A Brief Introduction to Text Summarization

Irene Li, Project Assistant Professor

LiLab, University of Tokyo

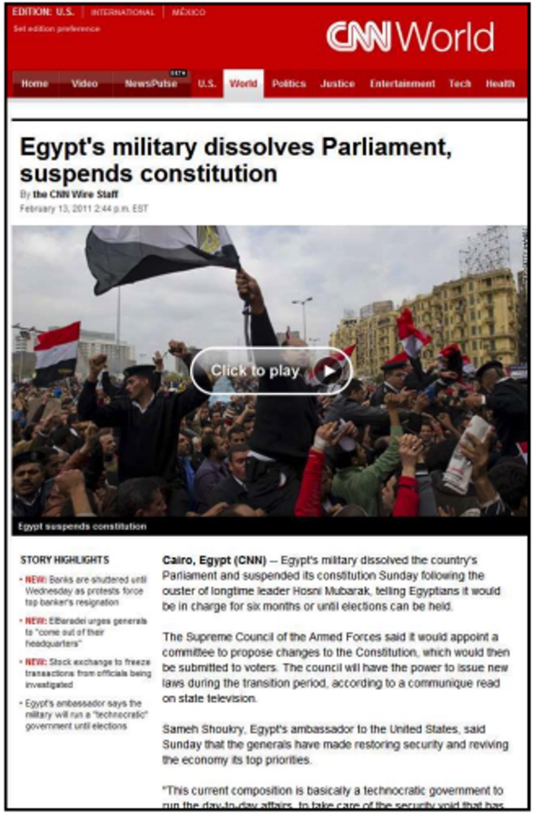
Outline

- Introduction
- Traditional and Recent Methods
- Evaluation Methods
- Extension

Introduction

Concepts

What is text summarization?



Egypt's military dissolves Parliament, suspends constitution

By **the CNN Wire Staff**
February 13, 2011 2:44 p.m. EST

Egypt suspends constitution

STORY HIGHLIGHTS

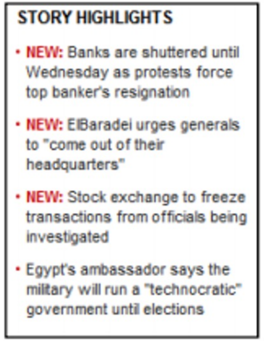
- **NEW:** Banks are shuttered until Wednesday as protests force top banker's resignation
- **NEW:** ElBaradei urges generals to "come out of their headquarters"
- **NEW:** Stock exchange to freeze transactions from officials being investigated
- Egypt's ambassador says the military will run a "technocratic" government until elections

Cairo, Egypt (CNN) -- Egypt's military dissolved the country's Parliament and suspended its constitution Sunday following the ouster of longtime leader Hosni Mubarak, telling Egyptians it would be in charge for six months or until elections can be held.

The Supreme Council of the Armed Forces said it would appoint a committee to propose changes to the Constitution, which would then be submitted to voters. The council will have the power to issue new laws during the transition period, according to a communique read on state television.

Sameh Shoukry, Egypt's ambassador to the United States, said Sunday that the generals have made restoring security and reviving the economy its top priorities.

"This current composition is basically a technocratic government to run the day-to-day affairs, to take care of the security void that has

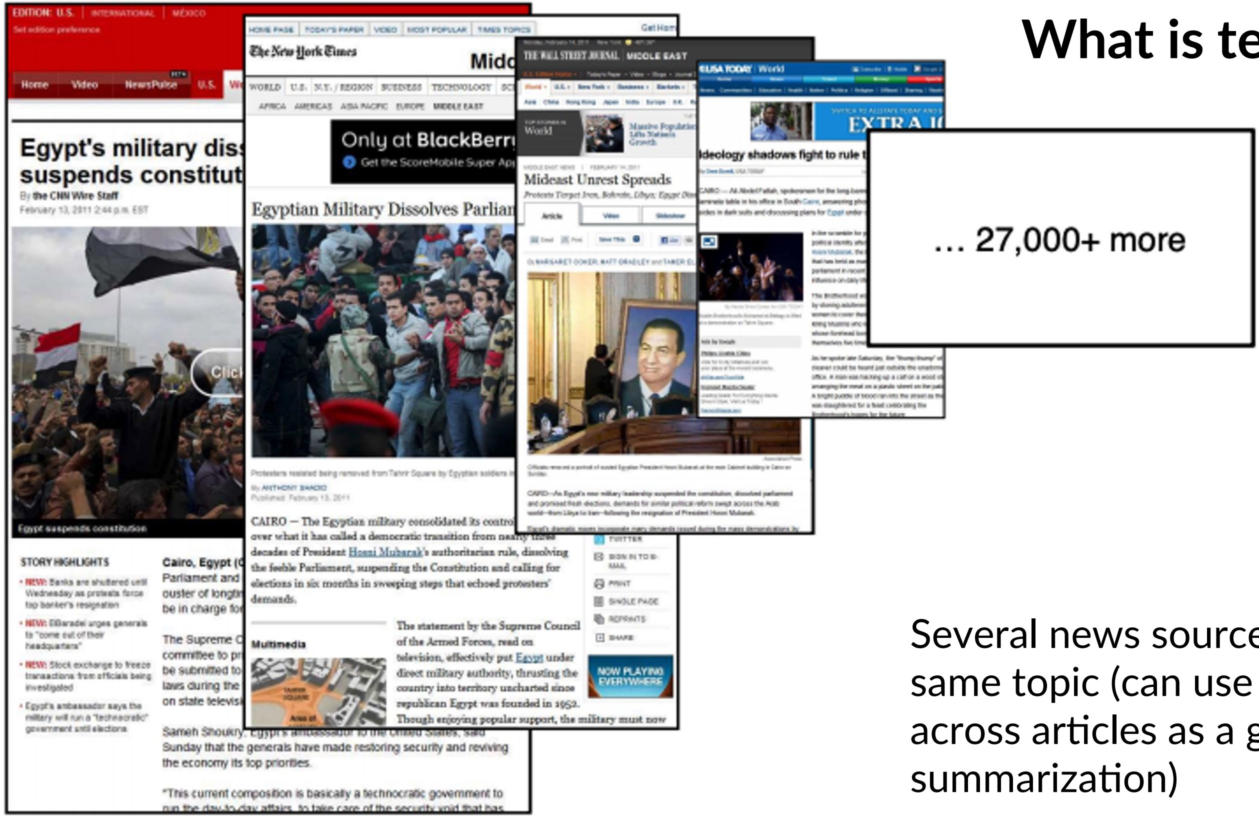


STORY HIGHLIGHTS

- **NEW:** Banks are shuttered until Wednesday as protests force top banker's resignation
- **NEW:** ElBaradei urges generals to "come out of their headquarters"
- **NEW:** Stock exchange to freeze transactions from officials being investigated
- Egypt's ambassador says the military will run a "technocratic" government until elections

News: Full document to a salient, non-redundant summary, i.e., in ~100 words.

