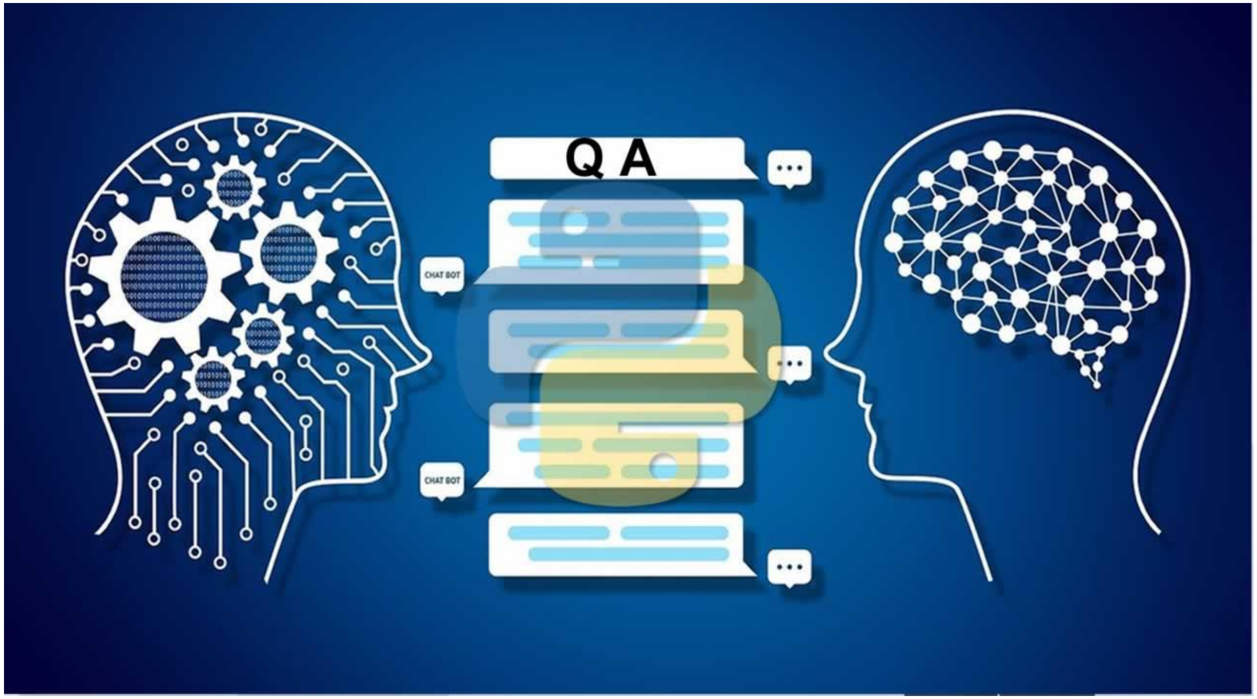
University of Ljubljana, Faculty of Computer and Information Science

Automatic summarization and question answering



Prof Dr Marko Robnik-Šikonja

Natural Language Processing, Edition 2023

Contents

- summarization
- question answering

- Literature:

Jurafsky and Martin, 3rd edition

Some slide taken from Jurafsky

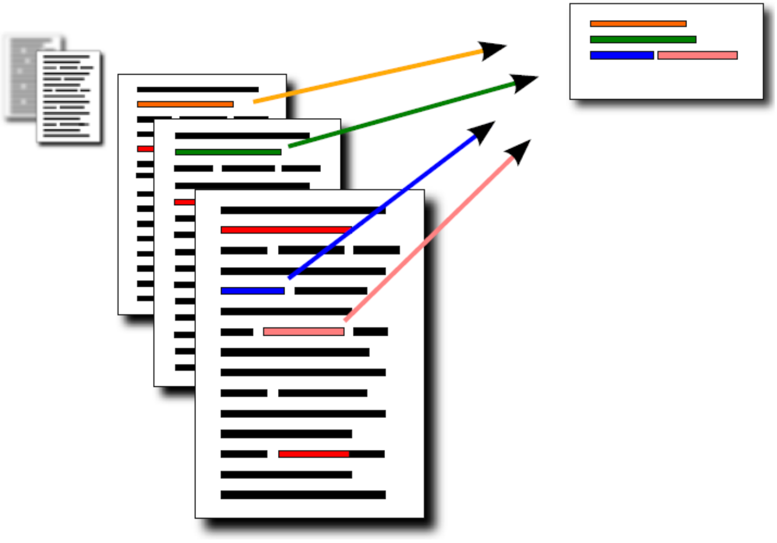
Text summarization

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

Illustration from The Economist

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?



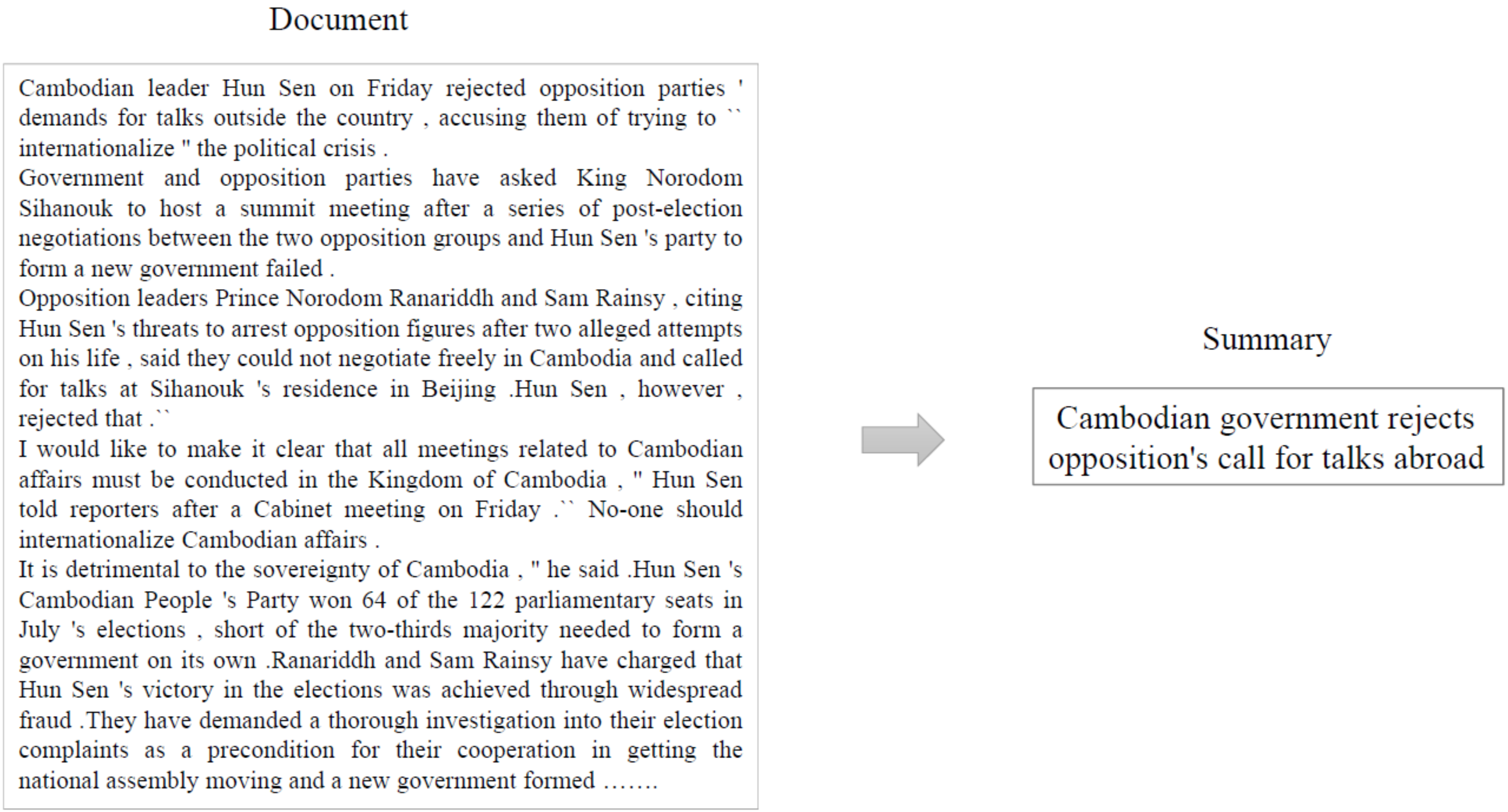
Abstract, outline, headline

- 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)



