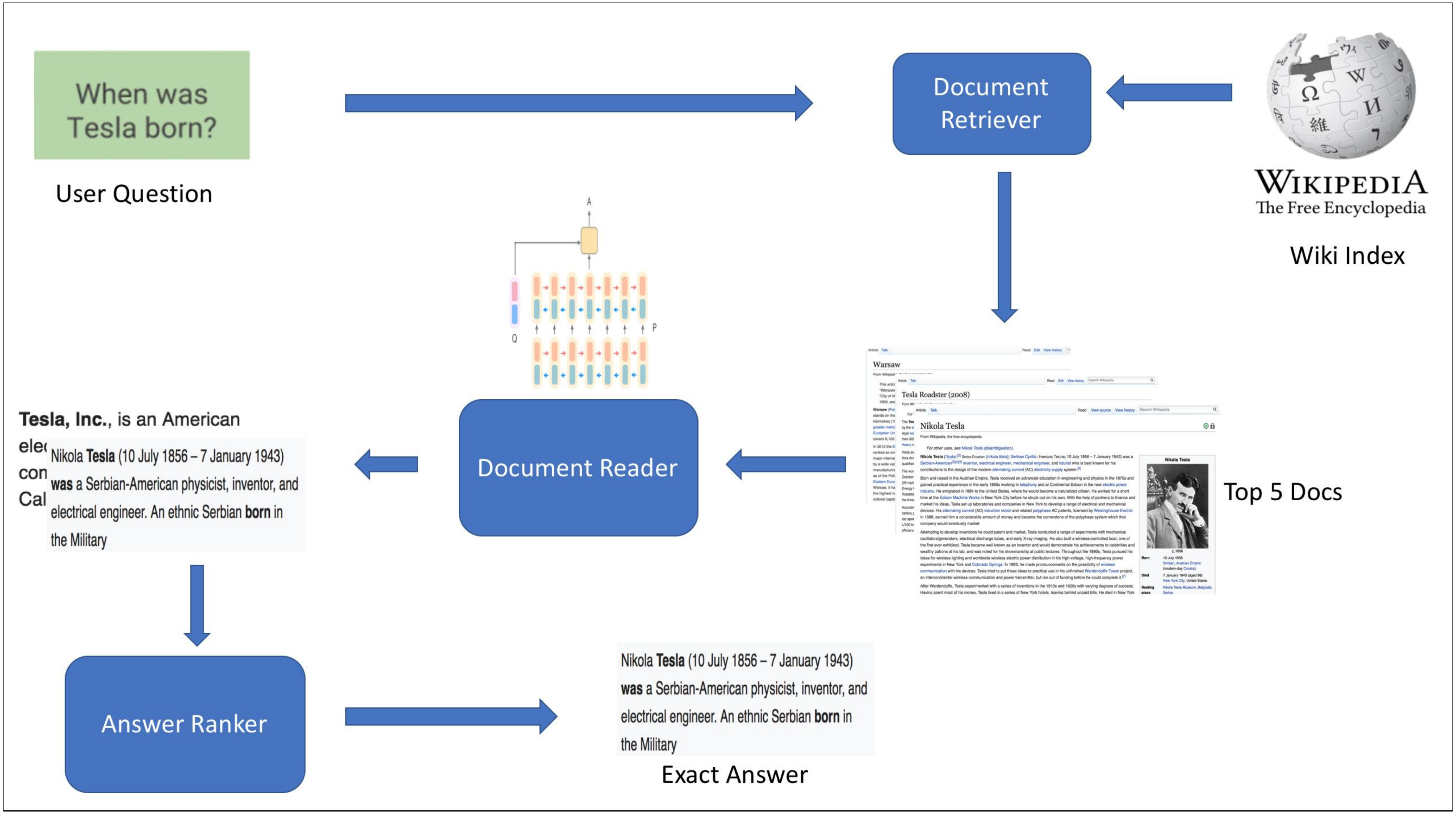
- Instead of asking LLM to generate an answer immediately, LlamaIndex:
  - retrieves information from your data sources first,
  - adds it to your question as context, and
  - asks the LLM to answer based on the enriched prompt.
- RAG overcomes the main weaknesses of the fine-tuning approach:
  - There’s no training involved, so it’s cheap.
  - Data is fetched only when you ask for them, so it’s always up to date.
  - LlamaIndex can show you the retrieved documents, so it’s more trustworthy.