What is text summarization?



Several news sources with articles on the same topic (can use overlapping info across articles as a good feature for summarization)

What is text summarization

Long Documents

Summarizer Model

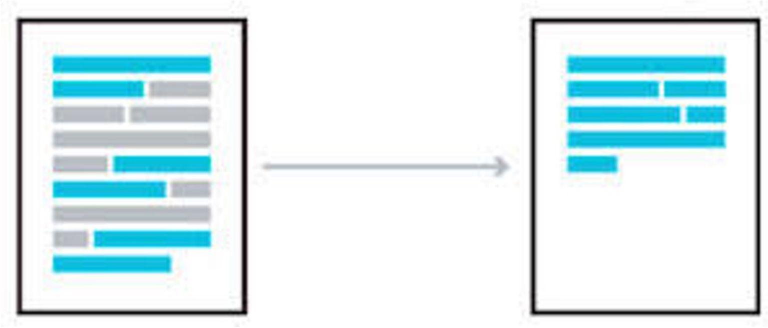
Shorter Text

Input could be various types: any document; dialogue; …

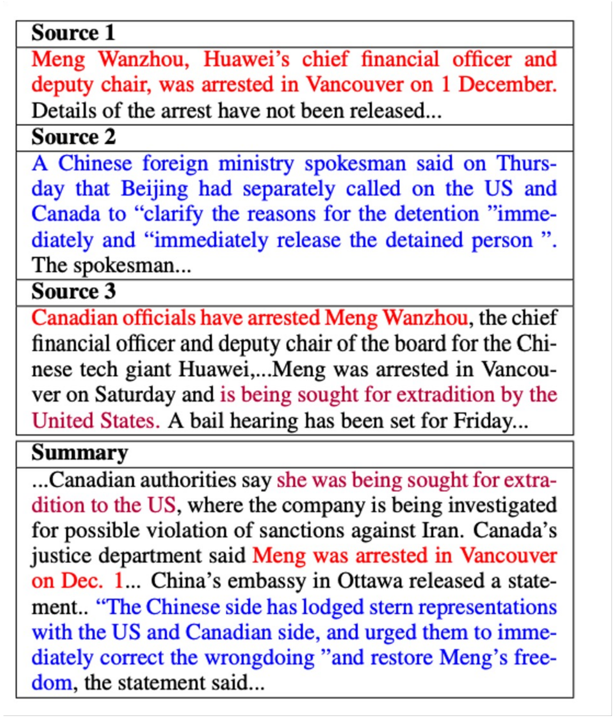
Types of Summarization:

- single-document vs multi-document;
- supervised vs unsupervised;
- abstractive vs extractive.

Single- vs Multi-



Single-document Summarization (SDS)



Source 1

Meng Wanzhou, Huawei’s chief financial officer and deputy chair, was arrested in Vancouver on 1 December. Details of the arrest have not been released...

Source 2

A Chinese foreign ministry spokesman said on Thursday that Beijing had separately called on the US and Canada to “clarify the reasons for the detention ”immediately and “immediately release the detained person ”. The spokesman...

Source 3

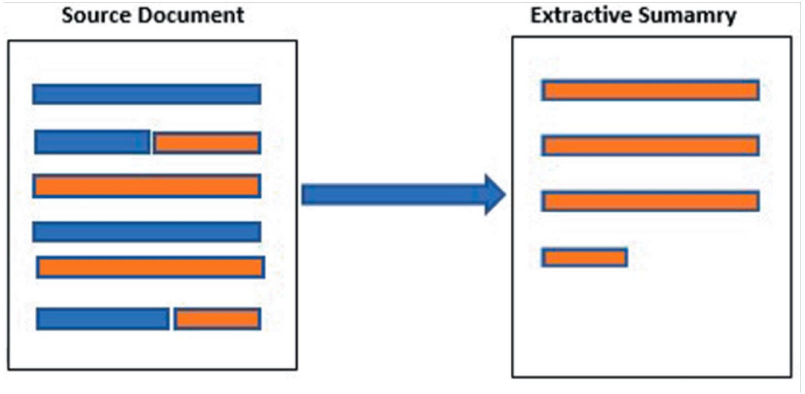
Canadian officials have arrested Meng Wanzhou, the chief financial officer and deputy chair of the board for the Chinese tech giant Huawei,...Meng was arrested in Vancouver on Saturday and is being sought for extradition by the United States. A bail hearing has been set for Friday...

Summary

...Canadian authorities say she was being sought for extradition to the US, where the company is being investigated for possible violation of sanctions against Iran. Canada’s justice department said Meng was arrested in Vancouver on Dec. 1... China’s embassy in Ottawa released a statement.. “The Chinese side has lodged stern representations with the US and Canadian side, and urged them to immediately correct the wrongdoing ”and restore Meng’s freedom, the statement said...

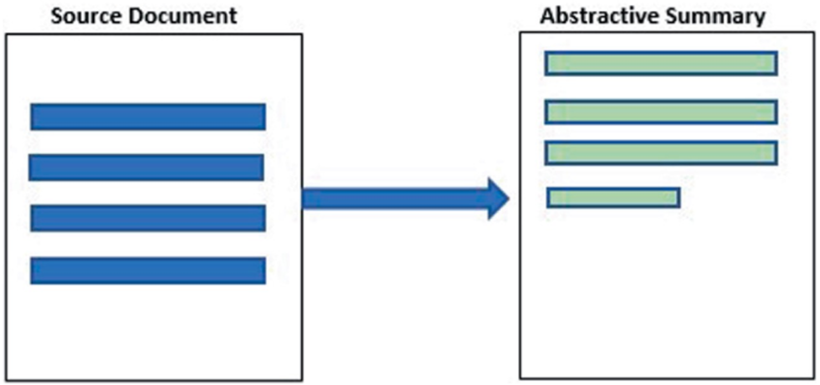
Multi-document Summarization (MDS)

Abstractive vs Extractive



Extractive

Use original vocabulary/sentence for the generated summary.



