



## N P T E L O N L I N E C E R T I F I C A T I O N C O U R S E S

# DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 31 : Applications: Question Answering-I



**PROF . PAWAN GOYAL**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## CONCEPTS COVERED

- What is Question Answering?
- Reading Comprehension
- Open Domain Question Answering

# Lots of Practical Applications



Where is the deepest lake in the world?



All

Maps

Images

News

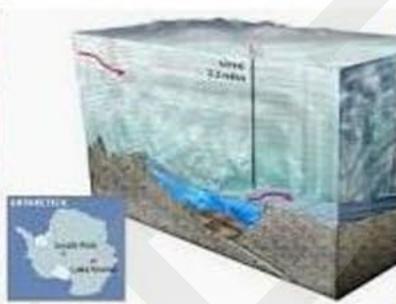
Videos

More

Settings

Tools

About 21,100,000 results (0.71 seconds)



## Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

<https://web.stanford.edu/class/cs224n/>

# Lots of Practical Applications

Google How can I protect myself from COVID-19? X |  

All Images News Shopping Videos More Settings Tools

The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

To help prevent the spread of COVID-19:

- Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
- Stay at least 6 feet (about 2 arm lengths) from others.
- Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
- Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Get vaccinated against COVID-19 when it's your turn.
- Avoid close contact with people who are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

 [Learn more on cdc.gov](#)

For informational purposes only. Consult your local medical authority for advice.

# What is Question Answering?

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

## Taxonomy

- What information source does a system build on?
  - ▶ A text passage, all Web documents, knowledge bases, tables, images, ...
- Question Type
  - ▶ Factoid vs non factoid, open-domain vs closed-domain, simple vs compositional, ...
- Answer Type
  - ▶ A short segment of text, a paragraph, a list, yes/no, ...

# Reading Comprehension

## *Reading Comprehension*

Comprehend a passage of text and answer questions about its content  
 $(P, Q) \rightarrow A$

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

Q: What language did Tesla study while in school?

A: German

# Reading Comprehension

## *Reading Comprehension*

Comprehend a passage of text and answer questions about its content  
 $(P, Q) \rightarrow A$

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

# Stanford Question Answering Dataset (SQuAD)



- 100k annotated (passage, question, answer) triples  

Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.  

This is a limitation— not all the questions can be answered in this way!
- SQuAD was for years the most popular reading comprehension dataset; it is “almost solved” today (though the underlying task is not,) and the state-of-the-art exceeds the estimated human performance.

---

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

What causes precipitation to fall?  
**gravity**

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

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

---

# SQuAD Evaluation

## Evaluation

exact match (0 or 1) and F1 (partial credit).

- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to each gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and served}

# SQuAD Evaluation

Q: What did Tesla do in December 1878?

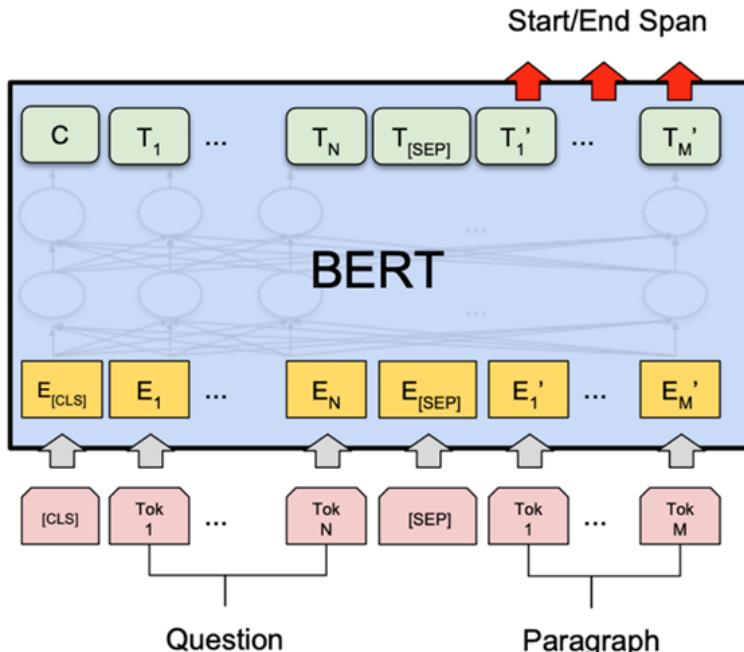
A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and served}

**Exact match:**  $\max\{0,0,0\} = 0$

**F1:**  $\max\{0.67, 0.67, 0.61\} = 0.67$

# Recap: BERT for Reading Comprehension



$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^\top \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^\top \mathbf{h}_i)$$

where  $\mathbf{h}_i$  is the hidden vector of  $c_i$ , returned by BERT

**Question** = Segment A  
**Passage** = Segment B  
**Answer** = predicting two endpoints in segment B



**Question:** How many parameters does BERT-large have?

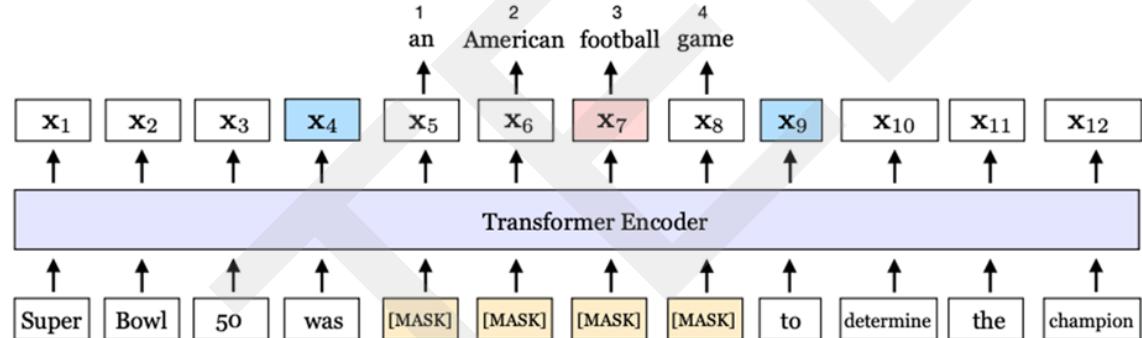
**Reference Text:** BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Image credit: <https://mccormickml.com/>

# Can we design better pre-training objectives?

The answer is yes!

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$



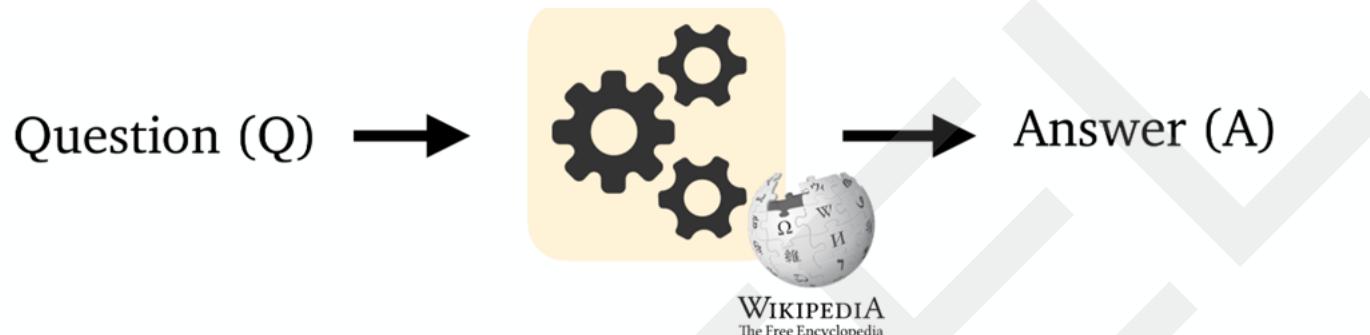
Two ideas:

- 1) masking contiguous spans of words instead of 15% random words
- 2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

(Joshi & Chen et al., 2020): SpanBERT: Improving Pre-training by Representing and Predicting Spans

# Open-Domain Question Answering



- Different from reading comprehension, we don't assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging but a more practical problem!

# Retriever-Reader Framework

**Input:** a large collection of documents  $D = D_1, \dots, D_N$  and  $Q$

**Output:** an answer string  $A$

- Retriever:  $f(\mathcal{D}, Q) \longrightarrow P_1, \dots, P_K$       K is pre-defined (e.g., 100)
- Reader:     $g(Q, \{P_1, \dots, P_K\}) \longrightarrow A$       A reading comprehension problem!

*or a dense  
retriever*

*DrQA (Chen et al., 2017)*

- **Retriever:** A standard TF-IDF information-retrieval sparse model (a fixed module)
- **Reader:** A neural Reading Comprehension Model

# Retriever-Reader Framework

- Fixed length blocks as **passages**
- Over 13M evidence **blocks *b***
- Each block has over 2000 possible **answer spans *s***

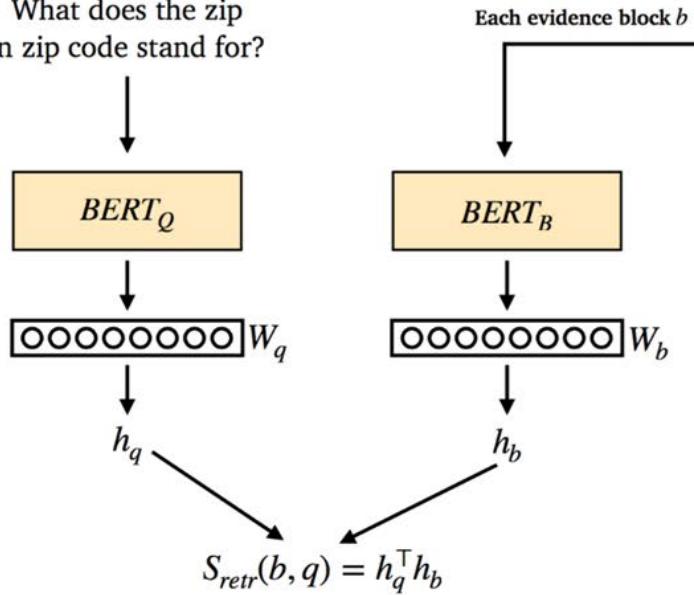


**Modeling**      
$$S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$$

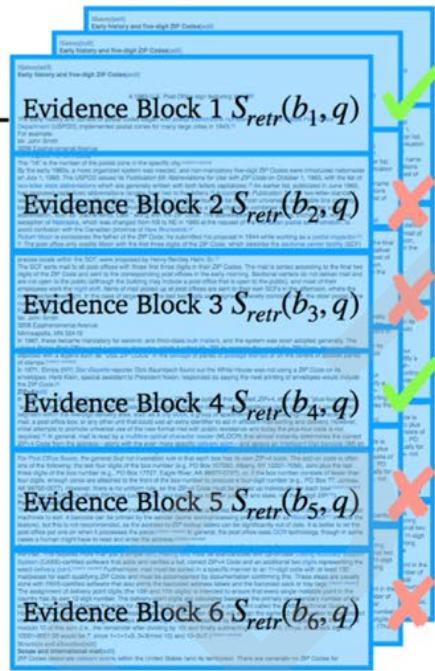
**Inference**      
$$a^* = \text{TEXT}(\operatorname{argmax}_{b,s} S(b, s, q))$$

# Retriever Component

Question  $q$   
What does the zip  
in zip code stand for?



All of Wikipedia: select top K

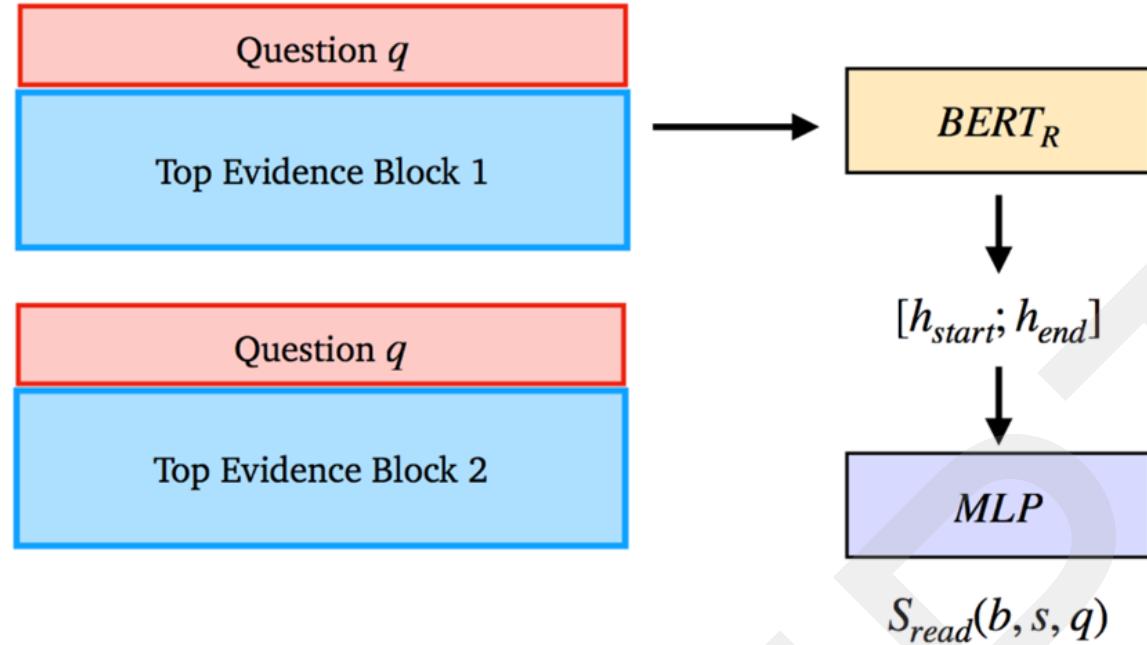


$$h_q = \mathbf{W}_q \text{BERT}_Q(q)[\text{CLS}]$$

$$h_b = \mathbf{W}_b \text{BERT}_B(b)[\text{CLS}]$$

$$S_{retr}(b, q) = h_q^\top h_b$$

# Reader Component

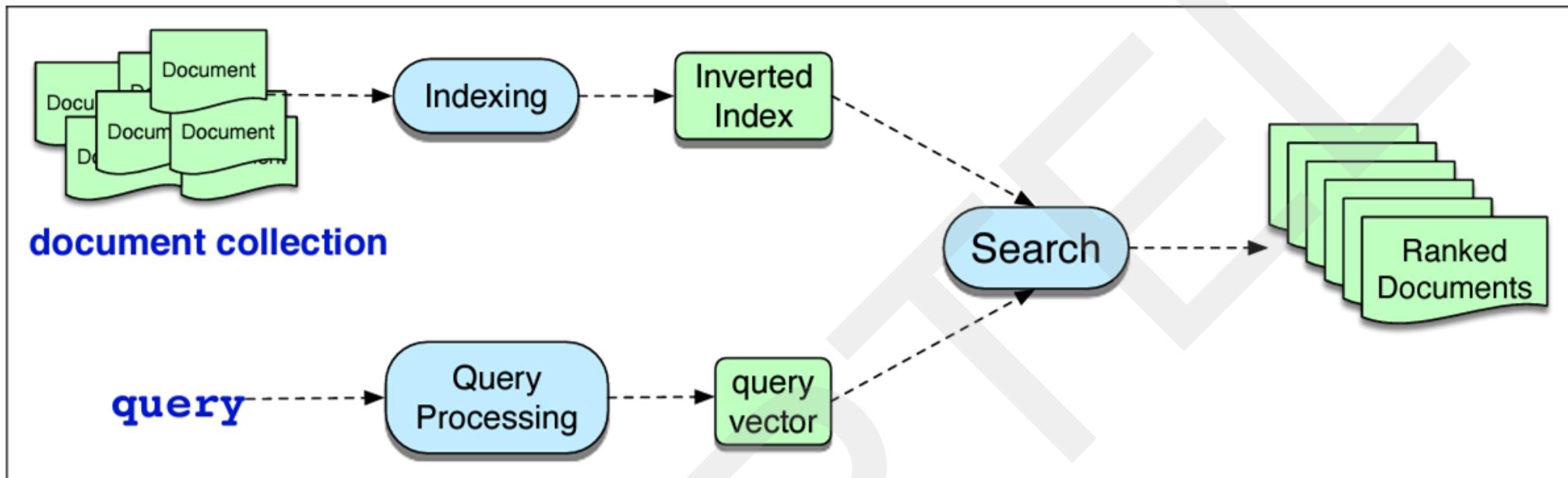


$$h_{start} = \text{BERT}_R(q, b)[\text{START}(s)]$$

$$h_{end} = \text{BERT}_R(q, b)[\text{END}(s)]$$

$$S_{read}(b, s, q) = \text{MLP}([h_{start}; h_{end}])$$

# Retriever: Standard TF-IDF based model



*term frequency-inverse document frequency (tf-idf)*

$$tf_{t,d} = \log_{10}(1 + \text{count}(t, d))$$

$$idf_t = \log_{10} \frac{N}{df_t}$$

$$tf - idf(t, d) = tf_{t,d} \cdot idf_t$$

# TF-IDF: A simplified formulation

## Document Scoring

We score document  $d$  by the cosine of its vector  $\mathbf{d}$  with the query vector  $q$ :

$$\text{score}(q, d) = \cos(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{|\mathbf{q}| \cdot |\mathbf{d}|}$$

$$\text{score}(q, d) = \sum_{t \in \mathbf{q}} \frac{\text{tf}_{t,q} \cdot \text{idf}_t}{\sqrt{\sum_{q_i \in q} \text{tf-idf}^2(q_i, q)}} \cdot \frac{\text{tf}_{t,d} \cdot \text{idf}_t}{\sqrt{\sum_{d_i \in d} \text{tf-idf}^2(d_i, d)}}$$