https://docs.llamaindex.ai/en/stable/

Problem statement

- Closed-domain QA
  - Questions and answers are referred to specific domains and thus exploit domain-specific knowledge
- Open-domain QA
  - Questions and answers are referred to general knowledge

Example of Open-Domain QA system



Nikola **Tesla** (10 July 1856 – 7 January 1943) **was** a Serbian-American physicist, inventor, and electrical engineer. An ethnic Serbian **born** in the Military

Nikola **Tesla** (10 July 1856 – 7 January 1943) **was** a Serbian-American physicist, inventor, and electrical engineer. An ethnic Serbian **born** in the Military

Terminology

- Evidence/context documents/passages
  - supporting documents for answering a question
- Reasoning
  - the ability to logically analyze multiple references
  - single-hop reasoning
    - when this process requires one step of reasoning
  - multi-hop reasoning
    - requires models to gather information from different parts of a text to answer a question
    - It includes
      - word matching
      - paraphrasing
      - synthesis
      - Inference
      - ambiguity

Terminology (2)

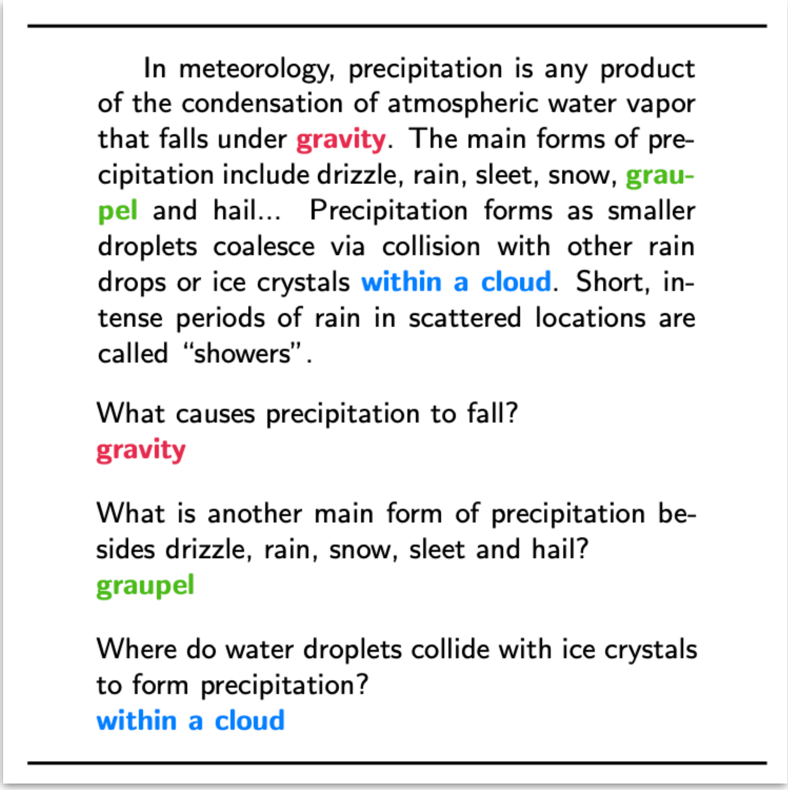
- Question
  - sentence that expresses what information is searched
  - characterized by a type (the purpose of the question) and a focus (the entity)
- Answer
  - defined by a type
    - a class of objects which are sought by the question
- Distant supervision
  - the system only knows what the answer is
    - do not know what supporting facts lead to it

Datasets: SQuAD

- SQuAD1.1-2.0 (The Stanford Question Answering Dataset) [1]
    - reading comprehension dataset consisting of questions from crowdworkers on a set of Wikipedia articles.
- questions are designed to be answered given a single paragraph as the context, and most of the questions can in fact be answered by matching the question with a single sentence in that paragraph.
- The answer consists in a span of text
- Size (SQuAD2.0)
    - 151054 question-answer pairs on 505 articles.