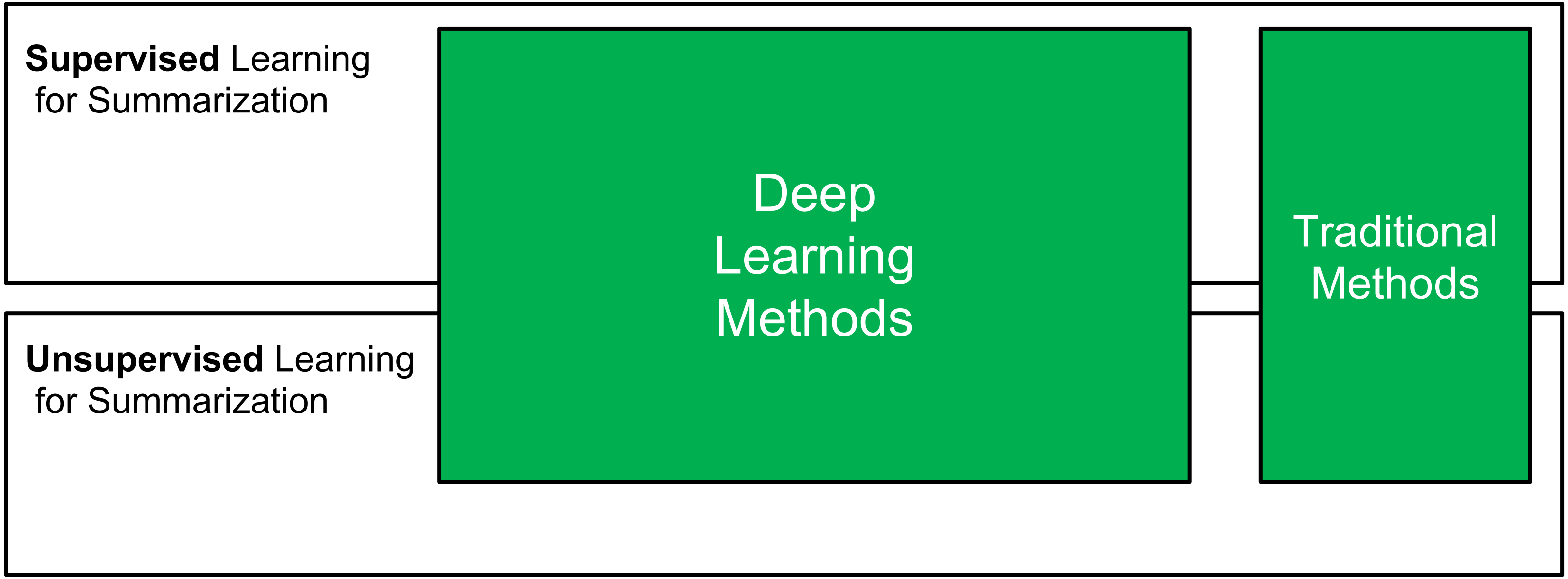
Abstractive

Use “novel” vocabulary for the generated summary, more creative.

❓ Categories

- Supervised / Unsupervised Summarization
- Extractive / Abstractive Summarization
- Deep Learning / not deep learning (traditional methods)