# A worked out example

Query						
word	cnt	tf	df	idf	tf-idf	n'lized = tf-idf/ q
sweet	1	1	3	0.125	0.125	0.383
nurse	0	0	2	0.301	0	0
love	1	1	2	0.301	0.301	0.924
how	0	0	1	0.602	0	0
sorrow	0	0	1	0.602	0	0
is	0	0	1	0.602	0	0

$|q| = \sqrt{.125^2 + .301^2} = .326$

Document 1					Document 2					
word	cnt	tf	tf-idf	n'lized	× q	cnt	tf	tf-idf	n'lized	× q
sweet	2	1.301	0.163	0.357	<b>0.137</b>	1	1.000	0.125	0.203	<b>0.0779</b>
nurse	1	1.000	0.301	0.661	0	0	0	0	0	0
love	1	1.000	0.301	0.661	<b>0.610</b>	0	0	0	0	0
how	0	0	0	0	0	0	0	0	0	0
sorrow	0	0	0	0	0	1	1.000	0.602	0.979	0
is	0	0	0	0	0	0	0	0	0	0

$|d_1| = \sqrt{.163^2 + .301^2 + .301^2} = .456$

$|d_2| = \sqrt{.125^2 + .602^2} = .615$

Cosine:  $\sum$  of column: **0.747**

Cosine:  $\sum$  of column: **0.0779**

# BM25 Weighting

$$\text{score}(q, d) = \sum_{t \in q} \underbrace{\log\left(\frac{N}{df_t}\right)}_{\text{IDF}} \underbrace{\frac{tf_{t,d}}{k\left(1 - b + b\left(\frac{|d|}{|d_{avg}|}\right)\right) + tf_{t,d}}}_{\text{weighted tf}}$$

BM25 adds two parameters: k, a knob that adjust the balance between term frequency and IDF, and b, which controls the importance of document length normalization.

**Default parameters:**

k1 = 1.2

b = 0.75

# Using a dense retriever

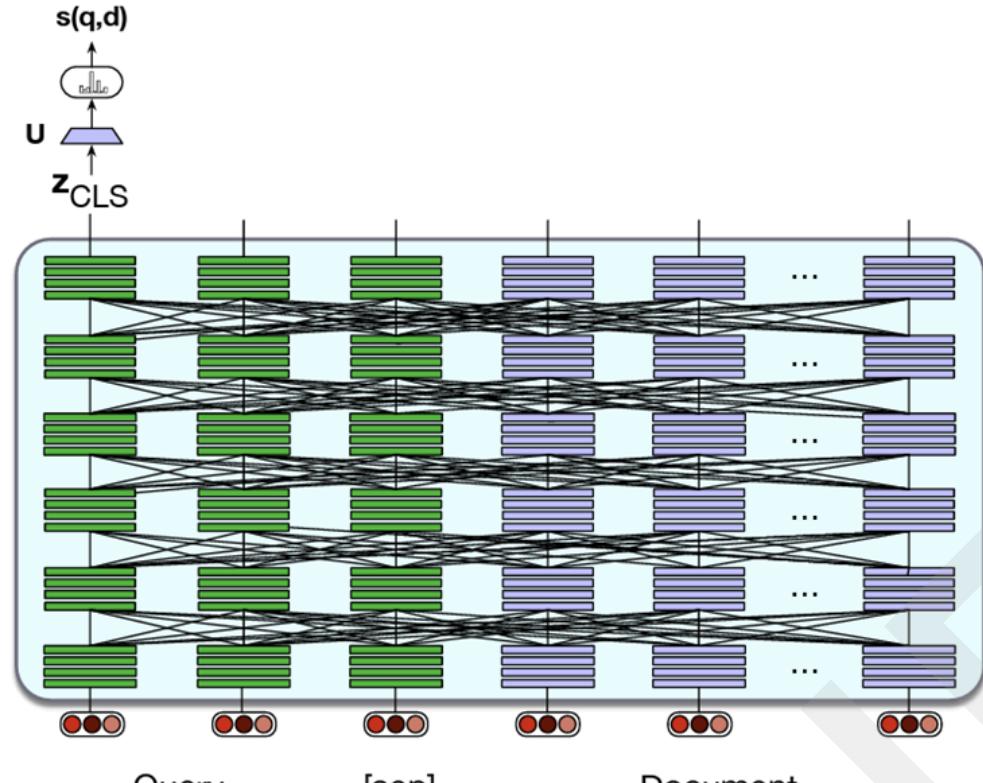
## *Issues with standard TF-IDF based retrieval approaches*

- They work only if there is exact overlap of words between the query and document.
- The user posing a query (or asking a question) needs to guess exactly what words the writer of the answer might have used.
- The user might decide to search for a *tragic love story* but Shakespeare writes instead about *star-crossed lovers*

## *Solution*

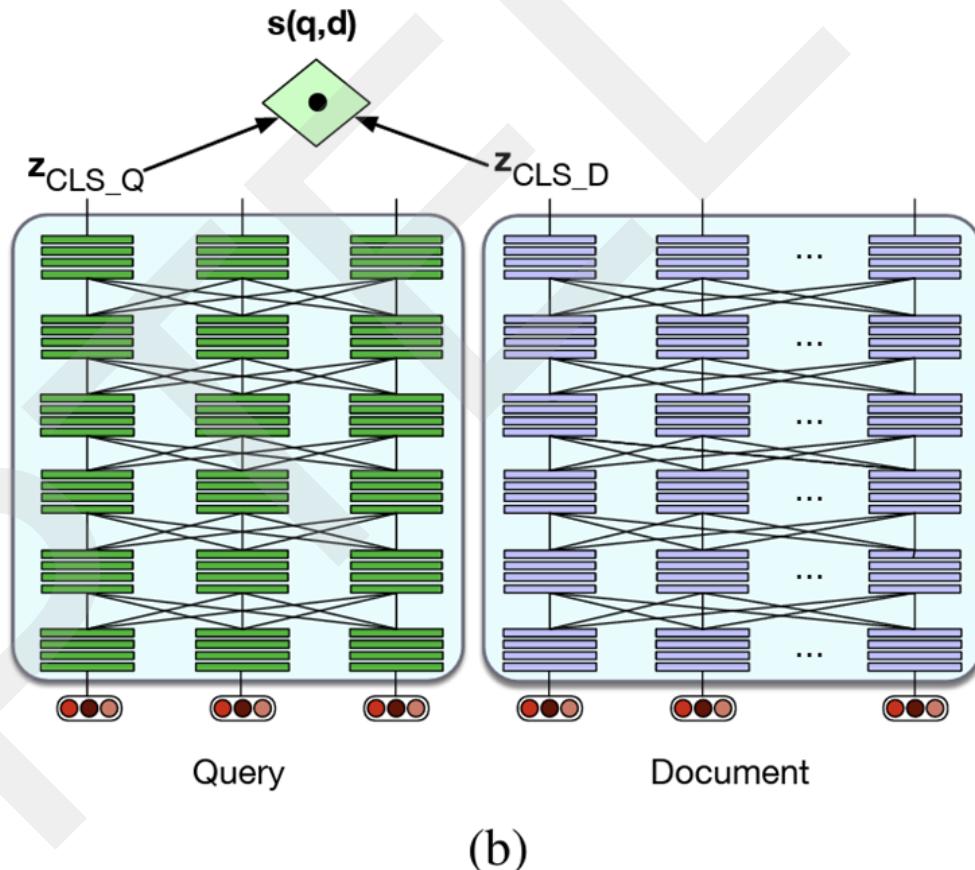
Use dense embeddings in the form of BERT!

# Dense retriever: Two options



(a)

(a) is too compute expensive



(b)

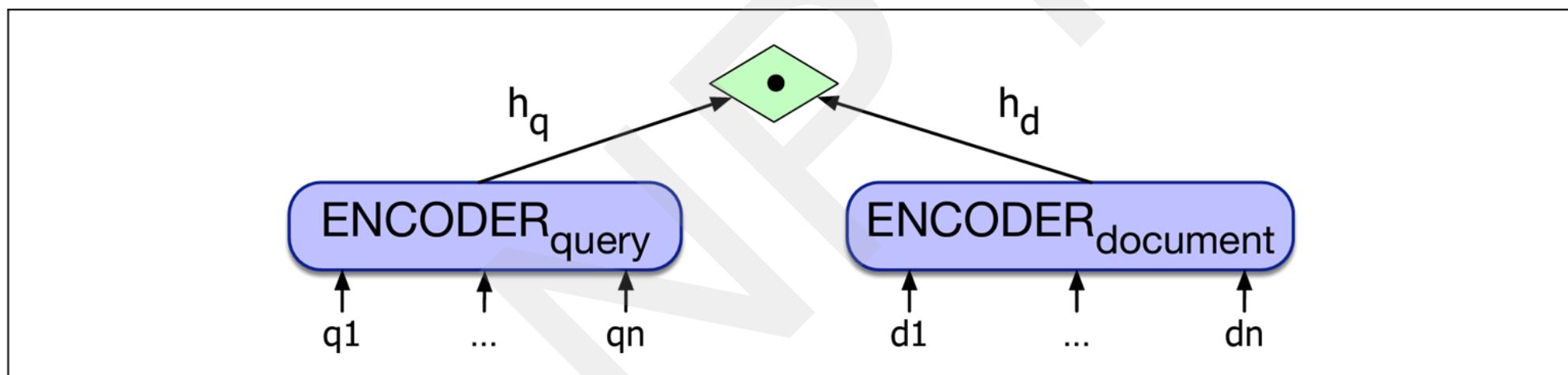
# Using a dense retriever: Bi-encoder

Two encoders,  $BERT_Q$  and  $BERT_D$  to encode the query and document, respectively.

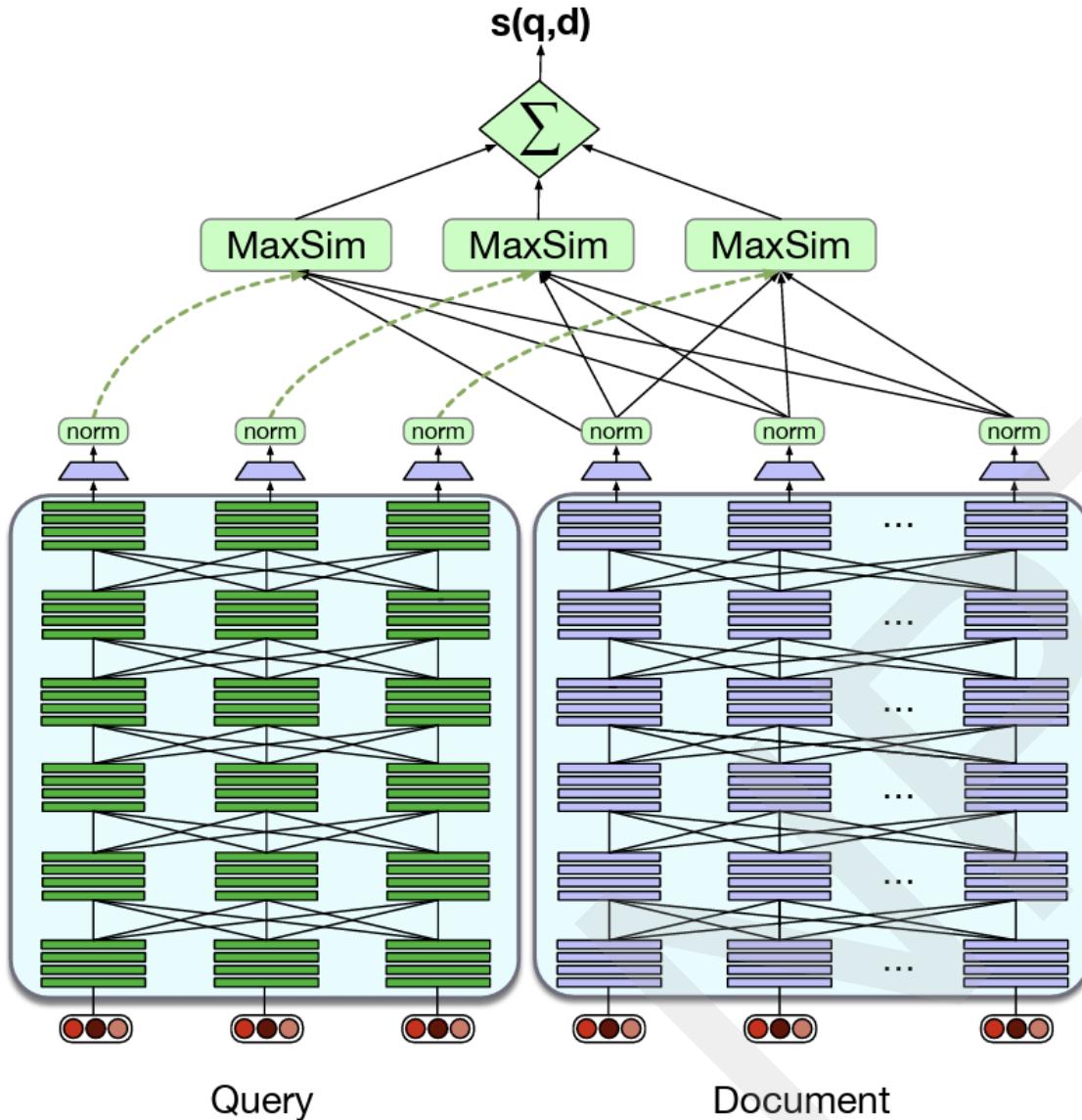
$$h_q = BERT_Q(q)[CLS]$$

$$h_d = BERT_D(d)[CLS]$$

$$score(d, q) = h_q \cdot h_d$$



# Using a dense retriever: ColBERT



A question  $q$  is tokenized as  $[q_1, \dots, q_N]$ , prepended with a [CLS] and a special [Q] token, and passed through BERT to get output vectors  $q = [q_1, \dots, q_N]$ . The passage  $d$  with tokens  $[d_1, \dots, d_m]$ , is processed similarly, including a [CLS] and special [D] token.

A linear layer is applied on top of  $d$  and  $q$  to control the output dimension, so as to keep the vectors small for storage efficiency, and vectors are rescaled to unit length, producing the final vector sequences  $E_q$  (length  $N$ ) and  $E_d$  (length  $m$ ). The ColBERT scoring mechanism is:

$$\text{score}(q, d) = \sum_{i=1}^N \max_{j=1}^m \mathbf{E}_{q_i} \cdot \mathbf{E}_{d_j}$$



# How to train these dense retrievers?

How to get training data?

What training objectives to use?