Datasets: SQuAD

- Size (SQuAD2.0)
  - 151054 question-answer pairs on 505 articles.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

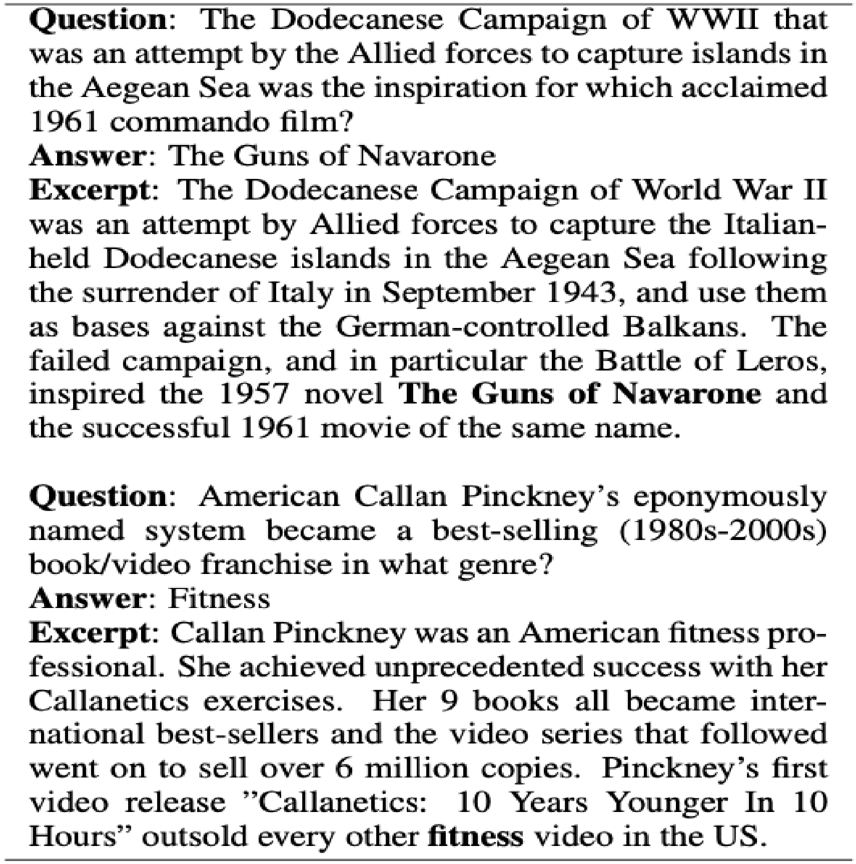
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Datasets: TriviaQA

- Open-domain dataset
- It includes various sources
- Wikipedia and more general Web search results
- The documents are not guaranteed to contain all facts needed to answer the question (only a source of distant supervision)
- For Web search it assumes that the documents that contain the correct answer are highly redundant, so, each question-answer-document tuple be an independent
  - whereas in Wikipedia most facts to be stated only once, so no questions are repeated

TriviaQA



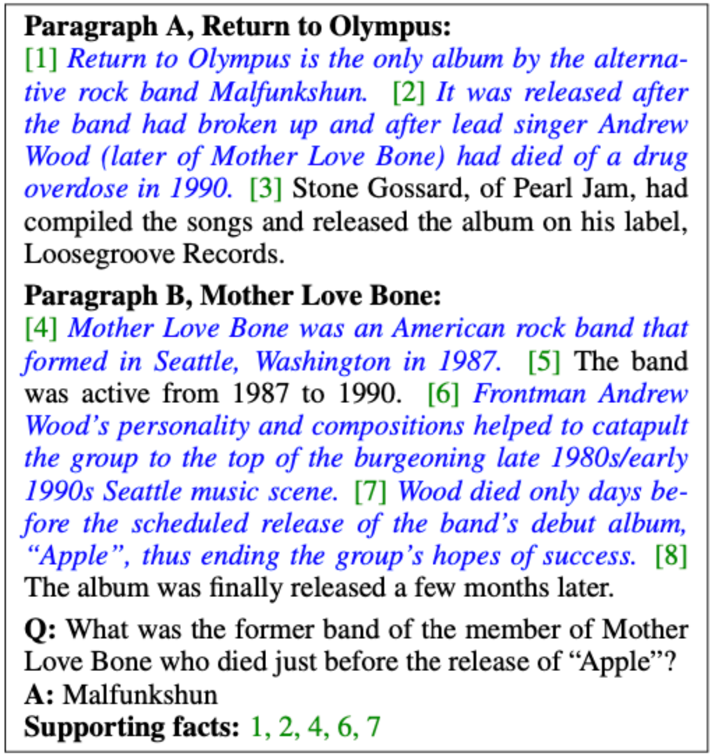
Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Datasets: HotpotQA

