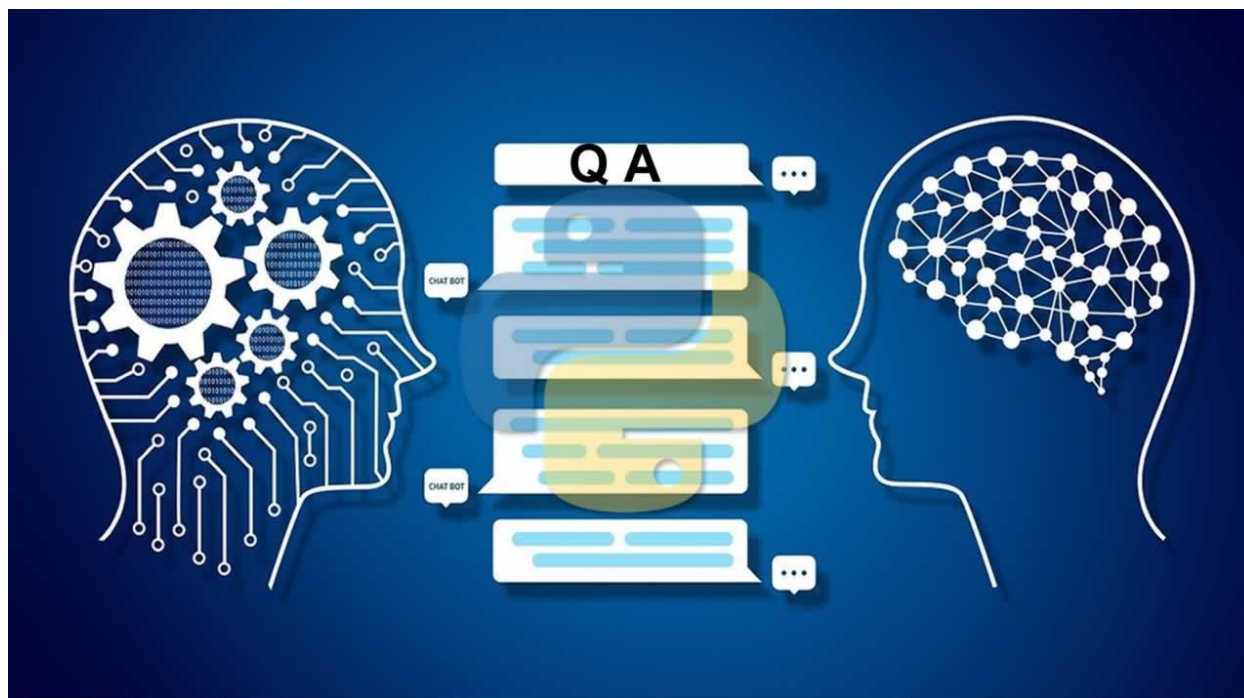


Automatic summarization and question answering



Prof Dr Marko Robnik-Šikonja

Natural Language Processing, Edition 2026

Contents

- summarization
- question answering

- Literature:

Jurafsky and Martin, 3rd edition

Some slides taken from Jurafsky

Text summarization

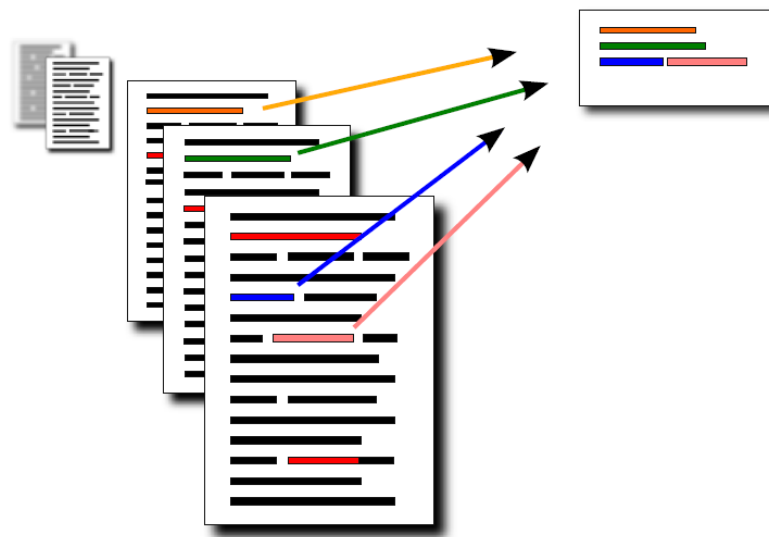
“It's not information overload. It's filter failure.”

Clay Shirky



Text summarization

- The goal of automatic text summarization is to automatically produce a succinct summary, preserving the most important information for a single document or a set of documents about the same topic (event).
- Neural text summarization uses the same seq2seq technology as MT.
- What are the differences and challenges?



Abstract, outline, headline

- An abstract is a concise summary placed at the beginning of a document, providing an overview of the main points and conclusions.
- An outline, on the other hand, is a structural plan that organizes the content of a document, outlining the main ideas and their hierarchical relationships.
- A headline is a brief, attention-grabbing statement or title that is typically used in journalism, advertising, or online content to capture the reader's interest and provide a concise summary of the main idea

Summarization applications

- outlines or abstracts or headlines of any document, article, etc
- summaries of email threads
- summaries of web commentaries
- action items from a meeting
- simplifying text by compressing sentences
- summarization for certain purpose, e.g., for certain customer, or from certain point of view
- generating headlines to replace click-bait headlines with informative ones

Single-Document Summarization (SDS)

Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize " the political crisis .

Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .

Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``

I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , " Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .

It is detrimental to the sovereignty of Cambodia , " he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed

Summary



Cambodian government rejects opposition's call for talks abroad

Multiple-Document Summarization (MDS)

Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American.

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said.

⋮

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 56 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday.

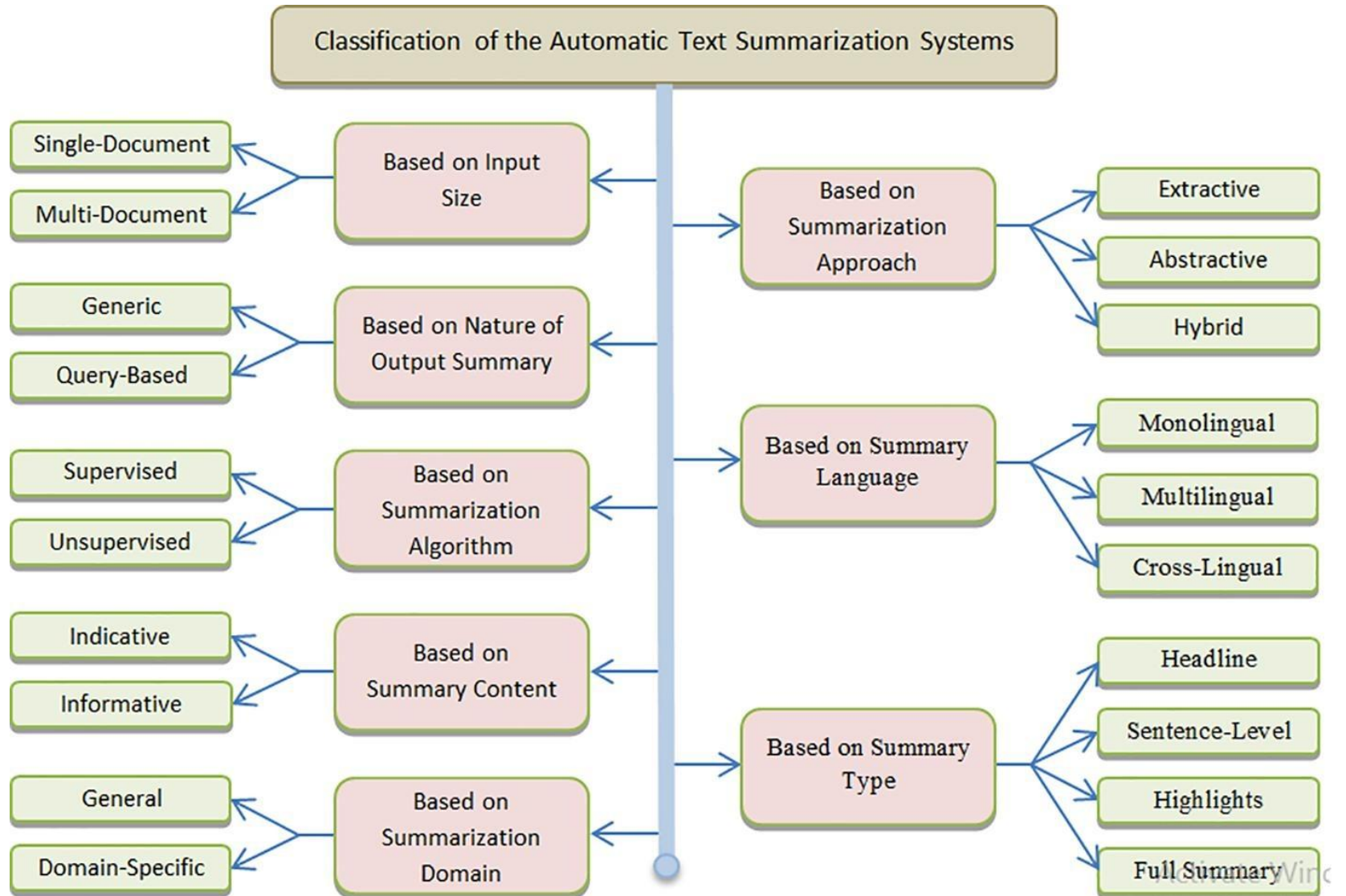
Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining the "true identities" of the passengers who used the stolen passports. Investigators were trying to determine the path of the plane by analysing civilian and military radar data while ships and aircraft from seven countries scouring the seas around Malaysia and south of Vietnam.

Text summarization categorization

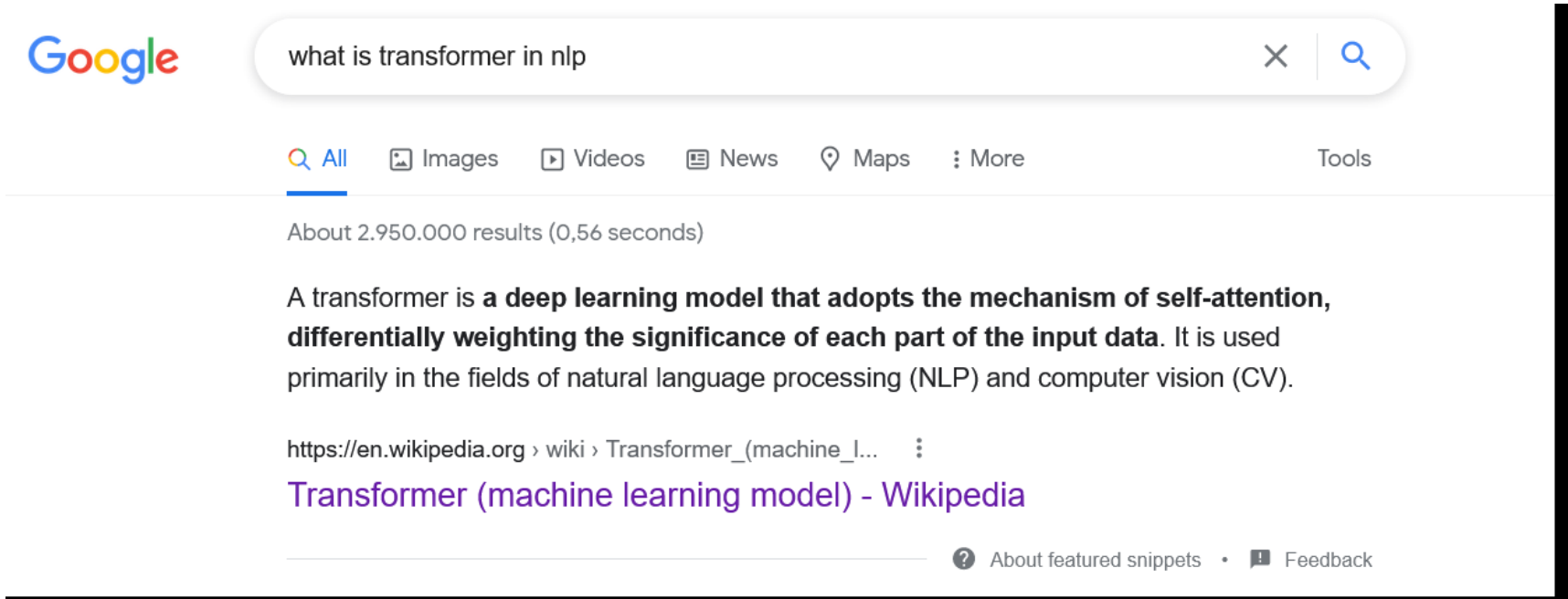
- Input:
 - Single-Document Summarization (SDS)
 - Multi-Document Summarization (MDS)
- Output:
 - Extractive:
 - The generated summary is a selection of relevant sentences from the source text in a copy-paste fashion (problem: redundancy).
 - Compressive (outdated):
 - Summary is constructed from compressed sentences, typically based on the dependency-trees (preserves original dependency relations) and extraction of some rooted subtrees; each subtree corresponds to a compressed sentence.
 - Abstractive:
 - The generated summary is a new cohesive text not necessarily present in the original source.
- Focus
 - Generic
 - Query based (also targeted or aspect-based)
 - Summarize a document with respect to an information need expressed in a user query.
 - A kind of complex question answering: Answer a question by summarizing a document that has the information to construct the answer
- Machine learning methods:
 - Supervised
 - Unsupervised

Text summarization categorization



Summarization for Question Answering: Snippets

- Create snippets summarizing a web page for a query
- Google: a short answer and link



The screenshot shows a Google search interface. The search bar contains the text "what is transformer in nlp". Below the search bar, there are tabs for "All", "Images", "Videos", "News", "Maps", and "More". The "All" tab is selected. Below the tabs, it says "About 2.950.000 results (0,56 seconds)". The main content area displays a snippet from Wikipedia: "A transformer is a **deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.** It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).". Below this snippet, there is a link: "https://en.wikipedia.org › wiki › Transformer_(machine_l...". The link is followed by the text "Transformer (machine learning model) - Wikipedia". At the bottom of the snippet, there are two links: "About featured snippets" and "Feedback".

