Neural Abstractive Summarization: Methods and Applications

Grigorios Tsoumakas

ARISTOTLE UNIVERSITY OF THESSALONIKI

Medoid AI

Agenda

Setting the Scene

Dealing with Long Documents

Bayesian Active Summarization

Controlling the Output’s Topic

Healthcare and Finance Apps

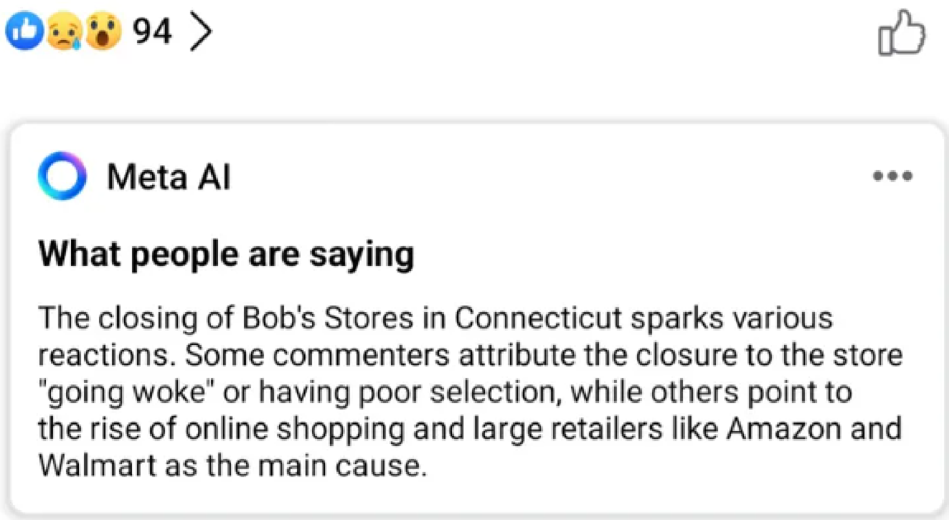
Setting the Scene

Automated Summarization vs Information Overload

Reduce reading time

Reduce cost and bias of human summarizers

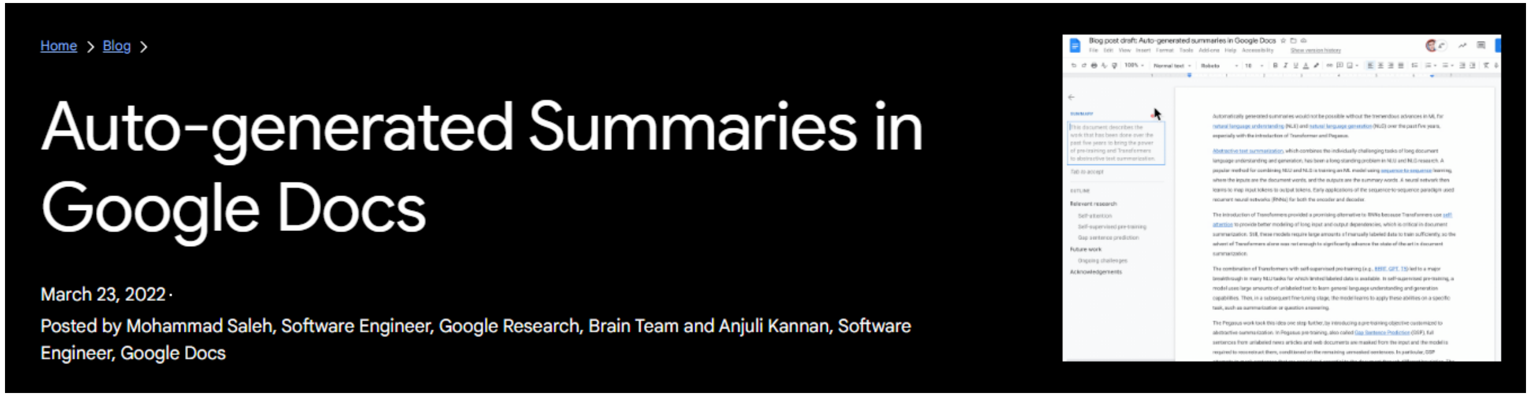
Improve downstream machine processing tasks