- Open-domain dataset
  - 113k Wikipedia-based question-answer pair
- It is a diverse and explainable question answering dataset
  - requires multi-hop reasoning, collected on the English Wikipedia
- The questions require finding and reasoning over multiple supporting documents to answer
- It introduce supporting facts
  - portions of text that “support” the reasoning in multiple documents
  - overcome the limitations of distant supervision this way
  - the models can make explainable predictions

HotpotQA



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

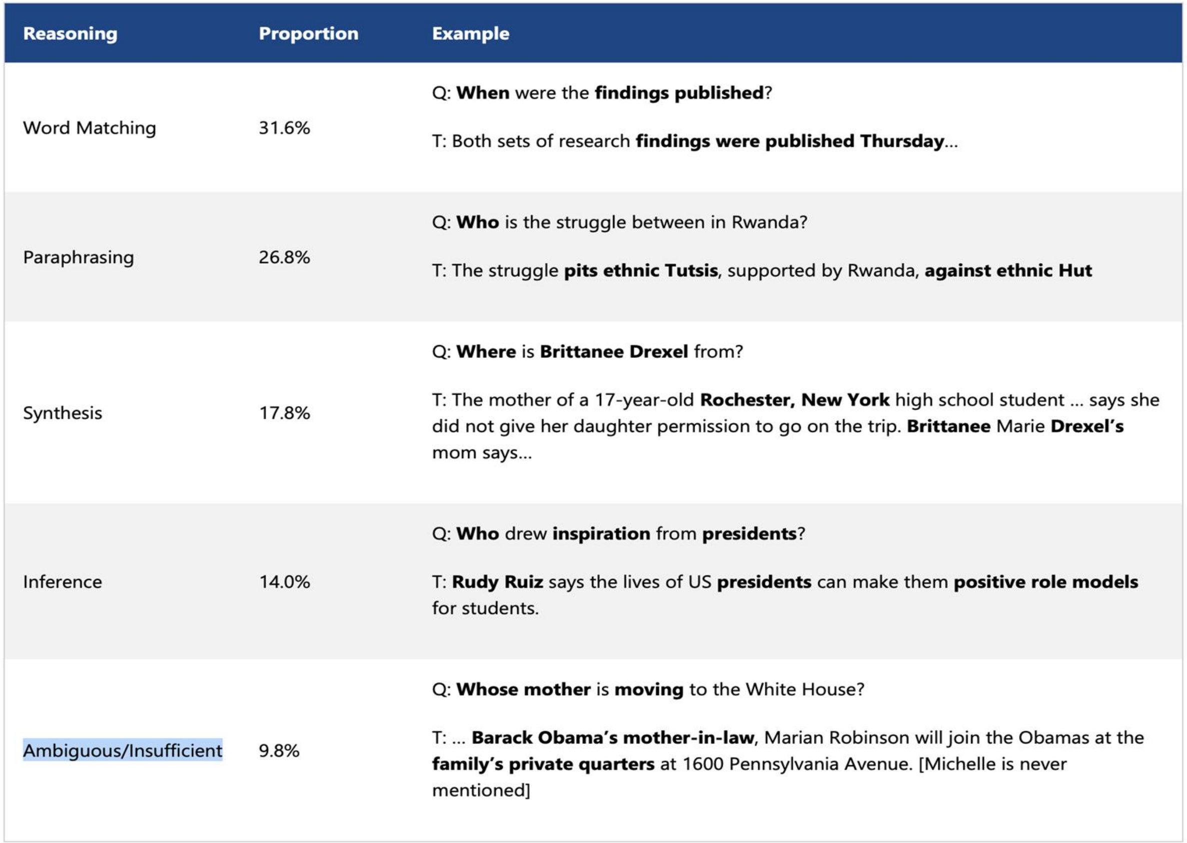
Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s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

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

Datasets: NewsQA

- It contains questions and answers generated by crowdworkers on CNN news articles.
  - Answers are spans of arbitrary length within an article, rather than single words or entities
  - Some questions have no answer in the corresponding article
- It encourages lexical and syntactic divergence between questions and answer
  - More challenging
- The authors assert that a significant proportion of questions requires reasoning beyond simple word-matching
- It contains 119,633 questions on 12,744 news