Google

what is transformer in nlp

× | 🔍

🔍 All 🖼️ Images 📺 Videos 📰 News 📍 Maps ⋮ More Tools

About 2.950.000 results (0,56 seconds)

A transformer is a **deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.** It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).

[https://en.wikipedia.org › wiki › Transformer_\(machine_l...](https://en.wikipedia.org › wiki › Transformer_(machine_l...)

Transformer (machine learning model) - Wikipedia

🔍 About featured snippets • 🗉 Feedback

Summarization for Question Answering: Multiple documents

- **Create answers** to complex questions summarizing multiple documents.
 - Instead of giving a snippet for each document
 - Create a cohesive answer that combines information from each document

Extractive & Abstractive summarization

- Extractive summarization:
 - create the summary from phrases or sentences in the source document(s)
- Abstractive summarization:
 - express the ideas in the source documents using (at least in part) different words

Summarization: common datasets

- Within single-document summarization, there are datasets with source documents of different lengths and styles:
 - Gigaword: first one or two sentences of a news article → headline (aka *sentence compression*)
 - LCSTS (Chinese microblogging): paragraph → sentence summary
 - NYT, CNN/DailyMail: news article → (multi)sentence summary
 - Wikihow: full how-to article → summary sentences
 - XSum: (Narayan et al., 2018), Newsroom: (Grusky et al., 2018): article → 1 sentence summary
 - BookSum (Kryściński et al, 2021) novels, plays and stories; includes abstractive, human written summaries on three levels: paragraph-, chapter-, and book-level.
- Slovene: STA news, Wikipedia, KAS-abstracts (Slovene and English)
- (outdated) List of summarization datasets, papers, and codebases: <https://github.com/mathsyouth/awesome-text-summarization>

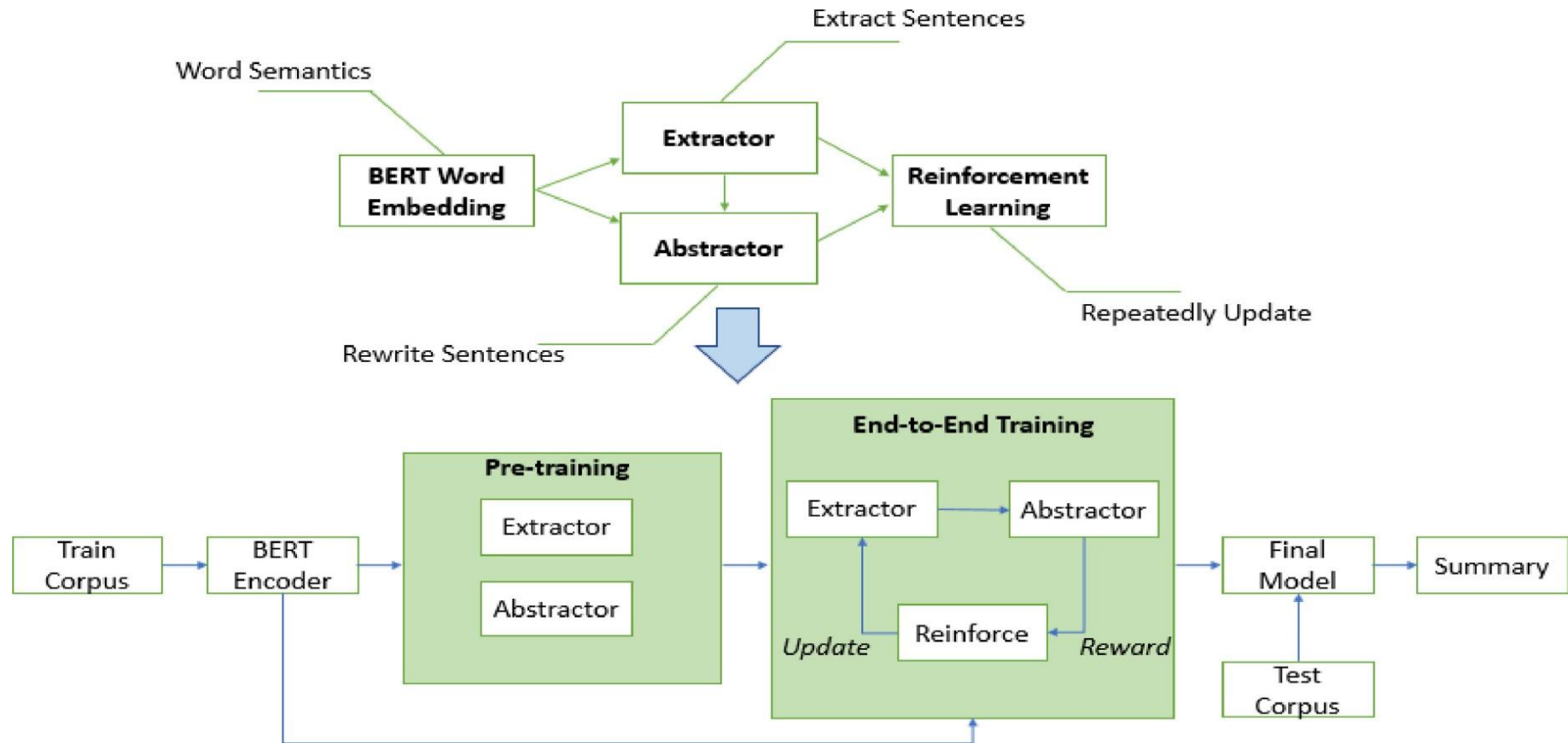
Sentence simplification: common datasets

- *Sentence simplification* is different but related task:
 - rewrite the source text in a simpler (sometimes shorter) way
 - Simple Wikipedia: standard Wikipedia sentence → simple version
 - Newsela: news article → four, increasingly simplified, versions

Neural summarization via Reinforcement Learning

- In 2017 Paulus et al published a “deep reinforced” summarization model
- Main idea: Use Reinforcement Learning (RL) to directly optimize ROUGE-L
- By contrast, standard maximum likelihood (ML) training can’t directly optimize ROUGE-L because it’s a non-differentiable function
- Interesting finding:
 - Using RL instead of ML achieved higher ROUGE scores, but lower human judgment scores

Combining extractive and abstractive summarization

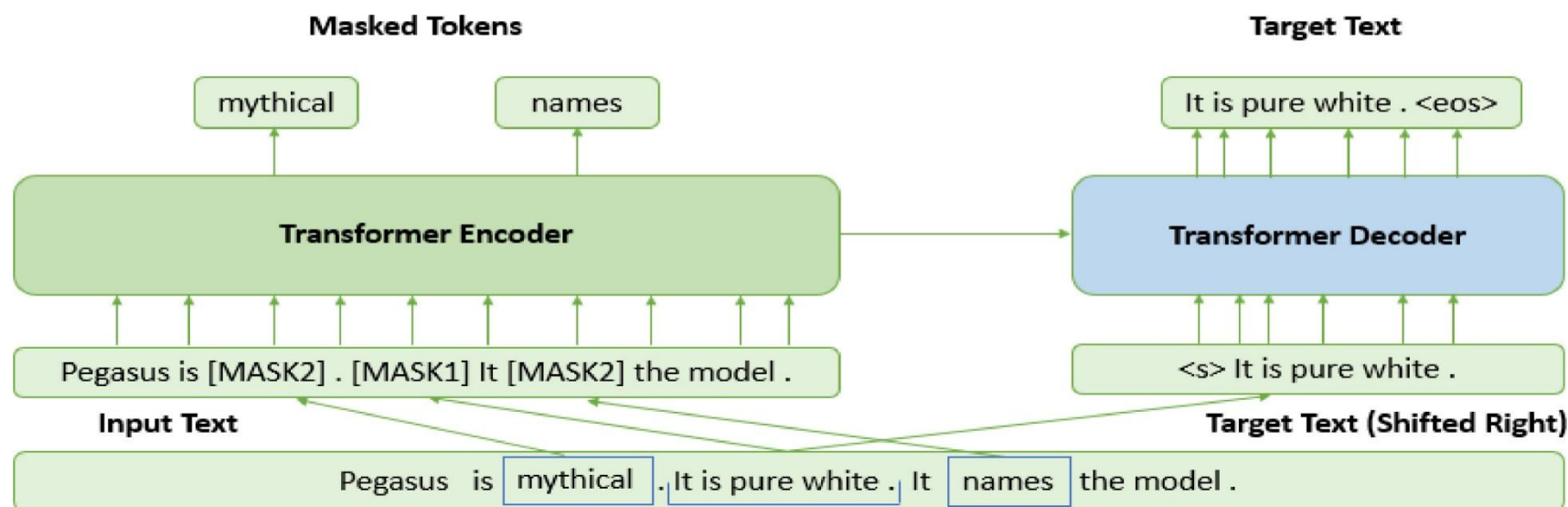


Wang et al, 2019. A text abstraction summary model based on BERT word embedding and reinforcement learning

Appl. Sci., 9 (2019), p. 4701, [10.3390/app9214701](https://doi.org/10.3390/app9214701)

Pegasus pretraining

- Transformer encoder-decoder model.
- Pre-trained on the objective of gap sentence generation.
- Masks important sentences from the input document to be generated as one output sequence from the remaining sentences.
- Chooses the sentences based on their importance (not randomly).
- Combines a masked language model and the gap sentence generation.