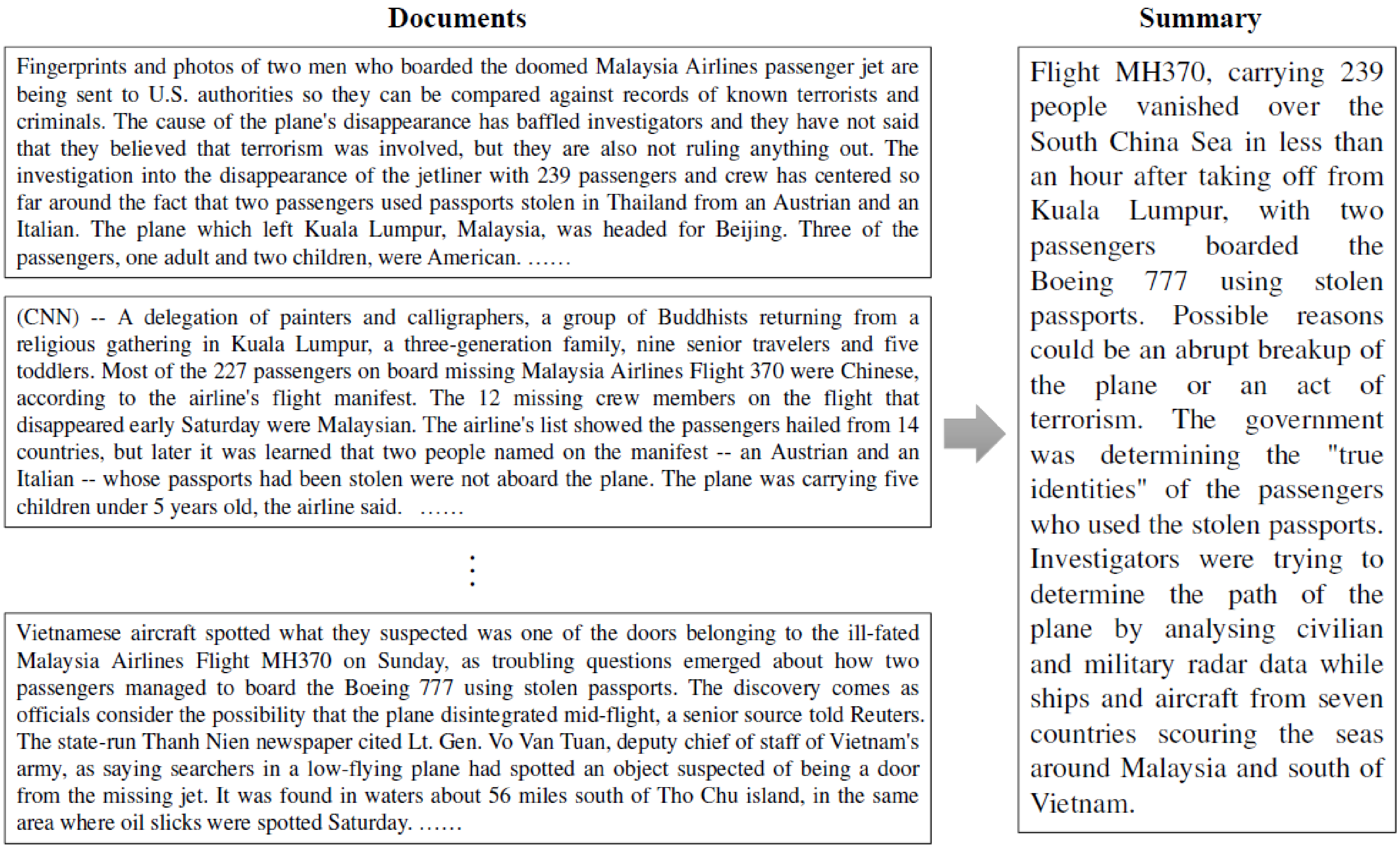
Document

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

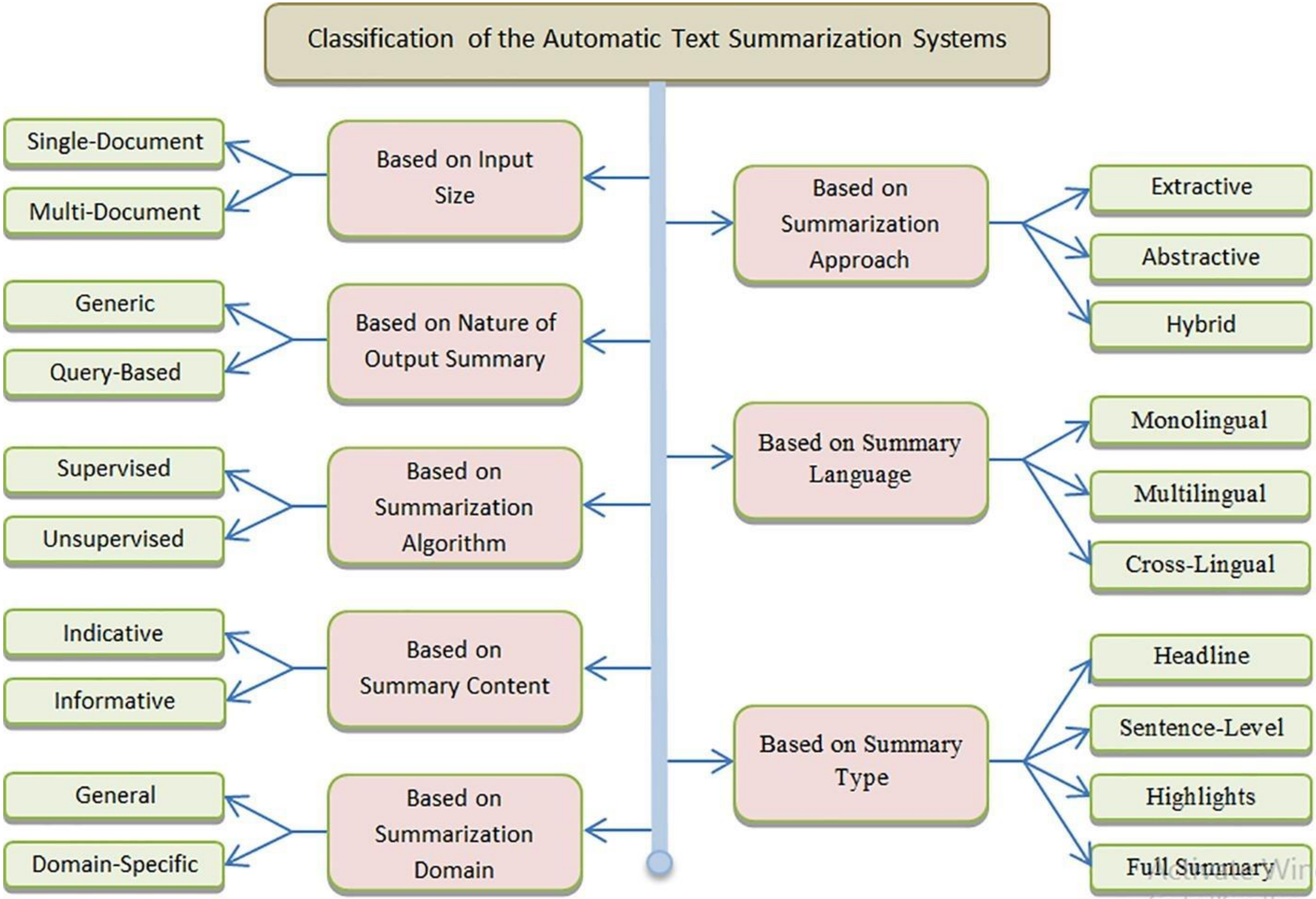
Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining 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
    - 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



Classification of the Automatic Text Summarization Systems

Based on Input Size
- Single-Document
- Multi-Document

Based on Nature of Output Summary
- Generic
- Query-Based

Based on Summarization Algorithm
- Supervised
- Unsupervised

Based on Summary Content
- Indicative
- Informative

Based on Summarization Domain
- General
- Domain-Specific

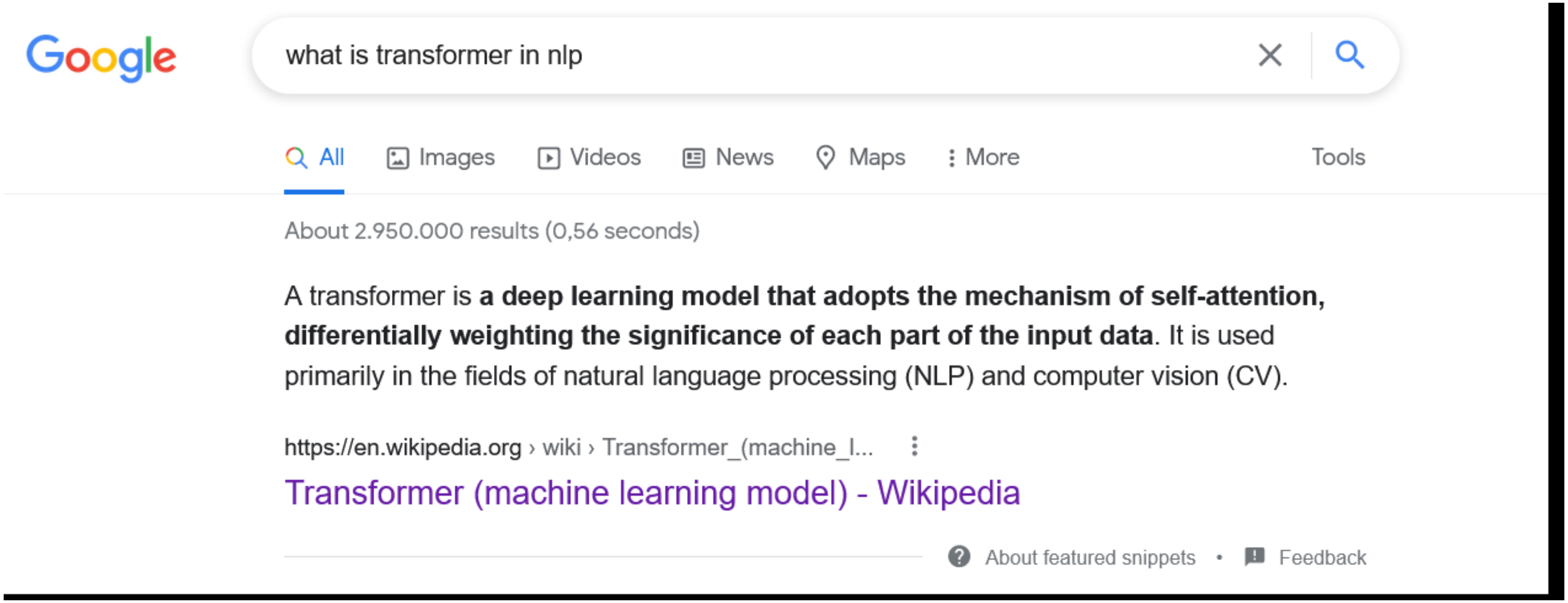
Based on Summarization Approach
- Extractive
- Abstractive
- Hybrid

Based on Summary Language
- Monolingual
- Multilingual
- Cross-Lingual

Based on Summary Type
- Headline
- Sentence-Level
- Highlights
- Full Summary

Summarization for Question Answering: Snippets

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



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...

Transformer (machine learning model) - Wikipedia

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

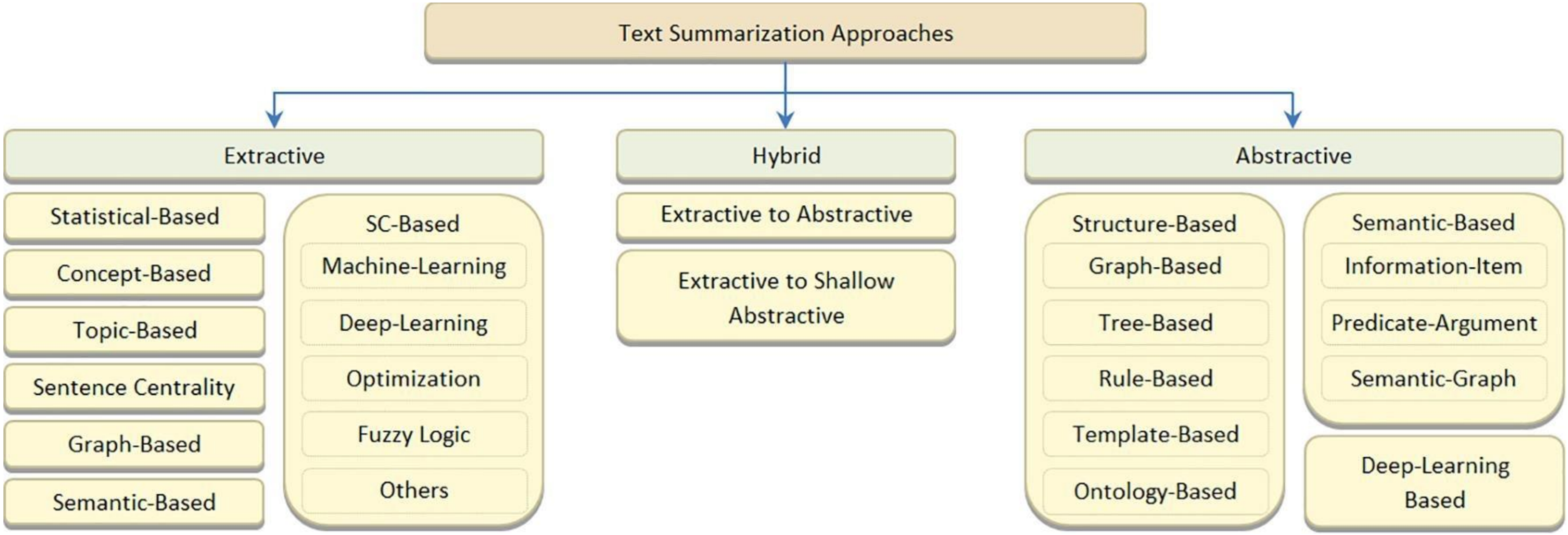
Summarization: common datasets

- 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 ali, 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)
- List of summarization datasets, papers, and codebases: https://github.com/mathsyouth/awesome-text-summarization

Sentence simplification: common datasets

- *Sentence simplification* is a 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 written for children

An overview of (historical) approaches



Pre-neural summarization

- Pre-neural summarization systems were mostly extractive
- Like pre-neural MT, they typically had a pipeline:
  - **Content selection**: choose some sentences to include
  - **Information ordering**: choose an ordering of those sentences
  - **Sentence realization**: clean up the sentences, edit the sequence of sentences (e.g. simplify, remove parts, fix continuity issues)

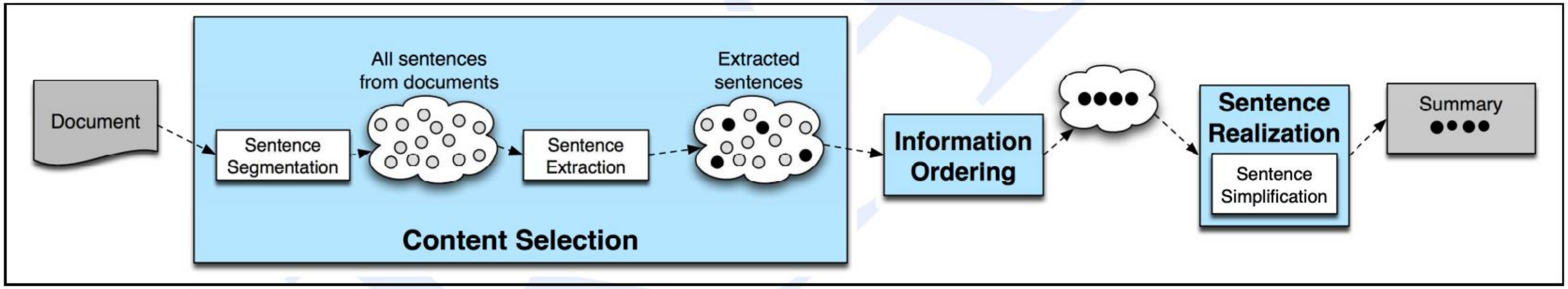


Figure 23.14 The basic architecture of a generic single document summarizer.

Diagram from Jurafsky and Martin: *Speech and Language Processing*, 2nd edition, 2009

Pre-neural **content selection** algorithms:

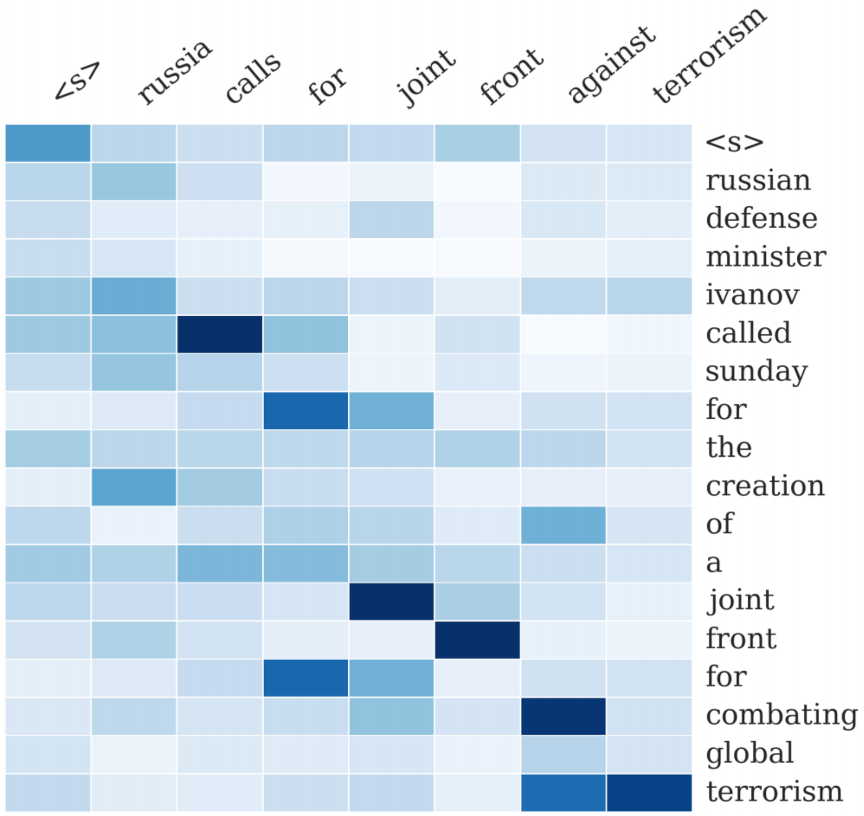
- Sentence scoring functions can be based on:
  - Presence of topic keywords, computed via e.g. tf-idf
  - Features such as where the sentence appears in the document
- Graph-based algorithms view the document as a set of sentences (nodes), with edges between each sentence pair
  - Edge weight is proportional to sentence similarity
  - Use graph algorithms to identify sentences which are *central* in the graph
- Supervised

