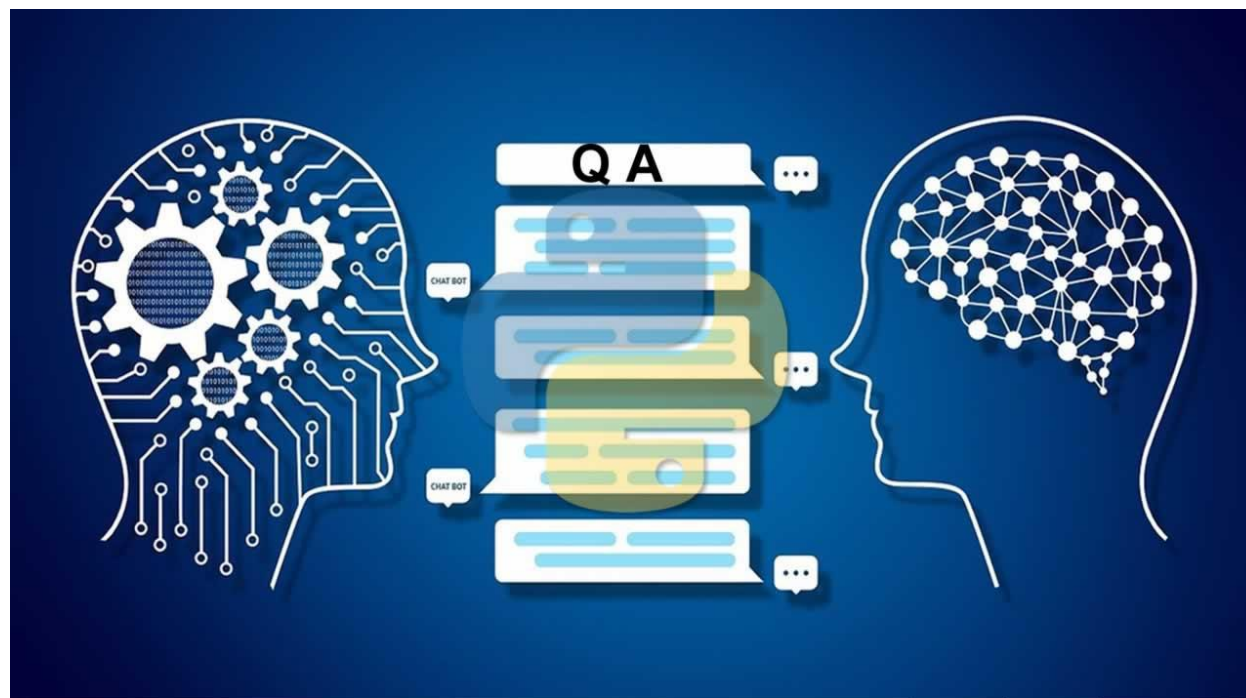


# Automatic summarization and question answering



Prof Dr Marko Robnik-Šikonja

Natural Language Processing, Edition 2022

# Contents

- summarization
- question answering

- Literature:

Jurafsky and Martin, 3<sup>rd</sup> edition

Some slide taken from Jurafsky

# Text summarization

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

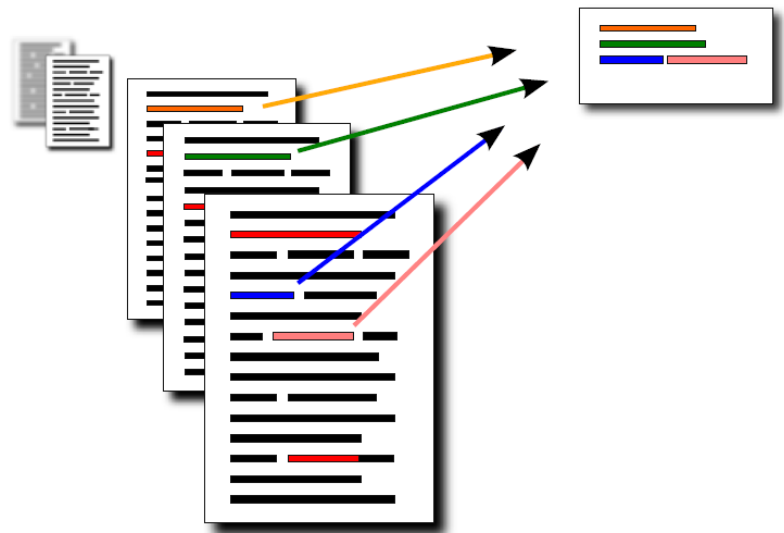
Clay Shirky



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?



# 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

# 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

- abstract
- outline
- headline

# 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:
    - 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



# Summarization for Question Answering: Snippets

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



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_learning_model))

[Transformer \(machine learning model\) - Wikipedia](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model))



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

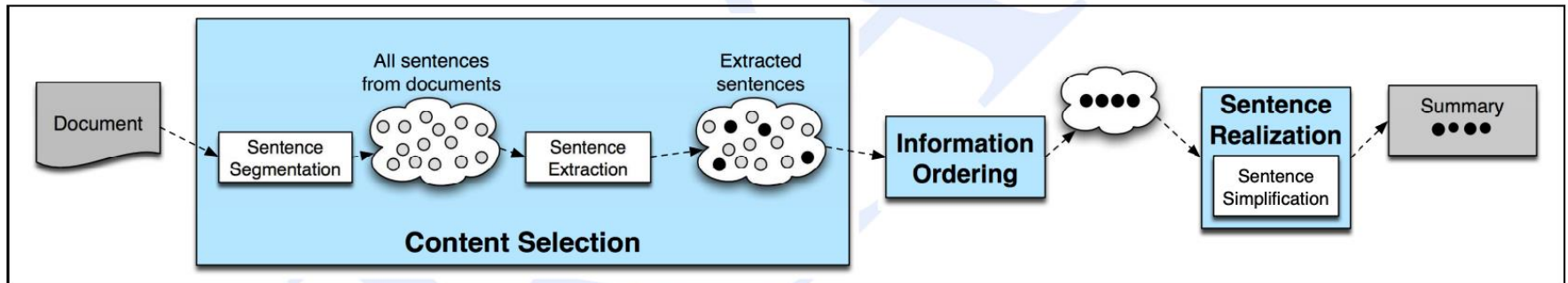
- Within single-document summarization, there are datasets with source documents of different lengths and styles:
  - Gigaword: first one or two sentences of a news article → headline (aka *sentence compression*)
  - LCSTS (Chinese microblogging): paragraph → sentence summary
  - NYT, CNN/DailyMail: 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
- 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 → version written for children