Cross-lingual summarization

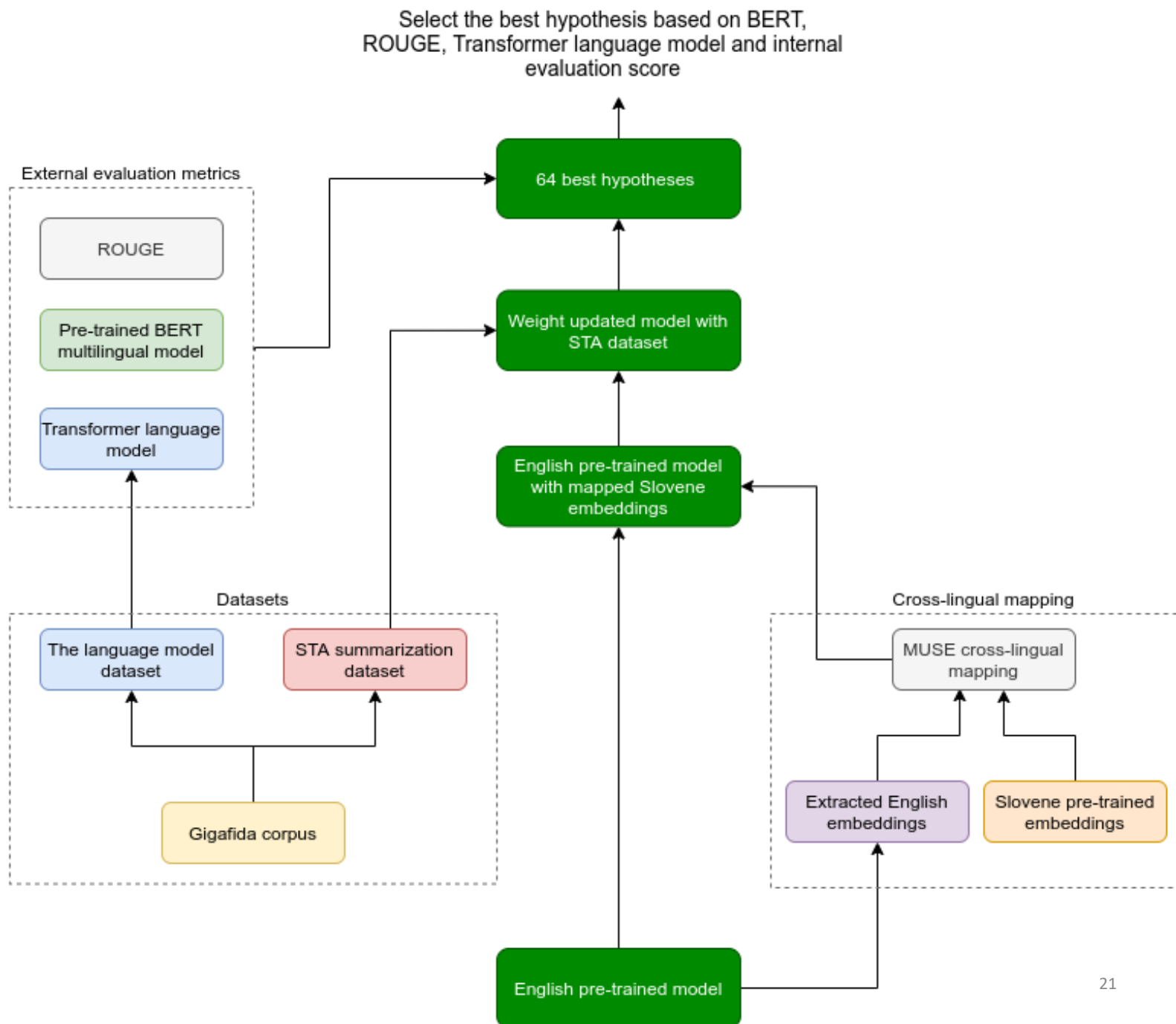
- Summarization into another language
 - E.g., Hindi paper is summarized into English
 - Basically the same approach
-
- Slovene dataset: KAS abstracts (Slovene theses, Slovene and English summaries)

Cross-lingual transfer of a summarizer

- Idea: use pretrained English model to summarize Slovene texts
- Two Slovene datasets
- STA news: 127,563 news with the first paragraph as a summary (length between 1,000 and 3,000 characters, no weather reports, no lists of events, etc.)
- Wikipedia corpus: 2,100 Slovene articles of sufficient length with the first paragraph as a summary

Žagar, A. and Robnik-Šikonja, M., 2022. Cross-lingual transfer of abstractive summarizer to less-resource language. *Journal of Intelligent Information Systems*, pp.1-21.

<https://arxiv.org/abs/2012.04307> .



Unsupervised summarization

- Mostly using sentence-based similarity measures to build a document graph
- Use graph centrality measures or node relevance measures such as PageRank
- Extract the most central sentences

Unsupervised Approach to Multilingual User Comments Summarization

Why?

- Readers are often interested in what others think

Problems

- A lot of irrelevant and deceiving comments
- Language is often informal and difficult to encode

Languages and Datasets

- Croatian (CroNews and CroComments)
- English (NYT Comments)
- German (DER STANDARD)

Methodology

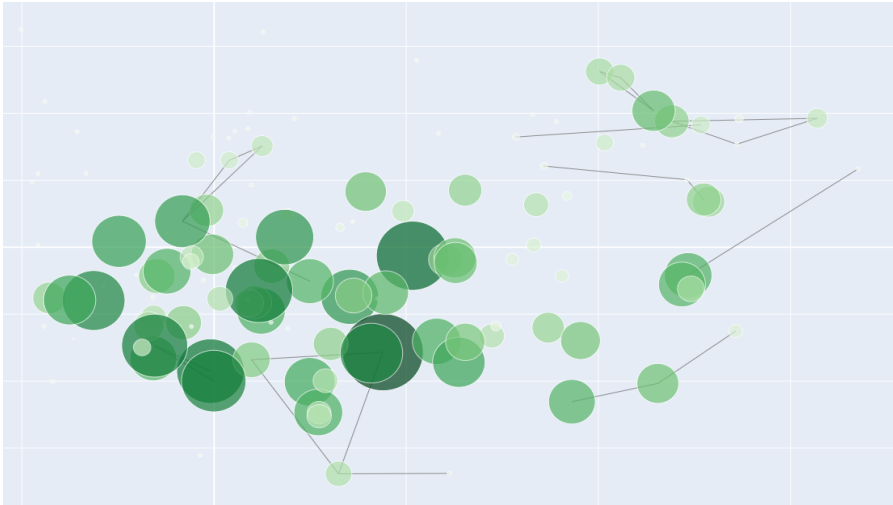
- Extractive approach based on graph-methods and clustering
- uses LaBSE sentence encoder

Aleš Žagar, Marko Robnik-Šikonja. (2021) Unsupervised Approach to Multilingual User Comments Summarization. [Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation](#)

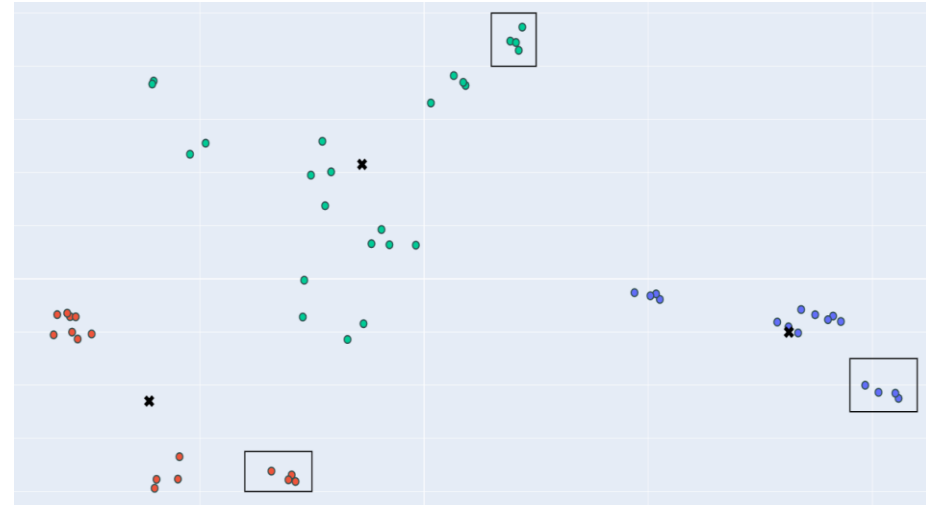
comment_id	text	score
22320520	Science is a necessary tool of understanding . Its use has proved to aid mankind in so many ways , but it should not politically . If climate change is real , then why does the USA need to be using millions of dollars to figure it out . The USA does not need to fund everything the world wants , that is the reason the US debt is so great . Let other countries spend their own money to figure it out , while we (the USA) focus on serious issues like the threat of nuclear war with North Korea , or illegal immigration .	0.017
22307395	We must share our scientific and knowledge bias . That is only the beginning . There is political and educational work to be done . But much of the population is against science , knowledge and education . In their minds , Obey Trumps Question .	0.017
22268176	Regarding the objections of Robert S. Young to the March on Science , I disagree heartily . First , scientists are now and have always been " caught up in the culture wars . " Simply reviewing the history of science and scientists , I ca n't imagine how any thoughtful person could see it otherwise . Second , " the wedge between scientists and a certain segment of the American electorate " could not possibly have been made deeper by the March for Science . Consider the people who disdain you and your work , Mr. Young . Do you think the March on Science might really change their attitudes one way or the other ? On the other hand , the March for Science and associated activism can help the nation as a whole better recognize that science matters to them - to their health , to their safety , to the storehouse of knowledge that their children and their children will inherit . Such activism also shines light on the fact that science and its benefits are under attack by the leaders of our current government . For my part , I briefly considered not traveling to Washington D.C. to march in the cold rain last Saturday , but then decided that doing so was a patriotic and moral duty . I marched .	0.018

Visual tools to investigate results

Graph-based



Clustering



Evaluation metric: ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE^a metrics compare an automatically produced summary against a reference or a set of references (human-produced) summary.

^aLin, Chin-Yew. ROUGE: a Package for Automatic Evaluation of Summaries. WAS 2004

- ROUGE-N: N-gram based co-occurrence statistics.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics.

$$ROUGE_N(X) = \frac{\sum_{S \in \{Ref \text{ Summaries}\}} \sum_{gram_n \in S} count_{match}(gram_n, X)}{\sum_{S \in \{Ref \text{ Summaries}\}} \sum_{gram_n \in S} count(gram_n)}$$

ROUGE

- Like BLEU, it's based on **n-gram overlap**.
- Differences:
 - ROUGE has no brevity penalty
 - ROUGE is based on **recall**, while BLEU is based on **precision**
 - Arguably, precision is more important for MT (then add brevity penalty to fix under-translation), and recall is more important for summarization (assuming you have a max length constraint)
 - However, often a F_1 (combination of precision and recall) version of ROUGE is reported anyway.
- BLEU is reported as a single number, which is combination of the precisions for $n=1,2,3,4$ n-grams
- ROUGE scores are reported separately for each n-gram
- The most commonly-reported ROUGE scores are:
 - ROUGE-1: unigram overlap
 - ROUGE-2: bigram overlap
 - ROUGE-L: longest common subsequence overlap
- A convenient Python implementation of ROUGE <https://github.com/pltrdy/rouge>

A ROUGE example:

- Q: “What is water spinach?”
- System output:
Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.
- Human Summaries
- Human 1:
Water spinach is a green leafy vegetable grown in the tropics.
- Human 2:
Water spinach is a semi-aquatic tropical plant grown as a vegetable.
- Human 3:
Water spinach is a commonly eaten leaf vegetable of Asia.
- $\text{ROUGE-2} = \frac{3+3+6}{10+9+9} = \frac{12}{28} = 0.43$

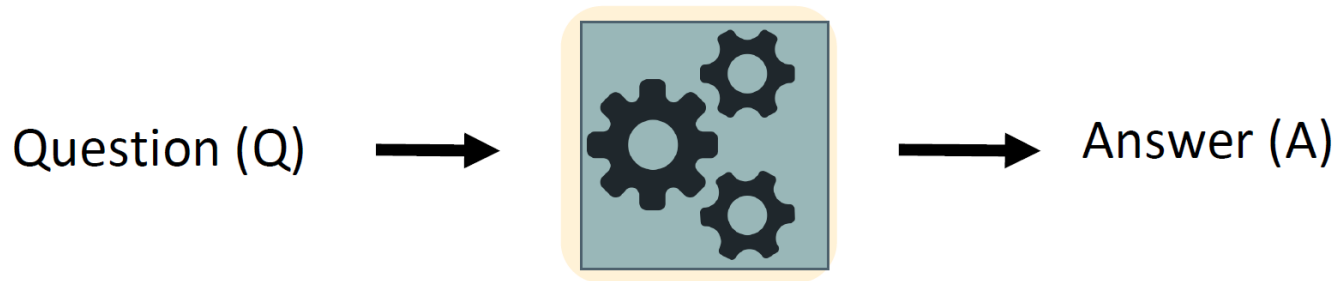
Summarization challenges

- Meaning representation and construction
- Long text abstractive summarization
- Large generative language models excel at summarization.
- They can be improved for specific domains or types of summarisations, e.g., query-focused summarisation, using finetuning on a suitable dataset

Question Answering

Question answering (QA)

- Question answering systems are designed to fill human information needs that might arise in situations like talking to a virtual assistant, interacting with a search engine, or querying a database



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

Question Answering

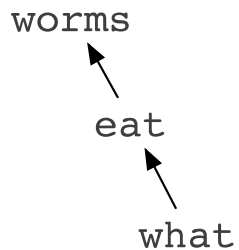
One of the oldest NLP tasks (punched card systems in 1961)

Question:

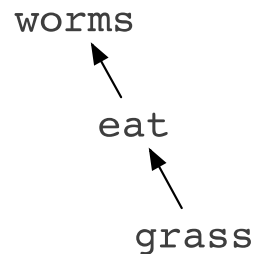
Potential Answers:

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

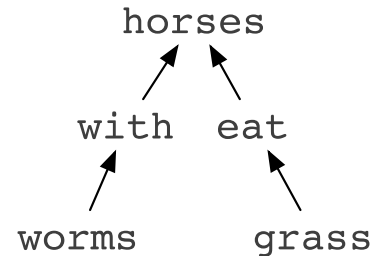
What do worms eat?