Categories



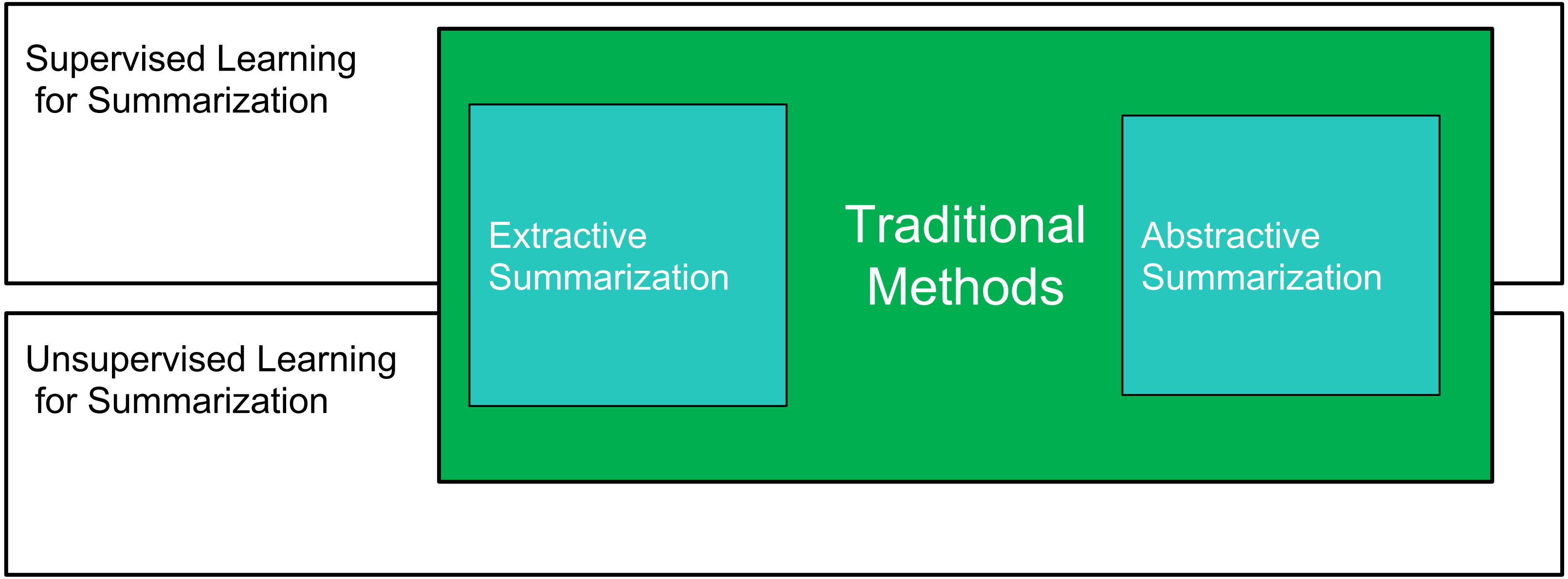
Supervised Learning for Summarization

Unsupervised Learning for Summarization

Deep Learning Methods

Traditional Methods

Categories



Supervised Learning for Summarization

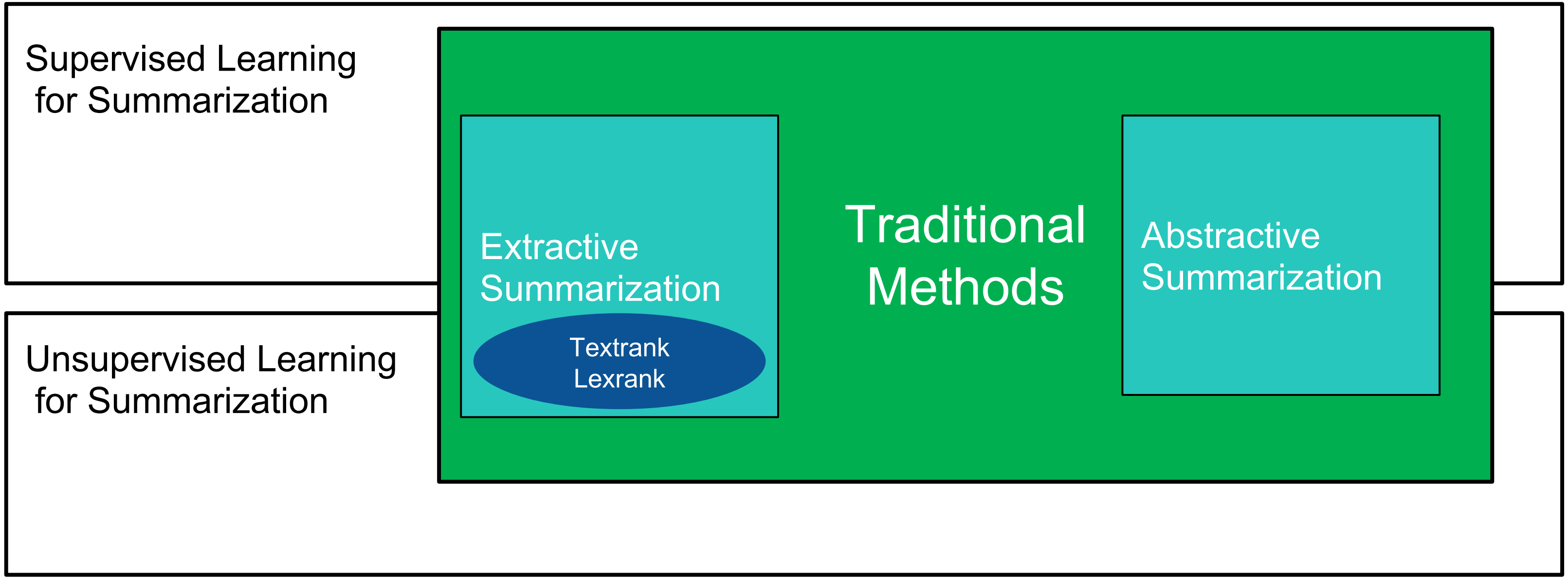
Unsupervised Learning for Summarization

Traditional Methods

Extractive Summarization

Abstractive Summarization

Categories



Supervised Learning for Summarization

Unsupervised Learning for Summarization

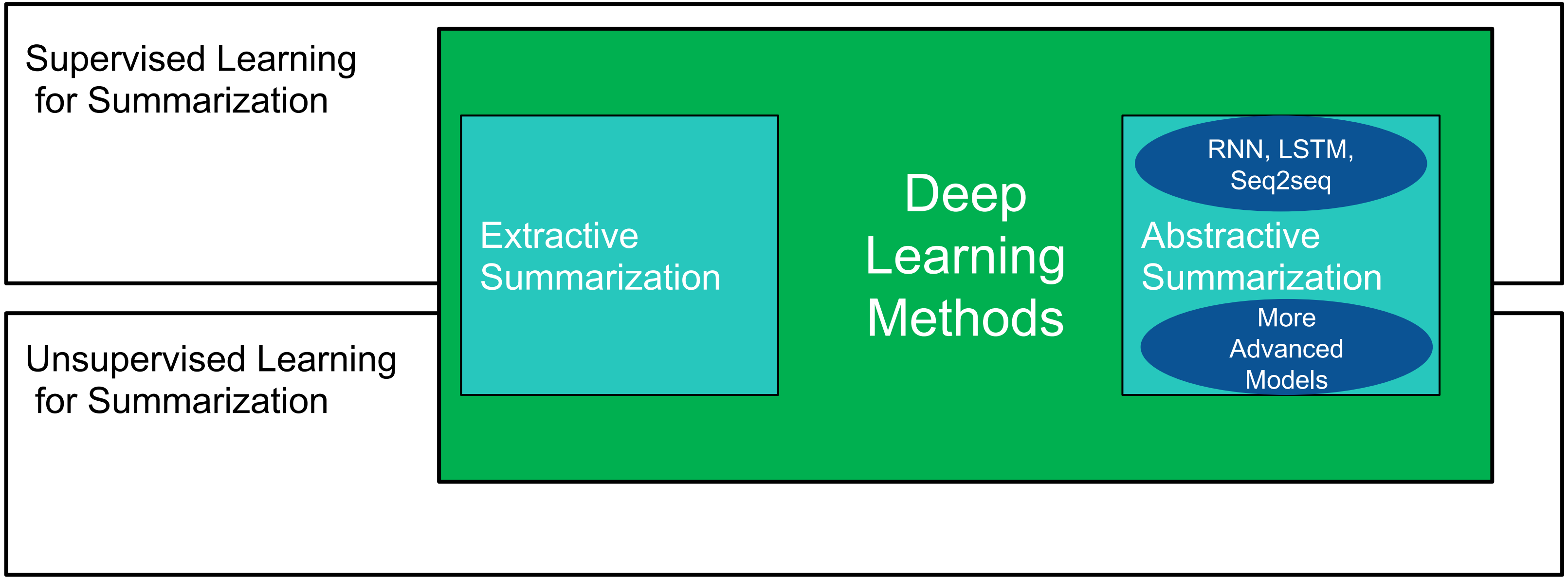
Traditional Methods

Extractive Summarization

Textrank
Lexrank

Abstractive Summarization

Categories



Supervised Learning for Summarization

Unsupervised Learning for Summarization

Extractive Summarization

Deep Learning Methods

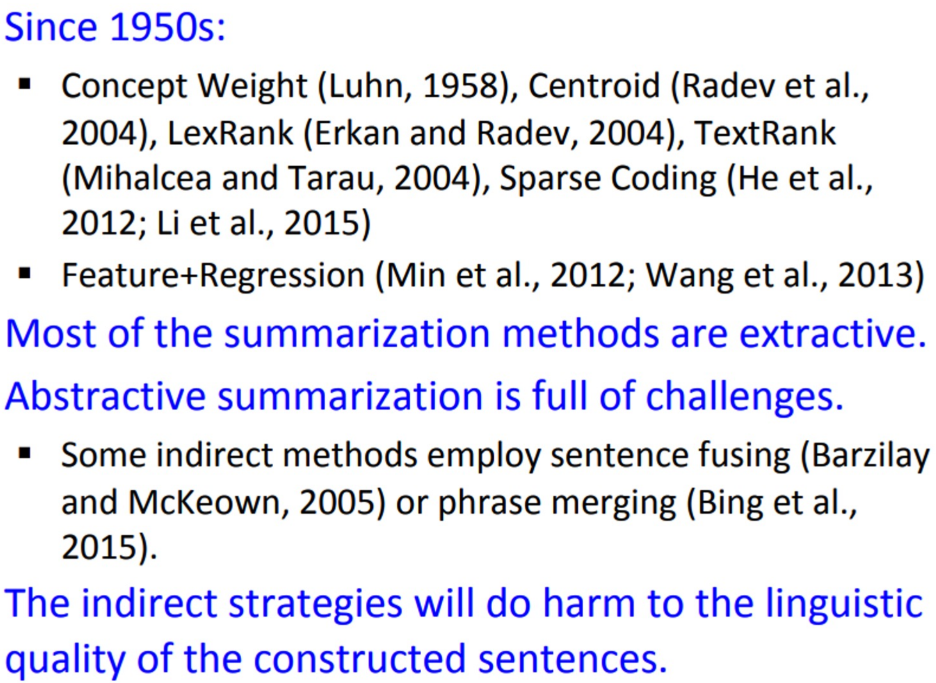
Abstractive Summarization

RNN, LSTM, Seq2seq

More Advanced Models

Methods

A brief history of summarization…



Since 1950s:

- Concept Weight (Luhn, 1958), Centroid (Radev et al., 2004), LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), Sparse Coding (He et al., 2012; Li et al., 2015)
- Feature+Regression (Min et al., 2012; Wang et al., 2013)

Most of the summarization methods are extractive.