# REFERENCES

- Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. [Chapter 14]



**THANK YOU**



## N P T E L O N L I N E C E R T I F I C A T I O N C O U R S E S

# DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 32 : Applications: Question Answering-II



**PROF . PAWAN GOYAL**

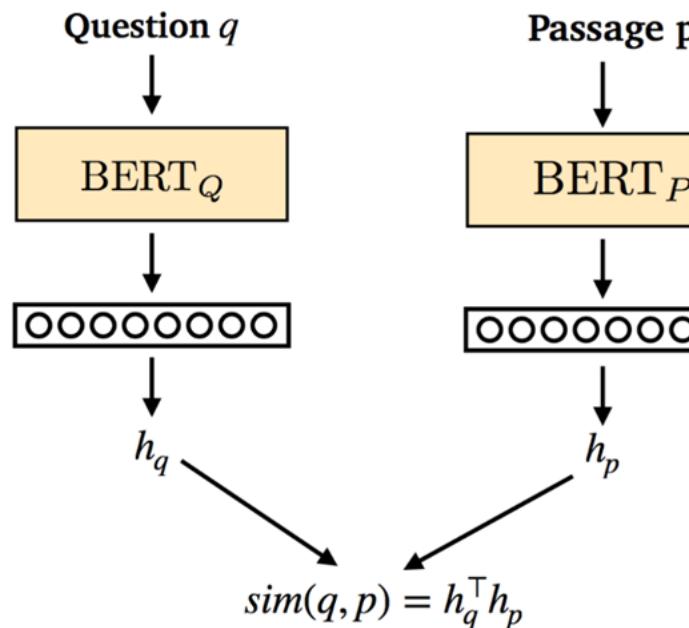
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

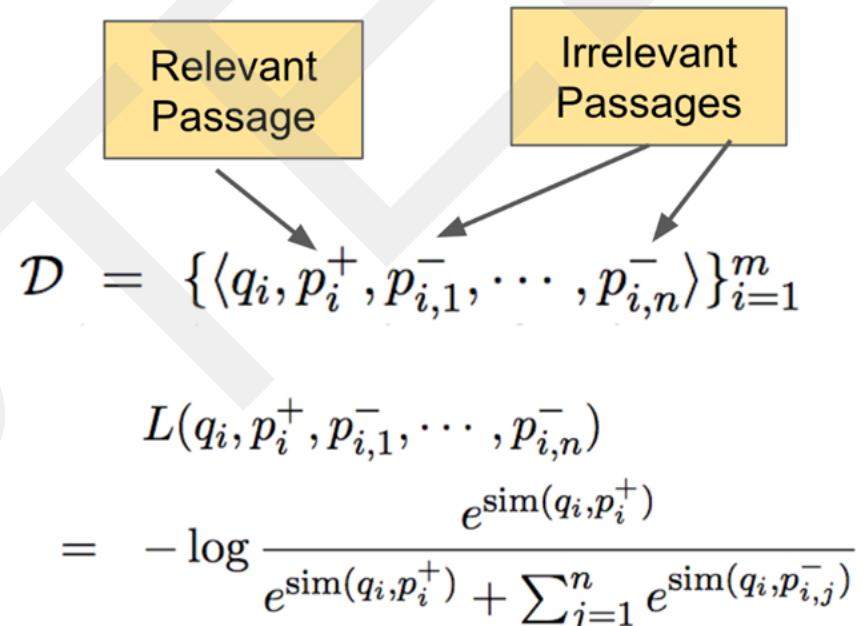
## CONCEPTS COVERED

- Open Domain QA: Training the Dense Retriever
- Generative QA
- Other QA Datasets
- Multilingual QA

# Training the dense retrievers

- We can train a dense retrieval only from a small number of QA pairs, ***without any pre-training***.





The diagram shows the training data generation and the corresponding loss function. The training data  $\mathcal{D}$  consists of  $m$  QA pairs, where each pair contains a relevant passage and  $n$  irrelevant passages. The loss function  $L$  is calculated as follows:

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^- \rangle\}_{i=1}^m$$

$$L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-)$$

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

# Training the dense retrievers

Some models just use a positive and a negative example with contrastive loss

$$\mathcal{L}(\theta, q) = \sum_{d_{\text{pos}} \in D_{\text{pos}}} \sum_{d_{\text{neg}} \in D_{\text{neg}}} \max(0, s(q, d_{\text{neg}}; \theta) - s(q, d_{\text{pos}}; \theta))$$

# How to select the positive and negative passages

- **Positives**

- Provided in the reading comprehension datasets
- Passages of high BM25 scores that contain the answer string

*Relevance-guided supervision*

- **Negatives**

- Random passages from the corpus
- Top passages returned by BM25 which DO NOT contain the answer but match most question tokens
- Positive passages of **OTHER** questions

In-batch negatives

Hard negatives

# Retrieval Free Approaches

- Can we use **pre-trained language models** to act as “knowledge storage” ?
- Can we query the LMs to obtain the answer directly?
- The LMs were **pre-trained** on Wikipedia (and other textual corpora) so *they should be able to memorize a fair amount of information*

# GPT-2 Zero-shot QA

- Evaluated on Natural Questions and **no training** at all
- Accuracy of 63.1%** on the 1% of questions it is most confident in

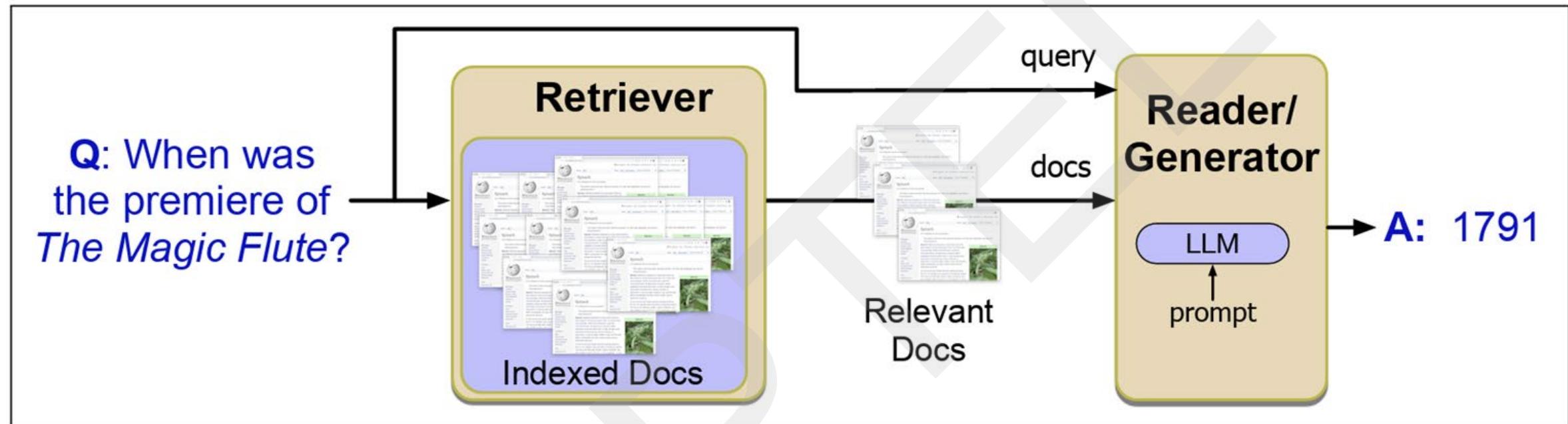
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%

# Generative QA: Fine-tunes T5

- Feed in input passage and question, use decoder to output answer
- Example: UnifiedQA (Khashabi et al. 2020), trained on many different datasets
  - Format each dataset into input/output format
  - Base model: T5

	<b>Dataset</b>	SQuAD 1.1
EX	<b>Input</b>	At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
	<b>Output</b>	16,000 rpm
AB	<b>Dataset</b>	NarrativeQA
	<b>Input</b>	What does a drink from narcissus's spring cause the drinker to do? \n 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.'' ...
MC	<b>Output</b>	fall in love with themselves
	<b>Dataset</b>	ARC-challenge
	<b>Input</b>	What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar
MC	<b>Output</b>	sugar
	<b>Dataset</b>	MCTest
	<b>Input</b>	Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...
YN	<b>Output</b>	The big kid
	<b>Dataset</b>	BoolQ
	<b>Input</b>	Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
YN	<b>Output</b>	no

# Retriever-Reader framework can benefit from Generation as well



We take a large pretrained language model, give it the set of retrieved passages and other text as its prompt, and autoregressively generate a new answer token by token (**retrieval-augmented generation**)

# Other Datasets: Natural Questions

- Real anonymized, aggregated queries issued to the Google search engine of 8 words or more.
- An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates
- a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no answer is present.

Question: **What color was john wilkes booth's hair**

Wikipedia Page: [John Wilkes Booth](#)

Long answer: Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had **jet-black** hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

Short answer: **jet-black**

# More complex: HotpotQA

Reasoning Type	%	Example(s)
Inferring the <b>bridge entity</b> to complete the 2nd-hop question (Type I)	42	<p><b>Paragraph A:</b> The 2015 Diamond Head Classic was a college basketball tournament ... <i>Buddy Hield was named the tournament's MVP.</i></p> <p><b>Paragraph B:</b> <i>Chavano Rainier "Buddy" Hield</i> is a Bahamian professional basketball player for the <b>Sacramento Kings</b> of the NBA...</p> <p><b>Q:</b> Which team does the player named 2015 Diamond Head Classic's MVP play for?</p>
Comparing two entities (Comparison)	27	<p><b>Paragraph A:</b> LostAlone were a British rock band ... consisted of <i>Steven Battelle, Alan Williamson, and Mark Gibson...</i></p> <p><b>Paragraph B:</b> Guster is an American alternative rock band ... Founding members <i>Adam Gardner, Ryan Miller, and Brian Rosenworcel</i> began...</p> <p><b>Q:</b> Did LostAlone and Guster have the same number of members? (<b>yes</b>)</p>
Locating the <b>answer entity</b> by checking multiple properties (Type II)	15	<p><b>Paragraph A:</b> Several <i>current and former members of the Pittsburgh Pirates</i> ... John Milner, <b>Dave Parker</b>, and Rod Scurry...</p> <p><b>Paragraph B:</b> <b>David Gene Parker</b>, <i>nicknamed "The Cobra"</i>, is an American former player in Major League Baseball...</p> <p><b>Q:</b> Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"?</p>
Inferring about the property of an entity in question through a <b>bridge entity</b> (Type III)	6	<p><b>Paragraph A:</b> <i>Marine Tactical Air Command Squadron 28</i> is a United States Marine Corps aviation command and control unit based at <b>Marine Corps Air Station Cherry Point</b>...</p> <p><b>Paragraph B:</b> <b>Marine Corps Air Station Cherry Point</b> ... is a United States Marine Corps airfield located in <b>Havelock, North Carolina</b>, USA ...</p> <p><b>Q:</b> What city is the Marine Air Control Group 28 located in?</p>

# Multilingual QA

- Most Open-Retrieval QA datasets are in English

Data Type	Language
Questions	English
Answers	English
Text Corpus	English

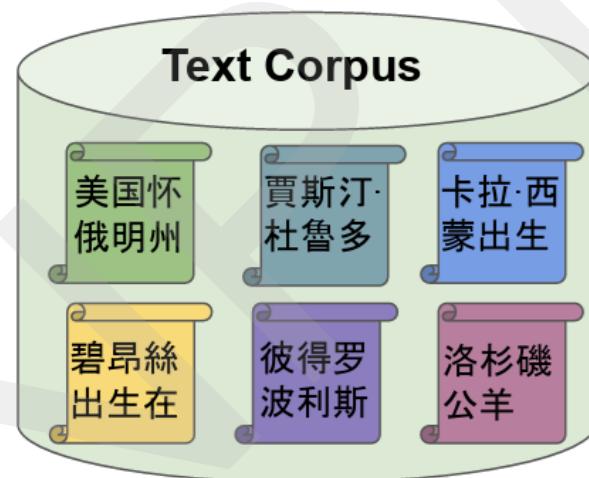
- Can we support *questions* in another language
- Can we search against a *corpus* in another language?

# Multilingual QA

- Two related problem settings:
  - 1) Multilingual QA

Data Type	Language
Questions	Chinese
Answers	Chinese
Text Corpus	Chinese

Q: “碧昂丝在哪里长大?”



碧昂丝出生在  
德州休士顿的  
公园广场医院

A: “休斯顿”

# Multilingual QA

- Two related problem settings:
  - 1) Multilingual QA
  - 2) Crosslingual QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English

# Multilingual Open Retrieval QA

## → Approach 1: Zero-Shot Transfer

- ◆ Choose a multilingual encoder (e.g. XLM-R)
- ◆ Finetune it on an English-language QA dataset (e.g. SQuAD)
- ◆ Transfer the encoder to a new language (e.g. Tamil)

## → Problem: zero-shot transfer usually doesn't work great

- ◆ Just giving a few target-language examples helps a lot

Task	Model	$k = 0$	score	$\Delta$	score	$\Delta$	score	$\Delta$
			$k = 2$		$k = 4$		$k = 6$	
XQUAD	MBERT	45.62	48.12	2.50	48.66	3.04	49.34	3.72
	XLM-R	53.68	53.73	0.05	53.84	0.17	54.76	1.08

# Multilingual Open Retrieval QA

## → Approach 2: Translation-Based Adaptation

### ◆ “Translate-Test”

- Translate question to English
- Apply an off-the-shelf English QA system against an English text corpus
- Translate the answer into the language of your choice
- Suffers from error propagation from MT systems + QA system
- Answers must be found in an English corpus
  - Leads to anglocentric QA systems

# Multilingual Open Retrieval QA

## → Approach 2: Translation-Based Adaptation

### ◆ “Translate-Train”

- Translate full training data (questions, answers, and text corpus/passages) to target language
- Train model in target language
- At test time, run open-retrieval QA system on translated text corpus
- Requires translating full text corpus (e.g. English Wikipedia)
- Text corpus (and training data) are noisy due to MT errors

# Multilingual Open Retrieval QA

→ Approach 3: multilingual retriever-generator (Asai et al. 2021)



One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

# Multilingual QA Datasets

- Multilingual Machine Reading Comprehension
  - ◆ XQuAD (Artetxe et al. 2020)
    - Based on 1.1K SQuAD question-answer-passage triples
    - Each professionally translated into 10 languages
  - ◆ MLQA (Lewis et al. 2020)
    - ~5K samples in each of 6 languages + English

# Multilingual QA Datasets

- Multilingual Open-Retrieval QA
  - ◆ MKQA (Longpre et al. 2020)
    - 10K QA pairs from *Natural Questions* (Kwiatkowski et al. 2020) are translated into 26 languages
    - Assumes answer can be found from English Wikipedia
  - ◆ TyDi QA (Clark et al. 2020)
    - 200K QA pairs are collected *naturally* in 11 languages
    - Text corpus is each language's native Wikipedia

# Other Modalities: Tabular QA

**Question:** What was the reported mainline RPM for American Airlines in 2017?

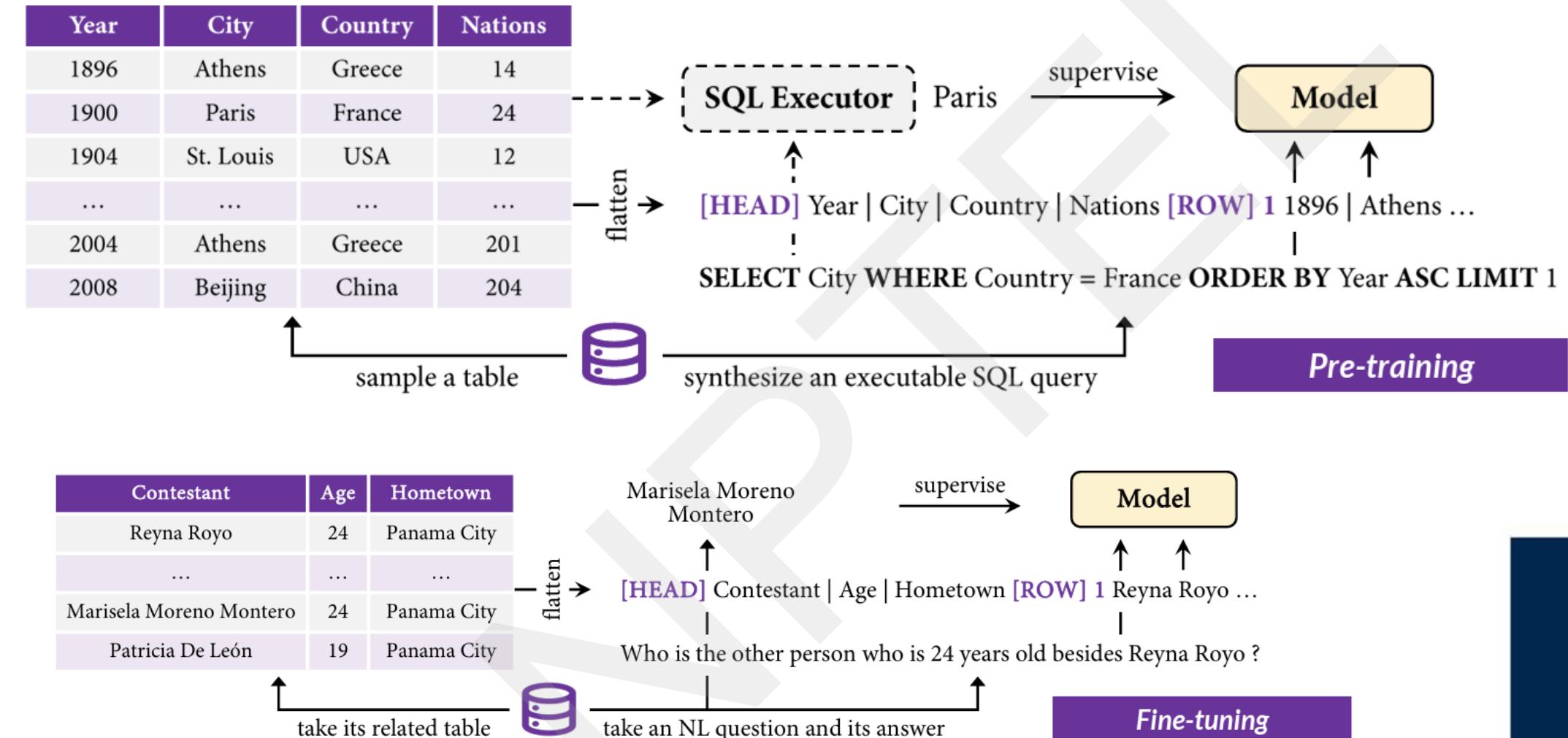
Dataset		Closed -domain	Question Type
<b>Table -only</b>	WTQ★ [36]	Yes	Factoid
	SQA‡★ [20]	Yes	Factoid
	WikiSQL★ [50]	Yes	Factoid
	Spider [47]	Yes	Factoid
	HiTab [8]	Yes	Factoid
	AIT-QA†★ [23]	Yes	Factoid
<b>Non -table -only</b>	FeTaQA[32]	Yes	Free form
	FinQA [7]	Yes	Factoid
	TAT-QA [52]	Yes	Factoid
	HybridQA [6]	Yes	Factoid
	TabMCQ [22]	Yes	Multiple choice
	GeoTSQA [27]	Yes	Multiple choice
	OTTQA [5]	No	Factoid
	NQ-tables [17]	No	Factoid

Table 1.