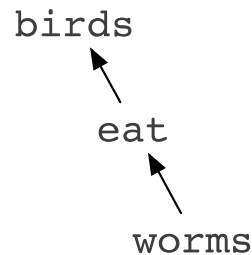
Worms eat grass



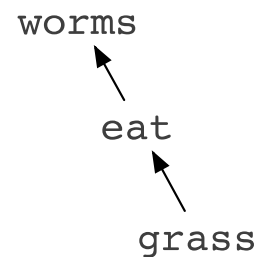
Horses with worms eat grass



Birds eat worms



Grass is eaten by worms

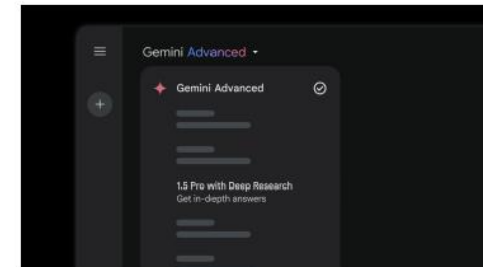
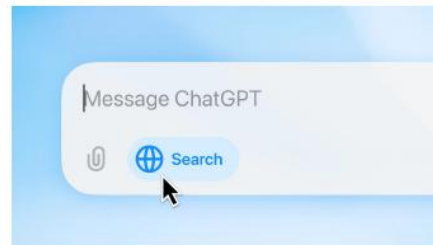


QA taxonomy

- 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, simple vs compositional, ..
- Answer type
A short segment of text, a paragraph, a list, yes/no, ...
- Knowledge store (LLMs-based)
 - Closed book – “parametric” QA (knowledge in the parameters of LLM)
 - Open book – external source (reading comprehension, RAG)

Many practical applications

- Many products, companies, focusing on QA
- Search engines
 - snippets
 - short and long answers
- LLMs
- ...



Question Answering: IBM's Watson

- Won Jeopardy! on February 16, 2011!



IBM Watson defeated two of Jeopardy's greatest champions in 2011

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL



Bram Stoker

Assistants



Wolfram Alpha



how many calories are in two slices of banana cream pie?



 Examples  Random

Assuming any type of pie, banana cream | Use **pie, banana cream, prepared from recipe** or **pie, banana cream, no-bake type, prepared from mix** instead

Input interpretation:

pie	amount	2 slices	total calories
	type	banana cream	

Average result:

[Show details](#)

702 Cal (dietary Calories)

Generative LLMs trained for instruction following

- state-of-the-art in 2022: multilingual, multimodal, extracts from web, RAG
- In 2022



na podlagi zadnjih novic iz Slovenije ustvari informativno poročilo o trenutnem političnem stanju v državi



Searched 4 sites

The current political situation in Slovenia has seen significant changes, with the recent parliamentary elections marking a notable shift in the country's political landscape. The opposition liberal Freedom Movement, which only launched earlier this year, has defeated the incumbent populist Prime Minister Janez Janša and his Slovenian Democratic Party. This election outcome reflects a broader discontent with Janša's leadership, which has been criticized for authoritarian tendencies and undermining democratic norms.

Generative LLMs trained for instruction following

- state-of-the-art in 2026: multilingual, multimodal, extracts from web, RAG, reasoning, more analytical, longer answer
- Only summary:

Slovenia is currently experiencing a politically fluid and potentially unstable transition period. The most likely short-term outcome is an attempt by Janez Janša to assemble a center-right governing coalition. However, the durability of such a government would depend on:

 - cohesion among smaller coalition partners,
 - the parliamentary behavior of Resni.ca,
 - and management of institutional disputes within the National Assembly.

The main risks are:

 - a narrow parliamentary majority,
 - intensified political polarization,
 - and the possibility of renewed elections if coalition negotiations collapse.

At the same time, the relatively high voter turnout and active coalition bargaining indicate that Slovenia's democratic institutions remain functional and politically competitive despite the instability.

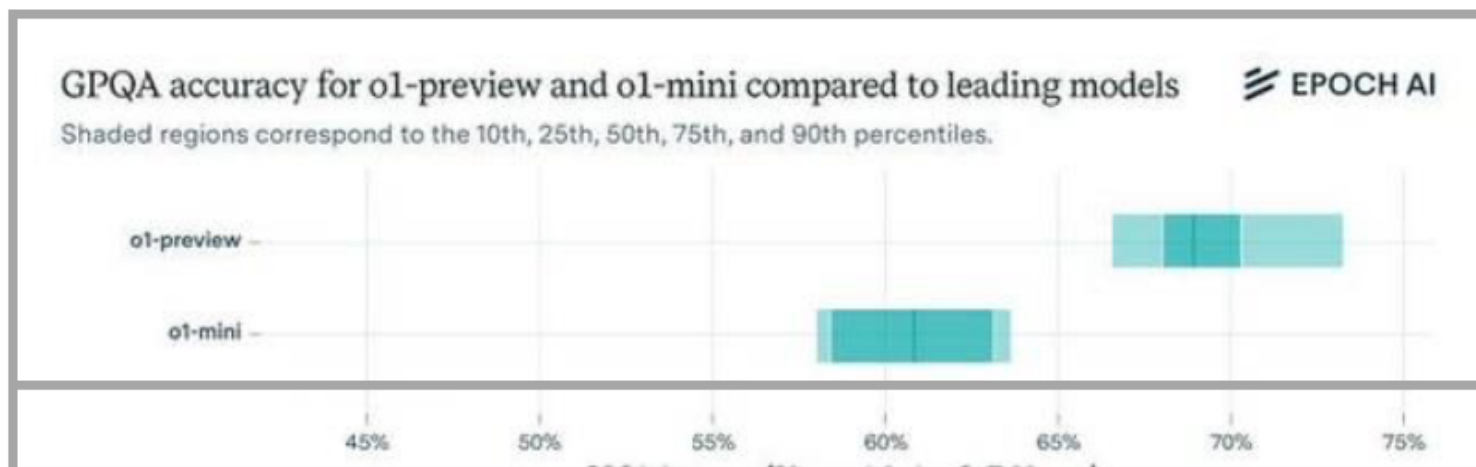
2023-2025: 'Expert level' QA tasks

GPQA (PhD level science questions)

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?

- A) 2
- B) 16
- C) 8
- D) 4



Expert level QA tasks

“Humanity’s last exam”

📖 Ecology

Question:

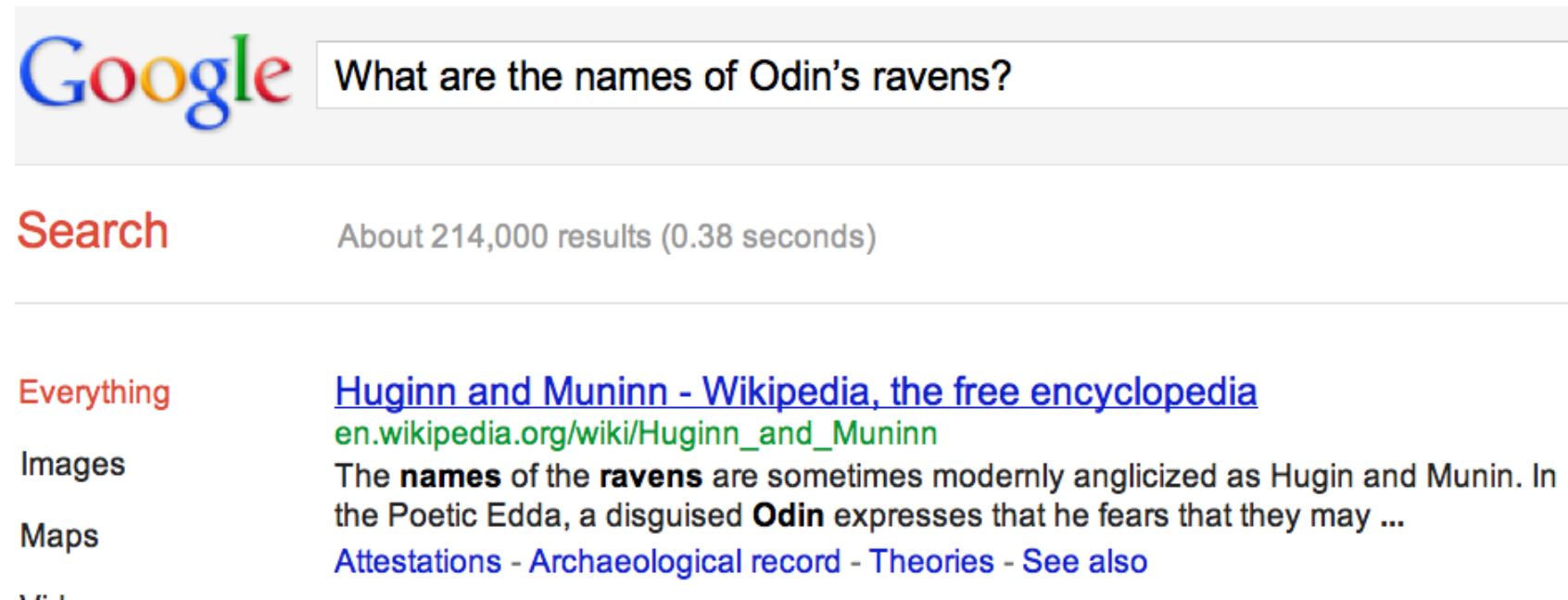
Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Model	Accuracy (%) ↑
 GPT-4o	3.1
 Grok-2	3.9
 Claude 3.5 Sonnet	4.8
 Gemini Thinking	7.2
 o1	8.8
 DeepSeek-R1*	8.6
 o3-mini (medium)*	11.1
 o3-mini (high)*	14.0

Types of Questions in Modern Systems

- Factoid questions
 - *Who wrote “The Universal Declaration of Human Rights”?*
 - *How many calories are there in two slices of apple pie?*
 - *What is the average age of the onset of autism?*
 - *Where is Apple Computer based?*
- Complex (narrative) questions:
 - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
 - *What do scholars think about Jefferson’s position on dealing with pirates?*

Many questions can already be answered by web search



- In fact, web search is more and more using LLM-based answers
- For the future development of information sources, and quality of the web, this is controversial

IR-based Question Answering



Where is the Louvre Museum located?

Search

About 904,000 results (0.30 seconds)

Everything

Best guess for Louvre Museum Location is **Paris, France**

Mentioned on at least 7 websites including [wikipedia.org](#), [answers.com](#) and [east-buc.k12.ia.us](#) - [Show sources](#) - [Feedback](#)

Images

Maps

[Musée du Louvre - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Musée_du_Louvre](#)

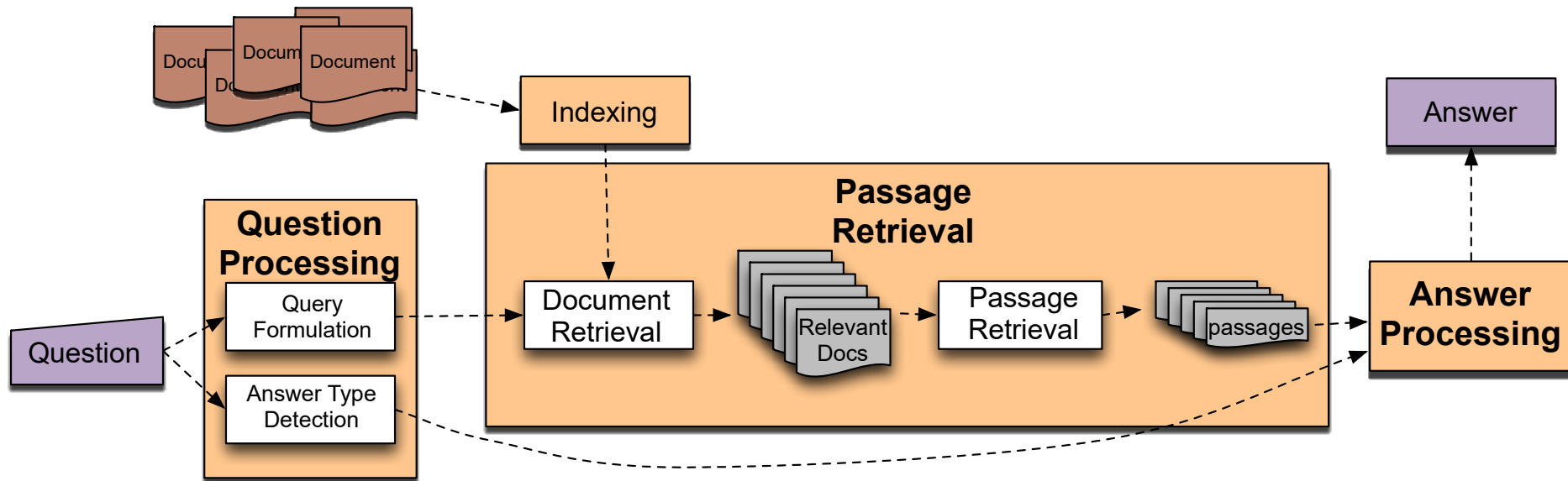
Videos

Musée du **Louvre** is **located** in Paris. **Location** within Paris. Established, 1793. **Location**, **Palais Royal**, Musée du **Louvre**, **75001 Paris, France**. Type, Art **museum** ...