Meta AI

What people are saying

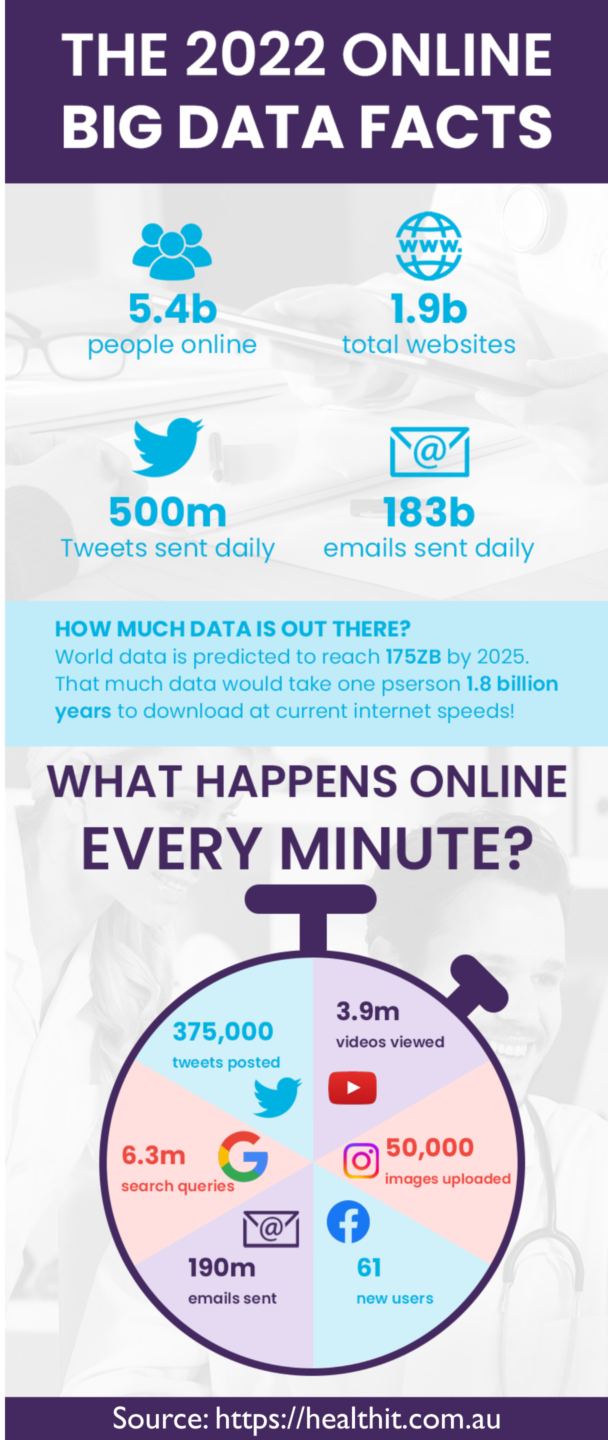
The closing of Bob's Stores in Connecticut sparks various reactions. Some commenters attribute the closure to the store "going woke" or having poor selection, while others point to the rise of online shopping and large retailers like Amazon and Walmart as the main cause.



Auto-generated Summaries in Google Docs

March 23, 2022 ·

Posted by Mohammad Saleh, Software Engineer, Google Research, Brain Team and Anjuli Kannan, Software Engineer, Google Docs



THE 2022 ONLINE BIG DATA FACTS

5.4b people online

1.9b total websites

500m Tweets sent daily

183b emails sent daily

HOW MUCH DATA IS OUT THERE?
World data is predicted to reach 175ZB by 2025.
That much data would take one pserson 1.8 billion years to download at current internet speeds!

WHAT HAPPENS ONLINE EVERY MINUTE?