### Mainline

Year Ended December 31.	2017	2016	2015
Revenue passenger miles (millions) <sup>(a)</sup>	201,351	199,014	199,467
Available seat miles (millions) <sup>(b)</sup>	243,806	241,734	239,375
Passenger load factor (percent) <sup>(c)</sup>	82.6	82.3	83.3

# How to pretrain and fine-tune on Tables?



# REFERENCES

- Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. [Chapter 14]



**THANK YOU**



## N P T E L O N L I N E C E R T I F I C A T I O N C O U R S E S

# DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 33 : Applications: Dialogue Systems-I



**PROF . PAWAN GOYAL**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## CONCEPTS COVERED

- Conversational Agents
- Open-Domain Dialogs: Response by Generation
- Open Domain Dialogs: Evaluation

# Conversational Agents

(AKA Dialogue Systems AKA Dialogue Agents AKA Chatbots)

Personal Assistants on phones or other devices

SIRI, Alexa, Cortana, Google Assistant

Playing music, setting timers and clocks

Chatting for fun

Booking travel reservations

Clinical uses for mental health

# ELIZA: Weizenbaum (1966)

Men are all alike.

**IN WHAT WAY**

They're always bugging us about something or other.

**CAN YOU THINK OF A SPECIFIC EXAMPLE**

Well, my boyfriend made me come here.

**YOUR BOYFRIEND MADE YOU COME HERE**

He says I'm depressed much of the time.

**I AM SORRY TO HEAR YOU ARE DEPRESSED**

...

You are like my father in some ways.

**WHAT RESEMBLANCE DO YOU SEE**

You are not very aggressive but I think you don't want me to notice that.

**WHAT MAKES YOU THINK I AM NOT AGGRESSIVE**

You don't argue with me.

**WHY DO YOU THINK I DON'T ARGUE WITH YOU**

You are afraid of me.

**DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU**

My father is afraid of everybody.

**WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER**

Bullies.

**DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE**

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45

Will you sing me a song?



Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of “Sing Along with Mitch”?

Yes, I would love that.

OK, I'll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I'm a teacher, so I don't have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: “Baking is my favorite thing to do in the world.”

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It's so catchy. I can't get it out of my head.

explain chatgpt like i'm five

explain chatgpt like i'm five

Sure! ChatGPT is like a really smart robot friend that can understand and talk to people using words. It has learned a lot of things by reading lots and lots of books, articles, and websites. It knows about animals, science, history, and even how to cook yummy recipes!

But remember, even though ChatGPT is really clever, it's just a computer program and doesn't have feelings or experiences like humans do. So, it's important to ask it questions and have fun with it, but also remember that it doesn't know everything and might make mistakes.

okay now explain it like i'm an adult

Certainly! ChatGPT is an advanced language model created by OpenAI called GPT-3.5. It's designed to understand and generate human-like text based on the prompts or questions it receives. It has been trained on an enormous amount of text data, including books, articles, informative answers by drawing upon the knowledge it has acquired during training. It can assist with a wide range of topics, including science, history, mathematics, literature, and more.

However, it's important to note that ChatGPT is not a human and lacks genuine understanding or awareness. It operates purely on statistical patterns and doesn't possess real-world experiences or common sense reasoning. While it can generate impressive responses, it's always important to verify the information it provides independently, especially for critical or sensitive topics.

OpenAI continues to improve and refine its language models like ChatGPT over time, but they still have limitations. It's crucial to for general information and

Send a message >

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT Max24 Version

# Two kind of conversational agents

## 1. Chatbots

- mimic informal human chatting
- for fun, or even for therapy

## 2. (Task-based) Dialogue Agents

- interfaces to personal assistants
- cars, robots, appliances
- booking flights or restaurants

# Task-oriented Dialogue

Example from a Task-oriented Dialog		
Turn	User1	User2 (or System)
1	Hi, I need to catch a train from Stevenage to Cambridge.	I can help you with that. What day will you be traveling?
2	I need to leave on Wednesday and arrive no later than 21:45.	I have train TR4015 that leaves at 19:54, and arrives at 20:43. Is that to your liking?
3	How long will that trip take?	The travel time for that train is 49 minutes

*Belief States:  
<Slots, Values>*

*Policy:  
suggest\_entity*

*Requested  
Slot Values*

*Intent: Book  
Train*

# Open-Domain Dialogs

- Schema design and data-annotation for Belief states, intents and other features becomes much more complicated
- Hence, ***representation learning*** becomes more important for building open-domain dialog systems
  - Main approaches: Encoder/Decoder Architecture Design, Loss function design

*Focus of this lecture*

## Example from an Open-domain Dialog

Turn	User1	User2 (or System)
1	Mom , what are we going to make tonight ?	Curry and rice . What do you think ?
2	That sounds nice . I can peel the potatoes first .	Really ! Oh , honey , you are so sweet . I'll chop up this carrot
3	Mom , the water is boiling !	Honey , just turn down the stove .

# Challenges in Dialog Modeling

- Interpretation of the context
- Handling background knowledge
- Extracting useful features from context
- Learning from limited data



# Corpus-based Chatbots: *Two architectures for corpus-based chatbots*



## Response by retrieval

- Use information retrieval to grab a response (that is appropriate to the context) from some corpus

## Response by generation

- Use a language model or encoder-decoder to generate the response given the dialogue context
  - *Modern corpus-based chatbots are very data-intensive*
  - *They commonly require hundreds of millions or billions of words*

# What conversations to draw on?

Transcripts of telephone conversations between volunteers

- Switchboard corpus of American English telephone conversations

Movie dialogue

- Various corpora of movie subtitles

Hire human crowdworkers to have conversations

- Topical-Chat 11K crowdsourced conversations on 8 topics
- EMPATHETICDIALOGUES 25K crowdsourced conversations grounded in a situation where a speaker was feeling a specific emotion

Pseudo-conversations from public posts on social media

- Drawn from Twitter, **Reddit**, Weibo (微博), etc.
- Tend to be noisy; often used just as pre-training.

Recent Datasets

- ShareGPT: User-LLM Chatbot conversations
- OpenAssistant (Human annotated 161,443 messages in 35 languages)

Crucial to remove personally identifiable information (PII)

<https://web.stanford.edu/~jurafsky/slp3/>

# Response by retrieval: neural IR method

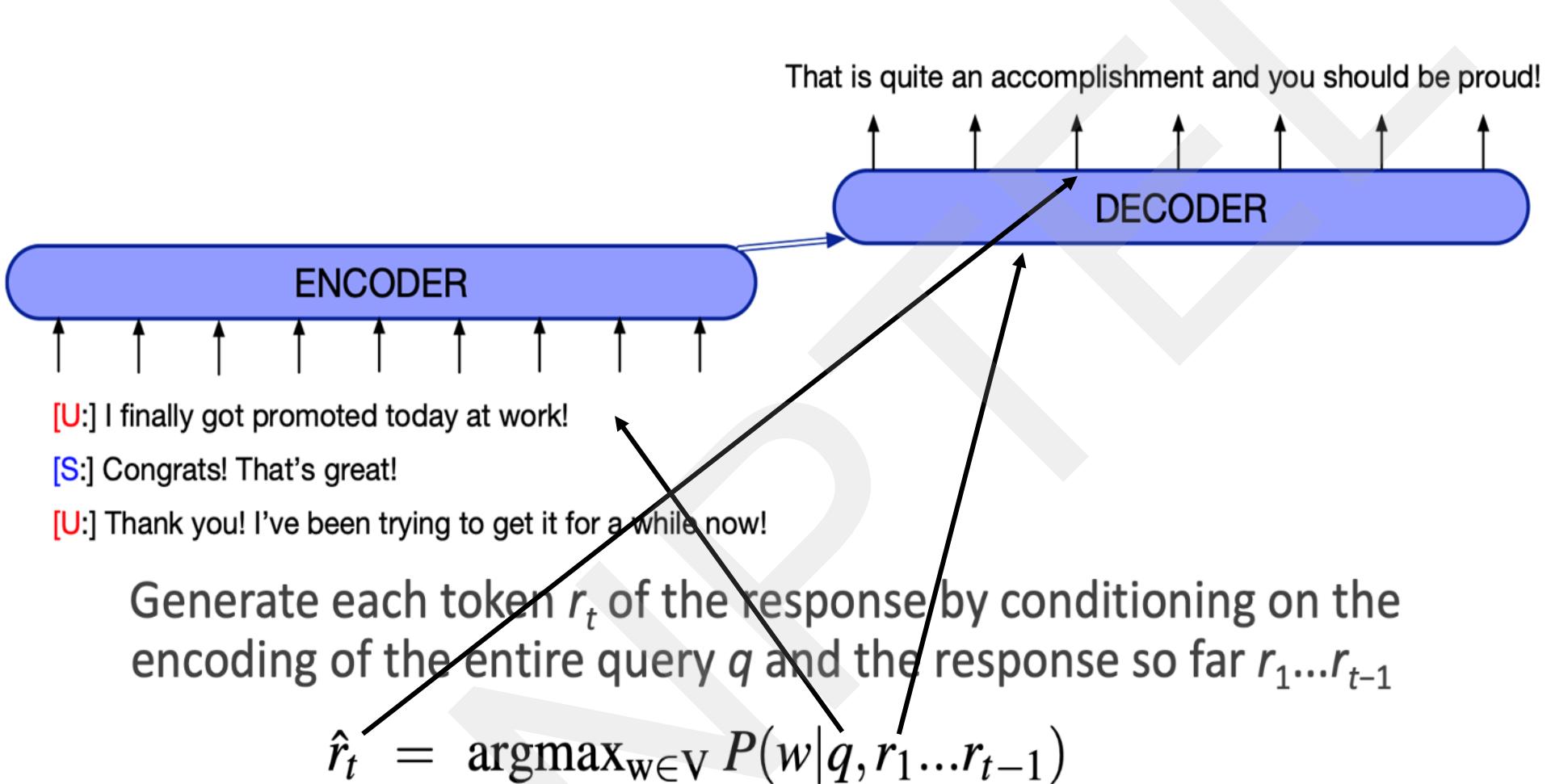
1. Given a user turn  $q$ , and a training corpus  $C$  of conversation
2. Find in  $C$  the turn  $r$  that is most similar (BERT dot product) to  $q$
3. Say  $r$

$$h_q = \text{BERT}_Q(q)[\text{CLS}]$$

$$h_r = \text{BERT}_R(r)[\text{CLS}]$$

$$\text{response}(q, C) = \underset{r \in C}{\operatorname{argmax}} h_q \cdot h_r$$

# Response by Generation



# Common problem with basic encoder-decoder: Neural chatbots can get repetitive and boring

- A: Where are you going?  
B: I'm going to the restroom.  
A: See you later.  
B: See you later.  
A: See you later.  
B: See you later.
- A: How old are you?  
B: I'm 16.  
A: 16?  
B: I don't know what you are talking about.  
A: You don't know what you are saying.  
B: I don't know what you are talking about.  
A: You don't know what you are saying.

# Fine-tune/Continue-pretrain on large dialog data

- Continue pre-training GPT-2 on conversations from Reddit
- Filter long utterances
- Filter non-English utterances
- Filter URLs
- Filter toxic comments
- Train on 147M dialog instances (1.8B words)
- “Human-level” response generation ability

DialoGPT

# Response by retrieving and refining knowledge



Can generate responses from informative text rather than dialogue

- To respond to turns like “Tell me something about Beijing”
  - Xiaoice collects sentences from public lectures and news articles.
  - And searches them using IR based on query expansion from user’s turn
- Can augment encoder-decoder model
  - use IR to retrieve passages from Wikipedia
  - concatenate each Wikipedia sentence to the dialogue context with a separator token.
  - Give as encoder context to the encoder-decoder model, which learns to incorporate text into its response

# Another common problem: Neural chatbots should have a personality

If we train on all of our data, our agent will be a mish-mash of personalities (e.g. Li et al. 2016)

---

*message* Where do you live now?  
*response* I live in Los Angeles.  
*message* In which city do you live now?  
*response* I live in Madrid.  
*message* In which country do you live now?  
*response* England, you?

---

We would like our agents to be consistent!

# PersonaChat Dataset

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi

[PERSON 2:] Hello ! How are you today ?

[PERSON 1:] I am good thank you , how are you.

[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.

[PERSON 1:] Nice ! How old are your children?

[PERSON 2:] I have four that range in age from 10 to 21. You?

[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

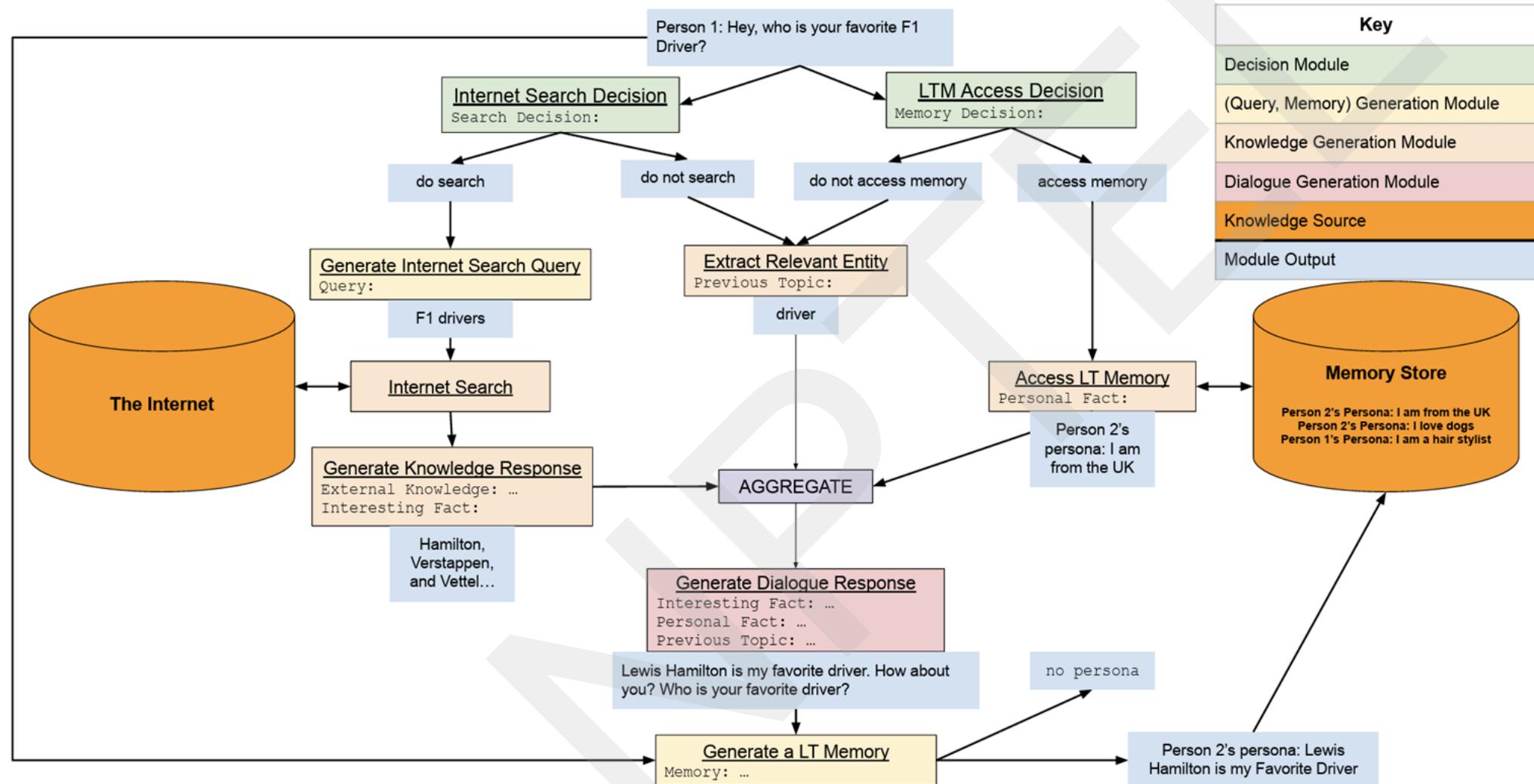
[PERSON 1:] And Cheetos at the moment!