News

[Louvre Palace - List of works in the Louvre - Category:Musée du Louvre](#)

Old IR-based Factoid QA is similar to RAG



Generative large language models

- Superior performance of very large models
- ChatGPT, GPT-4, Gemini
- LLaMa, Gemma, Qwen, GaMS

Relation Extraction

- Answers: Databases of Relations
 - born-in(“Emma Goldman”, “June 27 1869”)
 - author-of(“Cao Xue Qin”, “Dream of the Red Chamber”)
 - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

Whose granddaughter starred in E.T.?

(acted-in ?x “E.T.”)

(granddaughter-of ?x ?y)

Temporal Reasoning

- Relation databases
 - (and obituaries, biographical dictionaries, etc.)
- IBM Watson
 - ”In 1594 he took a job as a tax collector in Andalusia”
 - Candidates:
 - Thoreau is a bad answer (born in 1817)
 - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org:



The screenshot shows the geonames.org search interface. The search bar contains 'palo alto' and the country dropdown is set to 'all countries'. Below the search bar are buttons for 'search', 'show on map', and a link to 'advanced search'. The results section shows '459 records found for "palo alto"'. The table below lists the top three results.

	Name	Country	Feature class	Latitude	Longitude
1	Palo Alto Palo Al'to, Palo Alto, pa luo ao duo, paroaruto, Пало Алто, Пало Альто, פאלו, פאלו, パロアルト, 帕羅奧多	United States , California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2	Palo Alto Township Palo Alto Township	United States , Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3	Borough of Palo Alto	United States , Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21"	W 76° 10' 2"

Context and conversation in virtual assistants like Siri

- Coreference helps resolve ambiguities

U: “Book a table at Il Fornaio at 7:00 with **my mom**”

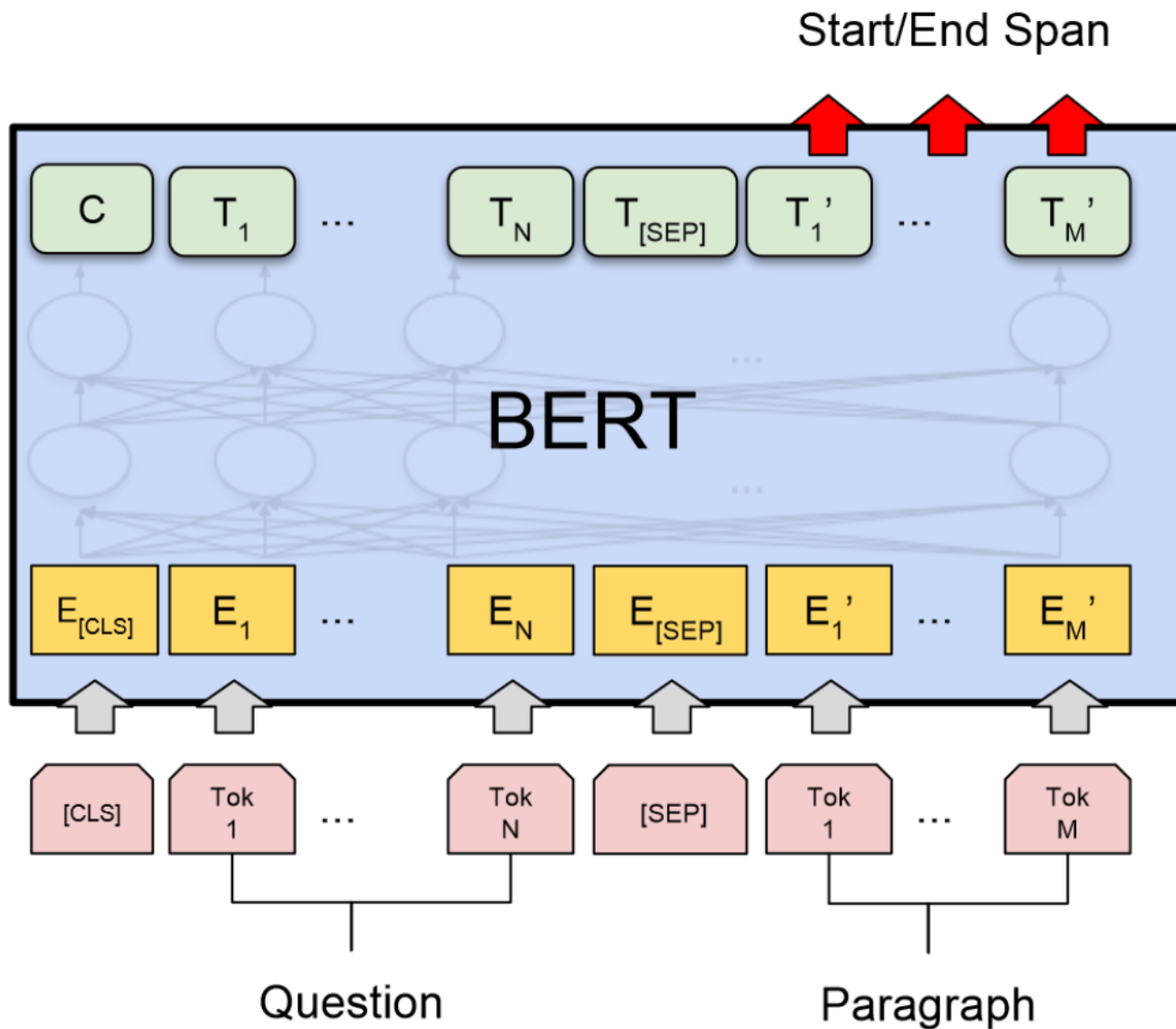
U: “Also send **her** an email reminder”

- Clarification questions:

U: “Chicago pizza”

S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”

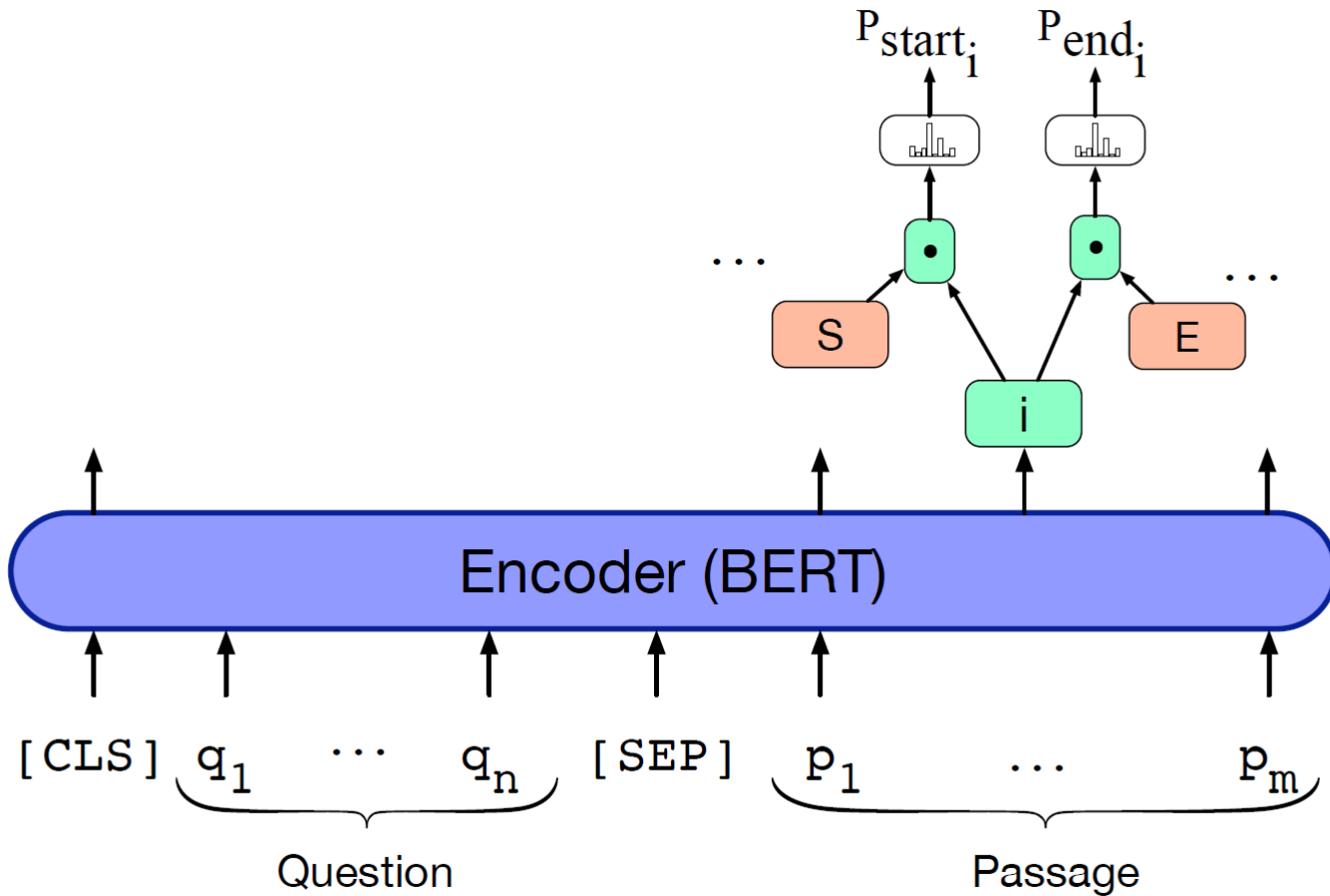
Questions and answers with BERT-like models



Factoid QA with BERT

- Answer Span Extraction
- span labelling: identifying in the passage a span (a continuous string of text) that constitutes an answer
- given a question q of n tokens $q_1, \dots, q_n$ and a passage p of m tokens $p_1, \dots, p_m$, the goal is to compute the probability $P(a, q, p)$ that each possible span a is the answer.

Factoid QA with BERT

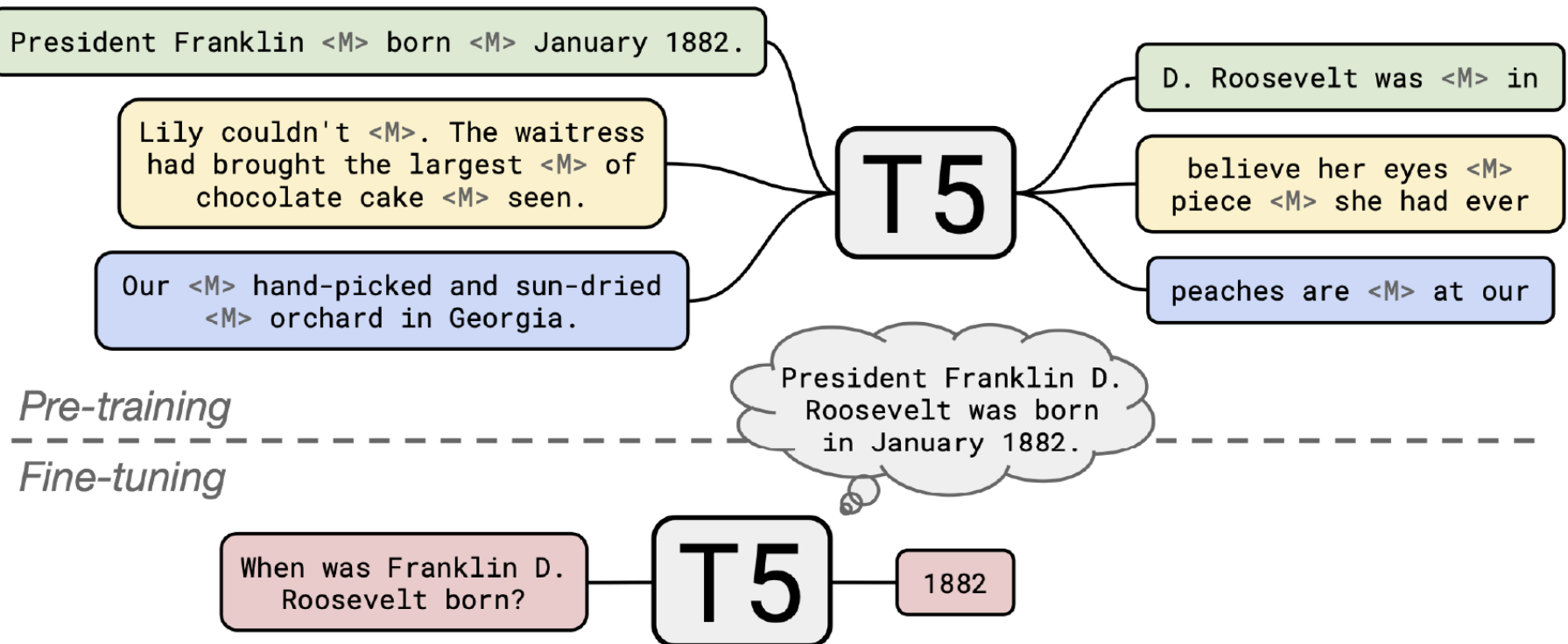


QA with language models

- a pretrained language model tries to answer a question solely from information stored in its parameters
- Even very large language models still suffer from certain problems in QA:
 - hallucinations
 - poor interpretability (might be addressed with e.g., chain-of-thought reasoning)
 - cannot give more context (e.g., a passage with the answer)

T5 model

- T5 learns to fill in masked spans of task (marked by <M>) by generating the missing spans (separated by <M>) in the decoder.
- It is then fine-tuned on QA datasets, given the question, without adding any additional context or passages.