Abstractive summarization is full of challenges.

- Some indirect methods employ sentence fusing (Barzilay and McKeown, 2005) or phrase merging (Bing et al., 2015).

The indirect strategies will do harm to the linguistic quality of the constructed sentences.

Adapted a slide page from Prof. Dragomir Radev

Classic Methods

TextRank

TextRank: Graph-based Ranking Algorithms

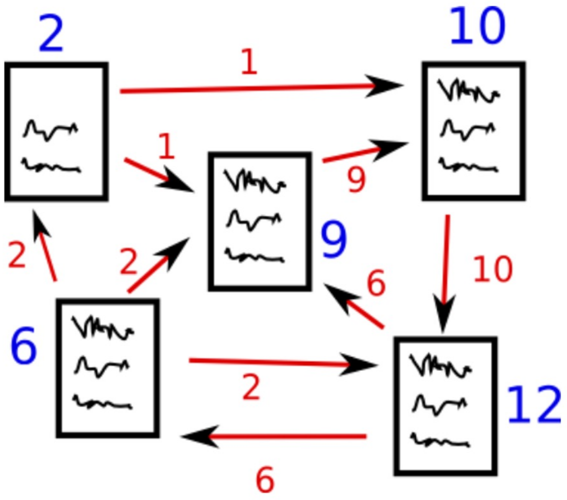
TextRank: Bringing Order into Texts (motivation)

Prerequisite: PageRank

An extractive method:

Select the most important/top pieces;

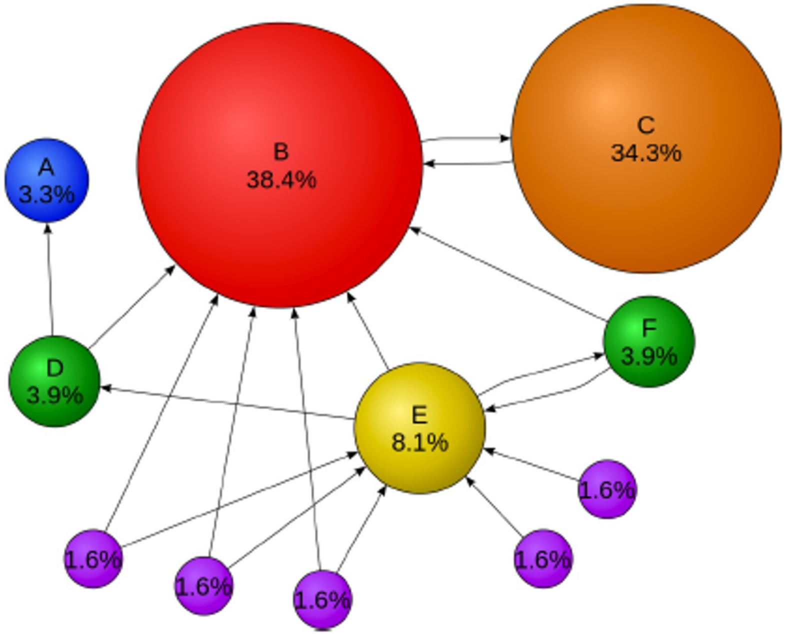
Work with “Graphs”.



Rada Mihalcea and Paul Tarau. *TextRank: Bringing Order into Text*. 2004, ACL

Source

Use PageRank to score the sentences in the graph…



- Rank the sentences with underlying assumption that “summary sentences” are similar to most other sentences

Differences between Lexrank and Textrank

TextRank was applied to summarization exactly as described, while LexRank combines the LexRank score with other features like sentence position and length.

TextRank was used for single document summarization, while LexRank has been applied to multi-document summarization.

How to implement Textrank?