375,000 tweets posted

3.9m videos viewed

6.3m search queries

50,000 images uploaded

190m emails sent

61 new users

Source: https://healthit.com.au

Extractive & Abstractive Summarization

EXTRACTIVE SUMMARY

During the United Nations General Assembly, Ukraine and climate change were high on the agenda

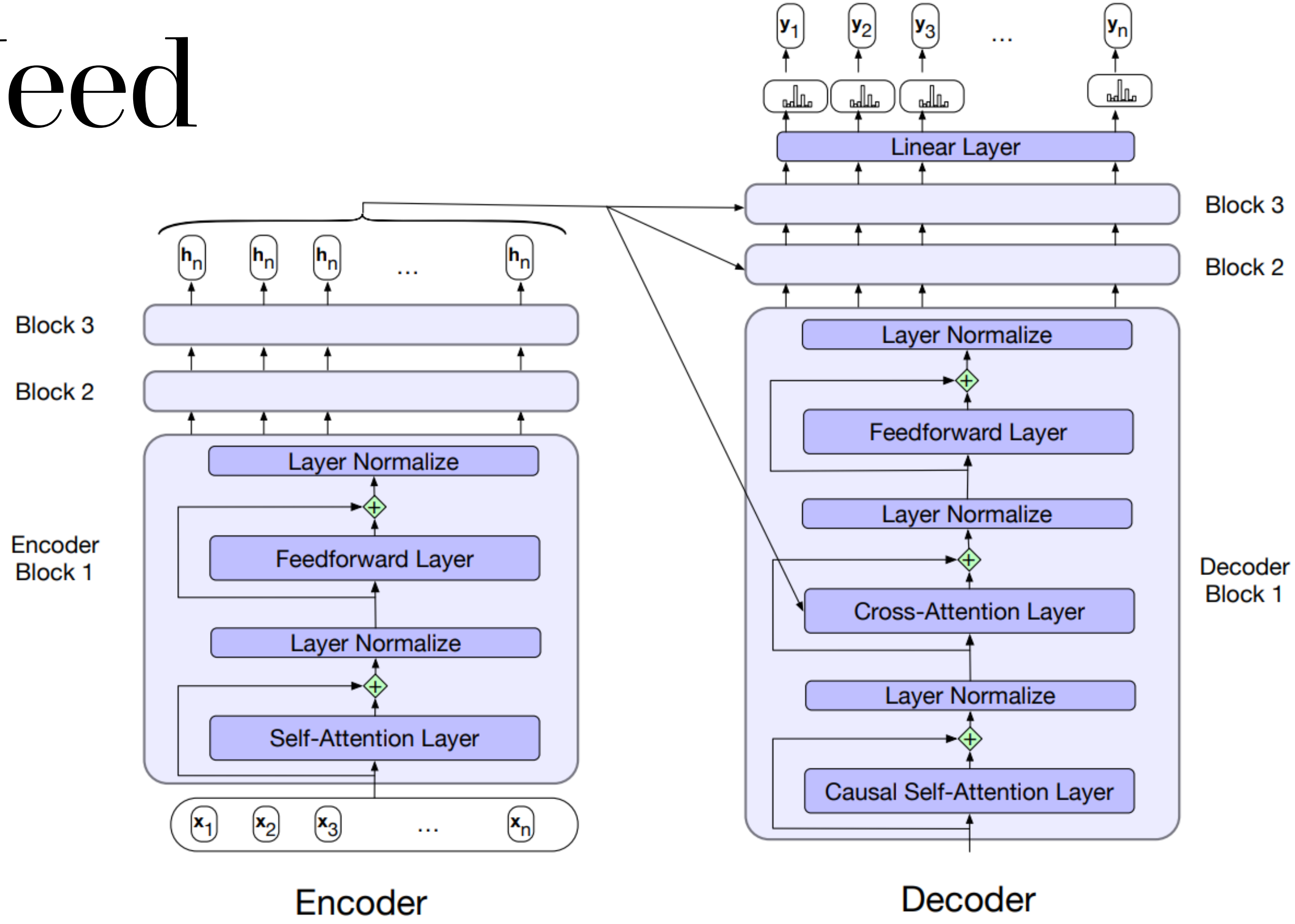
DOCUMENT

During the 77th session of the United Nations General Assembly, Russia's invasion of Ukraine and climate change were high on the agenda amid soaring prices for energy and food. DW, 21/09/2022