QA datasets: BoolQ

- BoolQ (Boolean Questions, Clark et al., 2019a) is a QA task where each example consists of a short passage and a yes/no question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer.
- Passage: Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.
- Question: is barq's root beer a pepsi product
- Answer: No

QA datasets: SQuAD

- SQuAD 2.0 (Stanford Question Answering Dataset) is a reading comprehension tasks. Crowd workers were employed to ask questions over a set of Wikipedia articles. They were then asked to annotate the questions with the text segment from the article that forms the answer. They also added ca. 50,000 unanswerable questions to the dataset based on Wikipedia articles.
- Article: Endangered Species Act
Paragraph: “. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”
- Question 1: “Which laws faced significant opposition?”
- Plausible Answer: later laws
- Question 2: “What was the name of the 1937 treaty?”
- Plausible Answer: Bald Eagle Protection Act

QA datasets: COPA

- COPA (Choice of Plausible Alternatives, Roemmele et al., 2011) is a causal reasoning task in which a system is given a premise sentence and must determine either the cause or effect of the premise from two possible choices. All examples are handcrafted and focus on topics from blogs and a photography-related encyclopedia.
- Premise: My body cast a shadow over the grass.
- Question: What's the CAUSE for this?
- Alternative 1: The sun was rising.
- Alternative 2: The grass was cut.
- Correct Alternative: 1

QA datasets: MultiRC

- MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al., 2018) is a QA task where each example consists of a context paragraph, a question about that paragraph, and a list of possible answers. The system must predict which answers are true and which are false. Each answer is independent from the others. The paragraphs are drawn from seven domains including news, fiction, and historical text.
- Paragraph: Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week
- Question: Did Susan's sick friend recover?
- Candidate answers: Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)

QA datasets: ReCoRD

- ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset, Zhang et al., 2018) is a multiple-choice QA task. Each example consists of a news article and a Cloze-style question about the article in which one entity is masked out. The system must predict the masked out entity from a list of possible entities in the provided passage, where the same entity may be expressed with multiple different surface forms, which are all considered correct. Articles are from CNN and Daily Mail.
- Paragraph: (CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood
- Query For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency
- Correct Entities: US

QA datasets: RiddleSense

- RiddleSense (Lin et al, 2021) is a multiple-choice question answering task containing riddle-style commonsense questions.
- Riddle: I have five fingers, but I am not alive. What am I?
- Answers: (A) piano (B) computer (C) glove (D) claw (E) hand
- Riddle: My life can be measured in hours. I serve by being devoured. Thin, I am quick; Fat, I am slow. Wind is my foe. What am I?
- Answers: (A) paper (B) candle (C) lamp (D) clock (E) worm

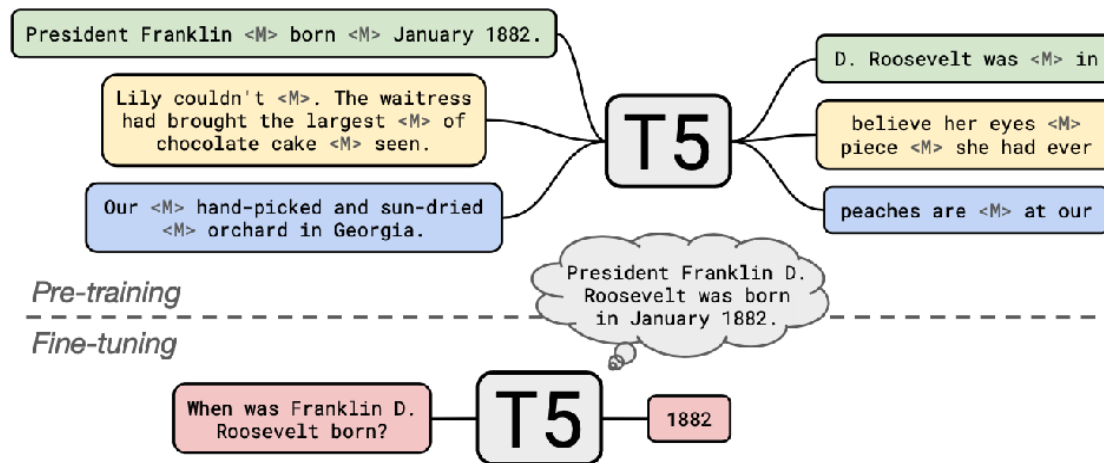
LLM-based QA

Closed book QA

- **Closed book** –answering the question with *no access to external information*
- Q&A done just by ‘predicting the next word’
Ljubljana is the capital of _____
- When does this work? Why do we care about this setting?

Large language models can do open-domain QA well

- Initially –for models like T5, people discovered LMs can directly answer questions
- Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?



LMs have remarkable knowledge about the world

- Closed-book QA then took on an important role as a *capability test*.
- Many popular benchmarks (ARC-C, MMLU, GPQA, etc) are partially knowledge tests

MMLU

Astronomy

What is true for a type-Ia supernova?

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

Scenario	Task	What	Who
NarrativeQA narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
NaturalQuestions (closed-book) natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries	web users
NaturalQuestions (open-book) natural_qa_openbook_longans	short-answer question answering	passages from Wikipedia, questions from search queries	web users
OpenbookQA openbookqa	multiple-choice question answering	elementary science	Amazon Mechanical Turk workers
MMLU (Massive Multitask Language Understanding) mmlu	multiple-choice question answering	math, science, history, etc.	various online sources
GSM8K (Grade School Math) gsm	numeric answer question answering	grade school math word problems	contractors on Upwork and Surge AI
MATH math_chain_of_thought	numeric answer question answering	math competitions (AMC, AIME, etc.)	problem setters
LegalBench legalbench	multiple-choice question answering	public legal and administrative documents, manually constructed questions	lawyers
MedQA med_qa	multiple-choice question answering	US medical licensing exams	problem setters
WMT 2014 wmt_14	machine translation	multilingual sentences	Europarl, news, Common Crawl, etc.

Is 'parametric knowledge' useful?

- Why is parametric knowledge a common capability test?
- Easy to collect, easy to automatically grade (GPQA, Humanity's last exam)
- Knowledge is useful, even if you are going to do things like summarization
- In practice, correlated with *other* model capabilities (Knowledge <-> scale)

How does a LM acquire parametric knowledge?

- Stanford University is located in _____
- We know ‘filling in the blank’ can lead to knowledge being acquired...
- But what kind of data do we need, and how much of it?
- **Recent works** –controlled synthetic experiments on how LLMs acquire knowledge
- **Empirical observations** –you need many exposures (~1000) or very large models to acquire knowledge

- SECTION 6: HOW TRAINING TIME AFFECTS MODEL CAPACITY.

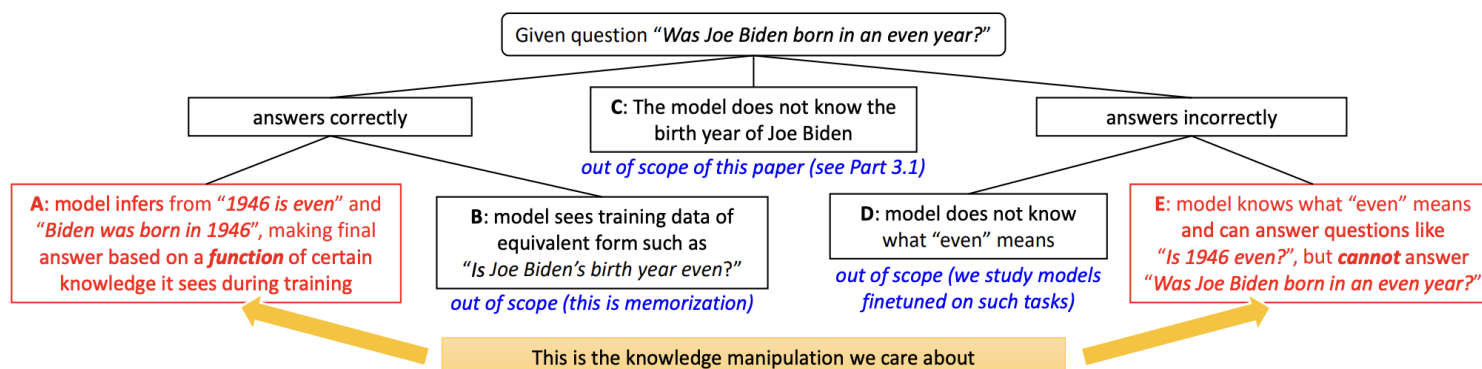
Achieving a 2bit/param capacity requires each knowledge piece to be visited 1000 times during training, termed ***1000-exposure*** to differentiate from traditional “1000-pass” terminology, as a single data pass can expose a knowledge piece 1000 times.⁵

- **RESULT 4**: With 100 exposures, an *undertrained* GPT2’s capacity ratio falls to 1bit/param.

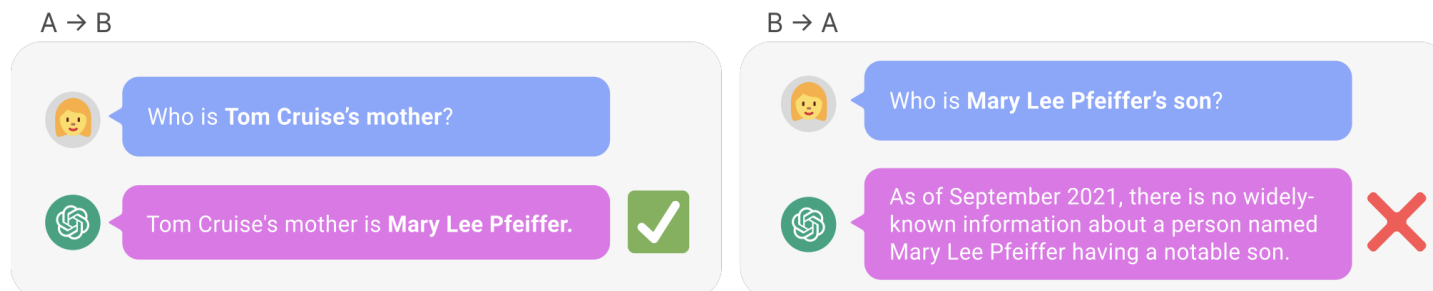
Remark 1.3. Another perspective on Result 4 is that *rare* knowledge, encountered only 100 times during training, is stored at a 1bit/param ratio.

Some challenges with parametric knowledge

- Acquiring knowledge *robustly* is difficult
- Entailed knowledge is hard to learn (without CoT)



- Reversal curse' constrains knowledge use



Hallucinations

- LLMs are *remarkably* good at many things – even closed-book QA
- .. But they aggressively make things up for things they don't know



Australian mayor readies world's first defamation lawsuit over ChatGPT content

By Byron Kaye

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare



Abstention – knowing what you know

- Humans don't know everything in the world
- ... but we generally know how to back off and communicate uncertainty
- Methods for abstention
- 'Backing off'
 - Estimate the *confidence* of a statement (by sampling or self-scoring)
 - Abstain whenever the confidence is too low (or remove details)

To go beyond abstaining...