No official implement, but there are some…[**Extractive**]

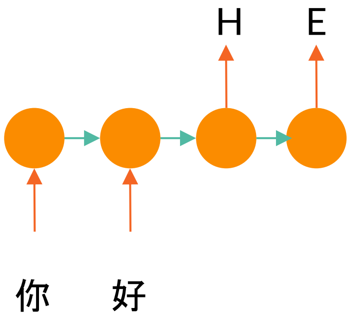
https://pypi.org/project/pytextrank/

https://pypi.org/project/textrank/

https://pypi.org/project/lexrank/ (Lexrank)

Deep Learning Methods

Neural Abstractive Summarization



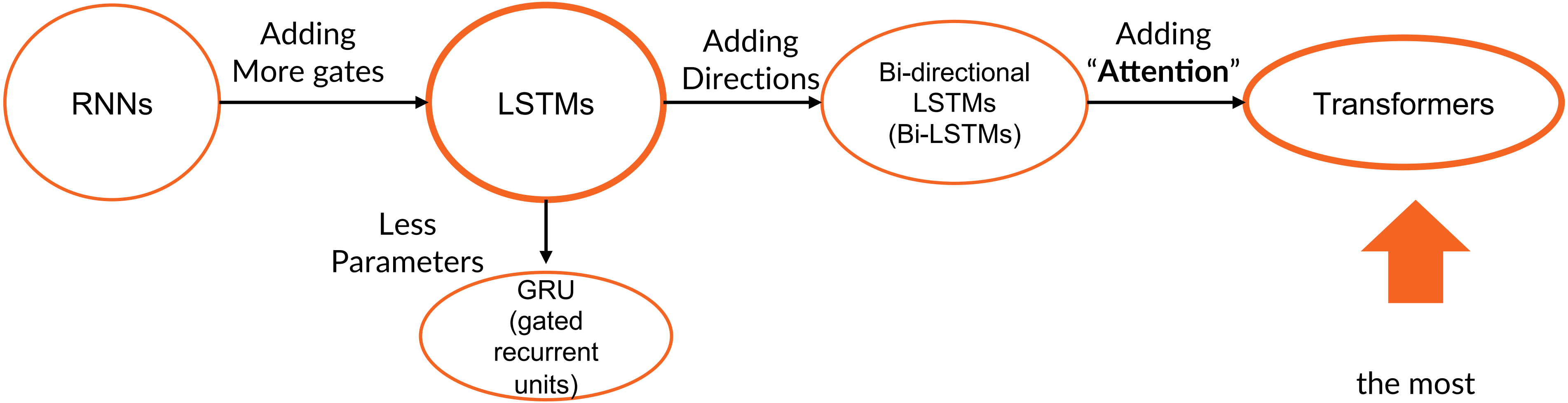
Seq2seq
Illustration

We consider summarization to be a **text generation** task.

Typical frameworks are based on **sequence-to-sequence** neural networks.

Such frameworks can also solve machine translation, question answering, and so on.

Seq2seq Frameworks: a brief history



RNNs

Adding
More gates

LSTMs

Adding
Directions

Bi-directional
LSTMs
(Bi-LSTMs)

Adding
“**Attention**”

Transformers

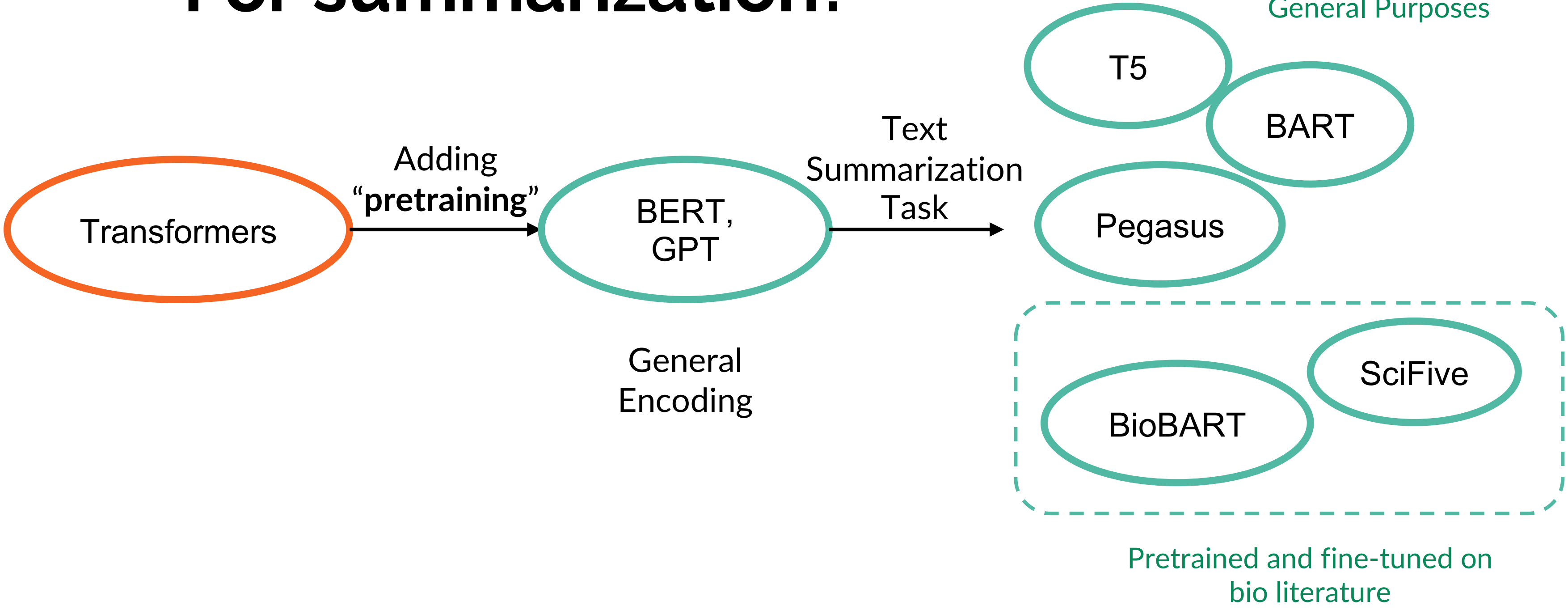
Less
Parameters

GRU
(gated
recurrent
units)

the most
advanced RNN
model.
(before LLMs)

<u>Before the LLMs came out…</u>
Roughly 2014~2021

For summarization?



BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model

🤖Based on BART, pretrained on biomedical corpus - 41 GB PMC articles.

  Masked Language Pretraining: 30% of the input tokens are masked.