Pre-neural supervised content selection

- Given:
  - A labeled training set of good summaries for each document
- Align:
  - The sentences in the document with sentences in the summary
- Extract features
  - position (first sentence?)
  - length of sentence
  - word informativeness, cue phrases
  - cohesion
- Train
  - A binary classifier (put sentence in summary? yes or no)
- Problems:
  - hard to get labeled training
  - alignment difficult
  - performance not better than unsupervised algorithms
- So in practice:
  - Unsupervised content selection was more common

Neural summarization developments

- 2015: Rush et al. publish the first seq2seq summarization paper
- Single-document abstractive summarization is a translation task!
- Thus we can apply standard seq2seq + attention NMT methods



Rush et al, 2015. *A Neural Attention Model for Abstractive Sentence Summarization*, https://arxiv.org/pdf/1509.00685.pdf

Neural summarization developments

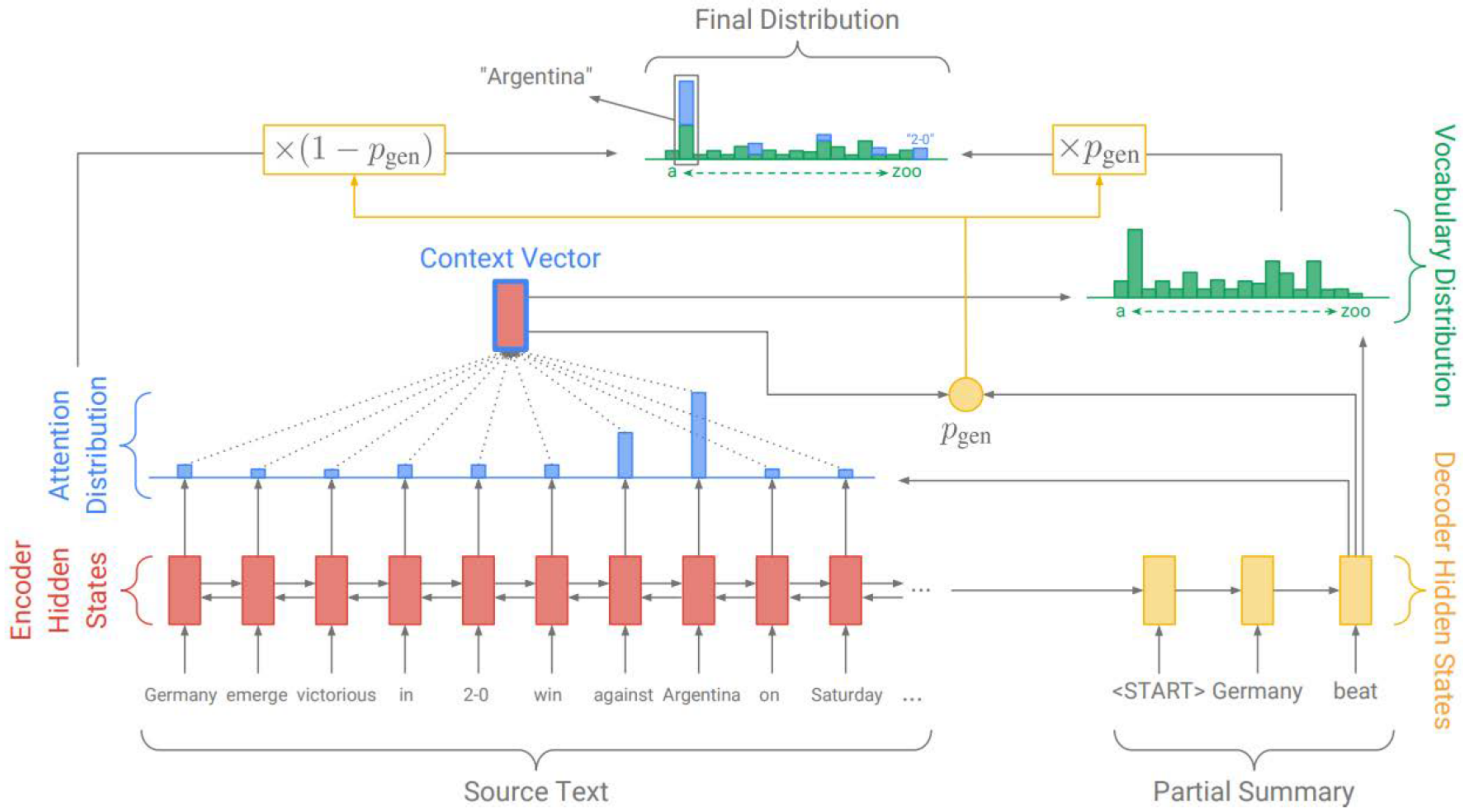
- Since 2015, there have been lots more developments!
- Making it easier to copy
- But also preventing too much copying!
- Hierarchical / multi-level attention
- More global / high-level content selection
- Using Reinforcement Learning to directly maximize ROUGE, or other discrete goals (e.g., length)
- Resurrecting pre-neural ideas (e.g., graph algorithms for content selection) and working them into neural systems
- …

- List of summarization datasets, papers, and codebases: https://github.com/mathsyouth/awesome-text-summarization
- Alomari et al., 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech & Language*, *71*, p.101276. https://doi.org/10.1016/j.csl.2021.101276
- Dong, 2018. *A Survey on Neural Network-Based Summarization Methods*, https://arxiv.org/pdf/1804.04589.pdf

Neural summarization: copy mechanisms

- Seq2seq + attention systems are good at writing fluent output, but bad at copying over details (like rare words) correctly
- Copy mechanisms use attention to enable a seq2seq system to easily copy words and phrases from the input to the output
- Clearly this is very useful for summarization
- Allowing both copying and generating gives us a hybrid extractive/abstractive approach

Neural summarization with pointer generator networks: copy mechanism



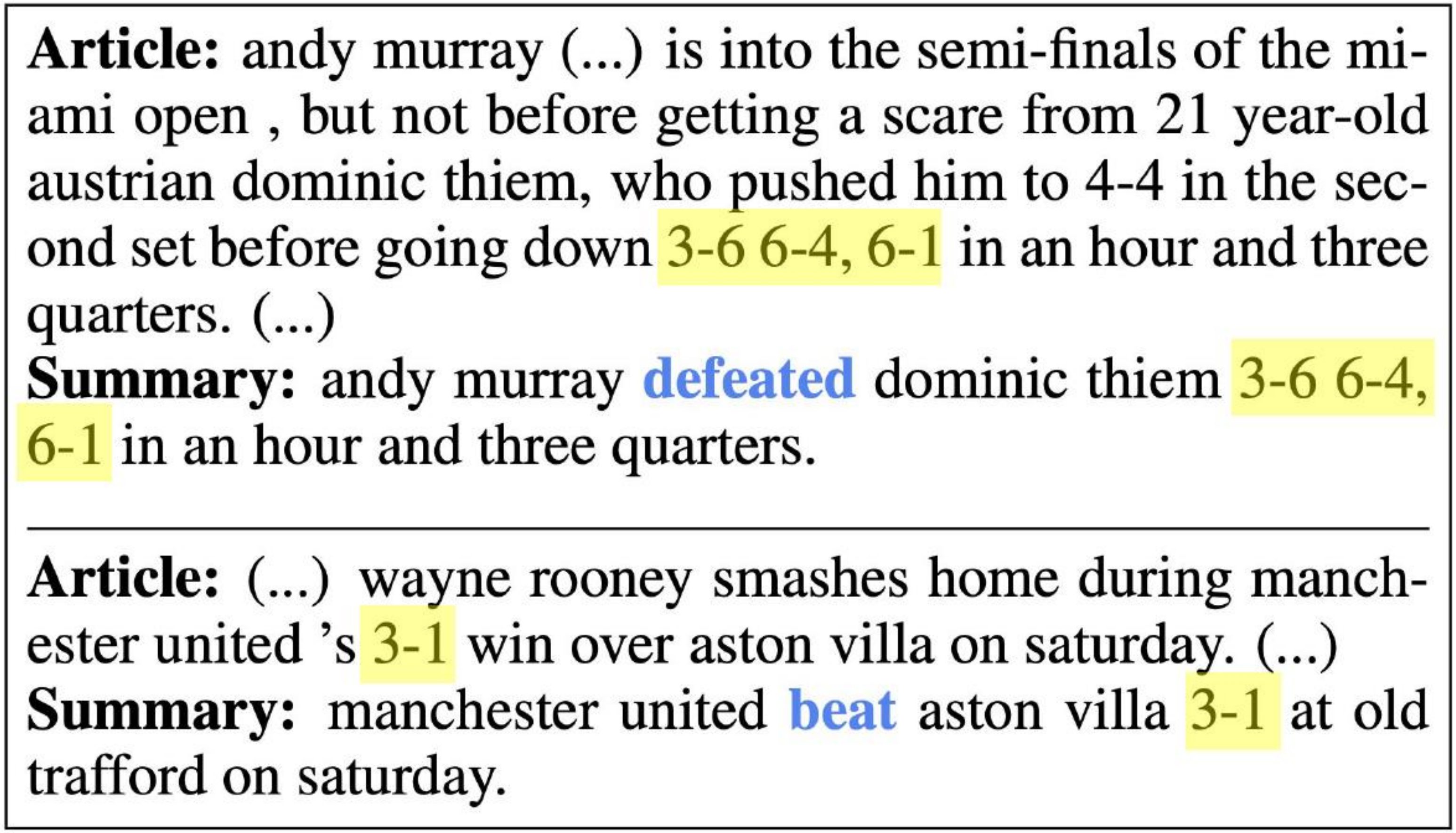
See et al, 2017, *Get To The Point: Summarization with Pointer-Generator Networks*, https://arxiv.org/pdf/1704.04368.pdf

One example of how to do a copying mechanism:

On each decoder step, calculate p_{gen}, the probability of *generating* the next word (rather than copying it). The final distribution is a mixture of the generation (aka "vocabulary") distribution, and the copying (i.e. attention) distribution:

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$$

Pointer Generator Networks: Example Output



Article: andy murray (...) is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. (...)

Summary: andy murray **defeated** dominic thiem 3-6 6-4, 6-1 in an hour and three quarters.

Article: (...) wayne rooney smashes home during manchester united 's 3-1 win over aston villa on saturday. (...)

Summary: manchester united **beat** aston villa 3-1 at old trafford on saturday.

Neural summarization: copy mechanisms

- Big problem with copying mechanisms:
  - They copy too much!
  - Mostly long phrases, sometimes even whole sentences
  - What *should* be an abstractive system collapses to a mostly extractive system.
- Another problem:
  - They're bad at overall content selection, especially if the input document is long
  - No overall strategy for selecting content

Neural summarization: better content selection

- Recall: pre-neural summarization had separate stages for **content selection** and **surface realization** (i.e. text generation)
- In a standard seq2seq+attention summarization system, these two stages are mixed in together
- On each step of the decoder (i.e. surface realization), we do word-level content selection (attention)
- This is bad: no *global* content selection strategy
- One solution: bottom-up summarization

Bottom-up summarization

- Content selection stage: Use a neural sequence-tagging model to tag words as *include* or *don't-include*
- Bottom-up attention stage: The seq2seq+attention system can't attend to words tagged *don't-include* (apply a mask)
- Simple but effective!
- Better overall content selection strategy
- Less copying of long sequences (i.e. more abstractive output)

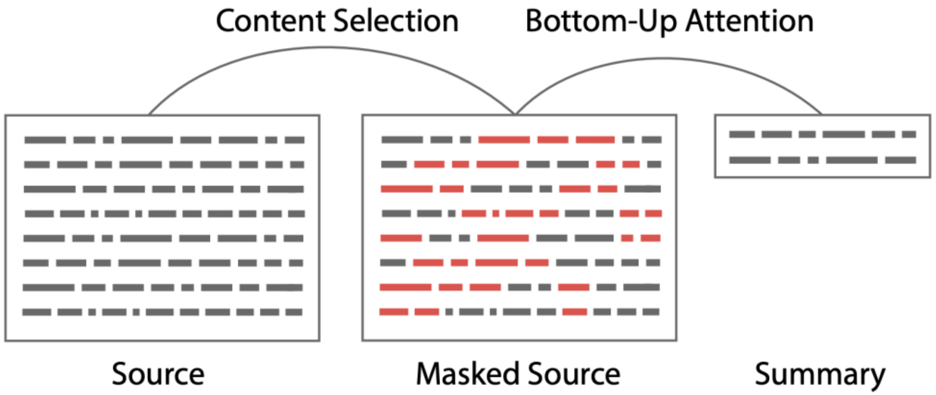


Figure 2: Overview of the selection and generation processes described throughout Section 4.