# 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.

# 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

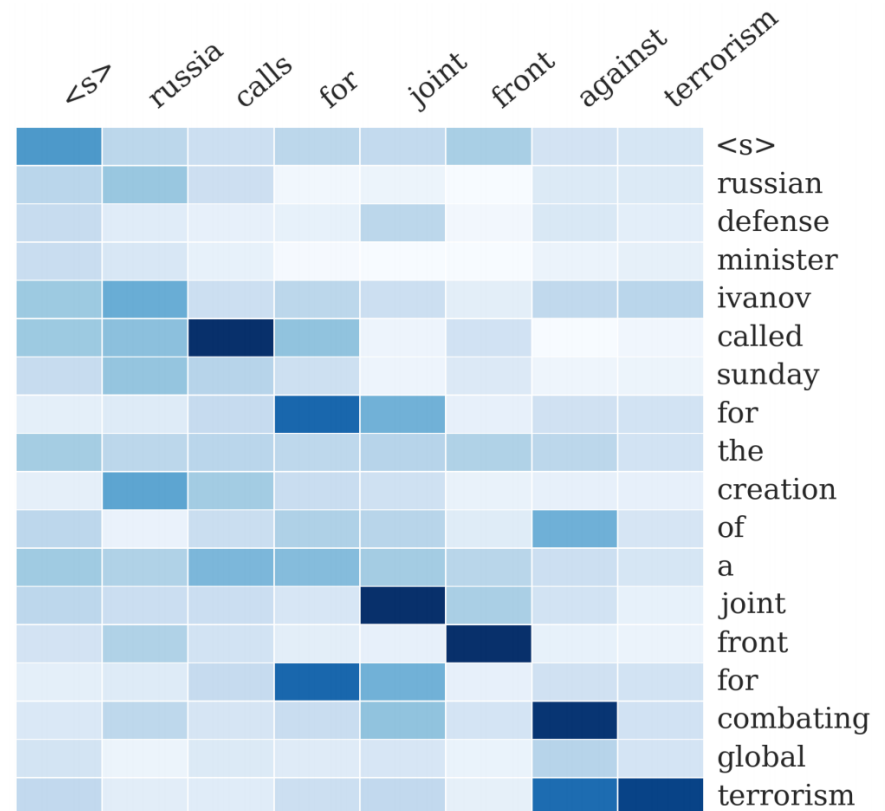
- Given:
  - A labeled training set of good summaries for each document
- Align:
  - The sentences in the document with sentences in the summary
- Extract features
  - position (first sentence?)
  - length of sentence
  - word informativeness, cue phrases
  - cohesion
- Train
  - A binary classifier (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

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



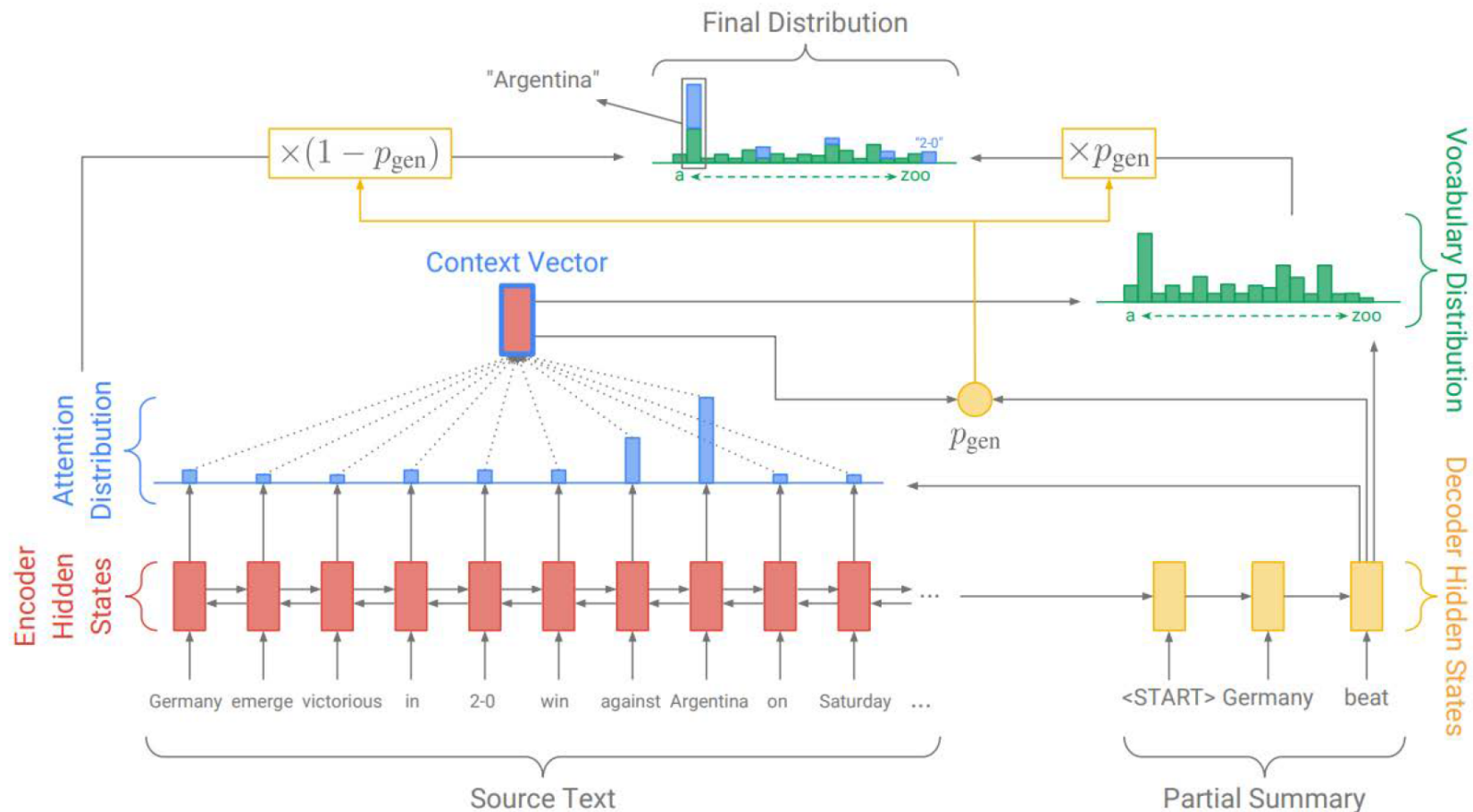
# Neural summarization developments

- 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>
- *A Survey on Neural Network-Based Summarization Methods*, Dong, 2018 <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: 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:

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

# 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)

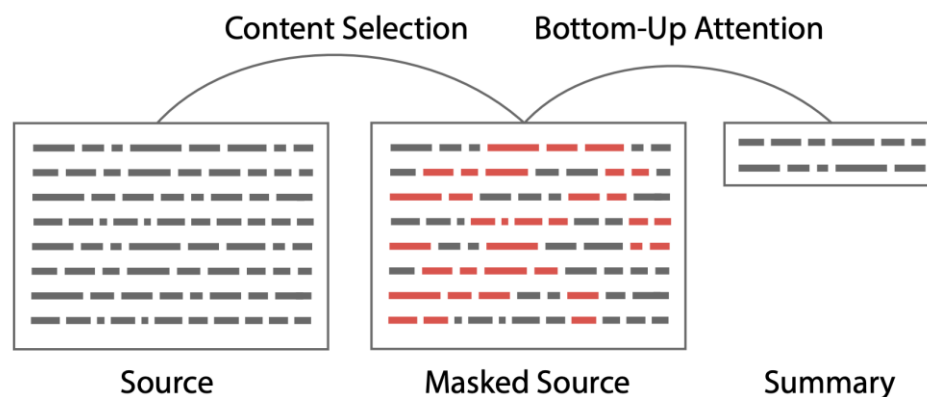


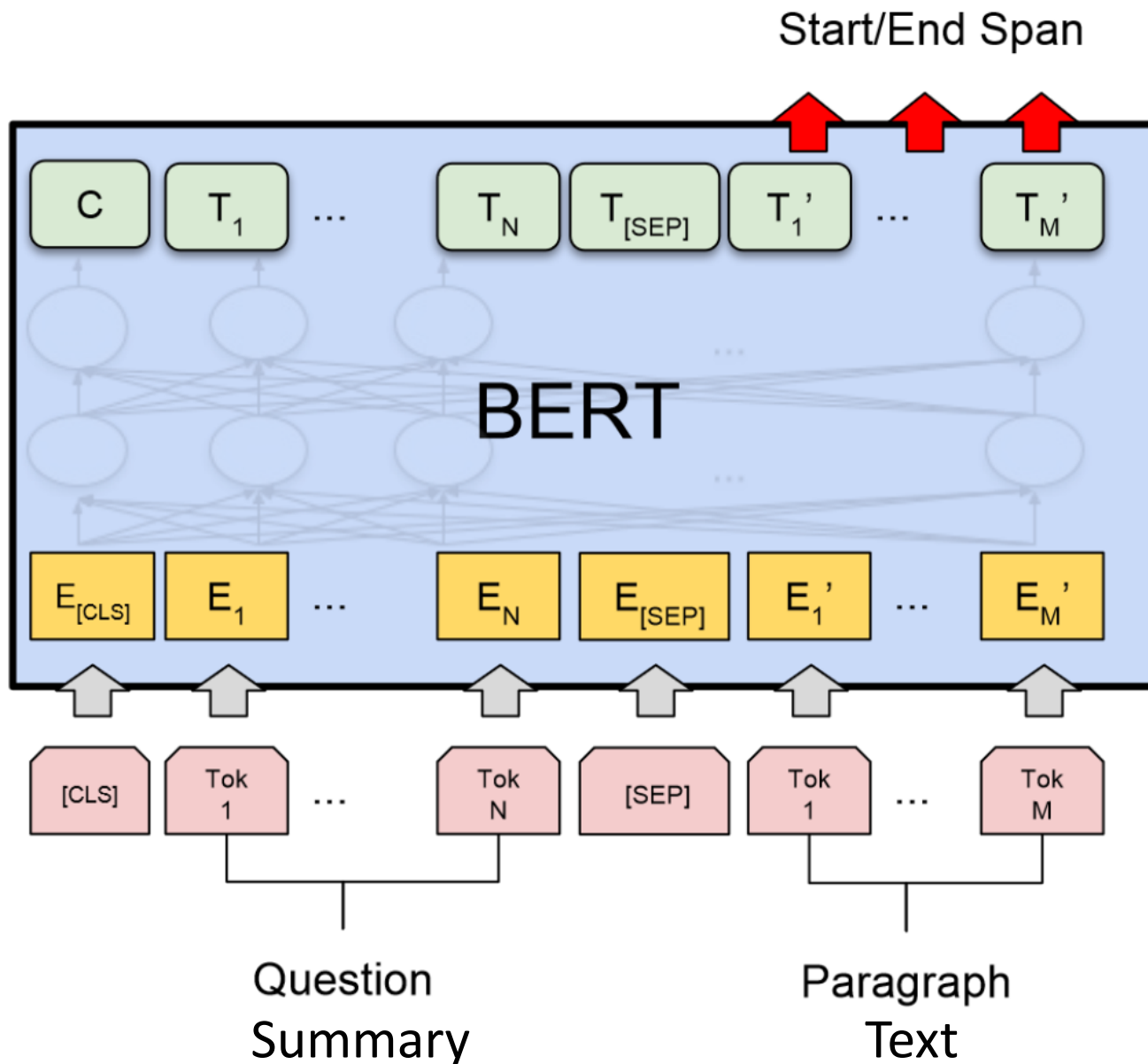
Figure 2: Overview of the selection and generation processes described throughout Section 4.



# 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

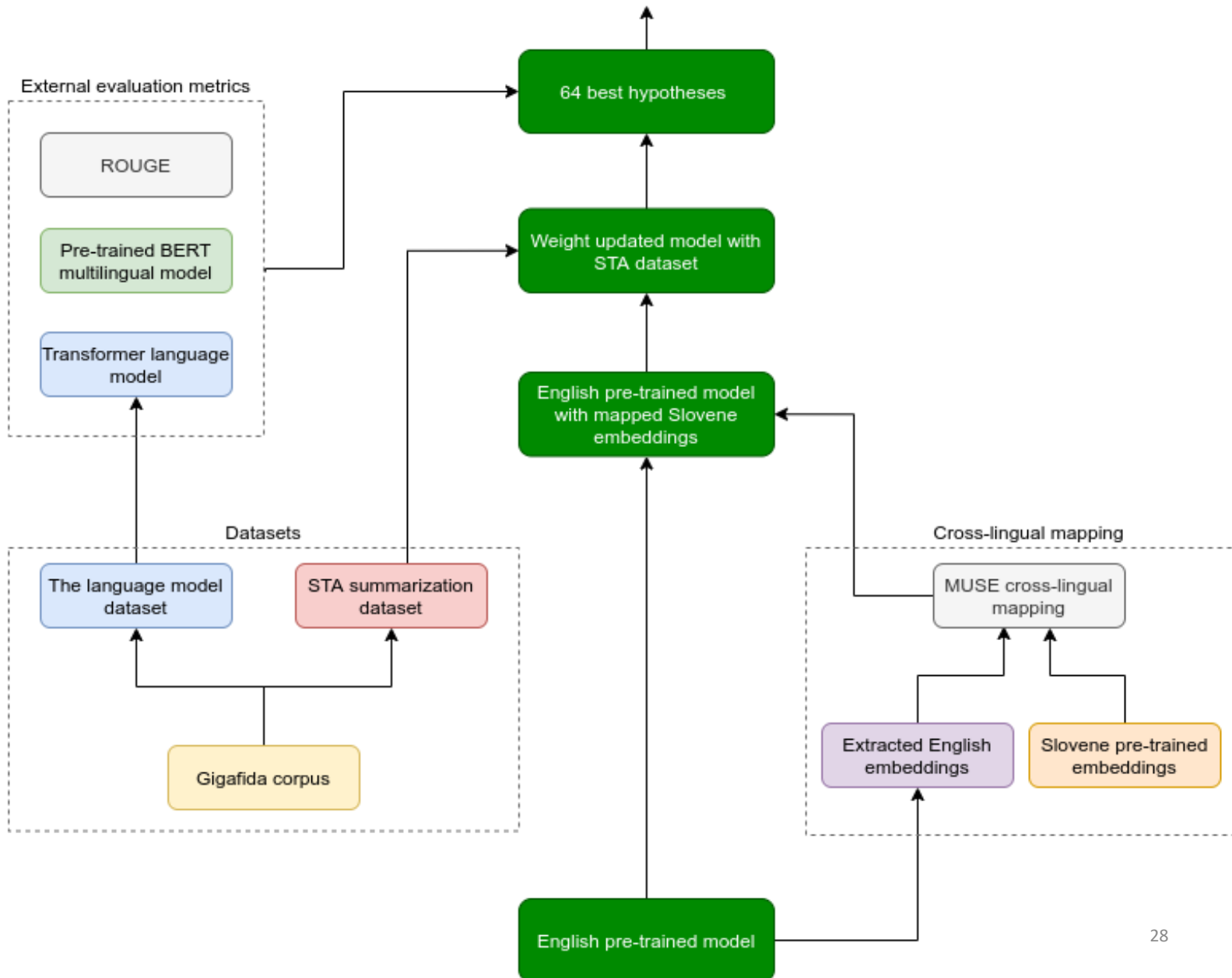
# Summarization / Questions and answers with BERT-like models



# Cross-lingual approach to summarization

- Aleš Žagar, Marko Robnik-Šikonja (2021) Cross-lingual Approach to Abstractive Summarization. <https://arxiv.org/abs/2012.04307> .
- 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

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



# 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

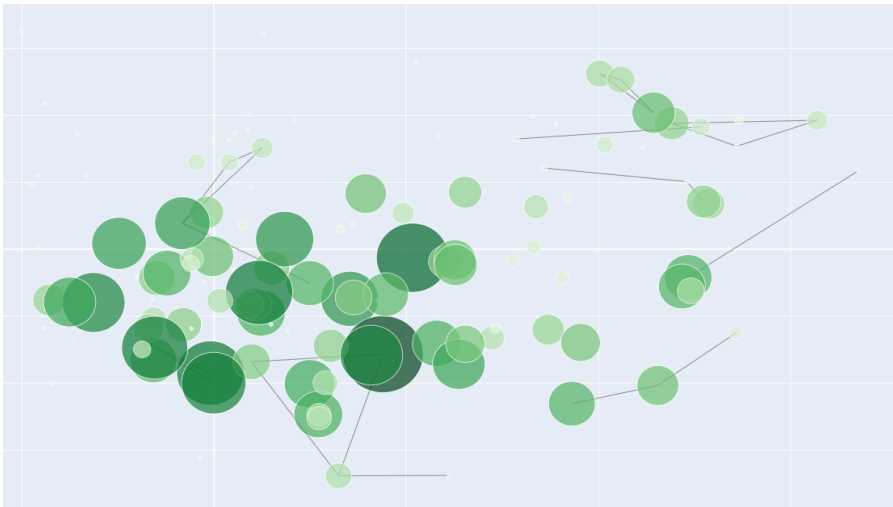
- Extractive approach based on graph-methods and clustering