NewsQA



Reasoning	Proportion	Example
Word Matching	31.6%	Q: **When** were the **findings published**? T: Both sets of research **findings were published Thursday**...
Paraphrasing	26.8%	Q: **Who** is the struggle between in Rwanda? T: The struggle **pits ethnic Tutsis**, supported by Rwanda, **against ethnic Hut**
Synthesis	17.8%	Q: **Where** is **Brittanee Drexel** from? T: The mother of a 17-year-old **Rochester, New York** high school student ... says she did not give her daughter permission to go on the trip. **Brittanee** Marie **Drexel's** mom says...
Inference	14.0%	Q: **Who** drew **inspiration** from **presidents**? T: **Rudy Ruiz** says the lives of US **presidents** can make them **positive role models** for students.
Ambiguous/Insufficient	9.8%	Q: **Whose mother** is **moving** to the White House? T: ... **Barack Obama's mother-in-law**, Marian Robinson will join the Obamas at the **family's private quarters** at 1600 Pennsylvania Avenue. [Michelle is never mentioned]

Other datasets

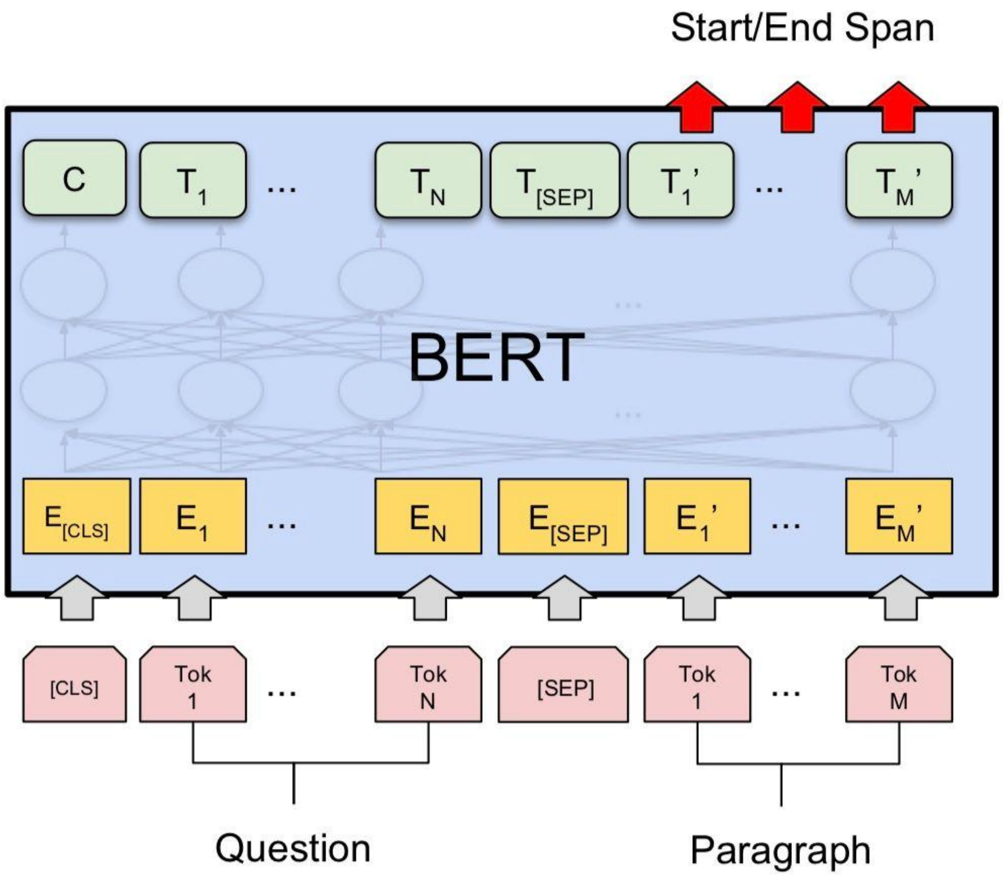
- TREC-QA
  - QA dataset with multiple-choice questions
    - up to 5 responses per question
- bAbI
  - textual QA benchmark composed of 20 different tasks
  - Each task is designed to test a different reasoning skill, such as deduction, induction, and coreference resolution
- WikiQA
  - set of question and sentence pairs, collected and annotated for research on open-domain question answering

N.B. The main competition is on SQuaD dataset, and typically the other dataset are employed in the training phase to increase the robustness of QA models.