[PERSON 2:] Good choice. Do you watch Game of Thrones?

[PERSON 1:] No, I do not have much time for TV.

[PERSON 2:] I usually spend my time painting: but, I love the show.

# BlenderBot3



# How to evaluate the Dialog Models?

Ref: They walked **to the grocery store** .  
Gen: **The woman went to the hardware store** .



Content Overlap Metrics



Model-based Metrics

Human Evaluations

# Content-Overlap Metrics

Ref: They walked **to the grocery store**.

Gen: **The woman went to the hardware store**.



- Compute a score that indicates the lexical similarity between *generated* and *gold-standard (human-written) text*
- Fast and efficient and widely used
- $N$ -gram overlap metrics (e.g., **BLEU**, ROUGE, METEOR, CIDEr, etc.)

# A Simple Failure Case

n-gram overlap metrics have no concept of semantic relatedness!

Score:

0.61

0.25

False negative

0

False positive

0.67

Reference answer

Heck yes !

Yes !

You know it !

Yup .

Heck no !

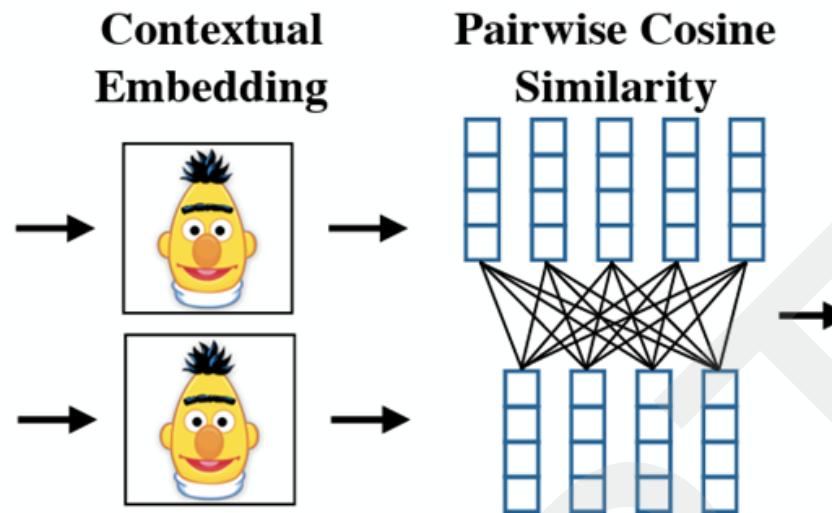
# Model-based metrics to capture more semantics

- Use **learned representations** of words and sentences to compute semantic similarity between generated and reference texts
- No more **n-gram bottleneck** because text units are represented as **embeddings!**
- The embeddings are **pretrained**, distance metrics used to measure the similarity can be **fixed**

# BERTScore

**Reference  $x$**   
*the weather is cold today*

**Candidate  $\hat{x}$**   
*it is freezing today*



## Maximum Similarity

	the	0.713	0.597	0.428	0.408	1.27
weather	-0.462	0.393	0.515	0.326	7.94	
is	-0.635	0.858	0.441	0.441	1.82	
cold	-0.479	0.454	0.796	0.343	7.90	
today	-0.347	0.361	0.307	0.913	8.88	
	it	is	freezing	today	idf weights	

## Importance Weighting (Optional)

$$R_{BERT} = \frac{(0.713 \times 1.27) + (0.515 \times 7.94) + \dots}{1.27 + 7.94 + 1.82 + 7.90 + 8.88}$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j , \quad P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j , \quad F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Importance weighting

$$\text{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^M \mathbb{I}[w \in x^{(i)}]$$

$$R_{BERT} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \text{idf}(x_i)}$$

# Dialog Evaluation is Tricky!

**Conversation History**



What are your plans for the weekend?

I'm thinking of going on a hike if the weather is nice. Do you want to join?

That sounds fun! I haven't gone hiking in awhile. Which trail were you thinking of?

**Candidate Responses**

GT?	Relevant?
Yes	Yes
No	Yes
No	No

**R1:** We should go to Red Rocks park, I heard they opened a new trail.

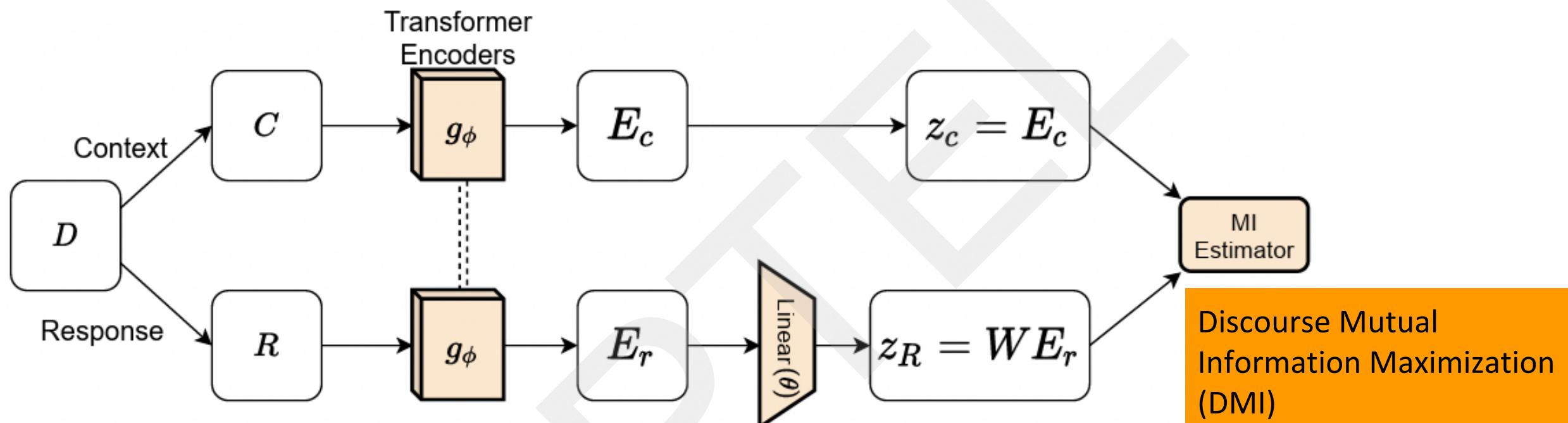
GT

**R2:** How about the summit trail at the nearby mountain? It has beautiful views.

**R3:** I'm going to the grocery store to buy some milk.

Even semantic similarity with the GT response may not suffice.  
*Need to evaluate how good is the response given the conversation history (reference-free)*

# Dialog Evaluation: Making use of the context



*We can train a model to give a high score to the good responses than the bad responses, given the context*

DMI: <https://arxiv.org/pdf/2112.05787>

# Human Evaluation

Overall or along some specific dimension:

- fluency
- coherence / consistency
- factuality and correctness
- commonsense
- style / formality
- grammaticality
- typicality
- redundancy

# REFERENCES

- Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. [Chapter 15]



**THANK YOU**



## N P T E L O N L I N E C E R T I F I C A T I O N C O U R S E S

# DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 34 : Applications: Dialogue Systems-II



**PROF . PAWAN GOYAL**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## CONCEPTS COVERED

- Task-oriented dialogue
- Frame-based dialogue agents (GUS)
- Dialogue State Architecture

# Task-oriented Dialogue

Example from a Task-oriented Dialog		
Turn	User1	User2 (or System)
1	Hi, I need to catch a train from Stevenage to Cambridge.	I can help you with that. What day will you be traveling?
2	I need to leave on Wednesday and arrive no later than 21:45.	I have train TR4015 that leaves at 19:54, and arrives at 20:43. Is that to your liking?
3	How long will that trip take?	The travel time for that train is 49 minutes

*Belief States:  
<Slots, Values>*

*Policy:  
suggest\_entity*

*Requested  
Slot Values*

*Intent: Book  
Train*

# Frame-based dialogue agents

Sometimes called "**task-based dialogue** agents"

- Systems that have the goal of helping a user solve a task like making a travel reservation or buying a product

## Architecture:

- First proposed in the GUS system of 1977
- A knowledge structure representing user intentions
- One or more **frames** (each consisting of **slots** with **values**)

## GUS: Genial Understannder System

# The Frame

A set of **slots**, to be filled with information of a given **type**

Each associated with a **question** to the user

Sometimes called a **domain ontology**

Slot	Type	Question
ORIGIN	city	"What city are you leaving from?
DEST	city	"Where are you going?
DEP DATE	date	"What day would you like to leave?
DEP TIME	time	"What time would you like to leave?
AIRLINE	line	"What is your preferred airline?

# Two basic architectures

## The **GUS** architecture

- Sometimes just called "**frame-based**" architecture
- Over 45 years old, but still used in most industrial task-based dialogue agents
- Bobrow, Daniel G., Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. "**GUS**, a frame-driven dialog system." *Artificial Intelligence* 8, 2:155-173.

## The **dialogue-state** architecture

- Extension of GUS
- More common in research systems
- Some aspects making their way into industrial systems

# Control structure for GUS frame architecture

System asks questions of user, filling any slots that user specifies

User might fill many slots at a time:

- I want a flight from San Francisco to Denver one way leaving after five p.m. on Tuesday.

When frame is filled, do database query

# GUS slots have condition-action rules attached

Some rules attached to the DESTINATION slot for the plane booking frame

1. Once the user has specified the destination
  - Enter that city as the default *StayLocation* for the hotel booking frame.
2. Once the user has specified DESTINATION DAY for a short trip
  - Automatically copy as ARRIVAL DAY.

# GUS systems have multiple frames

Frames like:

- Car or hotel reservations
- General route information
  - *Which airlines fly from Boston to San Francisco?*

Frame detection:

- System must detect which slot of which frame user is filling
- And switch dialogue control to that frame.

# GUS: Natural Language Understanding for filling dialog slots

## 1. Domain classification

weather? flight? alarm clock? ....

## 2. Intent Determination

Find a Movie, Show Flight, Remove Calendar Appt

## 3. Slot Filling

Extract the actual slots and fillers

# Natural Language Understanding for filling slots

Show me morning flights from Boston to SF on Tuesday.

DOMAIN:	AIR-TRAVEL
INTENT:	SHOW-FLIGHTS
ORIGIN-CITY:	Boston
ORIGIN-DATE:	Tuesday
ORIGIN-TIME:	morning
DEST-CITY:	San Francisco

# Natural Language Understanding for filling slots

Wake me tomorrow at six.

**DOMAIN:** ALARM-CLOCK

**INTENT:** SET-ALARM

**TIME:** 2017-07-01 0600-0800

# Dialogue-State or Belief-State Architecture



A more sophisticated version of the frame-based architecture

- Has dialogue acts, more ML, better generation

The basis for modern research systems

Slowly making its way into industrial systems

- Some aspects (ML for slot-understanding) already widely used industrially

# Components in a dialogue-state architecture

**NLU:** extracts slot fillers from the user's utterance using machine learning

**Dialogue state tracker:** maintains the current state of the dialogue (user's most recent dialogue act, set of slot-filler constraints from user)

**Dialogue policy:** decides what the system should do or say next

- GUS policy: ask questions until the frame was full then report back
- More sophisticated: know when to answer questions, when to ask a clarification question, etc.

**NLG:** produce more natural, less templated utterances

# Slot filling: Machine learning

Machine learning classifiers to map words to semantic frame-fillers

Given a set of labeled sentences

Input: "I want to fly to San Francisco on Monday please"

Output: Destination: SF

Depart-time: Monday

Build a classifier to map from one to the other

Requirements: Lots of labeled data

# Slot filling as sequence labeling: BIO tagging

## The BIO tagging paradigm

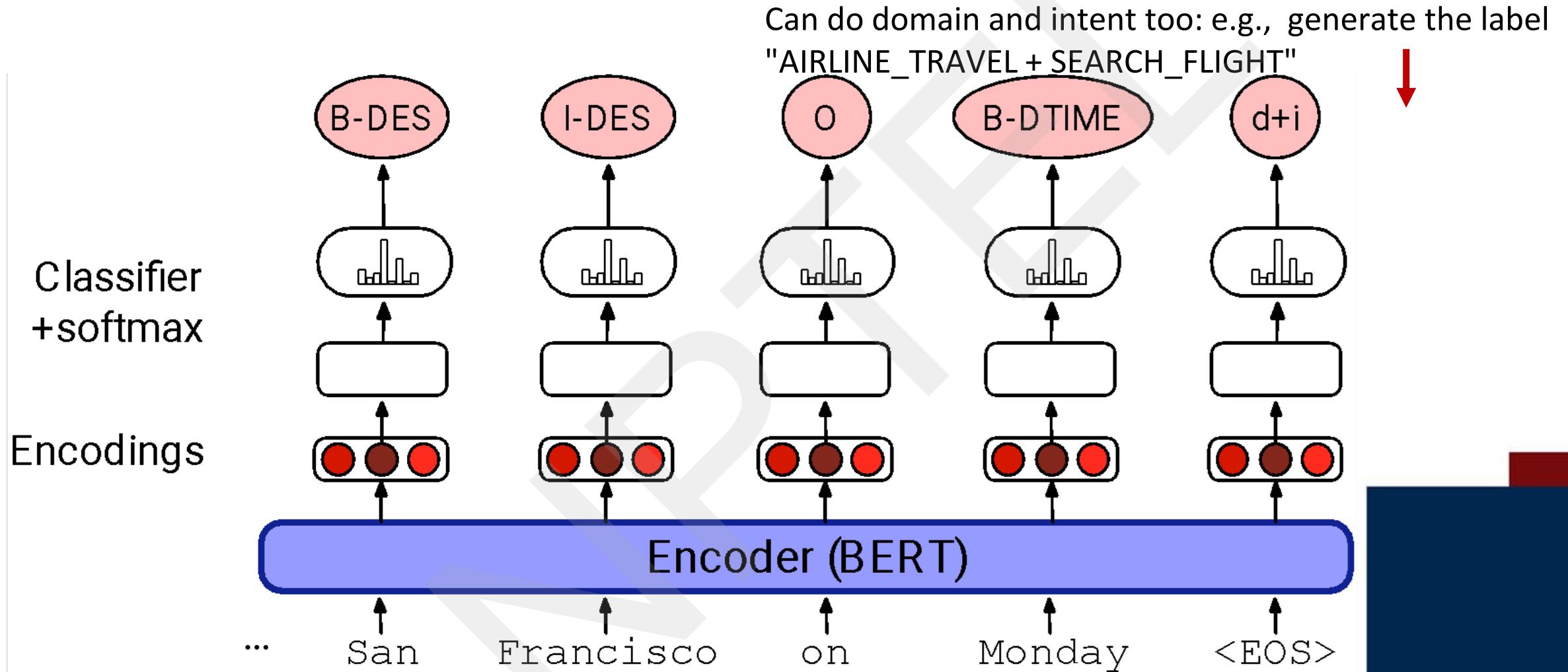
Idea: Train a classifier to label each input word with a tag that tells us what slot (if any) it fills

0	0	0	0	B-DES	I-DES	0	B-DEPTIME	I-DEPTIME	0	
I	want	to	fly	to	San	Francisco	on	Monday	afternoon	please

We create a *B* and *I* tag for each slot-type

And convert the training data to this format

# Slot filling using contextual embeddings



# Once we have the BIO tag of the sentence

0	0	0	0	0	B-DES	I-DES	0	B-DEPTIME	I-DEPTIME	0	0
I	want	to	fly	to	San	Francisco	on	Monday	afternoon	please	

- We can extract the filler string for each slot
- And then normalize it to the correct form in the ontology
- Like "SFO" for San Francisco
- Using homonym dictionaries (SF=SFO=San Francisco)

# The task of dialogue state tracking

User: I'm looking for a cheaper restaurant  
`inform(price=cheap)`

System: Sure. What kind - and where?

User: Thai food, somewhere downtown  
`inform(price=cheap, food=Thai, area=centre)`

System: The House serves cheap Thai food

User: Where is it?  
`inform(price=cheap, food=Thai, area=centre); request(address)`

System: The House is at 106 Regent Street

Example from Mrkšić, N., O Séaghdha, D., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. *ACL*.