Bottom-Up Abstractive Summarization, Gehrmann et al, 2018 https://arxiv.org/pdf/1808.10792v1.pdf

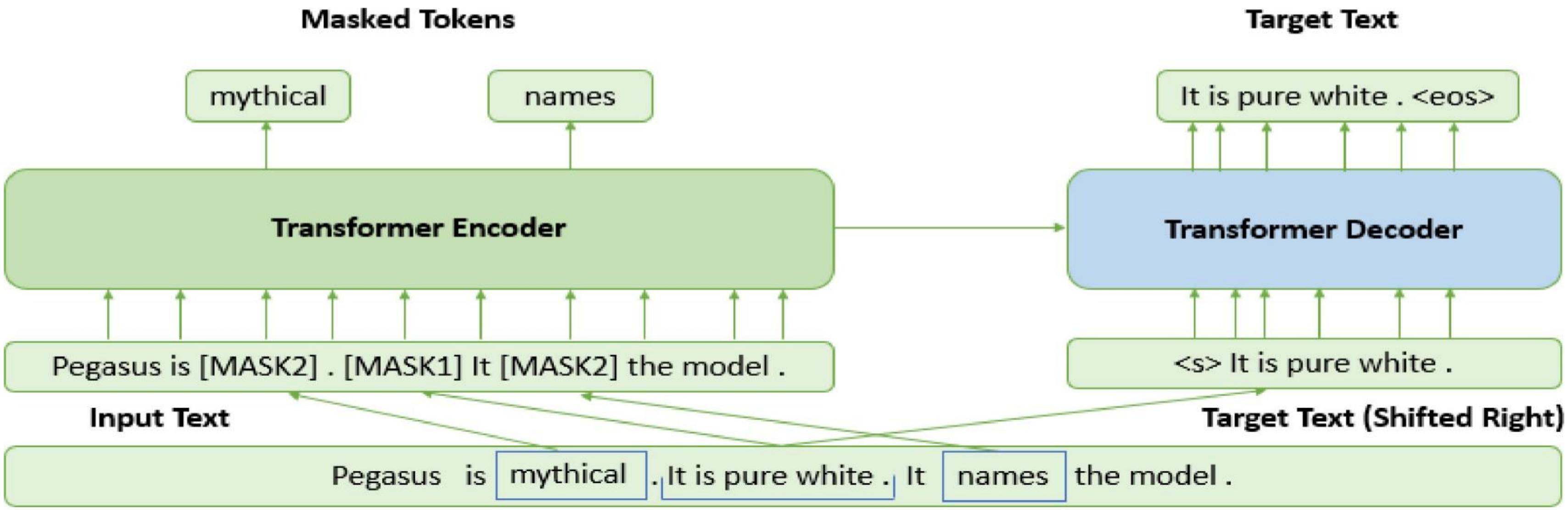
Neural summarization via Reinforcement Learning

- 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

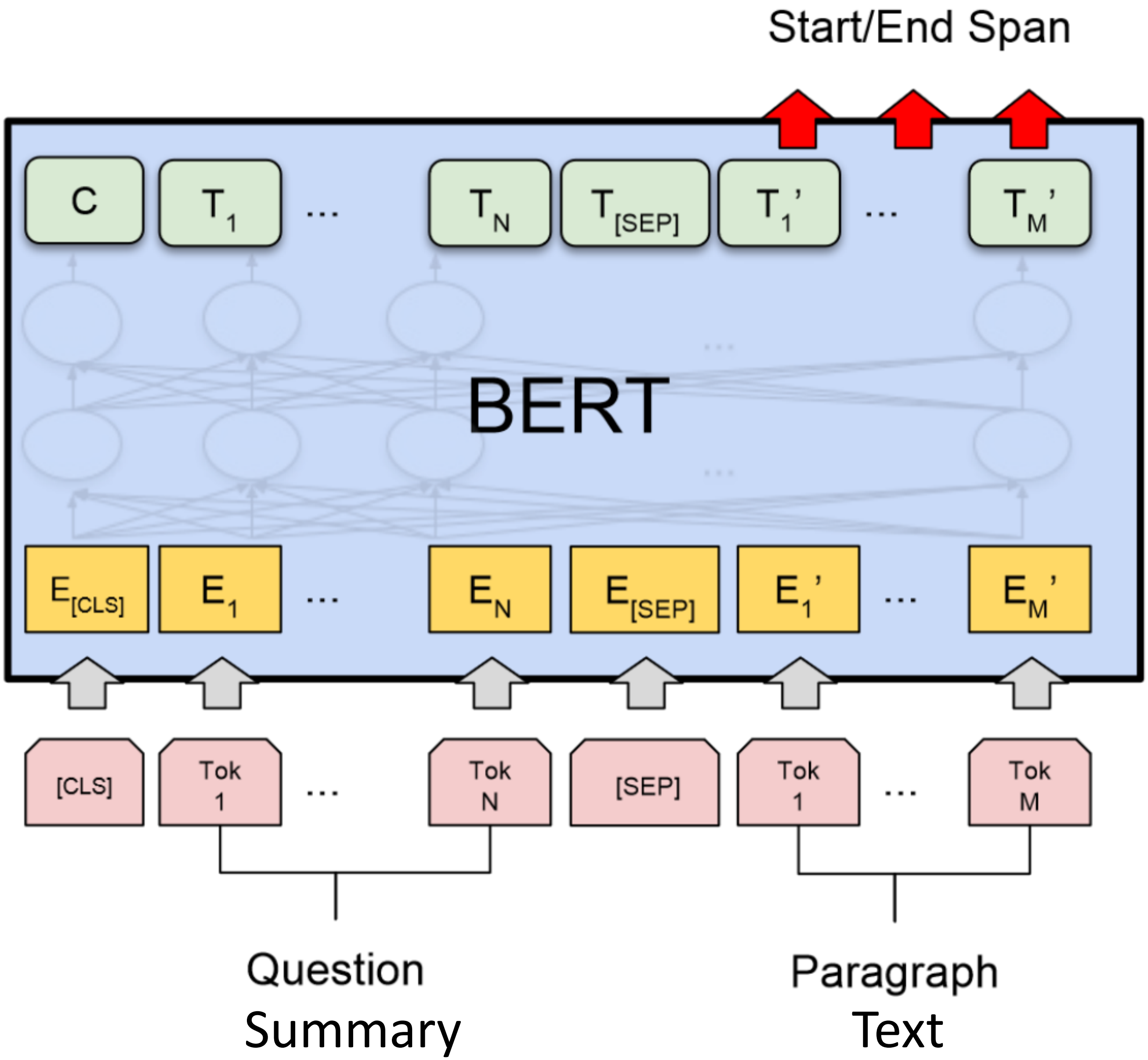
Deep Reinforced Model for Abstractive Summarization, Paulus et al, 2017 https://arxiv.org/pdf/1705.04304.pdf

Pegasus pretraining

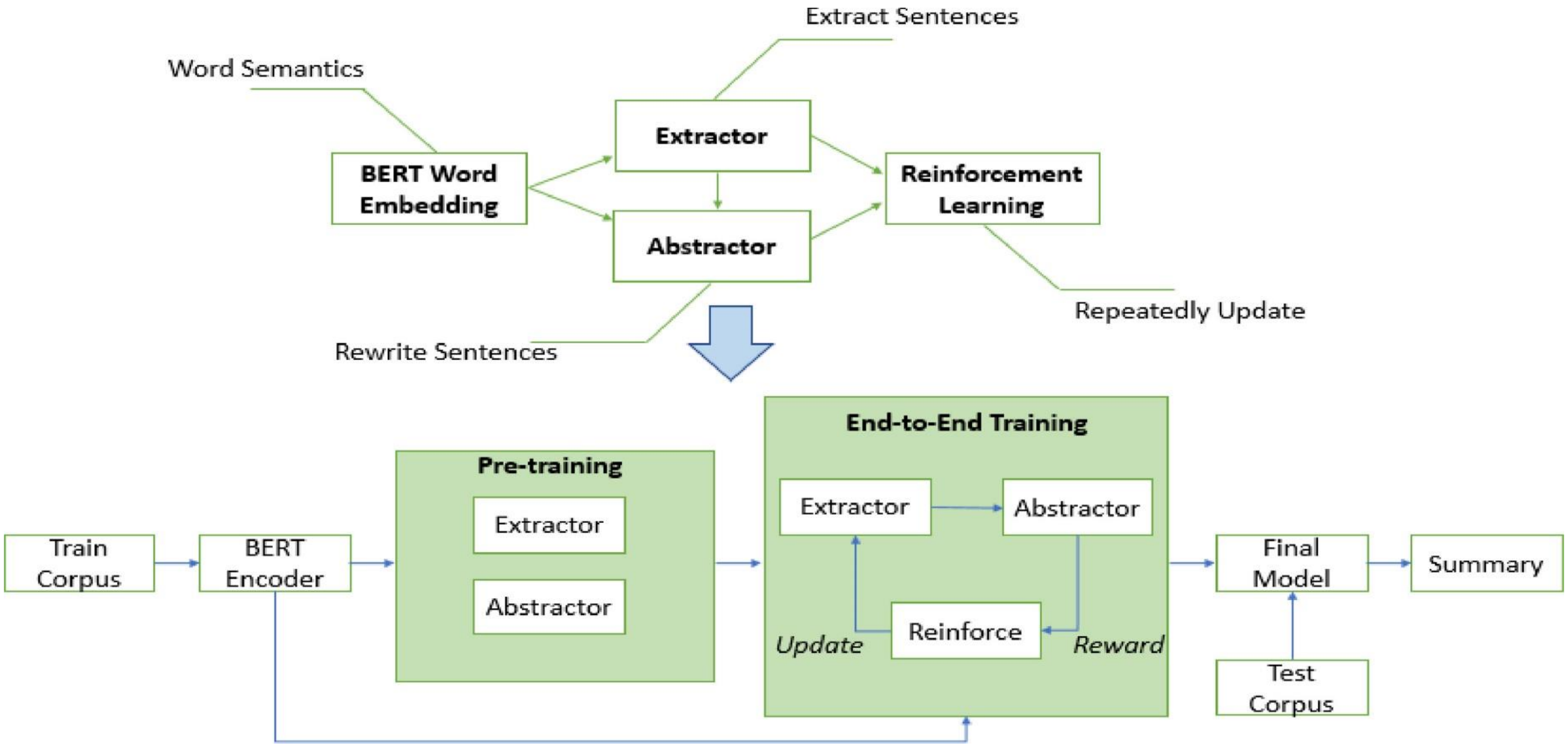
- transformer encoder-decoder model
- pre-trained on the objective of gap sentences generation
- mask important sentences from the input document to be generated as one output sequence from the remaining sentences.
- chooses the sentences based on their importance (not randomly)
- combines masked language model and the gap sentence generation


Masked Tokens
Target Text
mythical
names
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input Text
Target Text (Shifted Right)
Pegasus is mythical . It is pure white . It names the model .

Summarization / Questions and answers with BERT-like models



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

Cross-lingual summarization

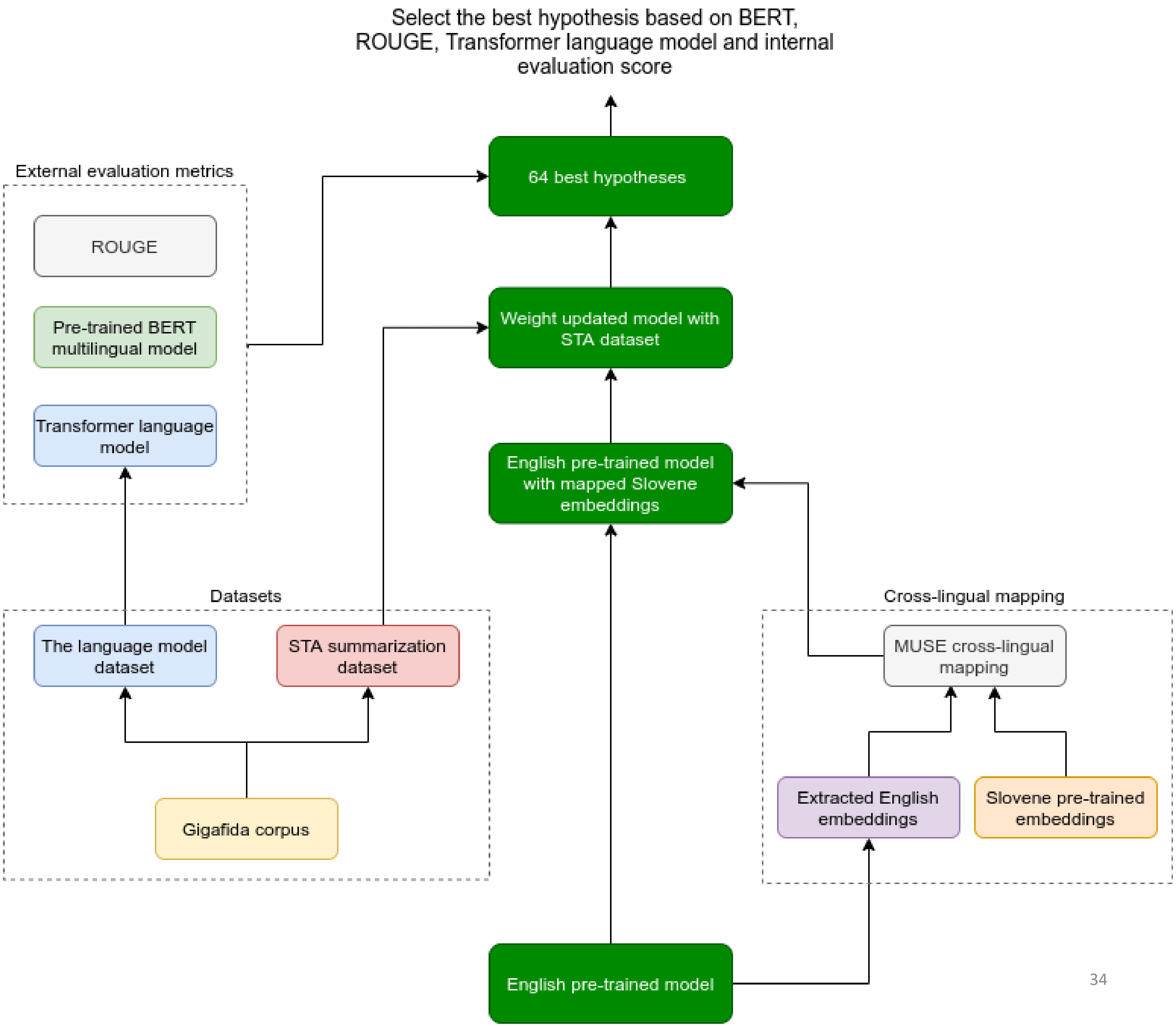
- Summarization into another language
- E.g., Hindi paper is summarized to English
- Basically the same approach

- Slovene dataset: KAS abstracts (Slovene theses, Slovene and English summaries)

Cross-lingual transfer of 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 articles of sufficient length

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 .