ABSTRACTIVE SUMMARY

Russia and climate change dominate UN General Assembly

Attention is All you Need



Source: "Speech and Language Processing (3rd ed. draft)"

Pretrained Models: PEGASUS

Architecture

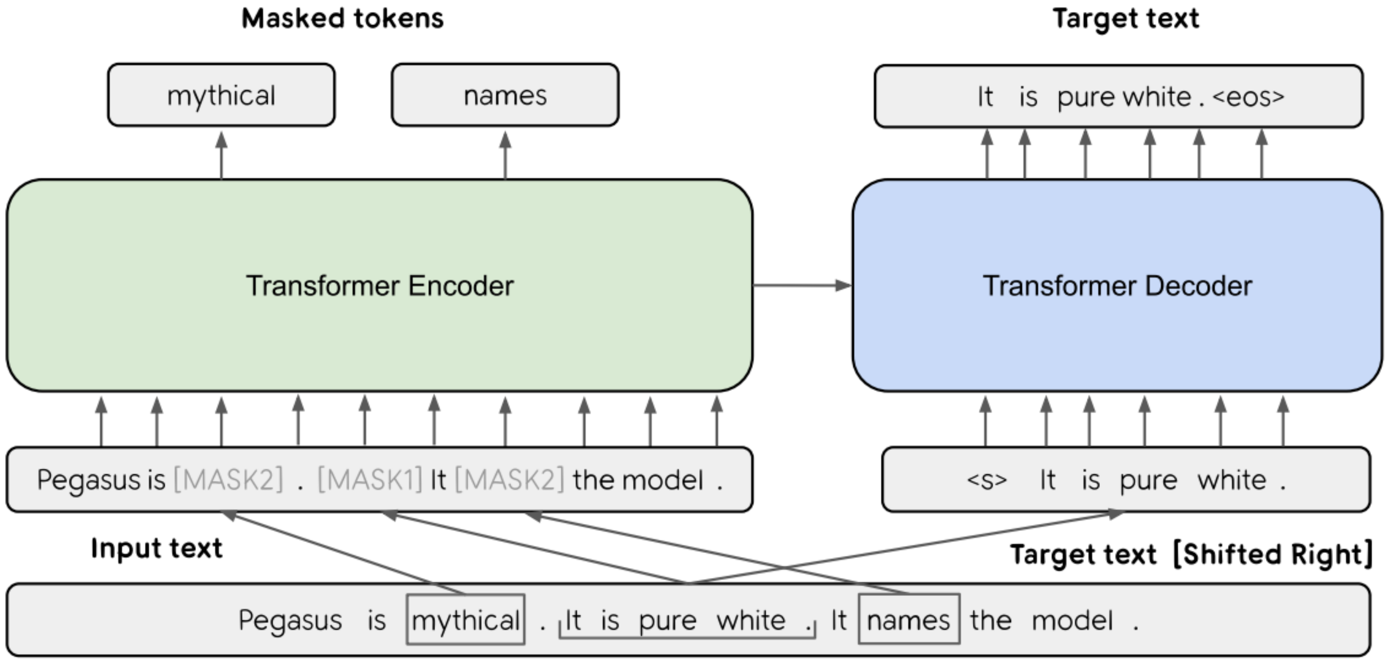
- Large: 16 layers, 1024 hidden layer size, 4096 feed-forward layer, 568M params

Data

- HugeNews: 1.5B articles (3.8TB) from news and news-like websites
- C4: 350M Web-pages (750GB)

Objective

- Gap Sentence Generation: masking 30% of sentences and concatenating them as summary



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

Source: "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization"