# Dialogue state tracking

I'd like Cantonese food near the Mission district.

→

inform(food=cantonese, area=mission) .

Dialogue act interpretation algorithm:

- 1-of-N supervised classification to choose inform
- Based on encodings of current sentence + prior dialogue acts

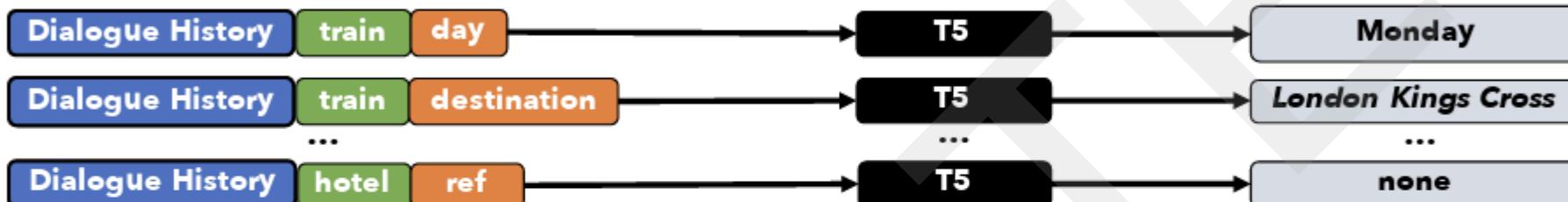
Simple dialogue state tracker:

- Run a slot-filler after each sentence

# Dialogue state tracking: Can use generative models



(a) Generation-based DST w/ Sequential Decoding



(b) Schema-Based Prompt DST w/ Independent Decoding



(c) Natural Language Augmented Prompt DST w/ Independent Decoding

# Natural Language Generation

NLG in information-state architecture modeled in two stages:

- **content planning** (what to say)
- **sentence realization** (how to say it)

We'll focus on sentence realization here.

# Sentence Realization

Assume content planning has been done by the dialogue policy

- Chosen the dialogue act to generate
- Chosen some attributes (slots and values) that the planner wants to say to the user
  - Either to give the user the answer, or as part of a confirmation strategy)

## 2 samples of Input and Output for Sentence Realizer

```
recommend(restaurant name= Au Midi, neighborhood = midtown,  
cuisine = french
```

- 1 Au Midi is in Midtown and serves French food.
- 2 There is a French restaurant in Midtown called Au Midi.

```
recommend(restaurant name= Loch Fyne, neighborhood = city  
centre, cuisine = seafood)
```

- 3 Loch Fyne is in the City Center and serves seafood food.
- 4 There is a seafood restaurant in the City Centre called Loch Fyne.

# Sentence Realization

Training data is hard to come by

- Don't see each restaurant in each situation

Common way to improve generalization:

- **Delexicalization:** replacing words in the training set that represent slot values with a generic placeholder token:

```
recommend(restaurant name= Au Midi, neighborhood = midtown,  
cuisine = french
```

- 1 Au Midi is in Midtown and serves French food.
- 2 There is a French restaurant in Midtown called Au Midi.

# Sentence Realization

```
recommend(restaurant name= Au Midi, neighborhood = midtown,  
cuisine = french
```

- 1 **restaurant\_name** is in **neighborhood** and serves **cuisine** food.
- 2 There is a **cuisine** restaurant in **neighborhood** called **restaurant name**.

Training data is hard to come by

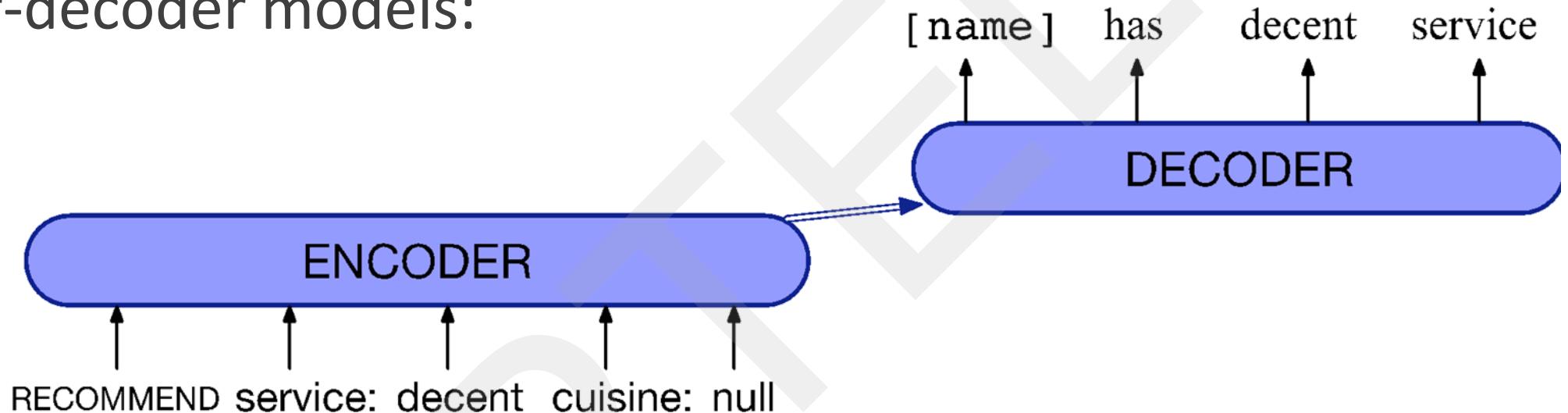
- Don't see each restaurant in each situation

Common way to improve generalization:

- **Delexicalization:** replacing words in the training set that represent slot values with a generic placeholder token:

# Sentence Realization: mapping from frames to delexicalized sentences

Encoder-decoder models:



Output:

restaurant\_name has decent service

Relexicalize to:

Au Midi has decent service

# Evaluation Metrics

## Evaluation of Delexicalized Response Generation Task:

- **Delexicalized-BLEU:** BLEU score on responses with slot values replaced by slot tags (e.g. [HOTEL\_AREA])
- **Inform:** Measures how often all the entities provided by the system are correct
- **Success:** Reflects how often the system is able to correctly answer all the requested attributes
- **Entity-F1:** F1-score for the entities returned to the user through the generated response

**Combined Score:** Used for selecting best checkpoint/hyperparameters or a model.

$$\text{Score} = \text{BLEU} + 0.5 (\text{Inform} + \text{Success})$$

# REFERENCES

- Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. [Chapter 15]



**THANK YOU**



## N P T E L O N L I N E C E R T I F I C A T I O N C O U R S E S

# DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 35 : Applications: Text Summarization



**PROF . PAWAN GOYAL**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

# CONCEPTS COVERED

- What is Text Summarization ✓
- Summary of methods used for summarization
- Domain Specific Example ✓
- Summarizing Long Documents ↗
- Evaluation ↘

# What is Text summarization?

- Automatically condensing the given document(s) to present the most relevant information in a concise and quickly- readable format.
- Extensively researched problem : From rule- based methods in the 1950s to Large Language Models (LLMs) in recent years.

# Various (kinds of) Text Summarization

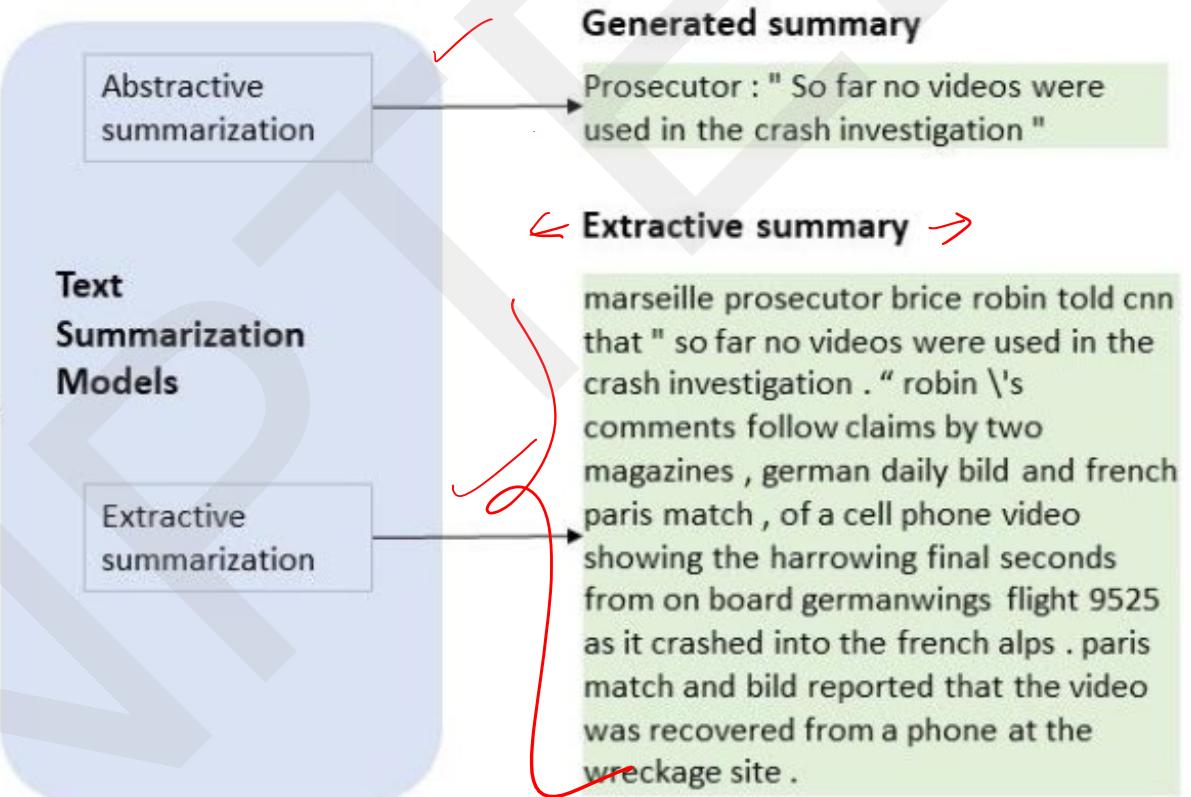
- Based on **output**:

*Extractive vs. Abstractive vs. Hybrid*

## Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane.

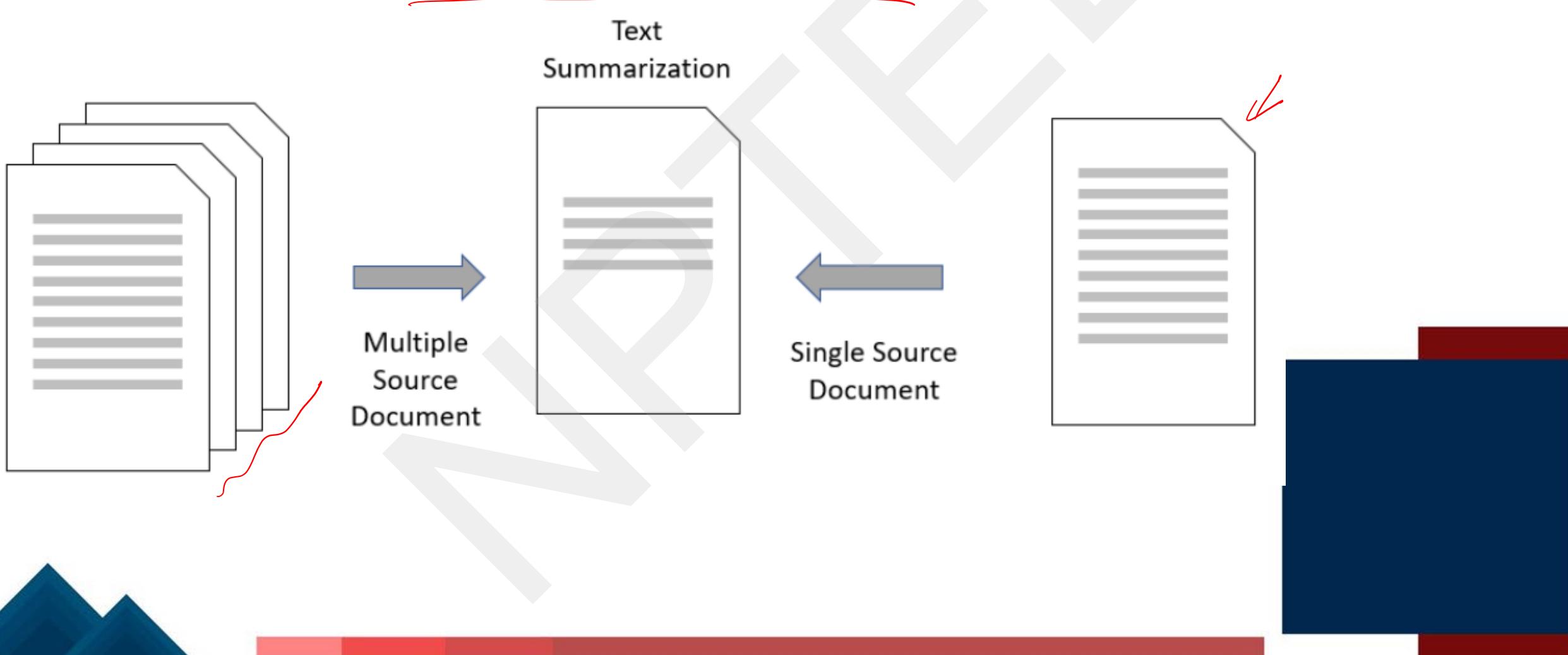
Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\ 's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...



*The length can be similar to extractive*

# Various (kinds of) Text Summarization

- Based on the number of input documents:  
*Single Document vs. Multi- Document Summarization*



# Various (kinds of) Text Summarization



- Based on content/genre:  
Generic vs. Domain-Specific



Live  
Law

2022 LiveLaw (SC) 234

IN THE SUPREME COURT OF INDIA  
CRIMINAL APPELLATE JURISDICTION  
INDIRA BANERJEE; V. RAMASUBRAMANIAN  
MARCH 02, 2022

CRIMINAL APPEAL No. .... OF 2022  
(Arising out of SLP (Crl.) No.717 of 2020)

KARAN SINGH VERSUS THE STATE OF UTTAR PRADESH & ORS.  
Criminal Trial - The prosecution is required to prove its case beyond reasonable doubt and not beyond all iota of doubt. (Para 46)

Criminal Trial - The court is not supposed to give undue importance to omissions, contradictions and discrepancies which do not go to the heart of the matter, and shake the basic version of the prosecution witness. [Referred to Rohit Kumar v. State of Haryana (2013) 14 SCC 434]

Summary: Appeal against High Court judgment upholding conviction of accused in a murder case - dismissed - The prosecution proved its case beyond reasonable doubt - The fact that the trial/appeal should have taken years and that other accused should have died during the appeal cannot be a ground for acquittal.

(Arising out of impugned final judgment and order dated 30-07-2018 in CRLA No. 1813/1983 passed by the High Court of Judicature at Allahabad)

For Petitioner(s) Mr. Ekanash Agarwal, Adv. Mr. Sanjeev Agarwal, AOR  
For Respondent(s) Mr. Sanjay Kumar Tyagi, AOR Mr. Prabhat Kumar Rai, Adv. Mr. Ajay Kumar Pandey, Adv. Mr. Jagat Singh Chahra, AOR

JUDGMENT

Indira Banerjee, J.

Leave granted.

2. This appeal is against a judgment and order dated 30-July 2018 passed by a Division Bench of the Allahabad High Court, dismissing Criminal Appeal No. 1813 of 1983 filed by the Appellant along with other accused persons, against a judgment and order of conviction dated 1-August 1983 passed by the VI Additional District and Sessions Judge, Shahjahanpur in Sessions Trial No. 268 of 1981, inter alia, convicting the Appellant of the offences under Section 302 read with Section 149, Section 307 read with Section 149 and Section 148 of the Indian Penal Code (hereinafter referred to as "the IPC") and sentencing him, inter alia, to life

COLGATE-PALMOLIVE COMPANY Consolidated Statements of Income For the years ended December 31, (Dollars in Millions Except Per Share Amounts)			
	2015	2014	2013
Net sales	16,034	\$ 17,277	\$ 17,420
Cost of sales	6,635	7,168	7,219
Gross profit	9,399	10,109	10,201
Selling, general and administrative expenses	5,464	5,982	6,223
Other (income) expense, net	62	570	422
Charge for Venezuela accounting change	1,084	—	—
Operating profit	2,789	3,557	3,556
Interest (income) advance, net	26	24	(9)
Order Code 96-334 Updated April 15, 2008	2,763	3,533	3,565
	1,215	1,194	1,155
	1,548	2,339	2,410
	164	159	169
	\$ 1,384	\$ 2,180	\$ 2,241
	\$ 1,53	\$ 2.38	\$ 2.41
	\$ 1.52	\$ 2.36	\$ 2.38

## Financial Results



CRS Report for Congress

Provisions of Special Rules in the House:  
An Example of a Typical Open Rule

Judy Schneider  
Specialist on Congress  
Government and Finance Division

This report includes a typical example of a simple open rule that the House Committee on Rules may report to govern House floor action on a bill that is not otherwise privileged for consideration. This resolution has been divided into five parts. See <http://www.crs.gov/pubs/works/guide/conc.html> for more information on legislative process.