Select the best hypothesis based on BERT, ROUGE, Transformer language model and internal evaluation score

External evaluation metrics
ROUGE
Pre-trained BERT multilingual model
Transformer language model
64 best hypotheses
Weight updated model with STA dataset
English pre-trained model with mapped Slovene embeddings
Datasets
The language model dataset
STA summarization dataset
Gigafida corpus
Cross-lingual mapping
MUSE cross-lingual mapping
Extracted English embeddings
Slovene pre-trained embeddings
English pre-trained model

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

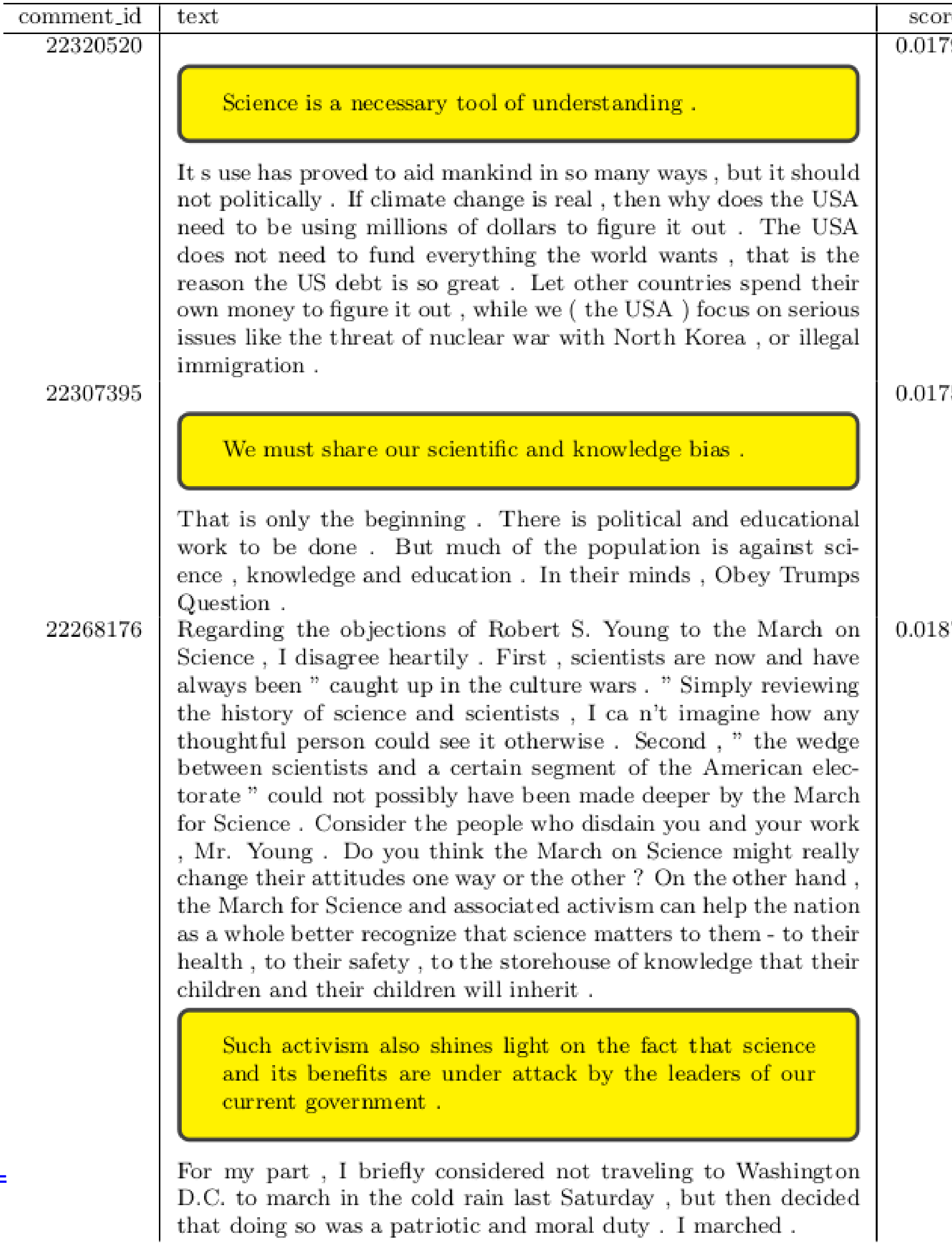
- A lot of irrelevant and deceiving comments
- Language is often informal and difficult to encode

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



comment_id	text	scor
22320520	Science is a necessary tool of understanding . It s 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

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

LaBSE sentence encoder

- LaBSE (Language-agnostic BERT Sentence Encoder)
- dual-encoder architecture, where source and target sentences (in different languages) are encoded separately using a shared BERT-based encoder
- pre-trained on masked language modeling and translated language modeling
- supports 109 languages
- allows finding similar sentences across different languages.
- loss

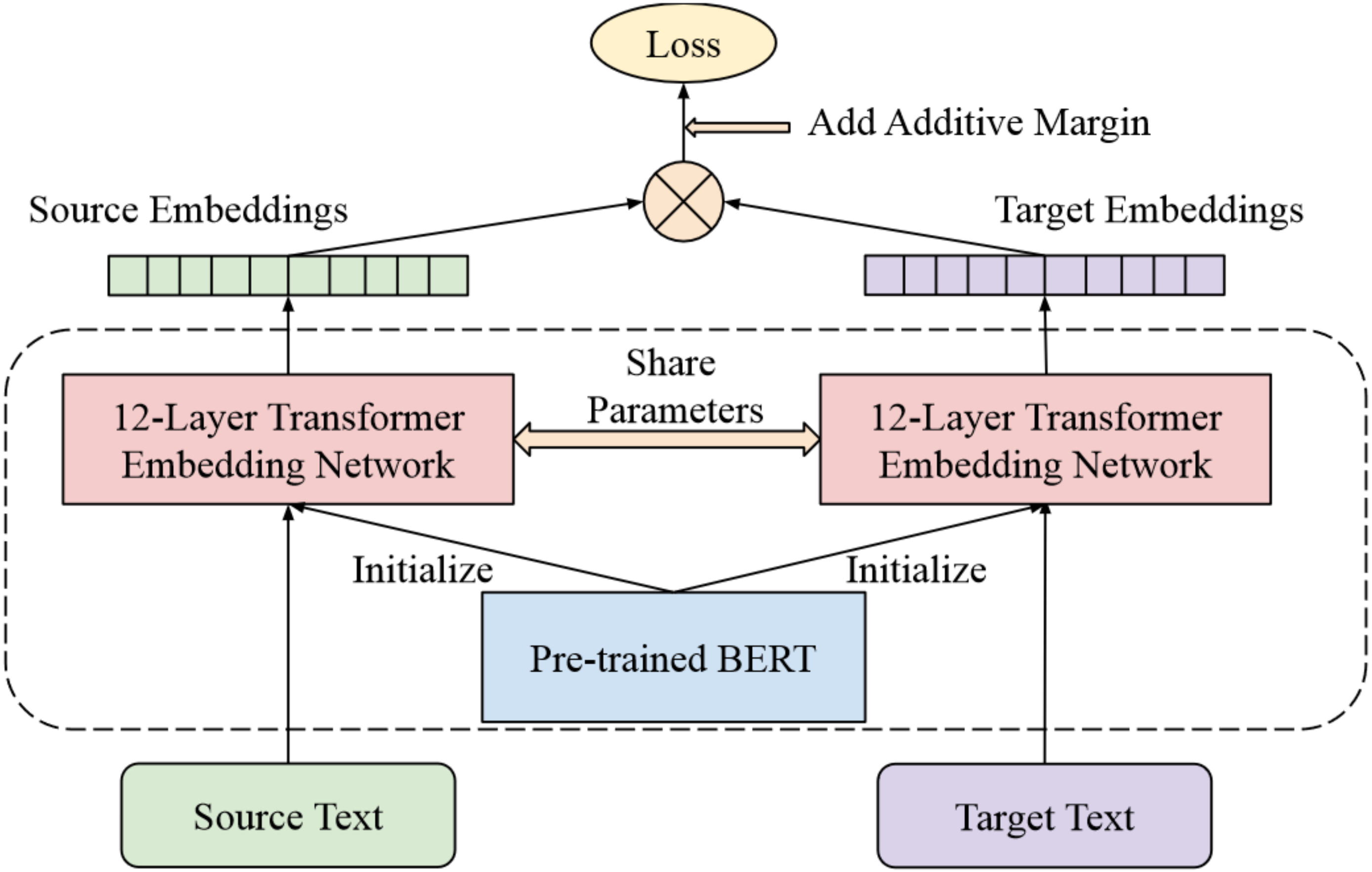
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\phi(x_i,y_i)}}{e^{\phi(x_i,y_i)} + \sum_{n=1,n\neq i}^{N} e^{\phi(x_i,y_n)}}$$

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W., 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 878-891).

https://arxiv.org/abs/2007.01852 https://tfhub.dev/google/LaBSE

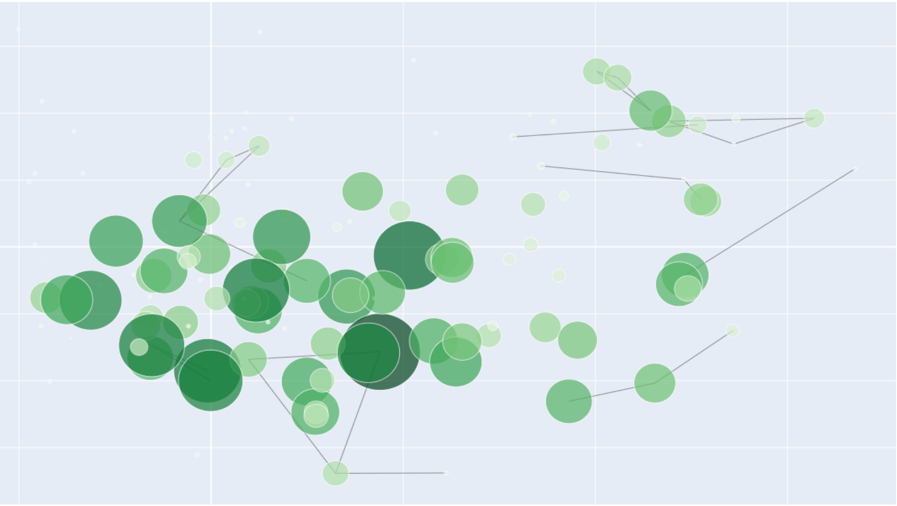
LaBSE architecture

- Dual encoder model with BERT based encoding modules.