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](#)

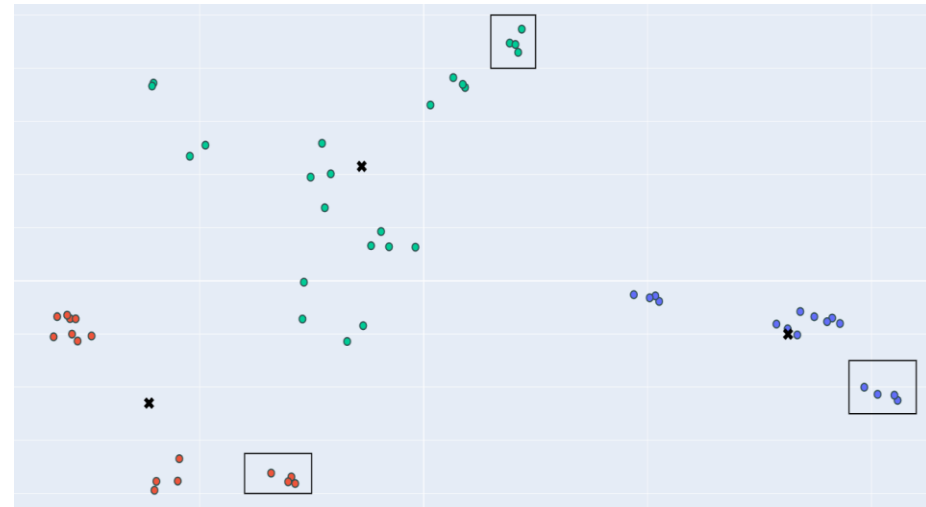
comment_id	text	score
22320520	<p>Science is a necessary tool of understanding .</p> <p>Its use has proved to aid mankind in so many ways , but it should not politically . If climate change is real , then why does the USA need to be using millions of dollars to figure it out . The USA does not need to 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 .</p>	0.017
22307395	<p>We must share our scientific and knowledge bias .</p> <p>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 .</p>	0.017
22268176	<p>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 .</p> <p>Such activism also shines light on the fact that science and its benefits are under attack by the leaders of our current government .</p> <p>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 .</p>	0.018

# Visual tools to investigate results

Graph-based



Clustering



# Evaluation metric: ROUGE

## ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE<sup>a</sup> metrics compare an automatically produced summary against a reference or a set of references (human-produced) summary.

---

<sup>a</sup>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.
- $\text{ROUGE-2} = \frac{3+3+6}{10+9+9} = \frac{12}{28} = 0.43$

# 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 (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)

Question:

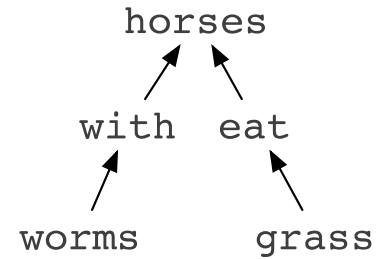
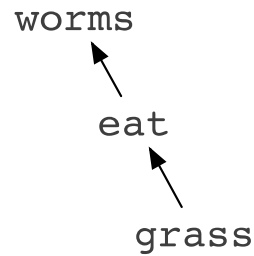
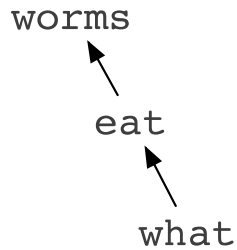
Potential Answers:

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

What do worms eat?

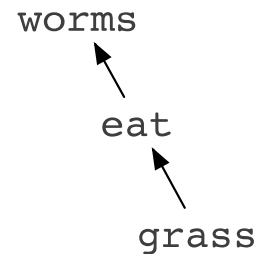
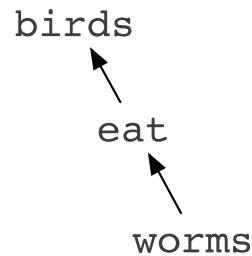
Worms eat grass

Horses with worms eat grass



Birds eat worms

Grass is eaten by worms



# 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





# 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)

# Types of Questions in Modern Systems

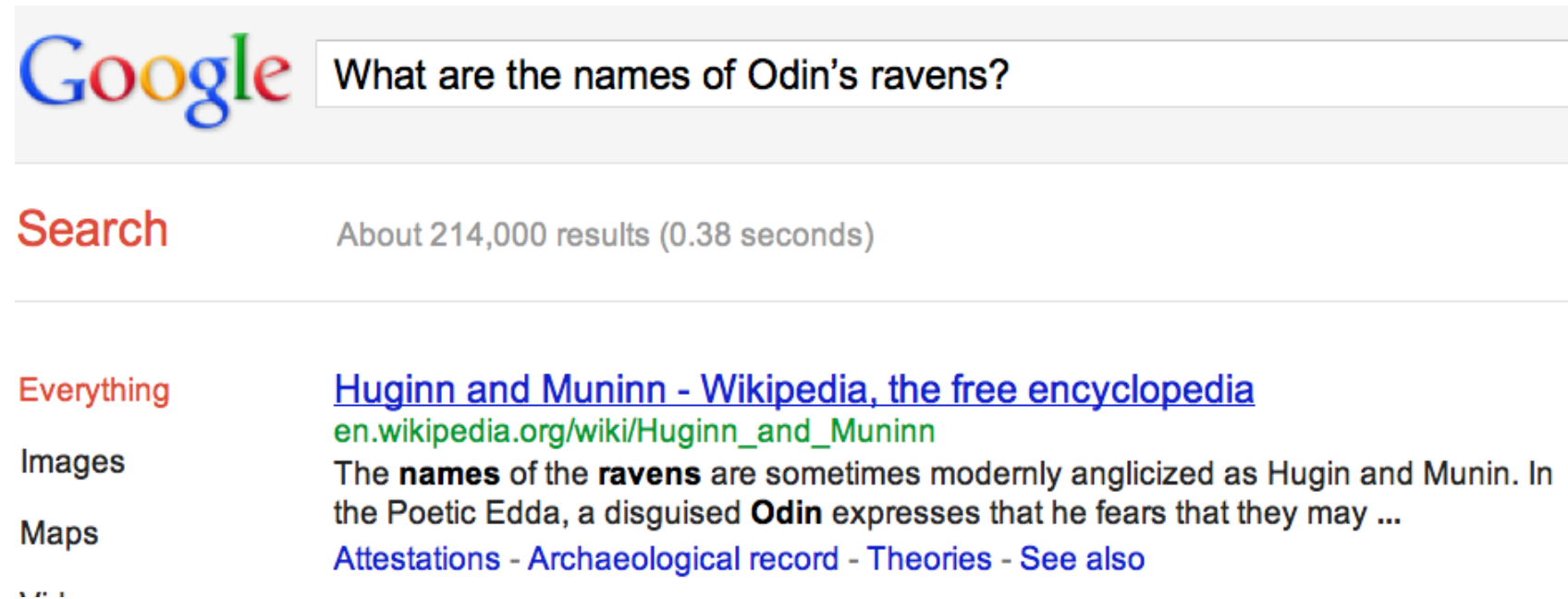
- Factoid questions
  - *Who wrote “The Universal Declaration of Human Rights”?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile 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