- Can we add new, parametric knowledge after pre-training?
- Goal: reliably back off when models don't know, but try to add knowledge when we can
- One way to update knowledge – continue to do pretraining with domain-specific data
- Unfortunately... requires tons of data (50+B tokens)
- Updating knowledge (with synthetic data in continued pretraining)
- Synthetic continued pretraining: Train on LLM-transformed data
- Goal–replicate the diversity of pretraining
 - Vary content (topics)
 - Vary style (how it's presented)
 - Data diversity for generalization
- Can significantly boost closed-book QA performance with synthetic data

Black-box LMs are impressive, but parametric memory has its limitations

- LLMs can store an impressive amount of information in their parameters
 - But LLMs can't memorize everything
 - The world changes over time
 - You might want it to use private documents that aren't on the web
- Also, black-box LLMs are opaque
 - Given a query, it produces an answer, but it's difficult to verify if the answer is correct

Open book QA - Reading comprehension

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

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

- Q: *What language did Tesla study while in school?*
- A: *German*

Why do we care about reading comprehension 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, e.g., information extraction

Reading comprehension – span based

- Archetypical reading comprehension dataset –**SQuAD**
- Major focus of Q&A and research for many years..

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3 , $1 \cdot 3$, $1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

Ground Truth Answers: itself itself itself itself itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?

Ground Truth Answers: composite number composite number composite number primes

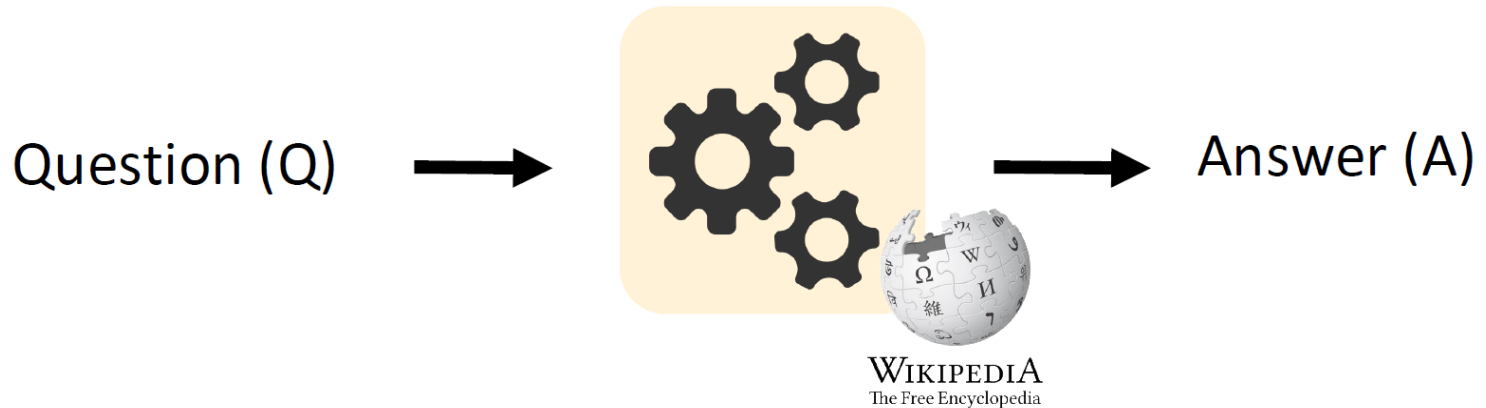
What theorem defines the main role of primes in number theory?

Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

QA datasets related to reading comprehension

- **TriviaQA:** Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss the question, but no human verification that the paragraph supports the answer to the question
- **Natural Questions:** Questions drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or NOT_PRESENT. Verified by human annotation.
- **HotpotQA.** Constructed questions to be answered from the whole of Wikipedia, which involve getting information from two pages to answer a multistep query:
Q: Which novel by the author of “Armada” will be adapted as a feature film by Steven Spielberg? A: *Ready Player One*

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 (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 and a more practical problem, in contrast to closed-domain systems that deal with questions under a specific domain (medicine, technical support).

Retrieval augmentation

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



Document
Retriever



Document
Reader



833,500



<https://github.com/facebookresearch/DrQA>

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Using retrieval to overcome LMs' shortcomings

- Instead of asking the LM to memorize everything, can we provide the LM with relevant and useful content just-in-time?
- Retrieval/search is a common mechanism for identifying such relevant information.
 - Dynamic: it's easy to update/add documents to your retrieval system
 - Interpretable: LM can generate pointers to retrieved documents that support human verification of its generations (citations)

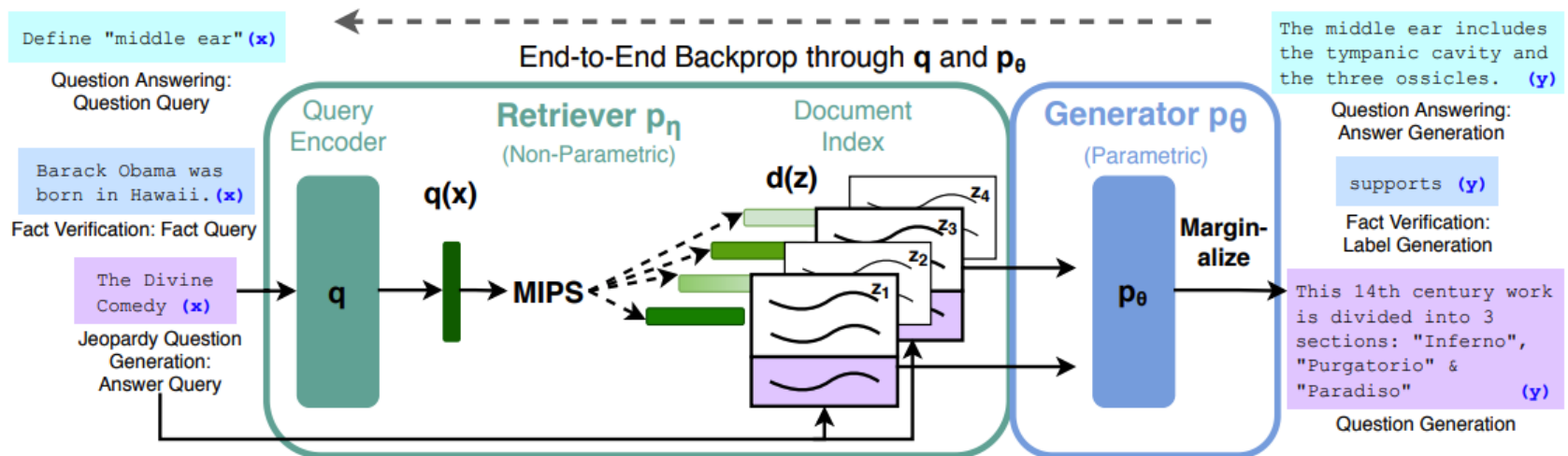


Retriever-reader framework

- Input: a large collection of documents $\mathcal{D} = D_1, D_2, \dots, D_N$ and Q
- Output: an answer string A
- Retriever: $f(\mathcal{D}, Q) \rightarrow P_1, \dots, P_K$
- Reader: $g(Q, \{P_1, \dots, P_K\}) \rightarrow A$
- Retriever = any neural retriever
- Reader = a neural reading comprehension model that we just learned
 - Could be trained on SQuAD and other distantly supervised QA datasets
 - Or a zero-shot LLM (ChatGPT etc)
- Often a reranker is added to rerank retrieved passages before passing a limited number of them to the reader

RAG and it's open problems

- Retrieval Augmented Generation (RAG) is very powerful!



- How well does RAG work, anyway? Is it more factual?

Different types of retrievers

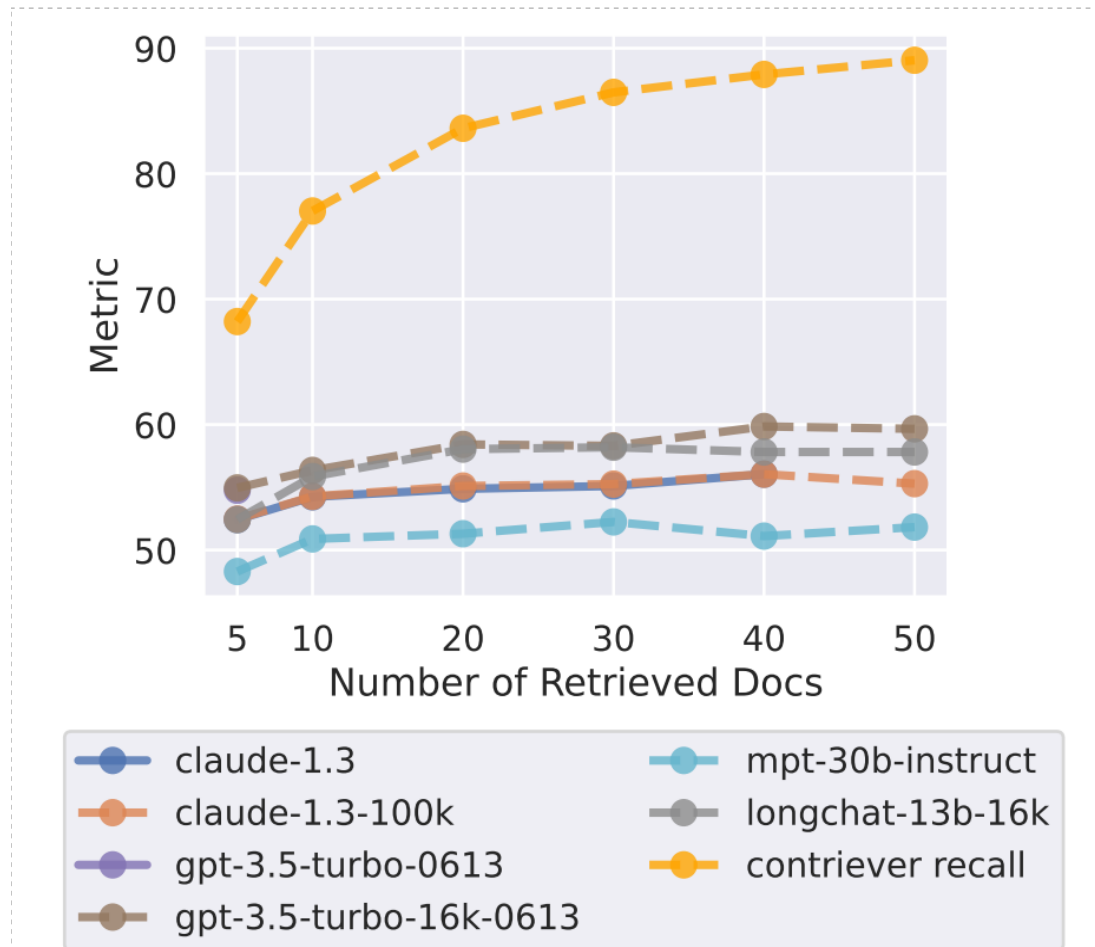
- How should we retrieve relevant passages for our retrieval system?
- Word-overlap (BM25)
- Vector retrieval (DPR, Sentence vectors)
- Other, neural systems (ColBERT)
- We discussed this in section 12-*Neural information retrieval*

Problem: How many documents can we use?

- The retriever is key – if we have to use only one document, then we have to get that right.
- Why not get lots of documents and pass them to the LM?
- LM's can't pay attention to the entire context
- The long-context problem bites you – LLMs do not pay attention to its context well!
- In practice: LLMs can't use many documents

LLMs can't use many documents

- Retriever performance (yellow) rises slowly to 90% recall
- RAG performance (other lines) saturate very quickly –after 10-20 documents.



Even more open domain: LLMs with web search



Where does Christopher D. Manning teach?



All

Chat

Images

Videos

News

Maps

More

Where does Christopher D. Manning teach?

Christopher D. Manning is a professor of computer science and linguistics at Stanford University [1][2]. He is also the Director of the Stanford Artificial Intelligence Laboratory [1], and has been teaching at Stanford since 1994.