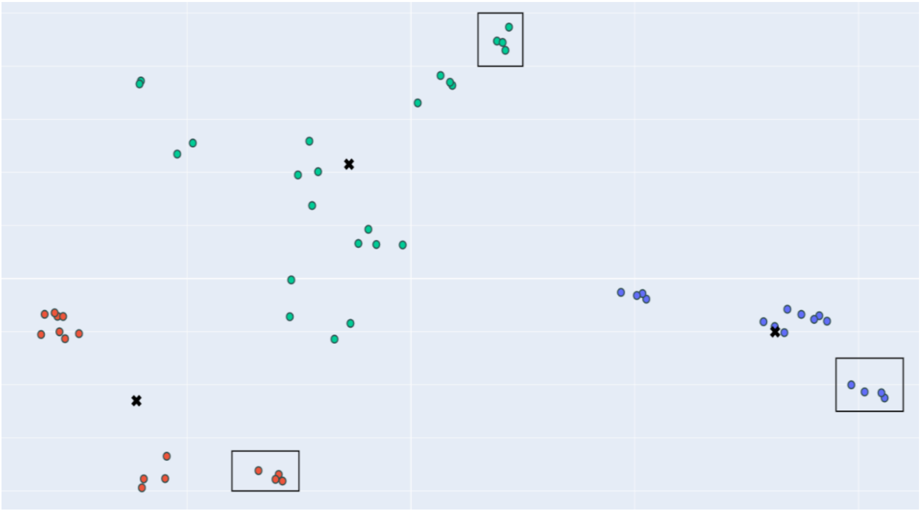


Visual tools to investigate results

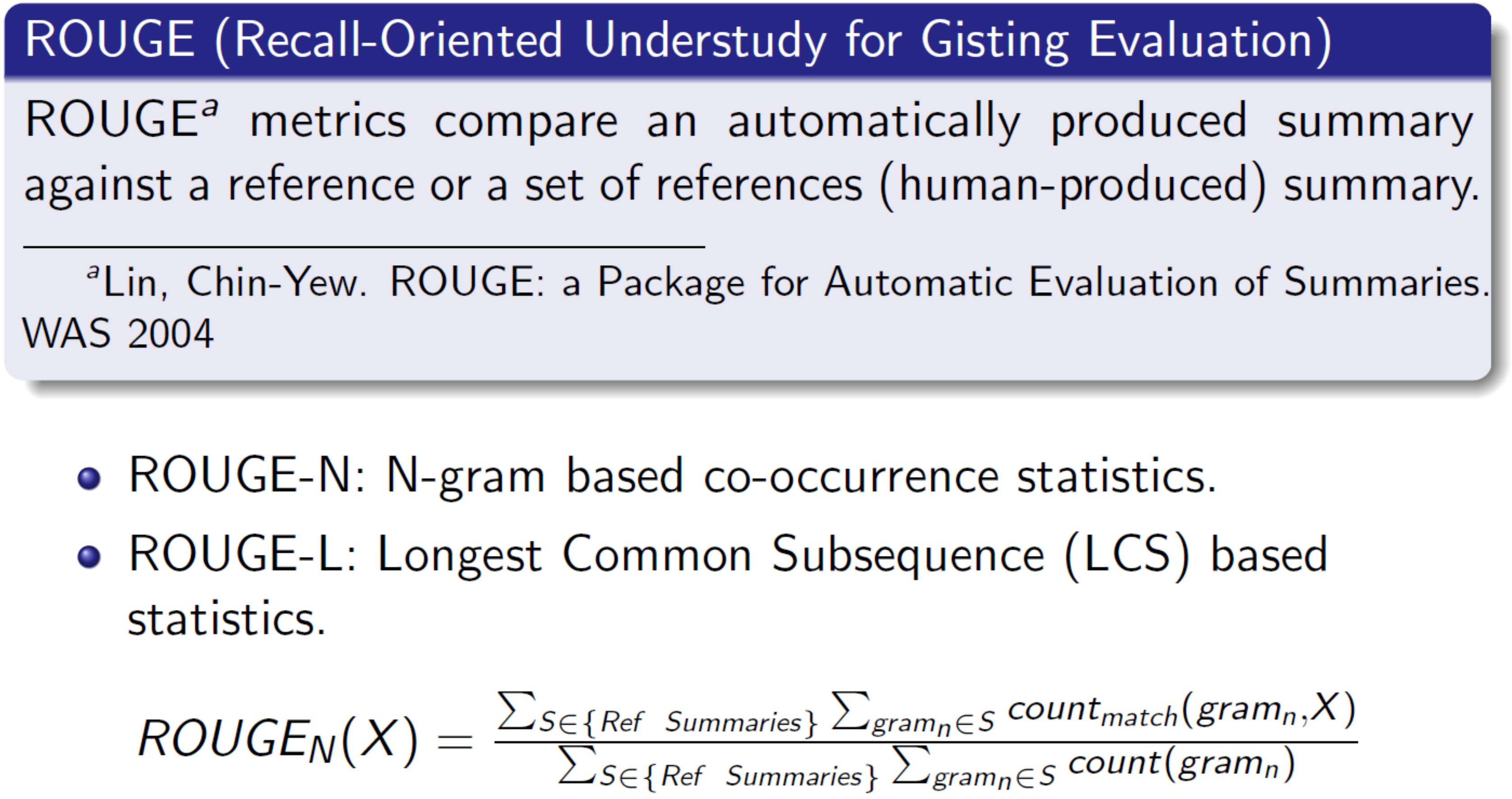
Graph-based



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.

[a] Lin, 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.

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

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.

- ROUGE-2 = $\frac{3+3+6}{10+9+9} = \frac{12}{28} = 0.43$

Evaluation metric: BERTScore

- idea: use pretrained BERT for matching tokens instead of ngrams
- calculate the token representations and similarity measures between tokens of two texts.
- use a pre-trained BERT model to generate the contextual token representations of the words in the candidate x and reference $\hat{x}$ sentences. In the next step, we calculate pairwise cosine similarity between the words and use greedy matching to maximize the similarity scores of recall, precision, and F1:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j,$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j,$$

$$F_{BERT} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}.$$

Summarization challenges

- Meaning representation and construction
- Long text abstractive summarization

¸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

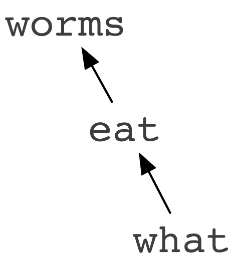
Question Answering

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

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

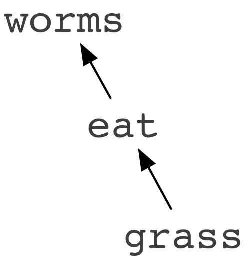
Question:

What do worms eat?

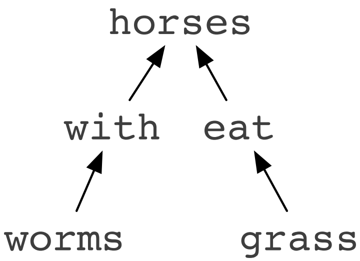


Potential Answers:

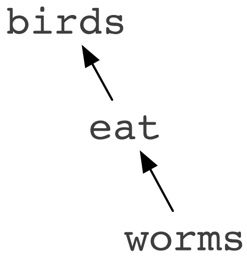
Worms eat grass



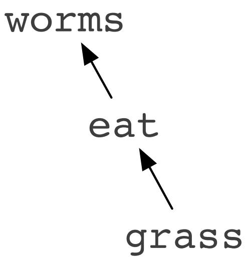
Horses with worms eat grass



Birds eat worms

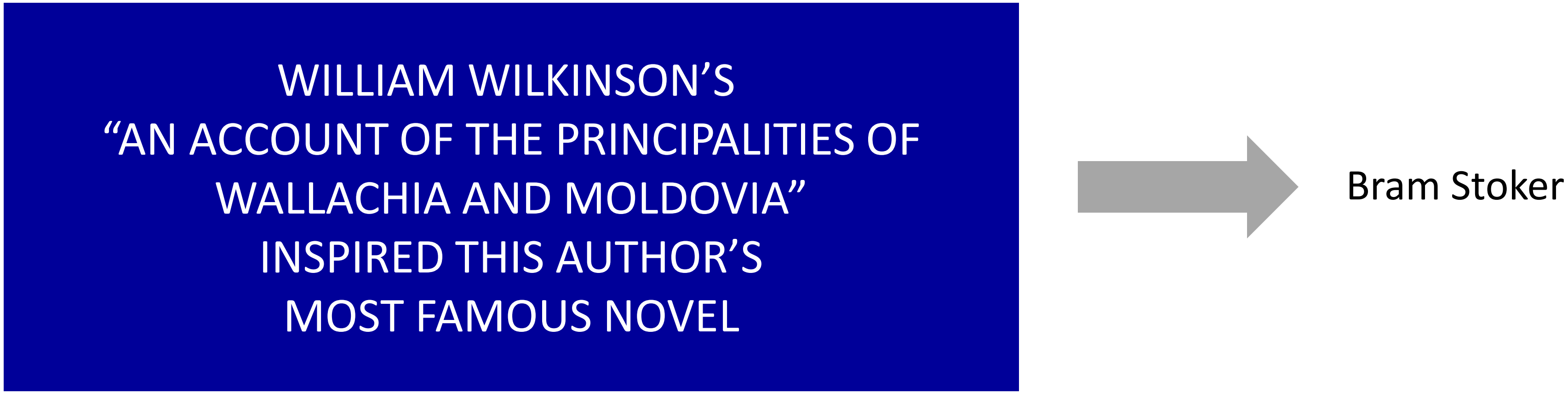


Grass is eaten by worms



Question Answering: IBM's Watson

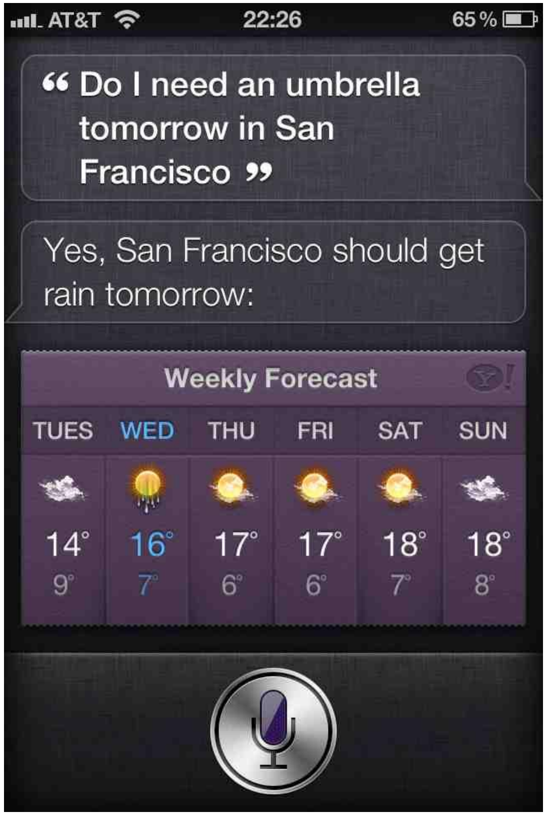
- Won Jeopardy on February 16, 2011!



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

Bram Stoker

Apple’s Siri



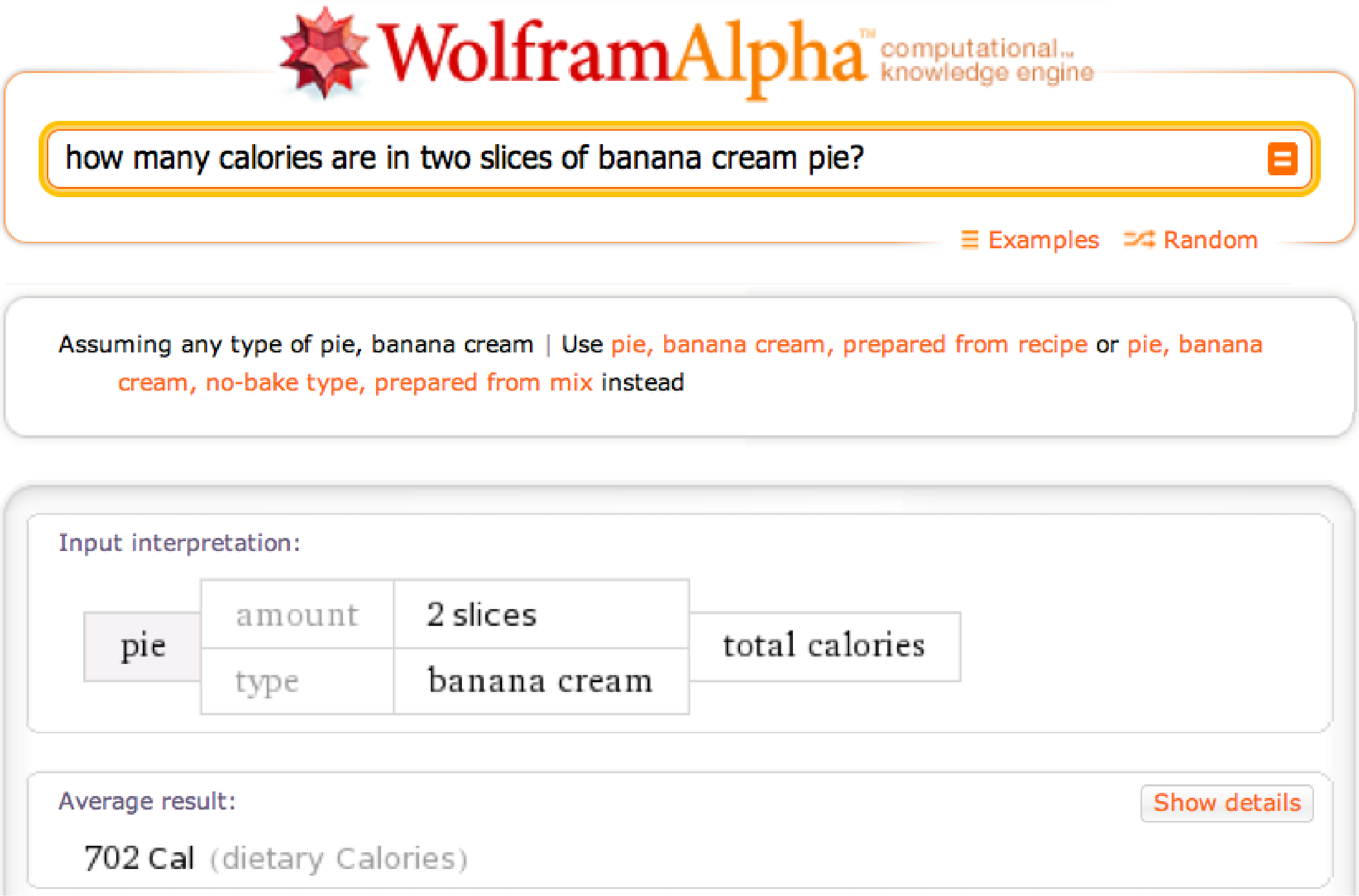
“ Do I need an umbrella tomorrow in San Francisco ”

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°

Wolfram Alpha



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

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:

702 Cal (dietary Calories)

Types of Questions in Modern Systems

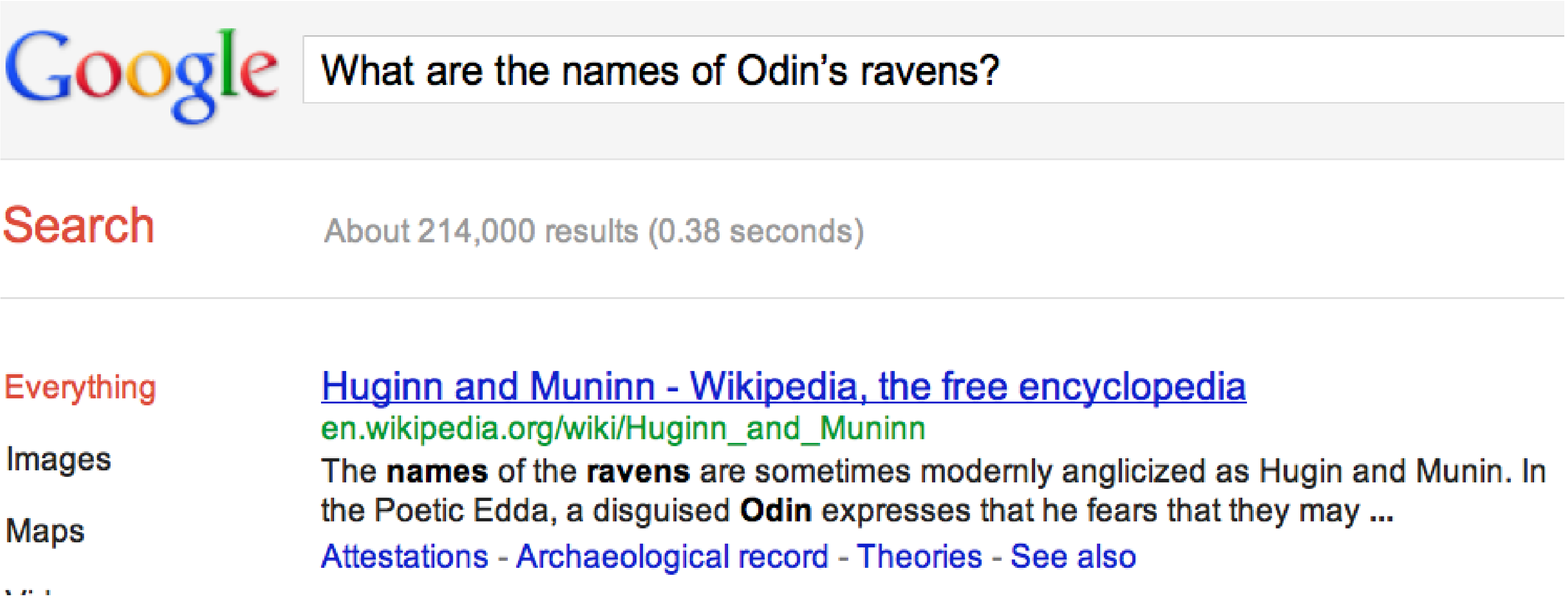
- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile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

Commercial systems: mainly factoid questions

Where is the Louvre Museum located?	In Paris, France
What’s the abbreviation for limited partnership?	L.P.
What are the names of Odin’s ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for Stanford University?	650-723-2300

Many questions can already be answered by web search

- a

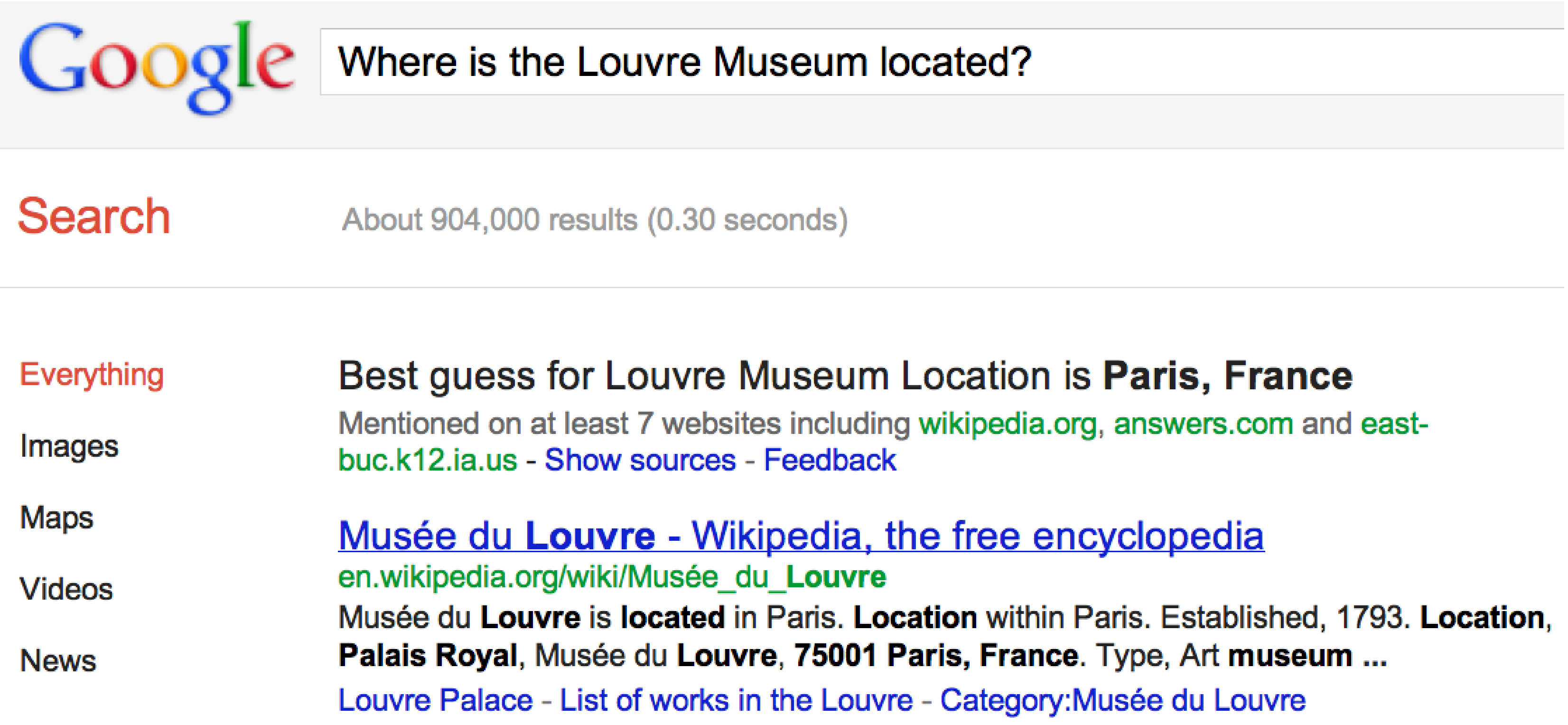


What are the names of Odin's ravens?

About 214,000 results (0.38 seconds)

Huginn and Muninn - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Huginn_and_Muninn
The **names** of the **ravens** are sometimes modernly anglicized as Hugin and Munin. In the Poetic Edda, a disguised **Odin** expresses that he fears that they may **...**
Attestations - Archaeological record - Theories - See also

IR-based Question Answering



Where is the Louvre Museum located?

About 904,000 results (0.30 seconds)

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

Musée du **Louvre** - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Musée_du_**Louvre**

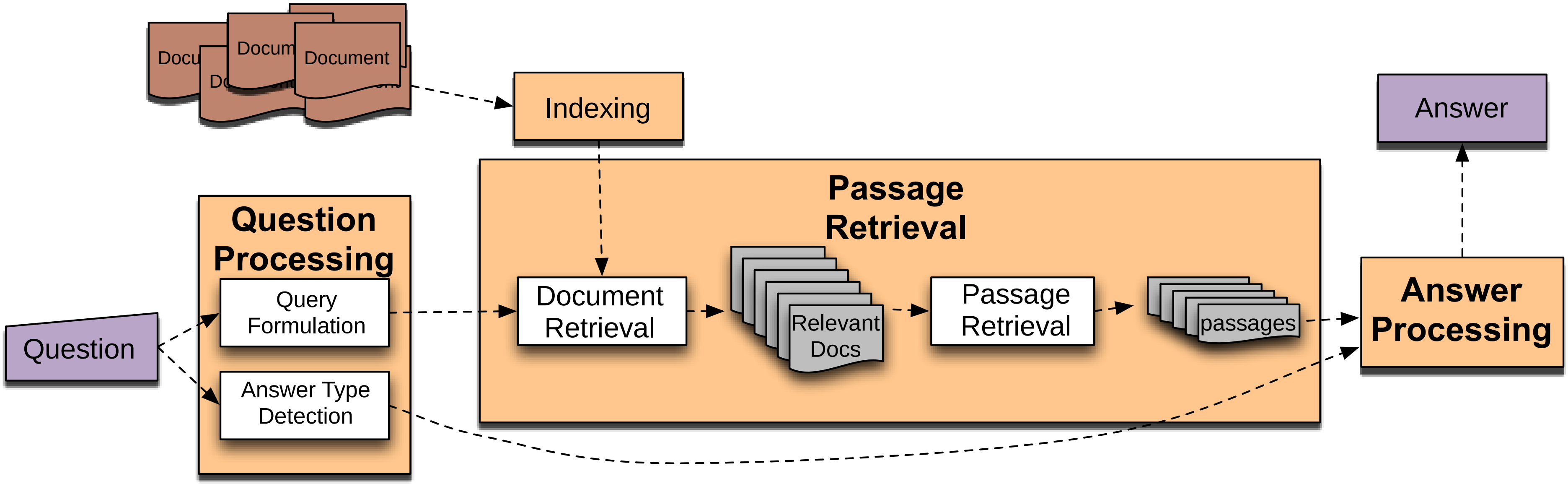
Musée du **Louvre** is **located** in Paris. **Location** within Paris. Established, 1793. **Location**, **Palais Royal**, Musée du **Louvre**, **75001 Paris, France**. Type, Art **museum ...**

Louvre Palace - List of works in the Louvre - Category:Musée du Louvre

Generative large language models

- Superior performance of very large models
- ChatGPT, GPT-4
- LLaMa, Alpaka, Koala

IR-based Factoid QA



Obtaining relevant context

- Still a relevant tasks, even for LLMs
- The context can constitute a part of the prompt to LLM
- Well-known approaches
  - BM25 (Best match 25)
  - DPR (Dense Passage Retrieval)

Ranking documents with BM25

- Okapi BM25 (Best match 25)
- uses bag-of-words document representation, works similarly to tf-idf weighting
- Given a query Q, with words q_1,..., q_n the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

- $f(q_i,D)$ is the number of times that q_i occurs in D,
- avgdl is the average document length in the text collection
- k_1 and b are parameters, usually chosen from $k_1 \in [\ 1.2\ ,\ 2.0\ ]$ and b = 0.75

IDF variant

- IDF (inverse document frequency) weights the query term q_i

$$\mathrm{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

- where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i

DPR retrieval

- BERT based passage retrieval
- ranks passages in the document collection relative to query q using dot product similarity
- BERT is additionally pretrained to maximize the similarity between q and correct passages and minimize the similarity between q and wrong passages using the loss:

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$
$$= -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}.$$

- passages and query are encoded with modified BERT (using the CLS token representation)
- works better than BM25

Karpukhin et al (2020) Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781.

Common Evaluation Metrics

1. *Accuracy* (does answer match gold-labeled answer?)
2. *Mean Reciprocal Rank*
   - For each query return a ranked list of M candidate answers.
   - Query score is 1/Rank of the first correct answer
     - *If first answer is correct: 1*
     - *else if second answer is correct: ½*
     - *else if third answer is correct: ⅓, etc.*
     - *Score is 0 if none of the M answers are correct*
   - Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

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

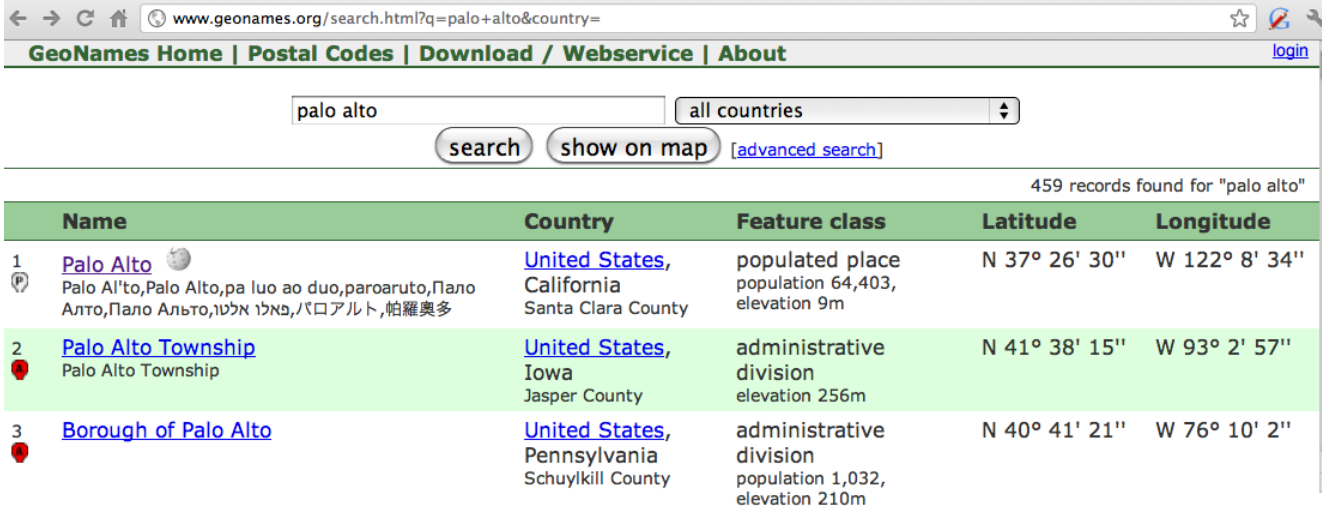
  ”In 1594 he took a job as a tax collector in Andalusia”