Google

**Search** About 214,000 results (0.38 seconds)

---

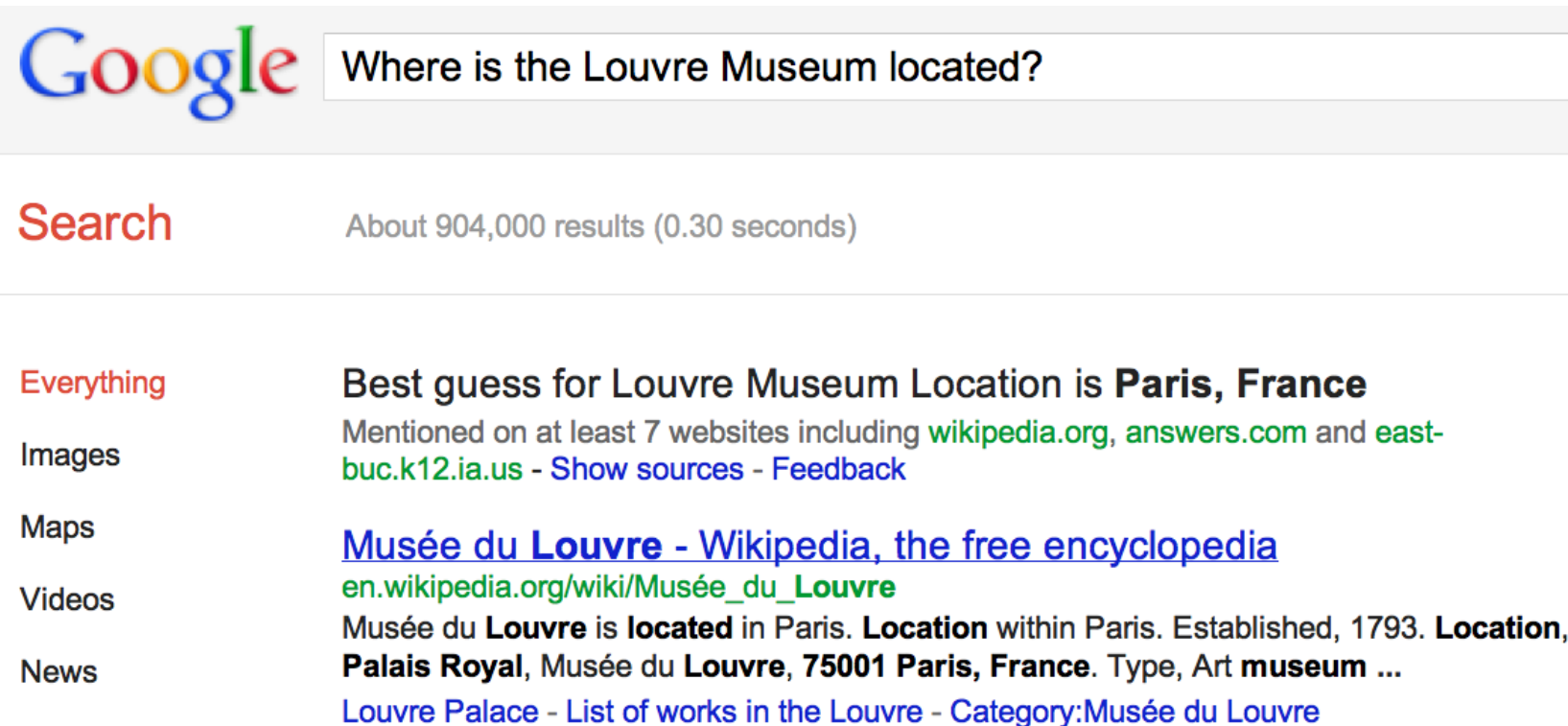
**Everything** [Huginn and Muninn - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Huginn_and_Muninn)  
[en.wikipedia.org/wiki/Huginn\\_and\\_Muninn](https://en.wikipedia.org/wiki/Huginn_and_Muninn)

**Images** 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 ...

**Maps** [Attestations](#) - [Archaeological record](#) - [Theories](#) - [See also](#)

...

# IR-based Question Answering



The image shows a screenshot of a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "Where is the Louvre Museum located?". Below the search bar, the word "Search" is displayed in red, followed by the text "About 904,000 results (0.30 seconds)". On the left side, there are navigation links for "Everything", "Images", "Maps", "Videos", and "News". The main content area displays the search results for the "Everything" tab. The top result is a "Best guess" for the question, stating "Best guess for Louvre Museum Location is Paris, France". Below this, it mentions that the answer is mentioned on at least 7 websites, including wikipedia.org, answers.com, and east-buc.k12.ia.us, with links to "Show sources" and "Feedback". The second result is a link to the Wikipedia page for "Musée du Louvre", with the URL en.wikipedia.org/wiki/Musée\_du\_Louvre. Below the link, there is a snippet of text: "Musée du Louvre is located in Paris. Location within Paris. Established, 1793. Location, Palais Royal, Musée du Louvre, 75001 Paris, France. Type, Art museum ...". At the bottom of the snippet is another link: "Louvre Palace - List of works in the Louvre - Category:Musée du Louvre".

Google

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

Videos

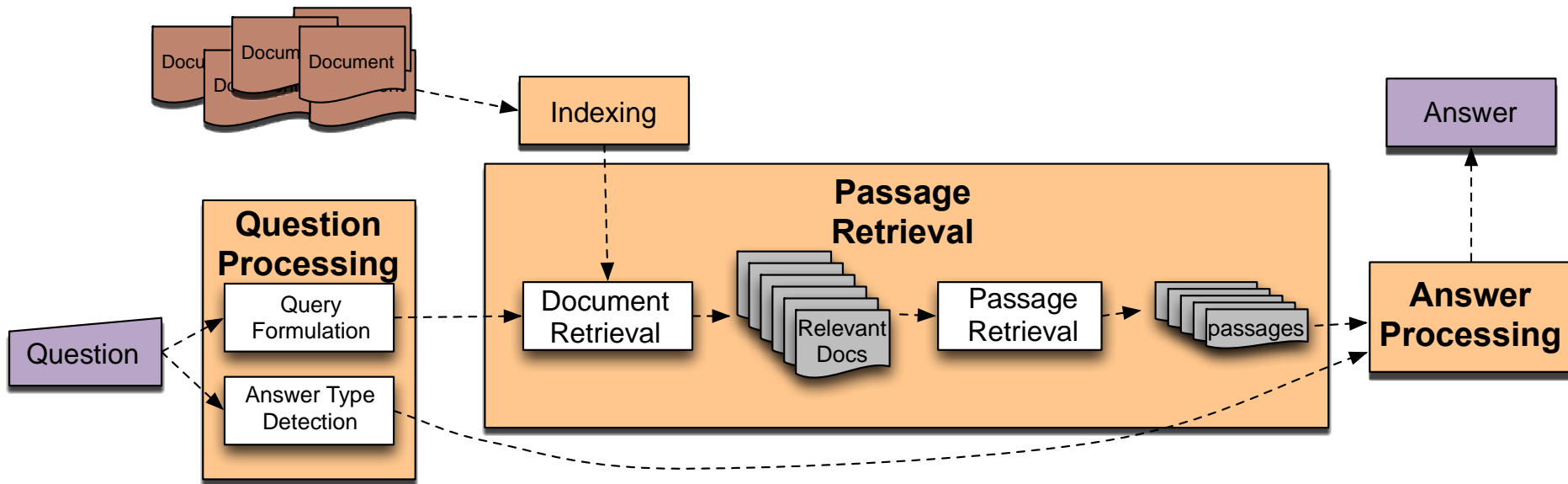
News

[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](#)

# IR-based Factoid QA



# IR-based Factoid QA

- **QUESTION PROCESSING**
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- **PASSAGE RETRIEVAL**
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- **ANSWER PROCESSING**
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