Pretrained Models: BART

Architecture

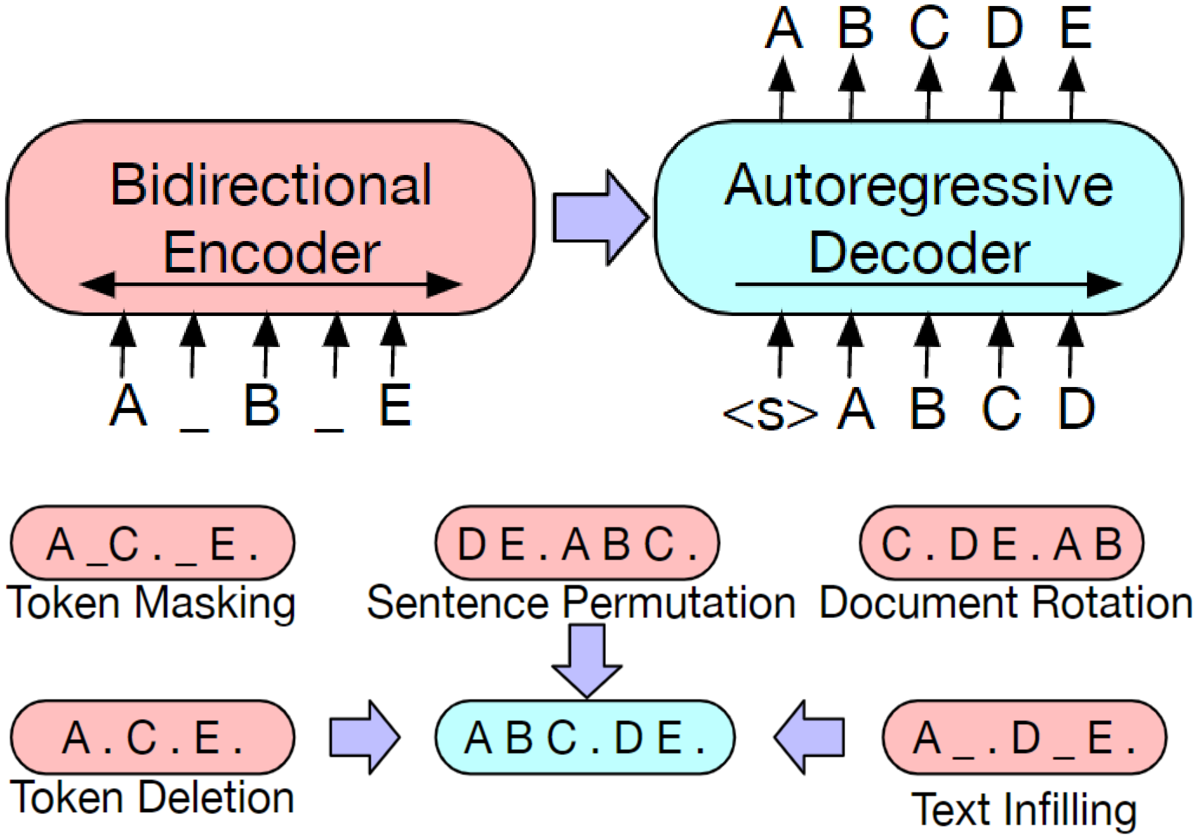
- Large: 12 layers, 1024 hidden layer size, 4096 feed-forward layer, 406M params

Data

- BookCorpus plus English Wikipedia (16Gb), CC-News (76Gb), OpenWebText (38Gb), Stories (31Gb)

Objective

- Input reconstruction



Dealing with Long Documents

The Challenge of Long Documents

Higher computational complexity

- Self-attention computation in transformers has $O(n^2)$ complexity with respect to n input tokens
- Typical capacity of PEGASUS and BART is 1024 tokens

Higher levels of noise

- Only a small fraction of a long doc is key to its narrative

Diverse key information in the summaries

- Difficult to capture, compared to single point of information in short documents