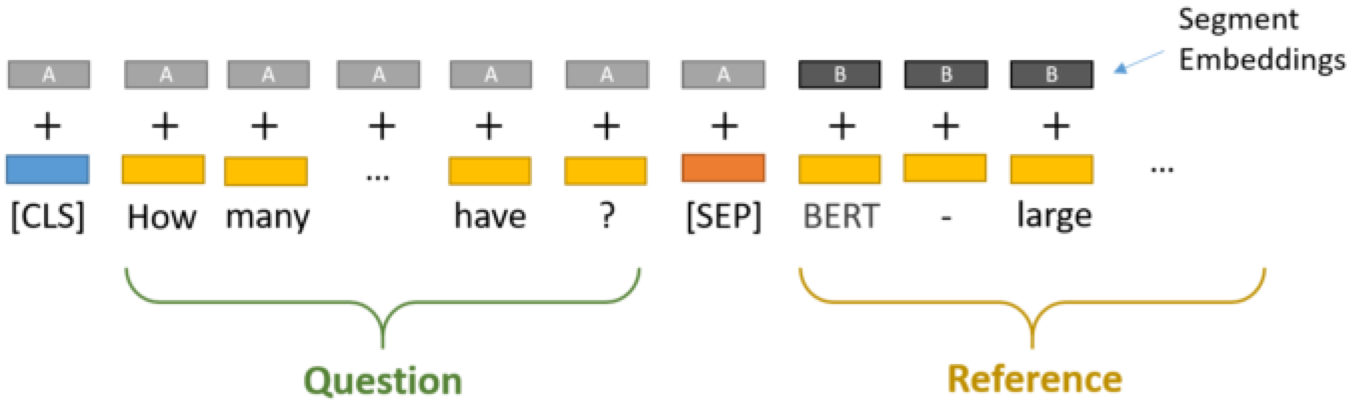
Evaluation metrics

- Exact match
  - percentage of predictions that match any one of the ground truth answers exactly
  - The exact match metric is a binary value that takes value 1 if the predicted answer and true answer are exactly equal (not counting punctuation and articles), zero otherwise.
- F1-Score
  - computed over the individual words in the prediction against those in the ground truth
  - Based on
    - Precision P: defined as the ratio between the length of the common subsequence and the predicted answer length
    - Recall R: defined as the ratio between the length of the common subsequence and the true answer length

BERT for QA


Start/End Span
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Question
Paragraph

BERT for QA



Question: How many parameters does BERT-large have?