  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:



www.geonames.org/search.html?q=palo+alto&country=

GeoNames Home | Postal Codes | Download / Webservice | About

login

palo alto

all countries

search

show on map

[advanced search]

459 records found for "palo alto"

	Name	Country	Feature class	Latitude	Longitude
1	Palo Alto Palo Al'to,Palo Alto,pa luo ao duo,paroaruto,Пало Алто,Пало Альто,פאלו אלטו,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:

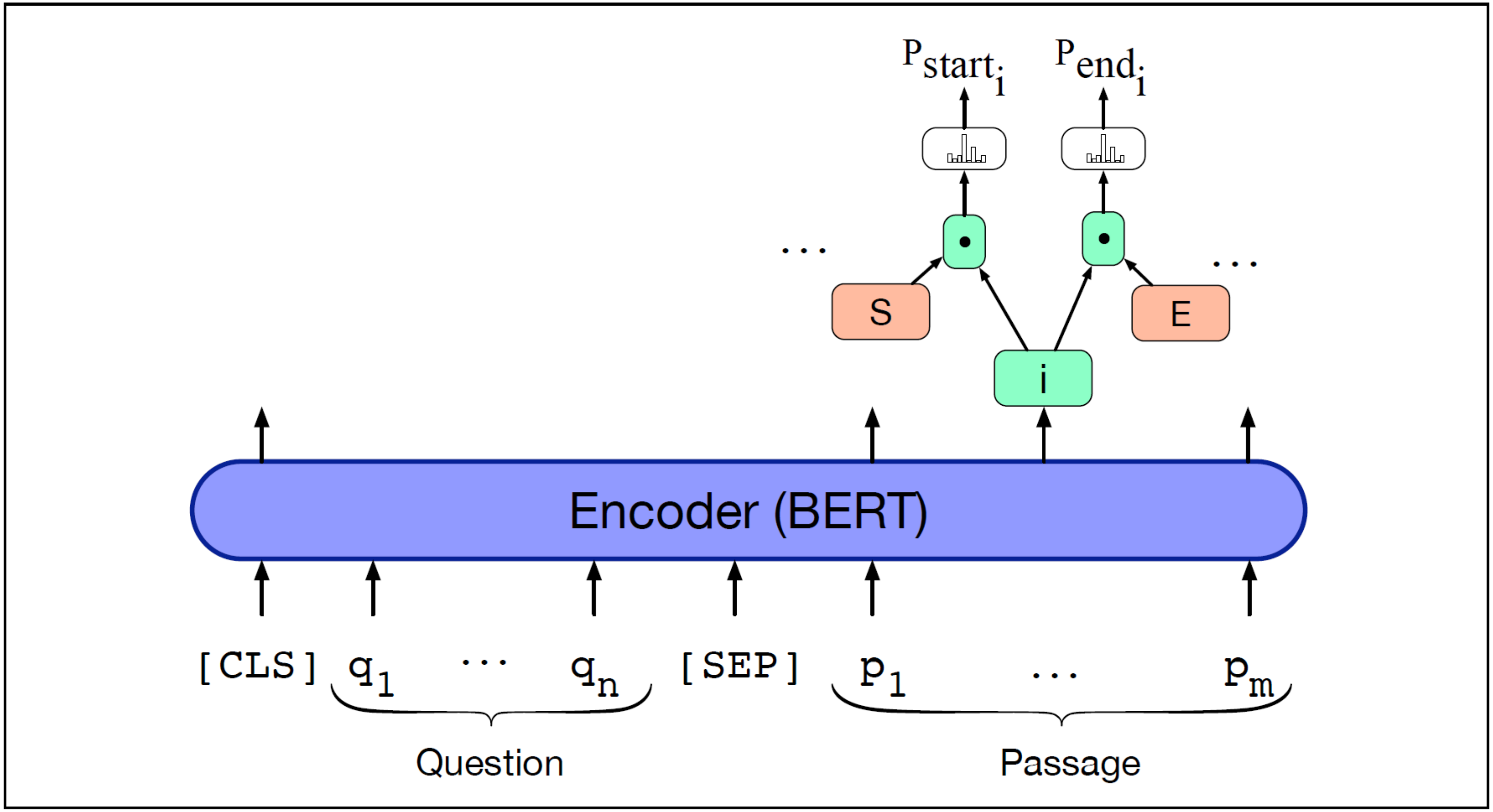
  U: “Chicago pizza”

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

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,… q_n and a passage *p* of *m* tokens p_1, … 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



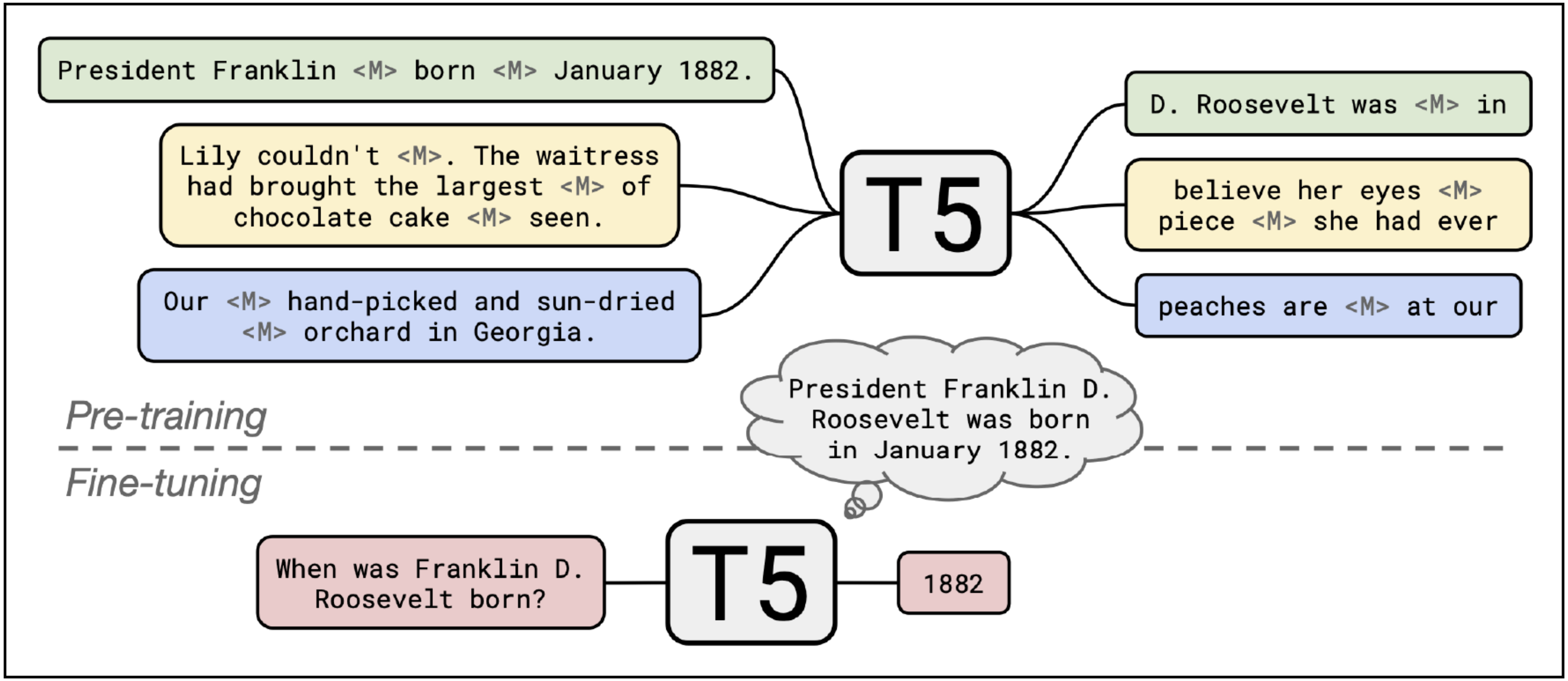
QA with language models

- a pretrained language model tries to answer a question solely from information stored in its parameters
- E.g., use the T5 language model, which is an encoder-decoder transformer model pretrained to fill in masked spans of task
- Even very large language models still suffer from certain problems in QA:
  - hallucinations
  - poor interpretability (addressed with chain-of-thought reasoning)
  - cannot give more context (e.g., a passage with the answer)

Roberts, A., C. Raffel, and N. Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? Proceedings of EMNLP 2020.

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.


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
Pre-training
Fine-tuning
President Franklin D. Roosevelt was born in January 1882.
When was Franklin D. Roosevelt born?
T5
1882

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 Electorial 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: WSC

- WSC (Winograd Schema Challenge, Levesque et al., 2012) is a coreference resolution task in which examples consist of a sentence with a pronoun and a list of noun phrases from the sentence. The system must determine the correct referrent of the pronoun from among the provided choices. Winograd schemas are designed to require everyday knowledge and commonsense reasoning to solve. The test examples are derived from fiction books.
- Text: Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.
- Coreference: False

QA datasets: WiC

- WiC (Word-in-Context, Pilehvar and Camacho-Collados, 2019) is a word sense disambiguation task cast as binary classification of sentence pairs. Given two text snippets and a polysemous word that appears in both sentences, the task is to determine whether the word is used with the same sense in both sentences. Sentences are drawn from WordNet, VerbNet, and Wiktionary.
- Context 1: Room and board.
- Context 2: He nailed boards across the windows.
- Sense match: False

QA datasets: CB

- CB (CommitmentBank, de Marneffe et al., 2019) is a corpus of short texts in which at least one sentence contains an embedded clause. Each of these embedded clauses is annotated with the degree to which it appears the person who wrote the text is committed to the truth of the clause. The resulting task framed as three-class textual entailment on examples that are drawn from the Wall Street Journal, fiction from the British National Corpus, and Switchboard. Each example consists of a premise containing an embedded clause and the corresponding hypothesis is the extraction of that clause. The inter-annotator agreement is above 80%.
- Text: B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?
- Hypothesis: they are setting a trend
- Entailment: Unknown

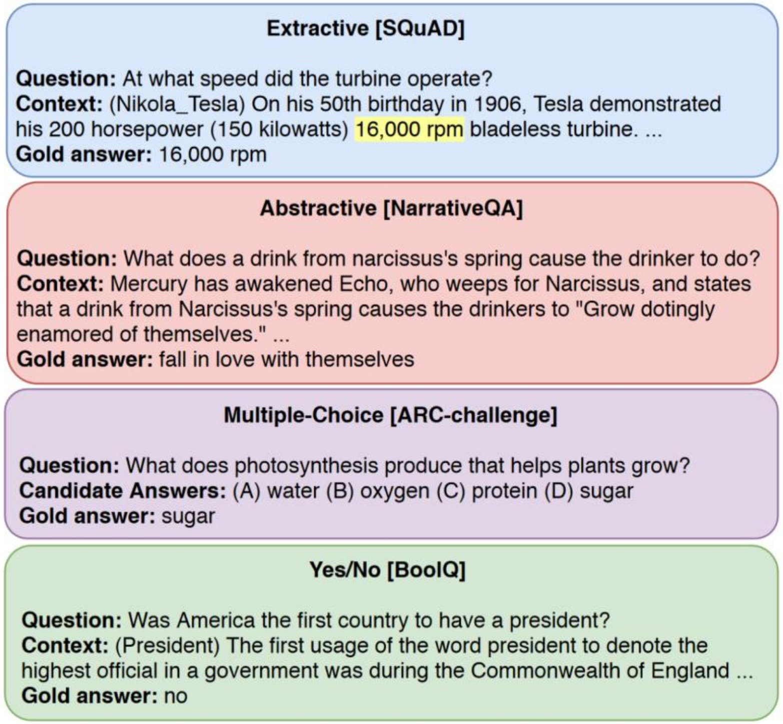
QA datasets: RiddleSense

- 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

Unified QA

- Use several types of questions in T5 model to generate answers: extractive, abstractive, multichoice, yes/no
- A model is trained on all types of questions,
- Finetuned tuned on a specific type of questions



Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Khashabi, Min, Khot, Sabharwal, Tafjord, Clark in
Hajishirzi. UnifiedQA: Crossing Format Boundaries With a Single QA System. V *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, str. 1896–1907, 2020.

Unified QA in Slovene

- Use partially human partially human translations of English QA datasets to Slovene (mostly taken from Slovene SuperGLUE benchmark)
- use SloT5 model and mT5 model
- quantitatively slightly worse than English model
- qualitative analysis:
  - the generated answers are mostly substrings or given choices in multiple-choice questions
  - models cannot paraphrase, rephrase or provide answers in the correct Slovene case
  - problems with multi-part questions requiring multiple answers that are not listed in the same place in the context
  - machine translations are not always grammatically correct or do not make it clear what the question is asking for
  - best performance on factoid questions that require a short answer

- Ulčar, M., and Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.932519
- Žagar, A., & Robnik-Šikonja, M. (2022). Slove ne SuperGLUE Benchmark: Translation and Evaluation. Proceedings of LREC 2022.
- Logar, K. and Robnik-Šikonja, M. (2022) Unified Question Answering in Slovene. Proceedings of IS 2022: Slovene Artificial Inteligence Conference, SCAI 2022.