	Input	Output
CNN	656	43
Daily Mail	693	52
PubMed	3,016	203
arXiv	4,938	220

Solutions for Long Documents

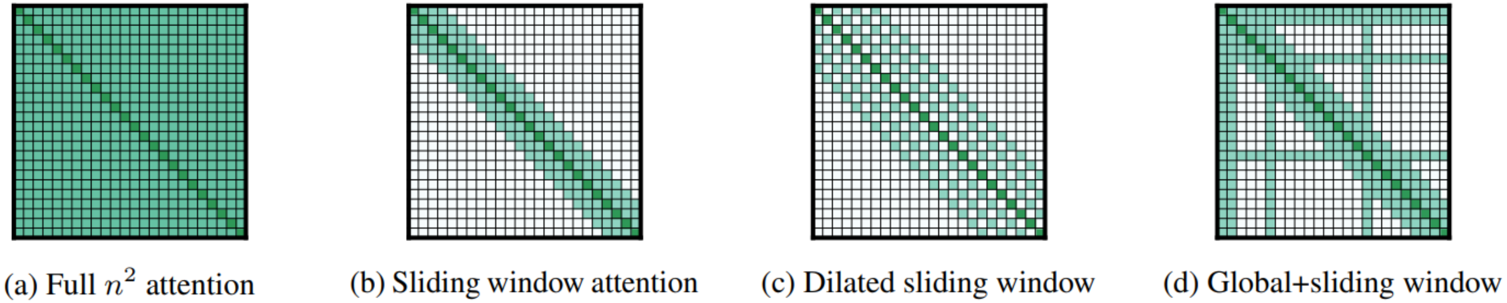
Truncation

Chunking

Sparse attention

- BigBird (Google)
- Longformer (Allen AI)

FlashAttention



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer, https://doi.org/10.48550/arxiv.2004.05150

Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

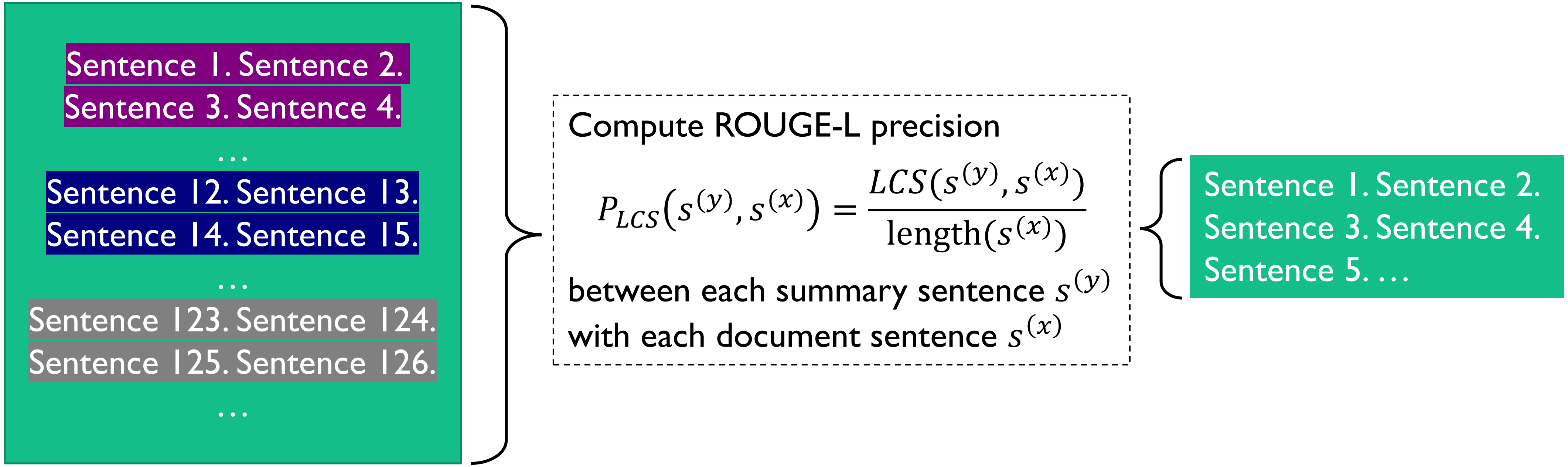
DANCER (Divide-ANd-ConquER)

Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
…
Sentence 12. Sentence 13.
Sentence 14. Sentence 15.
…
Sentence 123. Sentence 124.
Sentence 125. Sentence 126.
…

Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
Sentence 5. …

Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

DANCER (Divide-ANd-ConquER)



Sentence 1. Sentence 2. Sentence 3. Sentence 4.