# Knowledge-based approaches (e.g., Siri)

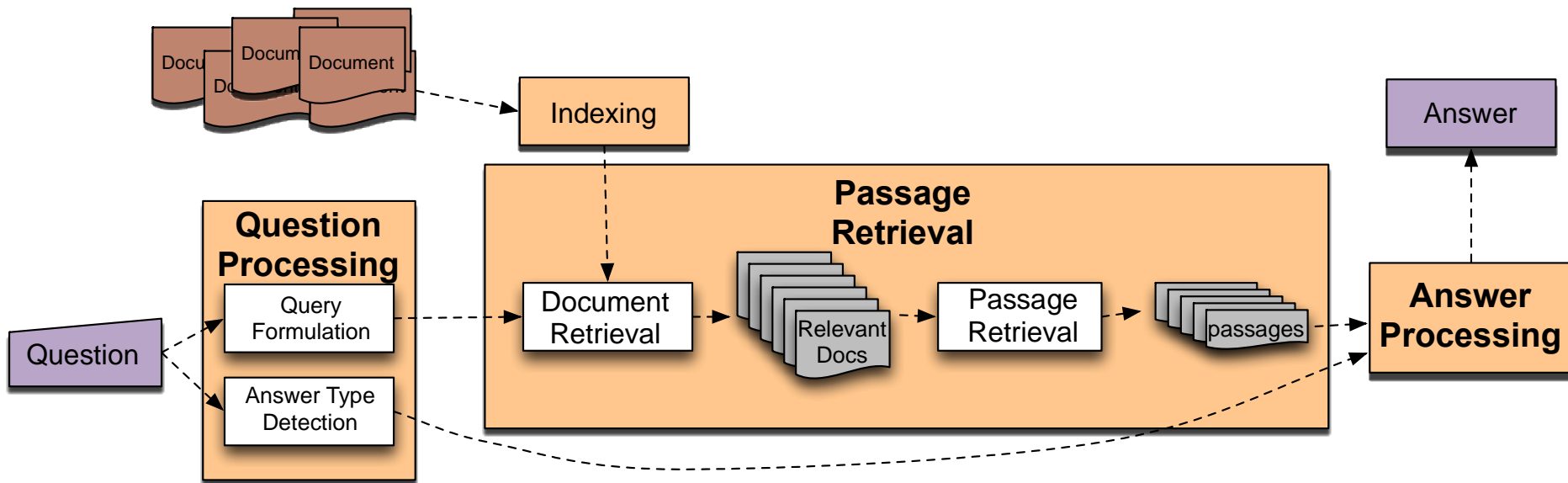
- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases



# Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

# Factoid Q/A



# Question Processing

## Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

# Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type: US state
- Query: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

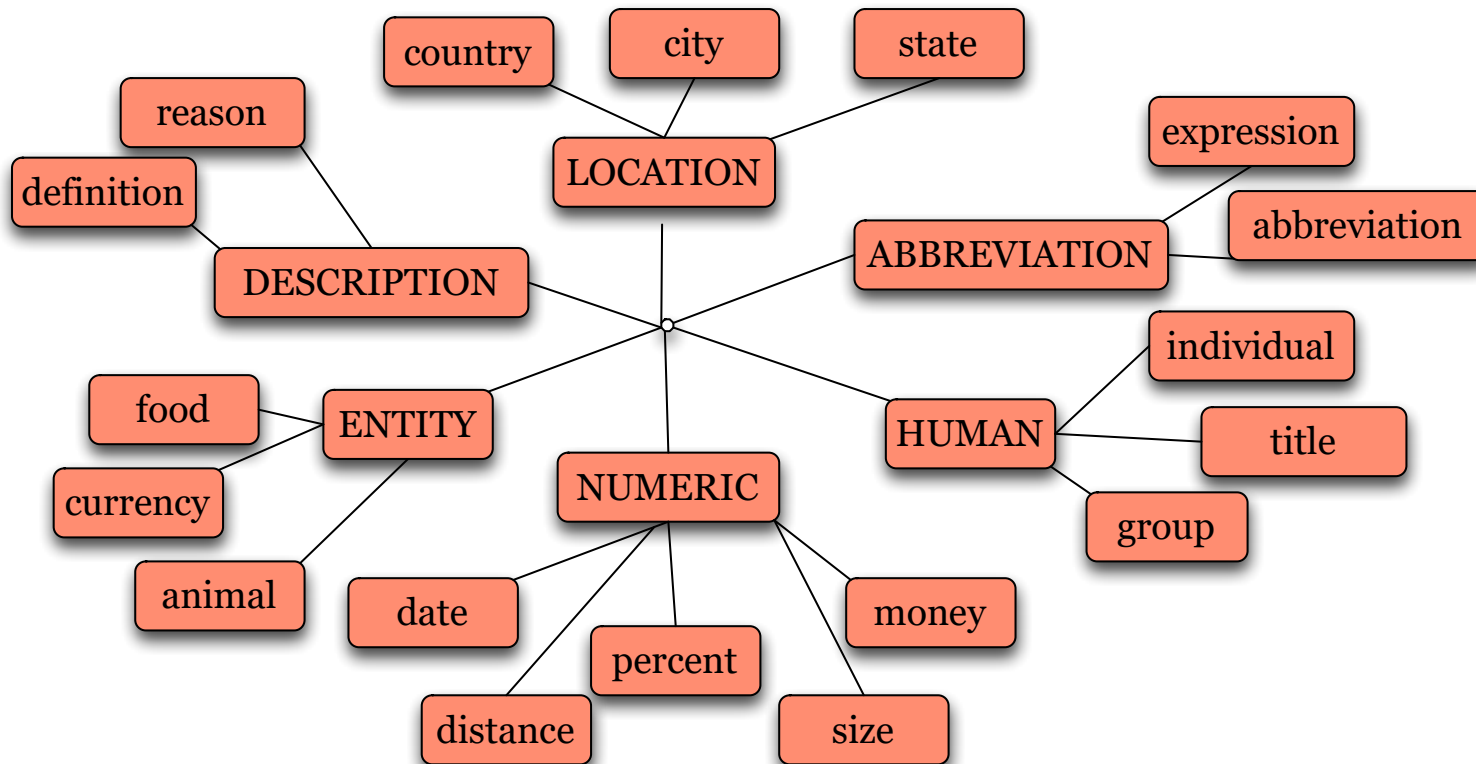
# Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*
  - PERSON
- *What Canadian city has the largest population?*
  - CITY.

# Answer Type Taxonomy

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain...
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency...

# Part of Li & Roth's Answer Type Taxonomy



# Answer Types

## ENTITY

animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?



# More Answer Types

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

# Answer types in Jeopardy

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

# Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

# Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is | was | are | were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:
  - (the headword of the first noun phrase after the wh-word)
  - Which **city** in China has the largest number of foreign financial companies?
  - What is the state **flower** of California?