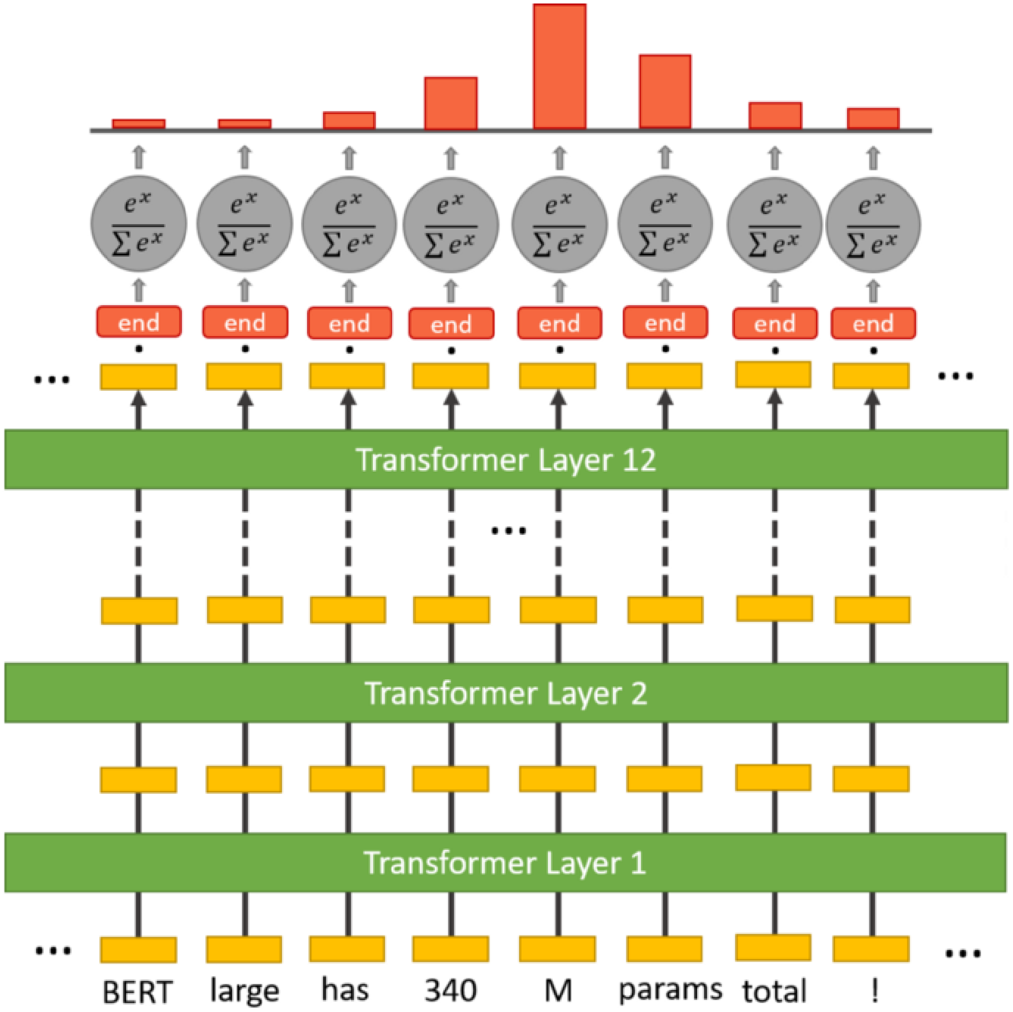
Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

BERT for QA

- Inputs
  - Token embeddings
    - a [CLS] token is added to the input word tokens at the beginning of the question
    - a [SEP] token is inserted at the end of both the question and the paragraph
  - Segment embeddings
    - A marker indicating Sentence A or Sentence B is added to each token to distinguish between sentences

https://medium.com/analytics-vidhya/question-answering-system-with-bert-ebe1130f8def

BERT for QA



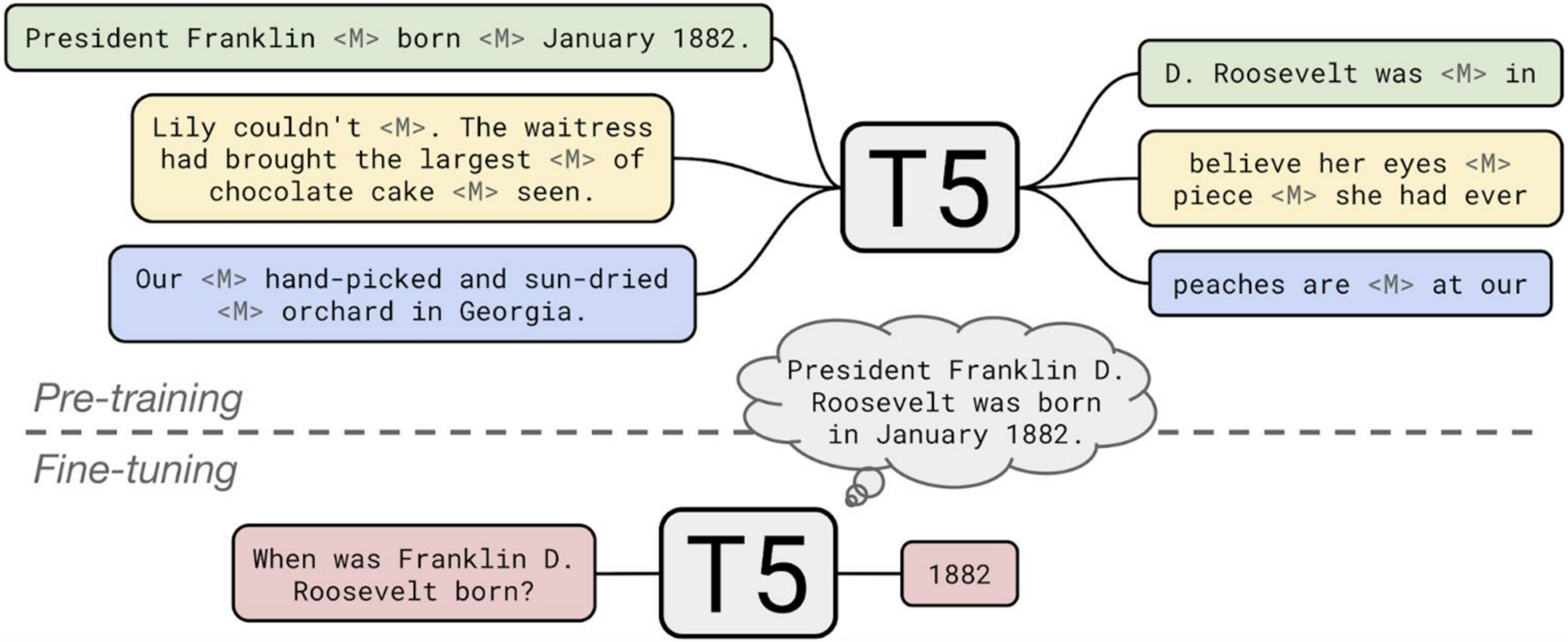
https://medium.com/analytics-vidhya/question-answering-system-with-bert-ebe1130f8def