…

Sentence 12. Sentence 13. Sentence 14. Sentence 15.

…

Sentence 123. Sentence 124. Sentence 125. Sentence 126.

…

Compute ROUGE-L precision

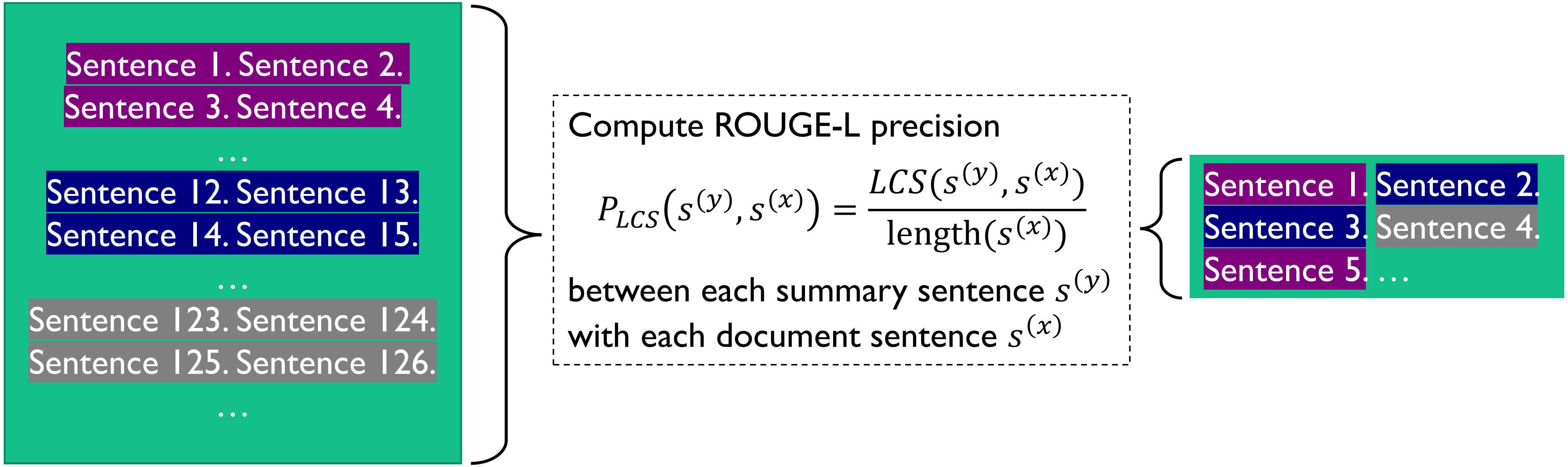
$$P_{LCS}\left(s^{(y)}, s^{(x)}\right) = \frac{LCS\left(s^{(y)}, s^{(x)}\right)}{\text{length}\left(s^{(x)}\right)}$$

between each summary sentence $s^{(y)}$ with each document sentence $s^{(x)}$

Sentence 1. Sentence 2. Sentence 3. Sentence 4. Sentence 5. …

Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

DANCER (Divide-ANd-ConquER)



Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
…
Sentence 12. Sentence 13.
Sentence 14. Sentence 15.
…
Sentence 123. Sentence 124.
Sentence 125. Sentence 126.
…