🤖For abstractive summarization (part):

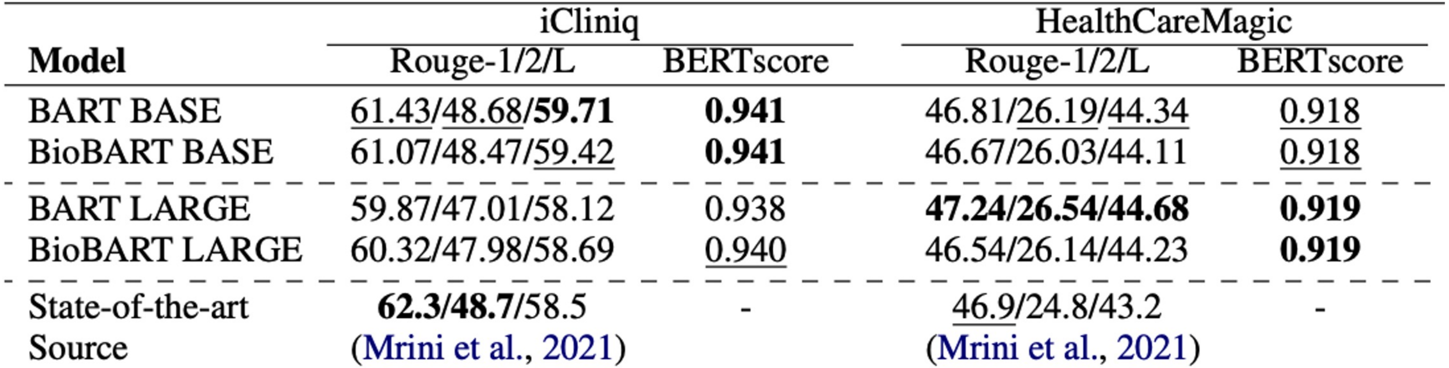
  Summary and dialogues between a patient and a doctor: iCliniq (31,062 samples) , HealthCareMagic (226,405 samples)

  Medical question summarization: MeQSum (1k+ questions)

Yuan, Hongyi et al. “BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model.” Workshop on Biomedical Natural Language Processing (2022).

Results (part)

Aligns with our own experiments...



Model	iCliniq		HealthCareMagic	
	Rouge-1/2/L	BERTscore	Rouge-1/2/L	BERTscore
BART BASE	61.43/48.68/**59.71**	**0.941**	46.81/26.19/44.34	0.918
BioBART BASE	61.07/48.47/59.42	**0.941**	46.67/26.03/44.11	0.918
BART LARGE	59.87/47.01/58.12	0.938	**47.24/26.54/44.68**	**0.919**
BioBART LARGE	60.32/47.98/58.69	0.940	46.54/26.14/44.23	**0.919**
State-of-the-art	**62.3/48.7**/58.5	-	46.9/24.8/43.2	-
Source	(Mrini et al., 2021)		(Mrini et al., 2021)	

BioBART performance on selected benchmark for text summarization.

Summarization Evaluation

Automatic Evaluation & Human Evaluation

Evaluation Methods

Automatic Evaluation

- **ROUGE,** BLEU

Human Evaluation

- Which perspectives?

Automatic Method: ROUGE

Recall-**O**riented **U**nderstudy for **G**isting **E**valuation

Motivation: how many are overlapped?

System Summary (output of our model): the cat was found under the bed

Reference Summary (ground truth): the cat was under the bed

Compute the precision and recall using the overlap.

System Summary: the cat was found under the bed

Reference Summary: the cat was under the bed

Recall in the context of ROUGE means how much of the overlapping content exist in the reference summary ?

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

$$Recall = \frac{6}{6} = 1.0$$

Calculate Precision

System Summary: the cat was found under the bed

Reference Summary: the cat was under the bed

Precision in the context of ROUGE means how much of the overlapping words exist in the system summary ?

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

$$Precision = \frac{6}{7} = 0.86$$

ROUGE: what to report

Precision

Recall

And **F1 Score** = 2*(Recall * Precision) / (Recall + Precision)

Another system summary

System Summary: the tiny little cat was found under the big funny bed

Reference Summary: the cat was under the bed

$$Recall = \frac{6}{6} = 1.0$$

$$Precision = \frac{6}{11} = 0.55$$

Rouge-N

ROUGE-1: unigram

ROUGE-2: bigram

ROUGE-3: trigram

…

Rouge-2

System Summary: the cat was found under the bed

Reference Summary: the cat was under the bed

System bigrams:
the cat, cat was, was found, found under, under the, the bed

Reference Summary:
the cat, cat was, was under, under the, the bed

Rouge-2 P,R, and F1

System bigrams:

the cat, cat was, was found, found under, under the, the bed

Reference Summary:

the cat, cat was, was under, under the, the bed

$$ROUGE2_{Precision} = \frac{4}{6} = 0.67$$

$$ROUGE2_{Recall} = \frac{4}{5} = 0.8$$

Python with Rouge

https://pypi.org/project/pyrouge/

Sample outputs:

```
------------------------------------------------------------
1 ROUGE-1 Average_R: 0.78378 (95%-conf.int. 0.78378 - 0.78378)
1 ROUGE-1 Average_P: 0.80556 (95%-conf.int. 0.80556 - 0.80556)
1 ROUGE-1 Average_F: 0.79452 (95%-conf.int. 0.79452 - 0.79452)
------------------------------------------------------------
1 ROUGE-2 Average_R: 0.69444 (95%-conf.int. 0.69444 - 0.69444)
1 ROUGE-2 Average_P: 0.71429 (95%-conf.int. 0.71429 - 0.71429)
1 ROUGE-2 Average_F: 0.70423 (95%-conf.int. 0.70423 - 0.70423)
------------------------------------------------------------
```

Practice: install pyrouge and have a try!

Is ROUGE always a smart way?

System Summary 1: the kitty was seen under the bed

System Summary 2: the cat was found under the bed

Reference Summary: the cat was seen under the bed

Human evaluation

Evaluation protocols:

- **relevance/informativeness (selection of important content from the source)**
- **consistency (factual alignment between the summary and the source)**
- **fluency (quality of individual sentences)**
- **coherence (collective quality of all sentences)**
- **non-redundancy (this is more used for multi-doc summarization)**

Human evaluation

Randomly choosing 100-200 samples (a reasonable number);

2-5 human judges (only 1 is not enough);

You can decide how to set the points (0/1 or 0-5);

Verify the agreement score.