1. Christopher Manning's Profile | Stanford Profiles

<https://profiles.stanford.edu/chris-manning>

2. Introduction to Information Retrieval: Manning, Christopher D ...

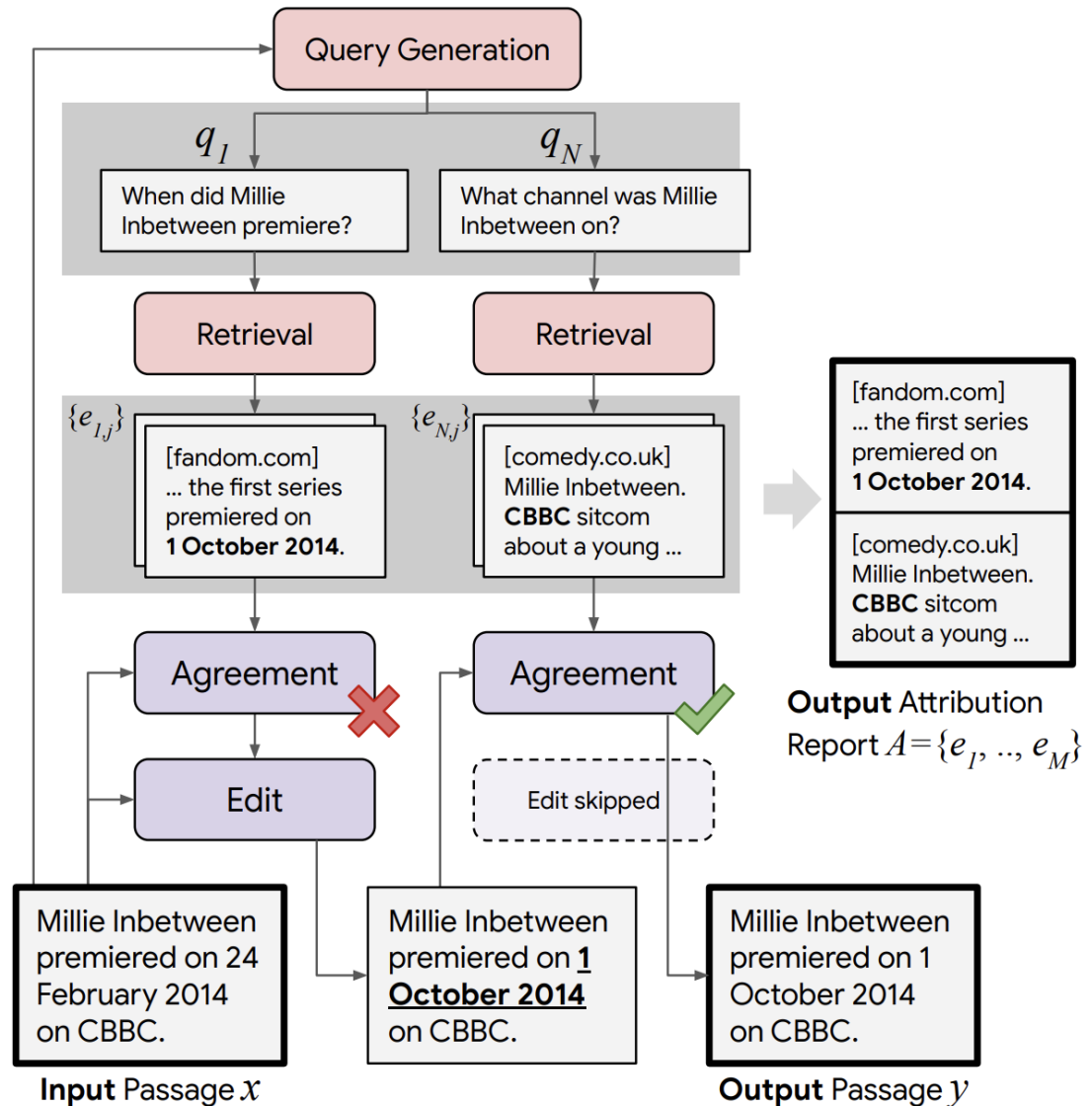
<https://www.amazon.com/Introduction-Information-Retrieval-Christopher-Manning/dp/0521865719>



Lots of systems that search and cite the web!

Systems architectures of such systems

- Uses QA to check for consistency between sources and responses
- Returns citations for each of these items using a LM.



Are LLM-based QA systems the answer?

- A unique benefit of RAG: citing your sources
- These citations themselves are generated by LLMs
- So the citations could *also be hallucinated*.. How often does that happen?
- Precision and recall for citations are both low (Liu et al, 2023)

<u>Citation Precision (%; ↑)</u>		<u>Citation Recall (%; ↑)</u>	
Average Over All Queries		Average Over All Queries	
Bing Chat	89.5	Bing Chat	58.7
NeevaAI	72.0	NeevaAI	67.6
perplexity.ai	72.7	perplexity.ai	68.7
YouChat	63.6	YouChat	11.1
Average	74.5	Average	51.5

LLM based QA recap

- QA is one of the top uses for a LLM
- QA as a capability evaluation, and for understanding knowledge
- QA as a system – RAG and internet retrieval
- Limitations: retrievers are not perfect, and models can't handle lots of documents

Query-focused summarization

- Produces a summary relevant to a given query
- Query can be a named entity, e.g., an organisation (a company)
- A special type of QA

QFS-Composer

- Task: query-focused summarisation
- Input: a query, a text to be summarized
- LLM analyses the input to determine the type of focus and decompose it into several questions
- For example, if the focus is a named entity, such as a person, or organization, it uses question generation (QG) module to generate questions based on the focused named entities and given text
- Use QA engine to answer questions
- Expand the initial query with answers and give it as an input to summarization LLM

Đuranović, V., & Šikonja, M. R. (2026). QFS-Composer: Query-focused summarization pipeline for less-resourced languages. *arXiv preprint arXiv:2604.10687*:

<https://arxiv.org/abs/2604.10687>

QAGS metric

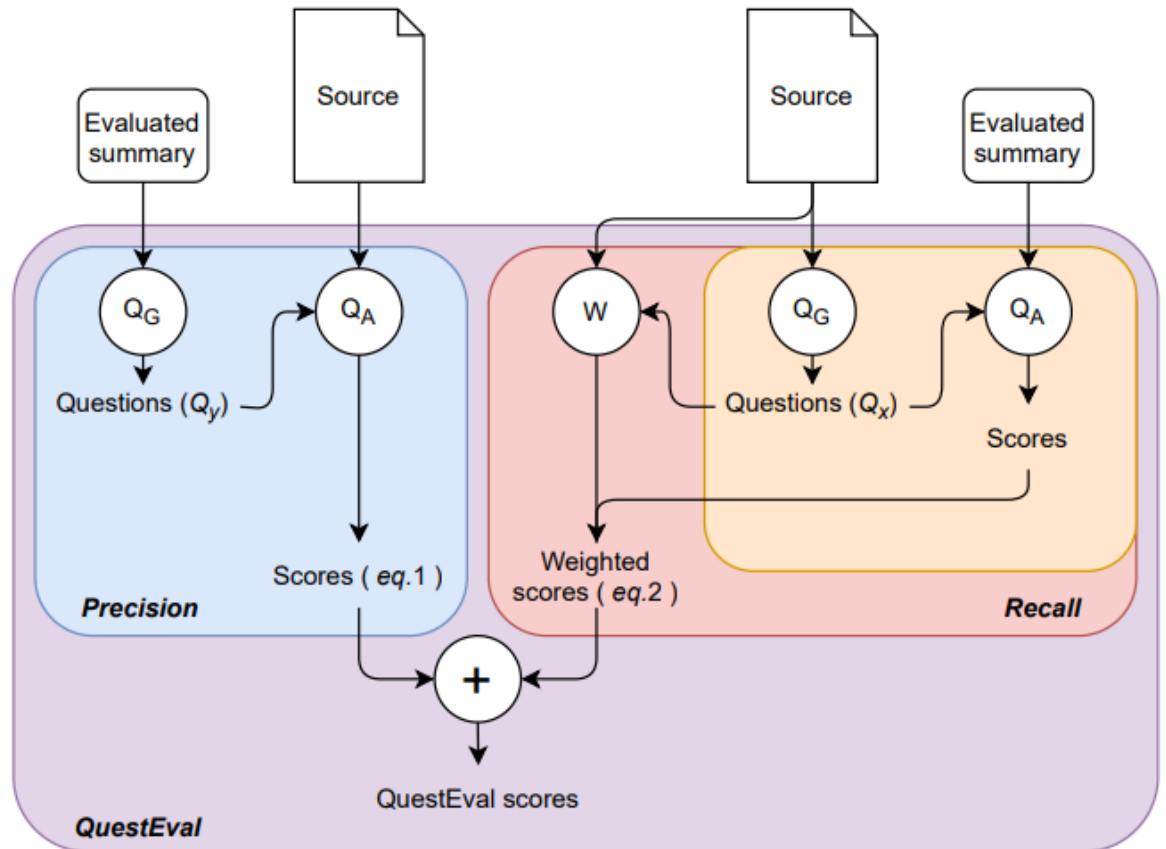
- Question answering and generation for summarization (QAGS)
- Measures how well a summary preserves facts from the original document.
- Based on the intuition that if we ask questions about a summary and its source, we will receive similar answers if the summary is factually consistent with the source.
- QAGS had significantly higher correlation for factual consistency with human judgment
- Needs a good question generator (QG) and question answering (QA) model

QAGS functioning

- QG component forms questions using extracted named entities and noun phrases from the summary as answer candidates.
- Filter out low-quality questions, such as removing duplicates.
- QA component generates two sets of answers to previously generated questions: one set of answers conditioned on the generated summary, and the other set observing the original text.
- Compute similarity between two answers to a single question in several ways: exact match score, token-level F_1 score, edit distance, and BERT score.
- Edit distance measures the number of characters between two strings that need to be changed in order for strings to match.

QuestEval metric

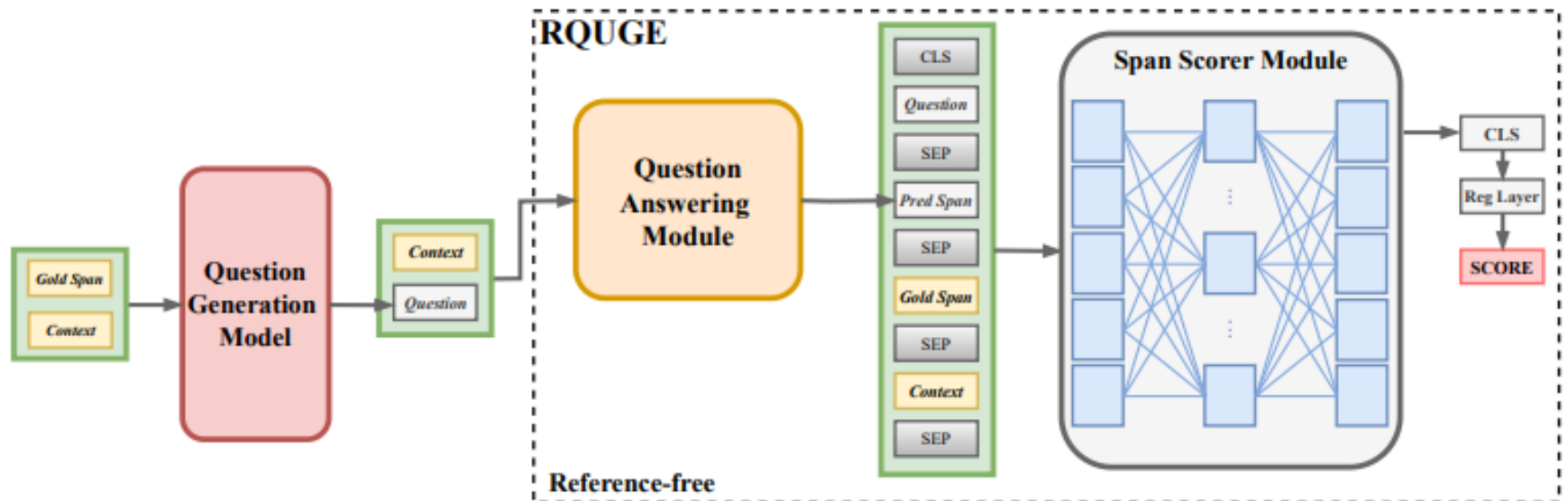
- A recall-oriented variant of QAGS
- Introduces also questions from the source document, weights the results



RQUGE score

- RQUGE (Reference-free QUestion Generation Evaluation)
- Evaluates generated questions
- Similar to ROUGE, but evaluates overlap of important spans
- Two components: QA and answer/span scorer.
- Given the candidate question q_c and the context D , the QA model computes the answer span a_c (position where the answer lies within the context)
- Span scorer compares the candidate answer a_c with the named entity or text segment selected as the answer a_r before the candidate question was generated, conditioned on the context
- This learned scoring mechanism enables automated selection of significant questions without requiring human-annotated references for every input. It serves as a proxy for human judgment as to which aspects of text are worth questioning.
- Weakness: strong dependence on QA component.

RQUGE score



Comparing scores for QG

Context: Today, Warsaw has some of the best medical facilities in Poland and East-Central Europe. The city is home to the Children's Memorial Health Institute (CMHI), ...

Reference Question: Where are some of the best medical facilities in East-Central Europe located?

Question ① Where is the Children's Memorial Health Institute located?

✗ BLEU:0.07 ✗ BERTScore:0.44 ✓ RQUGE: 1.0

Question ② In which city of Poland, can the best medical equipments be found?

✗ BLEU:0.12 ✗ BERTScore:0.52 ✓ RQUGE: 1.0

Question ③ Where are some of the best medical facilities in West-Central Europe located?

✗ BLEU:0.80 ✗ BERTScore:0.96 ✓ RQUGE: 0.0₉₅