BERT for QA

- Fine-tuning
  - For every token in the text, its final embedding is fed into the start token classifier
  - The start token classifier only has a single set of weights which applies to every word
  - Softmax is applied to the dot product between the output embeddings and the ‘start’ weights, producing a probability distribution over all of the words.
  - The word with the maximal probability of being the start token is returned.

https://medium.com/analytics-vidhya/question-answering-system-with-bert-ebe1130f8def

T5 for Generative QA

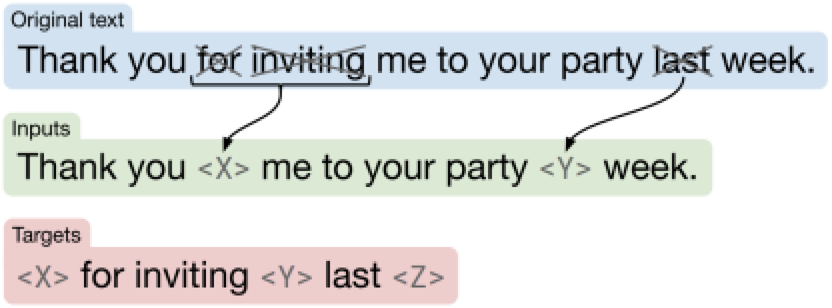


President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
President Franklin D. Roosevelt was born in January 1882.
Pre-training
Fine-tuning
When was Franklin D. Roosevelt born?
T5
1882

https://medium.com/analytics-vidhya/question-answering-system-with-bert-ebe1130f8def

T5 for Generative QA

- Pretraining
  - Similar denoising scheme to BART
- Input
  - Text with gaps
- Output
  - A series of phrases to fill those gaps
- Encoder
  - Question \n Passage
- Decoder
  - Answer



Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

T5 for Generative QA

- Text-to-Text Transfer Transformer (T5): key idea
  - treat every text processing problem as a “text-to-text” problem
    - i.e. taking text as input and producing new text as output
- T5 is based on transformer architecture
- For the QA task, the model is fed the question and its context and asked to generate the answer: it is trained with a maximum likelihood objective (cross-entropy loss) using “teacher forcing”
  - method for quickly and efficiently training recurrent neural network models that use the ground truth from a prior time step as input
- The authors verify the impact of different architectures (decoder-encoder vs decoder only), different pre-training objectives and training strategies, different model scaling

T-5 pre-training and fine-tuning example for QA task from https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

Acknowledgements and copyright license

- Copyright licence
  - Attribution + Noncommercial + NoDerivatives
- Acknowledgements
  - I would like to thank Dr. Moreno La Quatra, who collaborated to the writing and revision of the teaching content
- Affiliation
  - The author and his staff are currently members of the Database and Data Mining Group at Dipartimento di Automatica e Informatica (Politecnico di Torino) and of the SmartData interdepartmental centre
    - https://dbdmg.polito.it
    - https://smartdata.polito.it

Thank you!