The first part of the rule makes the bill in order for floor consideration by authorizing the speaker to call it up for consideration at any time during the course of the bill's life. Without this authority, a motion for the same purpose would not be in order; it would not be privileged to interrupt the regular daily order of business on the House floor.

The second part waives a reading of the bill. It also prevents general debate on the bill by setting the amount of time for the debate, by dividing control of this time, usually between the chairman and ranking minority member of the committee that reported the bill, and by requiring that all general debate be relevant to the subject of the bill.

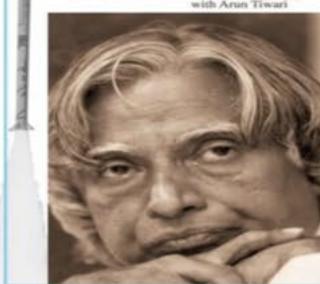
The third part merely states that the bill shall be read for amendment and that each Member may speak for five minutes on each amendment. By implication, this part also means that the bill is to be read for amendment one section at a time. Further, as each section is read, the chairman or ranking member of the committee that reported the bill may offer an amendment to the section. This part of the rule is designed to prevent the chairman from amending the bill in a way that would not be acceptable to the rest of the House.

The fourth part provides for the Committee of the Whole, after disposing of the last amendment, to transform itself back into the House, and report the bill to the House with whatever amendments the Committee of the Whole adopted. This provision eliminates the need for the House to vote on the bill as a whole, because the committee that reported the bill does not vote on the bill as a whole, and the committee may not actually amend the bill; it only makes recommendations to the House on amendments.

The fifth and final part of the rule expedites final House passage of the bill by precluding almost all debate in the House and all other actions except those necessary for the House to vote on the amendments the Committee of the Whole recommended and to

Congressional Research Service — The Library of Congress  
Prepared for Members and Committees of Congress

## APJ ABDUL KALAM WINGS OF FIRE An Autobiography with Arun Tiwari



## Pre-neural approaches

LexRank / Centroid method

*PageRank*

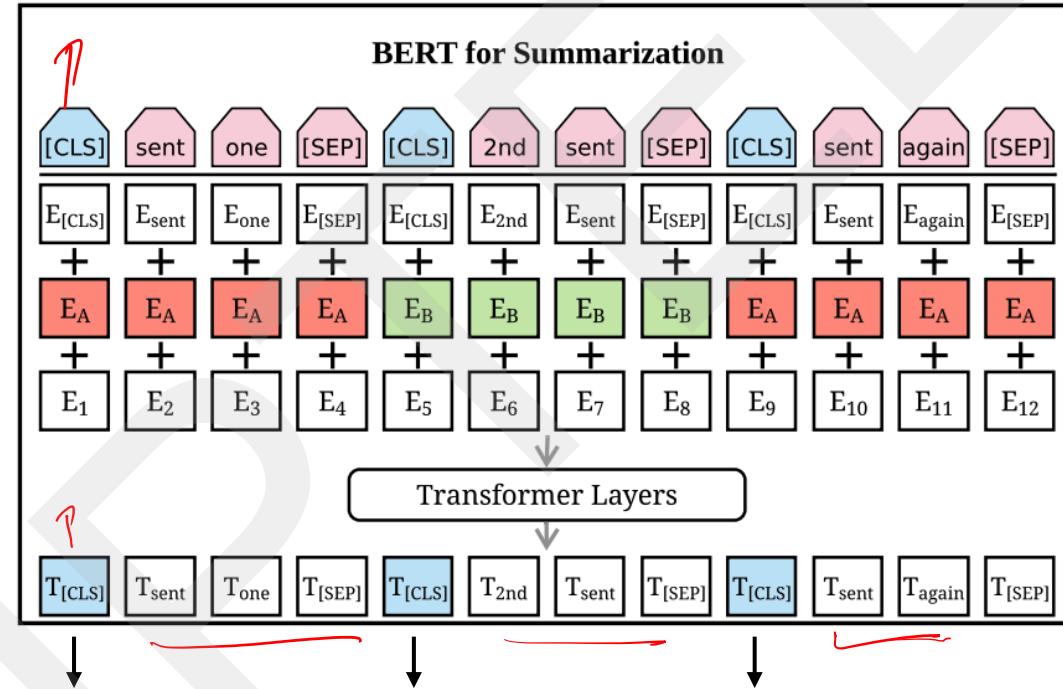
Integer linear programming

*S*

Maximum marginal relevance

*P*

## BERTSum [Liu and Lapata, 2019]



token embeddings

segment embeddings

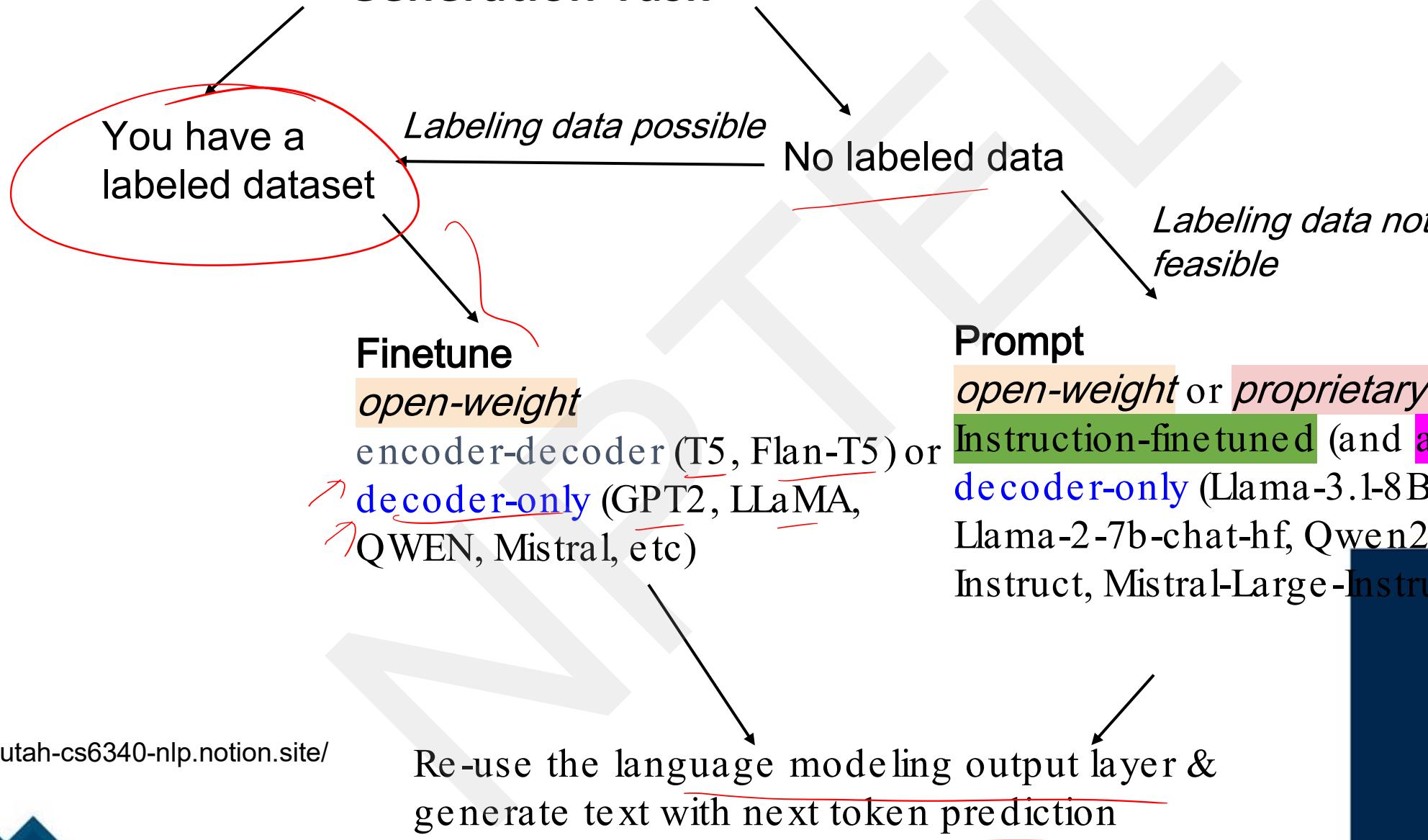
positional embeddings

### Binary classification:

0  $\Rightarrow$  Do not include in the summary

1  $\Rightarrow$  Include in the summary

# Generation Task

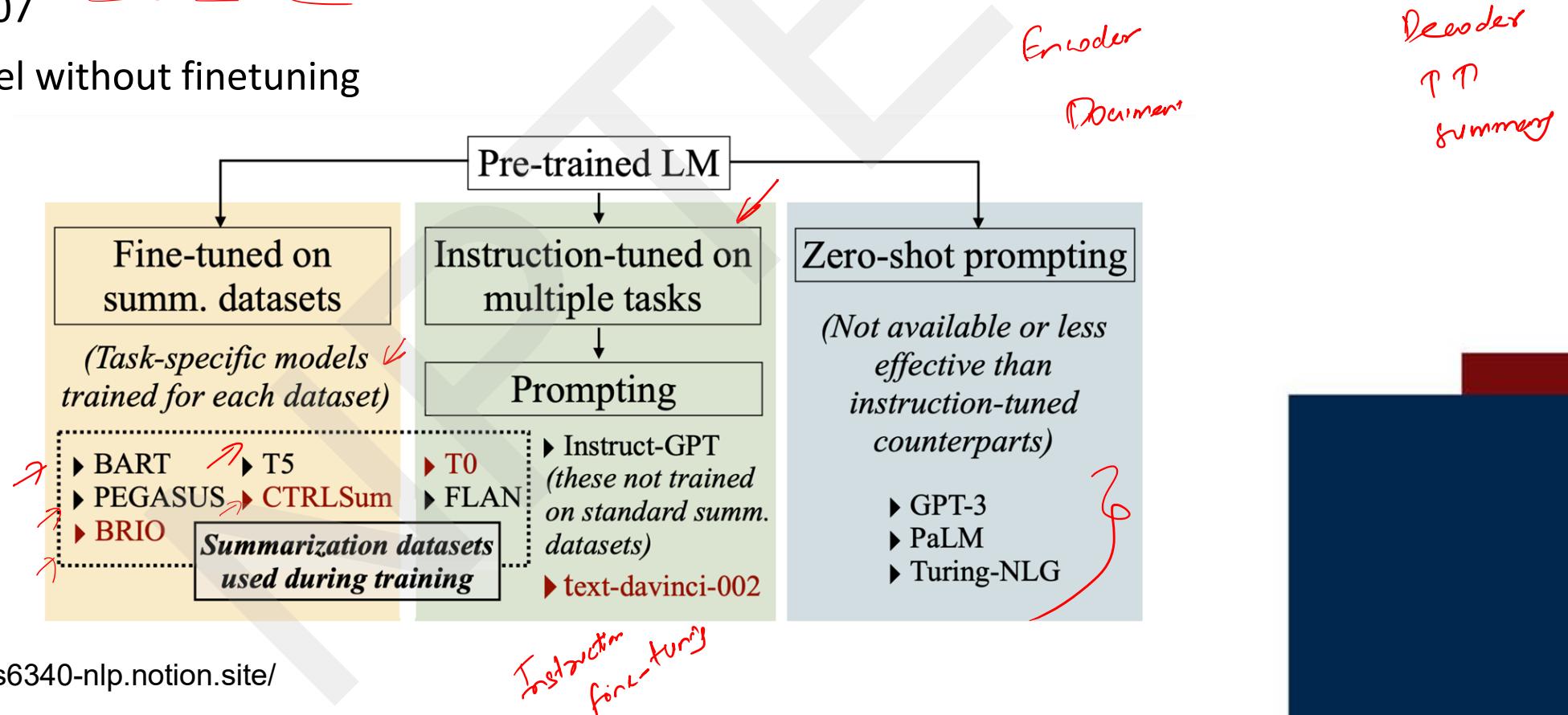


# Seq2seq abstractive summarization

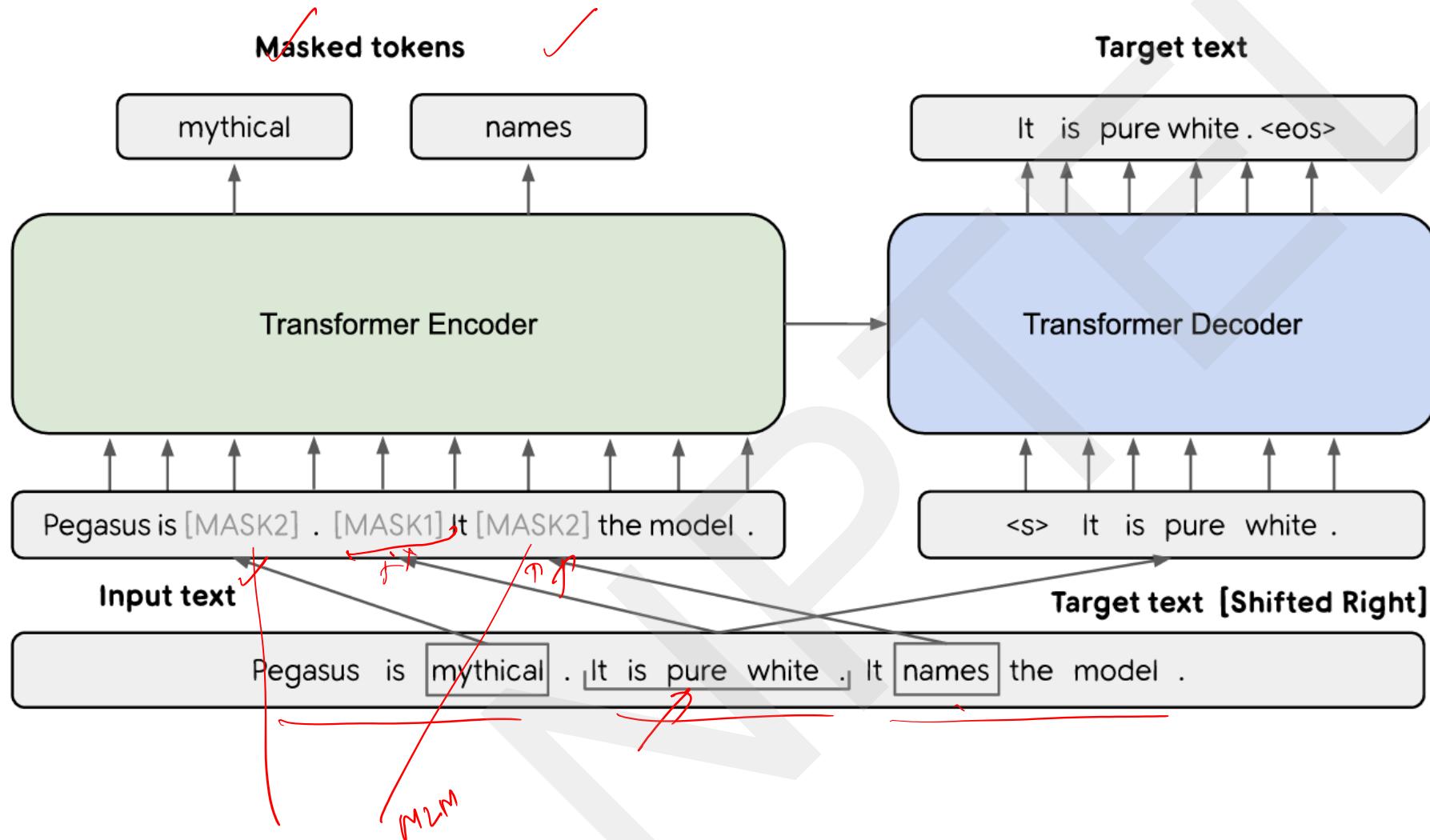
Input : Document(s) ↗  
 Out put : Summary ↗

Finetune a model like GPT2, T5, BART, or these days Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Mistral-Large-Instruct-2407

Or prompt a model without finetuning



# Pegasus Architecture (Pretraining)



Both **Gap Sentence Generation (GSG)** and **MLM** are applied simultaneously as pre-training objectives. Originally there are three sentences. One sentence is masked with [MASK1] and used as target generation text (GSG). The other two sentences remain in the input, but some tokens are randomly masked by [MASK2] (MLM)

# Selecting Gap Sentences to approximate a summary

We consider 3 primary strategies for selecting  $\underline{m}$  gap sentences without replacement from a document,  $D = \{x_i\}_n$ , comprised of  $n$  sentences:

**Random** Uniformly select  $\underline{m}$  sentences at random.

<https://github.com/google-research/pegasus>

**Lead** Select the first  $\underline{m}$  sentences.

**Principal** Select top- $\underline{m}$  scored sentences according to importance. As a proxy for importance we compute ROUGE1-F1 (Lin, 2004) between the sentence and the rest of the document,  $s_i = rouge(x_i, D \setminus \{x_i\}), \forall i$ .