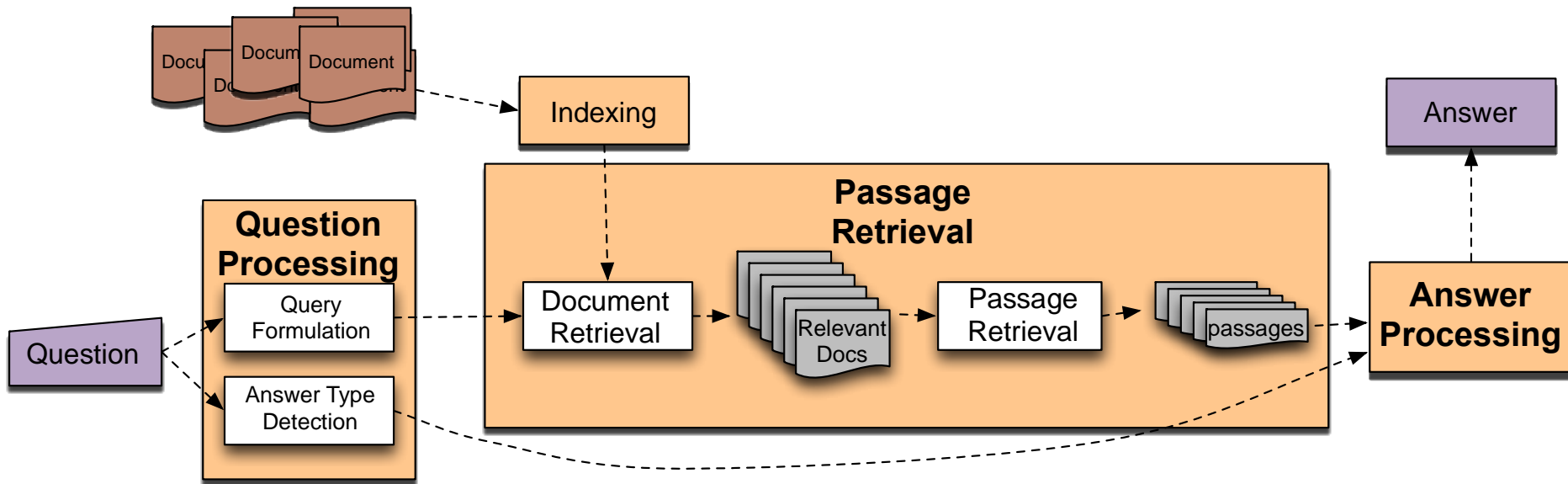
# Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

# Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

# Factoid Q/A



# Keyword Selection Algorithm

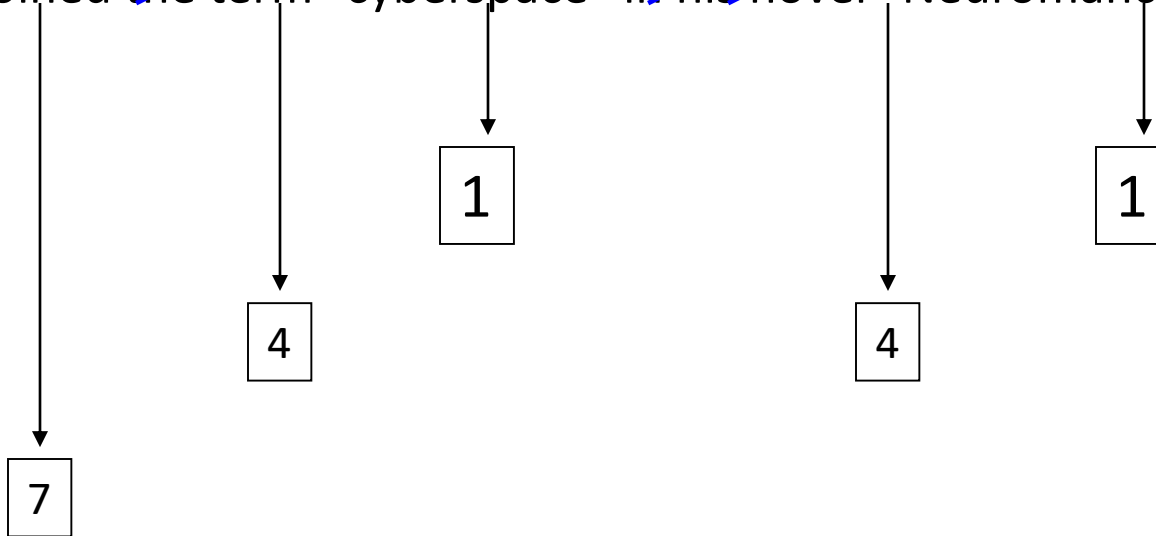
1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words



# Choosing keywords from the query

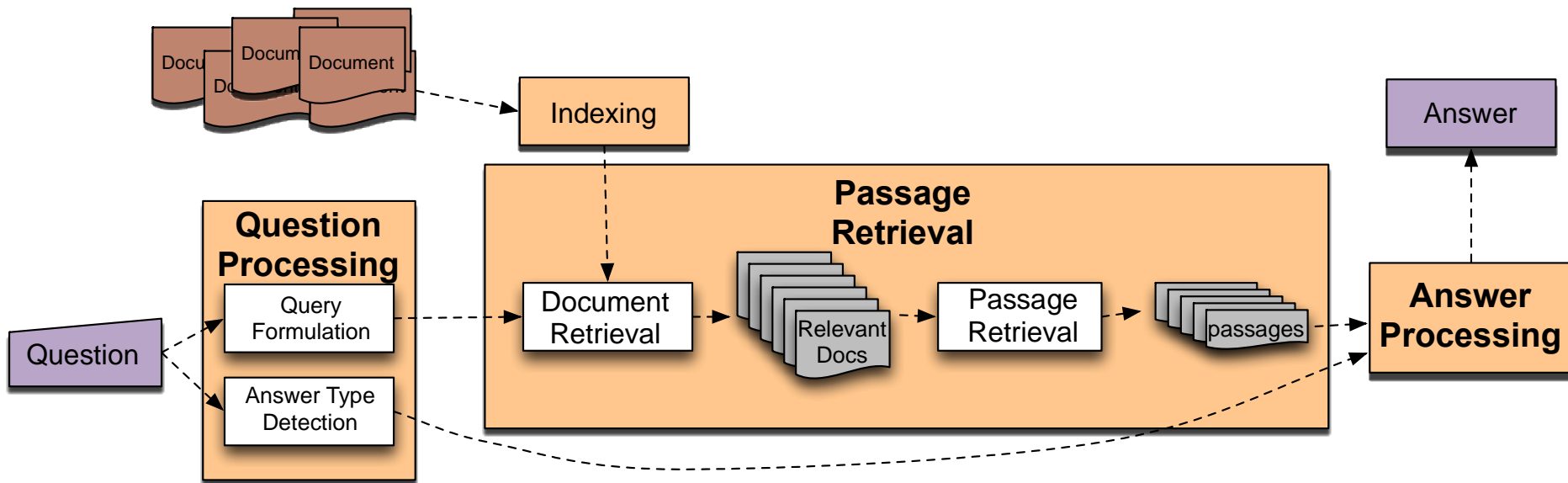
Slide from Mihai Surdeanu

~~Who~~ ~~coined~~ ~~the~~ ~~term~~ "cyberspace" in ~~his~~ ~~novel~~ "Neuromancer"?



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

# Factoid Q/A



# Passage Retrieval

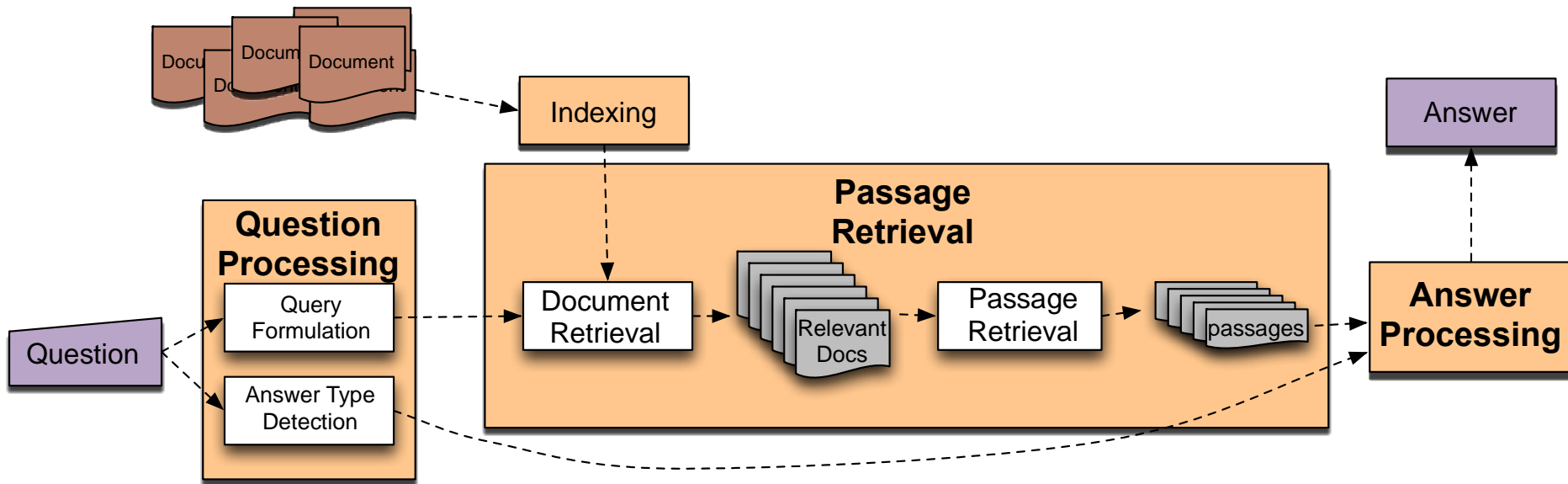
- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

# Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

# Factoid Q/A



# Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (PERSON)  
`Manmohan Singh`, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
  - How tall is Mt. Everest? (LENGTH)  
The official height of Mount Everest is `29035 feet`

## Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**

- Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

## Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**

- Passage:

The Marie biscuit is named after **Marie Alexandrovna**, the daughter of **Czar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**



# Use machine learning:

## Features for ranking candidate answers

**Answer type match:** Candidate contains a phrase with the correct answer type.

**Pattern match:** Regular expression pattern matches the candidate.

**Question keywords:** # of question keywords in the candidate.

**Keyword distance:** Distance in words between the candidate and query keywords

**Novelty factor:** A word in the candidate is not in the query.

**Apposition features:** The candidate is an appositive to question terms

**Punctuation location:** The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

**Sequences of question terms:** The length of the longest sequence of question terms that occurs in the candidate answer.

# Candidate Answer scoring in IBM Watson

- Each candidate answer gets scores from >50 components
  - (from unstructured text, semi-structured text, triple stores)
  - logical form (parse) match between question and candidate
  - passage source reliability
  - geospatial location
    - California is "southwest of Montana"
  - temporal relationships
  - taxonomic classification

# 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: 1/2*
- *else if third answer is correct: 1/3, 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**:



The screenshot shows a web browser window with the URL [www.geonames.org/search.html?q=palo+alto&country=](http://www.geonames.org/search.html?q=palo+alto&country=). The page title is "GeoNames Home | Postal Codes | Download / Webservice | About" and there is a "login" link. The search input field 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 search results show 459 records found for "palo alto". The table below lists the first three results.

	Name	Country	Feature class	Latitude	Longitude
1	<a href="#">Palo Alto</a> Palo Al'to, Palo Alto, pa luo ao duo, paraaruto, Пало Алто, Пало Альто, פאלו אלטו, パロアルト, 帕羅奧多	<a href="#">United States</a> , California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2	<a href="#">Palo Alto Township</a> Palo Alto Township	<a href="#">United States</a> , Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3	<a href="#">Borough of Palo Alto</a>	<a href="#">United States</a> , 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

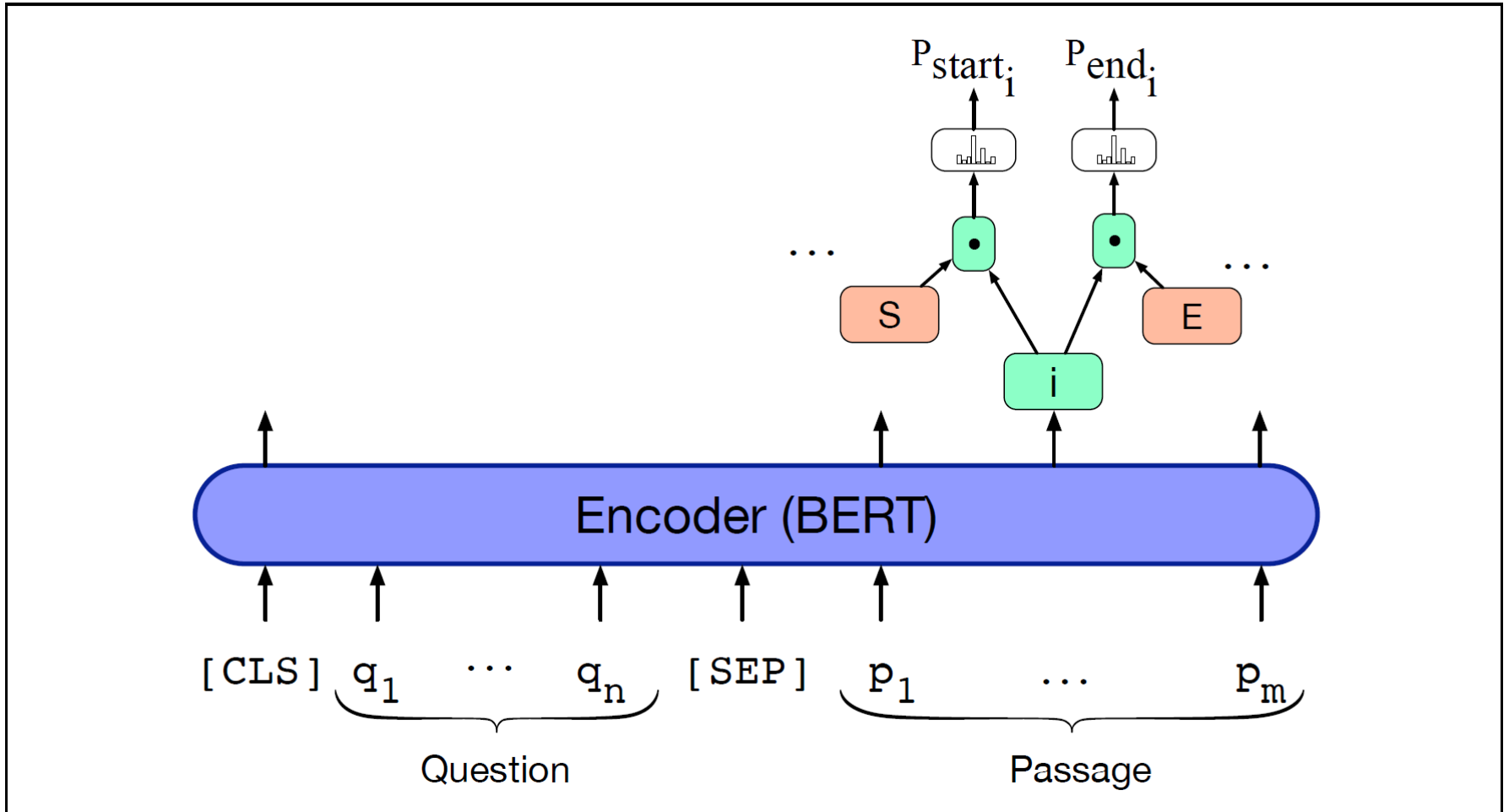
- Coreference helps resolve ambiguities
  - U: “Book a table at Il Fornaio at 7:00 with **my mom**”
  - U: “Also send **her** an email reminder”
- Clarification questions:
  - U: “Chicago pizza”
  - S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”

# 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  $q_1, \dots, q_n$  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
- E.g., use the T5 language model, which is an encoder-decoder transformer model pretrained to fill in masked spans of task
- Language modeling is not yet a complete solution for question answering;
  - not working quite as well,
  - poor interpretability
  - 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.

President Franklin <M> born <M> January 1882.

Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.

Our <M> hand-picked and sun-dried <M> orchard in Georgia.

T5

D. Roosevelt was <M> in

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

peaches are <M> at our

President Franklin D. Roosevelt was born in January 1882.

*Pre-training*

*Fine-tuning*

When was Franklin D. Roosevelt born?

T5

1882

# 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: 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 referent 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