Report in a table.

Extension

Open Questions

Some Open Questions

- Better Automatic Evaluation Methods?
    - BertScore: Evaluating Text

- Various Types? Data Hunger…
    - Scientific Summarization
    - Table Summarization
    - Web-data Summarization

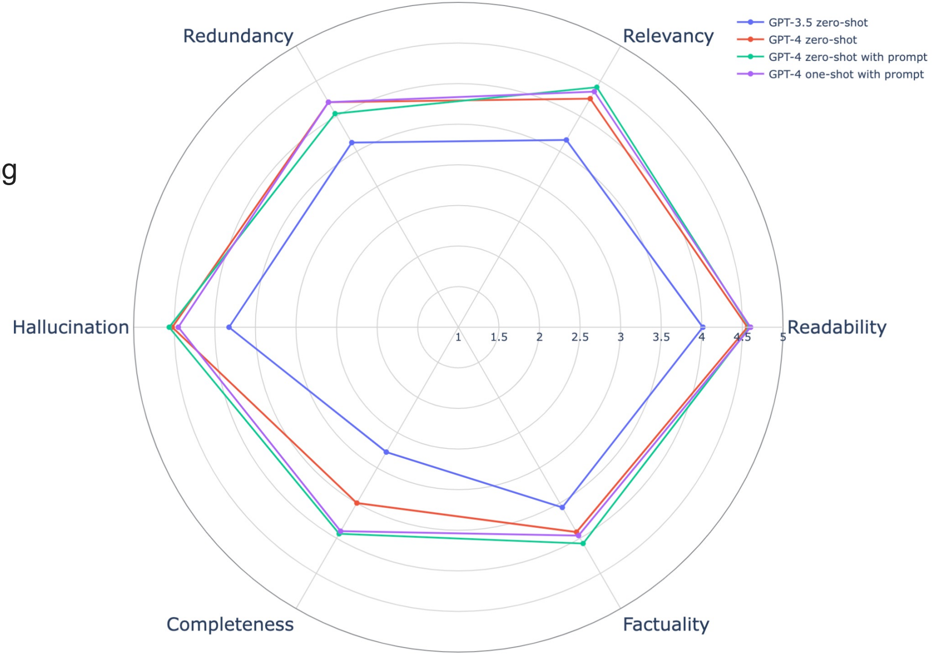
Evaluating LLMs for Text Generation

[Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts]

Fan Gao, Hang Jiang, Moritz Blum, Jinghui Lu, Yuang Jiang and Irene Li;

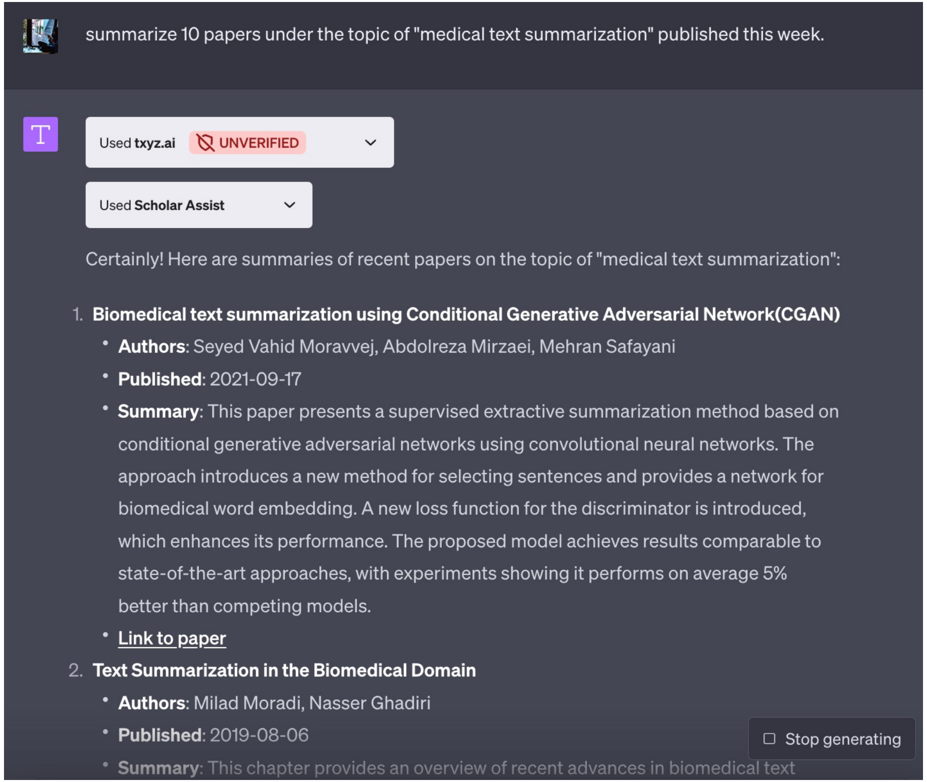
Preprint

- 📖 **Readability**
- 🎯 **Relevancy**
- 🔍 **Redundancy**
- 👻 **Hallucination**
- 📚 **Completeness**
- ✅ **Factuality**



With Large Language Models...

- Much powerful than what we expected;
- Challenging Tasks?
    - Legal domain?
    - Medical domain?
    - New evaluation perspectives?



summarize 10 papers under the topic of "medical text summarization" published this week.

Certainly! Here are summaries of recent papers on the topic of "medical text summarization":

1. **Biomedical text summarization using Conditional Generative Adversarial Network(CGAN)**
    - **Authors**: Seyed Vahid Moravvej, Abdolreza Mirzaei, Mehran Safayani
    - **Published**: 2021-09-17
    - **Summary**: This paper presents a supervised extractive summarization method based on conditional generative adversarial networks using convolutional neural networks. The approach introduces a new method for selecting sentences and provides a network for biomedical word embedding. A new loss function for the discriminator is introduced, which enhances its performance. The proposed model achieves results comparable to state-of-the-art approaches, with experiments showing it performs on average 5% better than competing models.
    - Link to paper
2. **Text Summarization in the Biomedical Domain**
    - **Authors**: Milad Moradi, Nasser Ghadiri
    - **Published**: 2019-08-06
    - **Summary**: This chapter provides an overview of recent advances in biomedical text

Thanks

https://www.li-lab.me/