<https://arxiv.org/pdf/1912.08777>

# Domain Specific Example: Summarizing ECTs



Earnings Calls Transcripts (ECTs) are long and unstructured documents with an **average length of 3K words**.

We collect publicly hosted transcripts from **The Motley Fool** to form our documents.

For this data collection, we targeted the ECTs of Russell 3000 Index companies posted between January 2019 and April 2022.

# Reference Summaries - Corresponding Reuters Articles



- QUARTERLY EARNINGS PER SHARE \$ 1.52.
- QUARTERLY TOTAL NET SALES \$97.28 BILLION VERSUS \$89.58 BILLION REPORTED LAST YEAR.
- BOARD OF DIRECTORS AUTHORIZED AN INCREASE OF \$90 BILLION TO THE EXISTING SHARE REPURCHASE PROGRAM.
- QUARTERLY IPHONE REVENUE \$50.57 BILLION VERSUS \$47.94 BILLION REPORTED LAST YEAR.
- Q2 EARNINGS PER SHARE VIEW \$1.43, REVENUE VIEW \$93.89 BILLION -- REFINITIV IBES DATA.

Experts- written **Reuters** articles corresponding to these earnings calls are collected as the reference summaries.

Summary articles consist of a concise set of **telegram-style bullet points** that precisely capture important metrics and numbers discussed in the calls.

3000 words  
↳

**Average length** of a summary is around 50 words.

# Long Document Summarization Problem

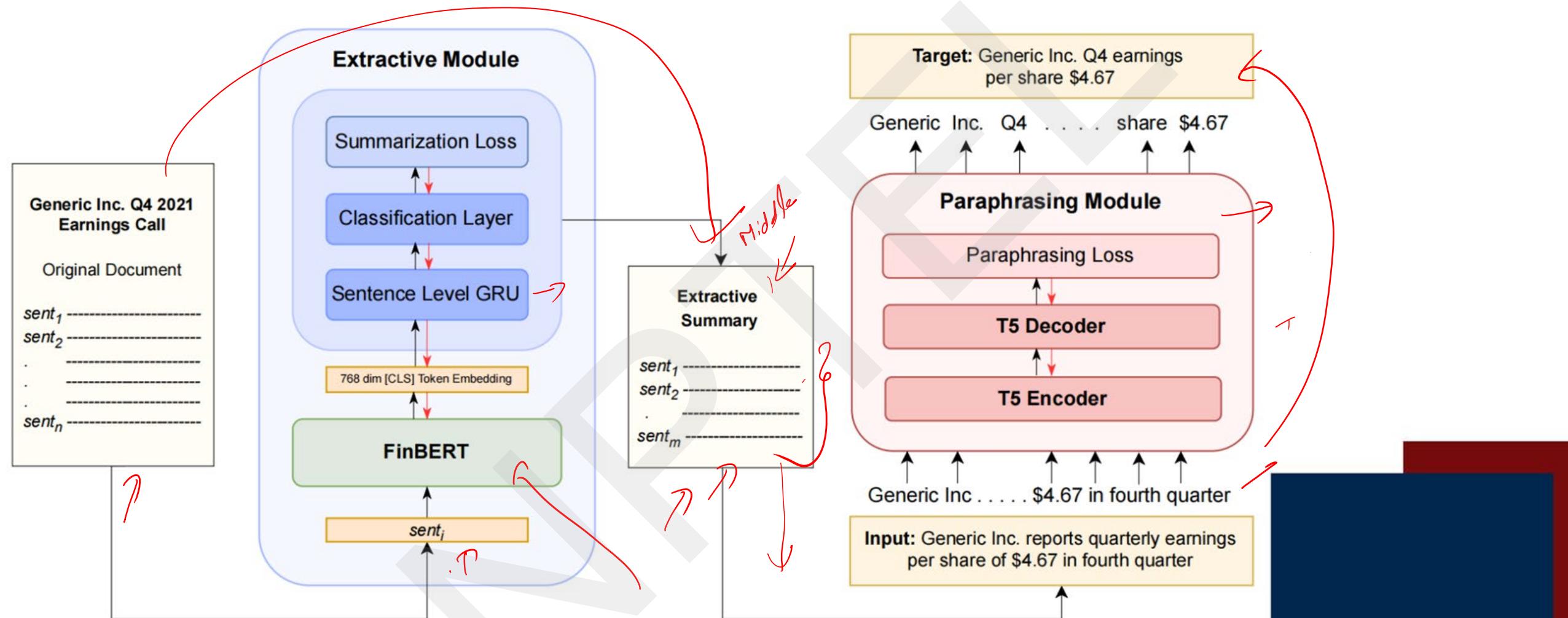
Dataset	# Docs.	Coverage	Density	Compression Ratio	# Tokens	
					Doc.	Summary
ARXIV/PUBMED ↗	346,187	0.87	3.94	31.17	5179.22	257.44
BILLSUM	23,455	-	4.12	13.64	1813.0	207.7
BIGPATENT ↗	1,341,362	0.86	2.38	36.84	3629.04	116.67
GOVREPORT	19,466	-	7.60	19.01	9409.4	553.4
BOOKSUM Chapters	12,293	0.78	1.69	15.97	5101.88	505.32
ECTSum ↗	2,425	0.85	2.43	103.67	2916.44	49.23

P

**Coverage** and **Density** quantify the extent to which a summary is derivative of the source text.

**Compression Ratio** represents the ratio between # Tokens in a document and its summary.

# Model Architecture: Extract-then-Paraphrase



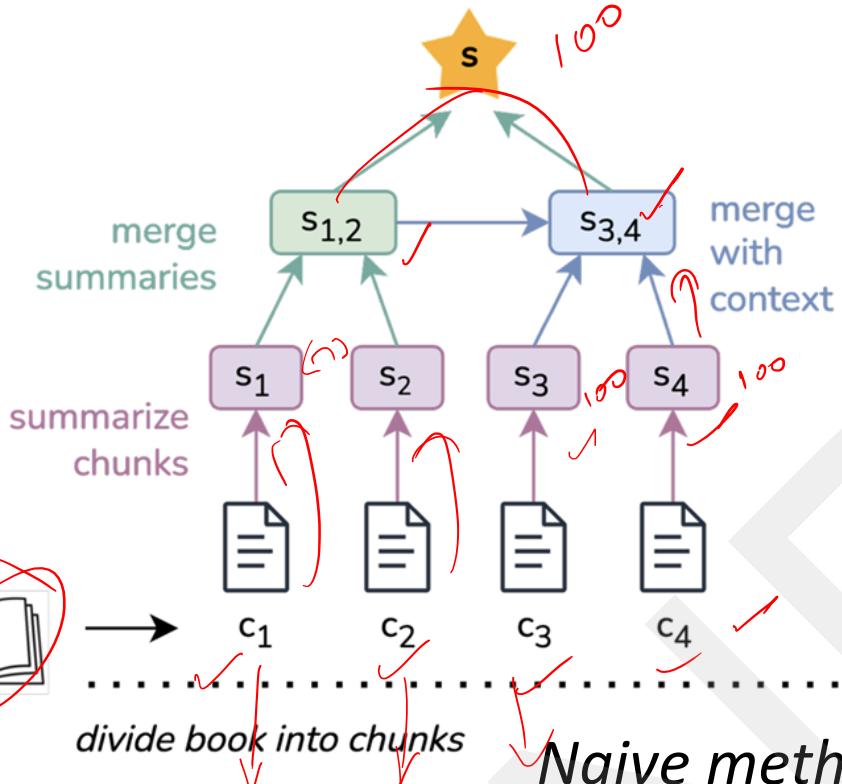
# Main Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Num-Prec.	SummaC <sub>CONV</sub>
<b>Unsupervised</b>						
LexRank (Erkan and Radev, 2004)	0.122	0.023	0.154	0.638	-	-
DSDR (He et al., 2012)	0.164	0.042	0.200	0.662	-	-
PacSum (Zheng and Lapata, 2019)	0.167	0.046	0.205	0.663	-	-
<b>Extractive</b>						
SummaRuNNer (Nallapati et al., 2017)	0.273	0.107	0.309	0.647	-	-
BertSumExt (Liu and Lapata, 2019)	0.307	0.118	0.324	0.667	-	-
MatchSum (Zhong et al., 2020)	0.314	0.126	0.335	0.679	-	-
<b>Abstractive</b>						
BART (Lewis et al., 2020)	0.327	0.153	0.361	0.692	0.594	0.431
Pegasus (Zhang et al., 2020b)	0.334	0.185	0.375	0.708	0.783	0.444
T5 (Raffel et al., 2020)	0.363	0.209	0.413	0.728	0.796	0.508
<b>Long Document Summarizers</b>						
BigBird (Zaheer et al., 2020)	0.344	0.252	0.400	0.716	0.844	0.452
LongT5 (Guo et al., 2021)	0.438	0.267	0.471	0.732	0.812	0.516
LED (Beltagy et al., 2020)	0.450	0.271	0.498	0.737	0.679	0.439
<b>Ours</b>						
ECT-BPS w/o Paraphrasing	0.313	0.137	0.351	0.714	-	-
<b>ECT-BPS</b>	<b>0.467</b>	<b>0.307</b>	<b>0.514</b>	<b>0.764</b>	<b>0.916</b>	<b>0.518</b>

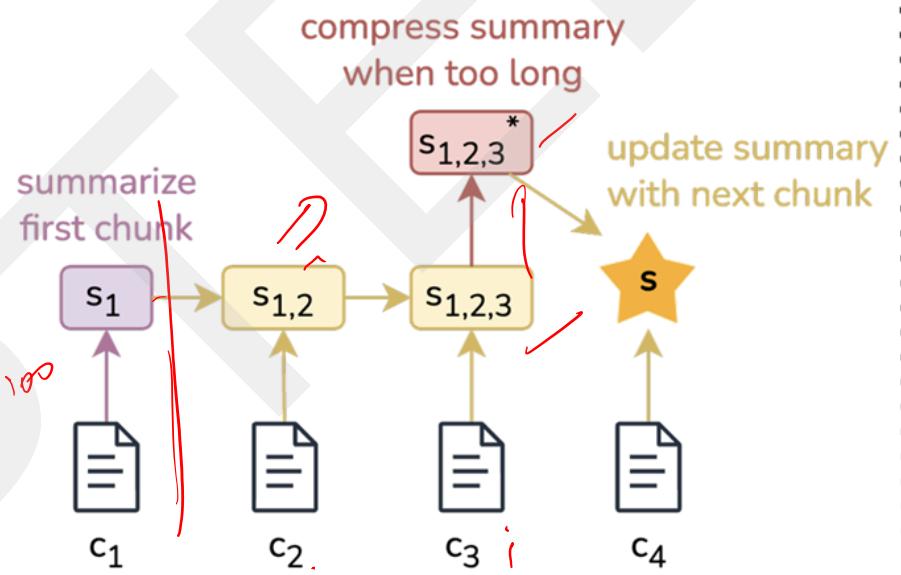
*Domain-specific  
Evaluation metrics*

# How to handle very long documents/books?

## hierarchical merging



## incremental updating



*Naive method is to generate summaries for each chunk and combine*

<https://arxiv.org/pdf/2310.00785>

# Evaluation: BLEU (Bilingual Evaluation Understudy)

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

$$\text{BLEU} = \min \left( 1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$

*Originally designed for MT evaluation, but is widely used for NLG evaluation*

# Evaluation: BLEU (Bilingual Evaluation Understudy)

SYSTEM A: Israeli officials 2-GRAM MATCH

responsibility of airport 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security 2-GRAM MATCH

Israeli officials are responsible 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

# ROUGE - a recall-based counterpart to BLEU

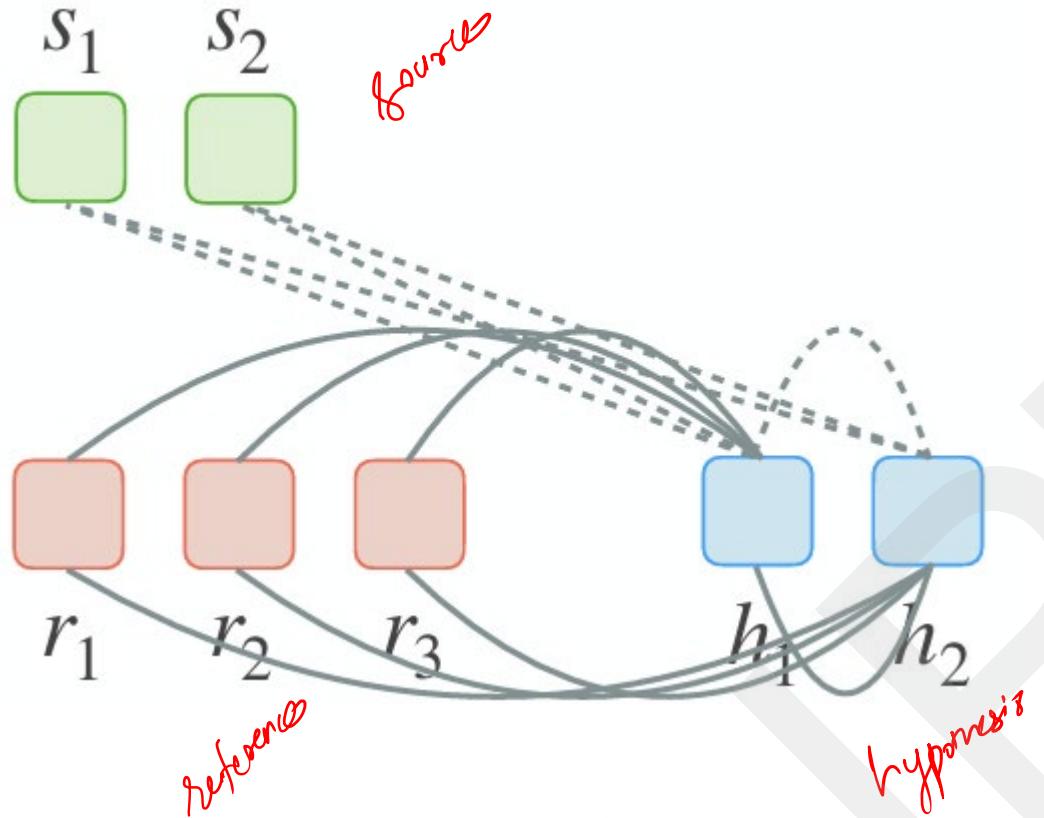
$$\text{ROUGE-n} = \frac{\text{Number of overlapping n-grams}}{\text{Total n-grams in the reference summary}}$$

ROUGE2 has been shown to correlate somewhat well with human judgments for multi-document *extractive* summarization tasks

## Example

- Reference summary (RS): The quick brown fox jumps over the lazy dog.
- Generated summary (GS): A quick brown fox leaped over a lazy dog.
- Bigrams in RS: [the quick, quick brown, brown fox, fox jumps, jumps over, over the, the lazy, lazy dog] ✓ 8
- Bigrams in GS: [a quick, quick brown, brown fox, fox leaped, leaped over, over a, a lazy, lazy dog]
- Overlapping Bigrams: [quick brown, brown fox, lazy dog] (3 overlapping bigrams)
- ROUGE-2 = Overlapping Bigrams / Bigrams in RS = 3 / 8

# BARTScore



$$\text{BARTSCORE} = \sum_{t=1}^m \omega_t \log p(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}, \theta)$$

we can use the weighted log probability of one text  $y$  given another text  $x$ . Weights can be *IDF* / *uniform*.

# BARTScore: flexible in various eval. scenarios

**Faithfulness ( $s \rightarrow h$ )**: from source document to hypothesis  $p(h|s, \theta)$ . This direction measures how likely it is that the hypothesis could be generated based on the source text.

**Precision ( $r \rightarrow h$ )**: from reference text to system-generated text  $p(h|r, \theta)$ . This direction assesses how likely the hypothesis could be constructed based on the gold reference

**Recall ( $h \rightarrow r$ )**: from system-generated text to reference text  $p(r|h, \theta)$ . This version quantifies how easily a gold reference could be generated by the hypothesis

$$\text{BARTSCORE} = \sum_{t=1}^m \omega_t \log p(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}, \theta)$$

we can use the weighted log probability of one text  $y$  given another text  $x$ . Weights can be *IDF / uniform*.

# REFERENCES

Various papers cited in the slides



**THANK YOU**