Compute ROUGE-L precision

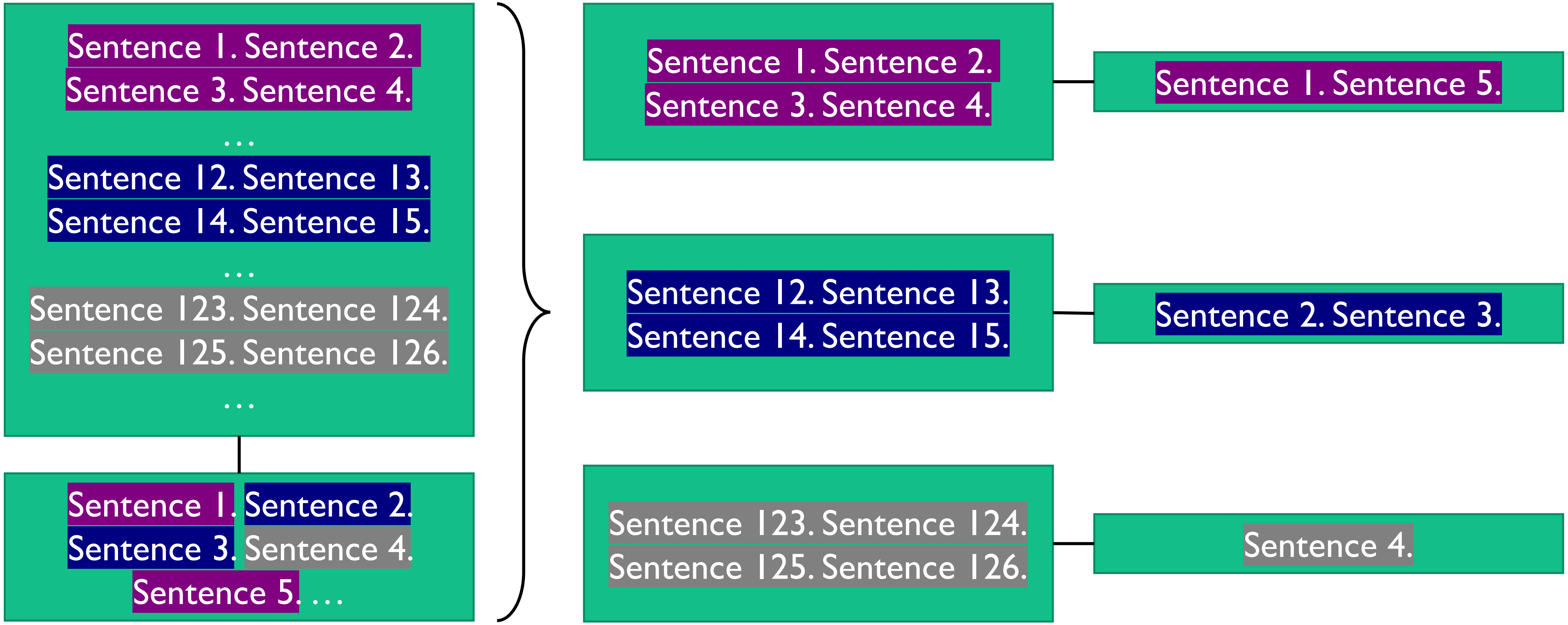
$$P_{LCS}\left(s^{(y)}, s^{(x)}\right) = \frac{LCS\left(s^{(y)}, s^{(x)}\right)}{\text{length}\left(s^{(x)}\right)}$$

between each summary sentence $s^{(y)}$
with each document sentence $s^{(x)}$

Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
Sentence 5. …

Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

DANCER (Divide-ANd-ConquER)



Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
…
Sentence 12. Sentence 13.
Sentence 14. Sentence 15.
…
Sentence 123. Sentence 124.
Sentence 125. Sentence 126.
…

Sentence 1. Sentence 2.
Sentence 3. Sentence 4.
Sentence 5. …

Sentence 1. Sentence 2.
Sentence 3. Sentence 4.

Sentence 1. Sentence 5.

Sentence 12. Sentence 13.
Sentence 14. Sentence 15.

Sentence 2. Sentence 3.

Sentence 123. Sentence 124.
Sentence 125. Sentence 126.

Sentence 4.

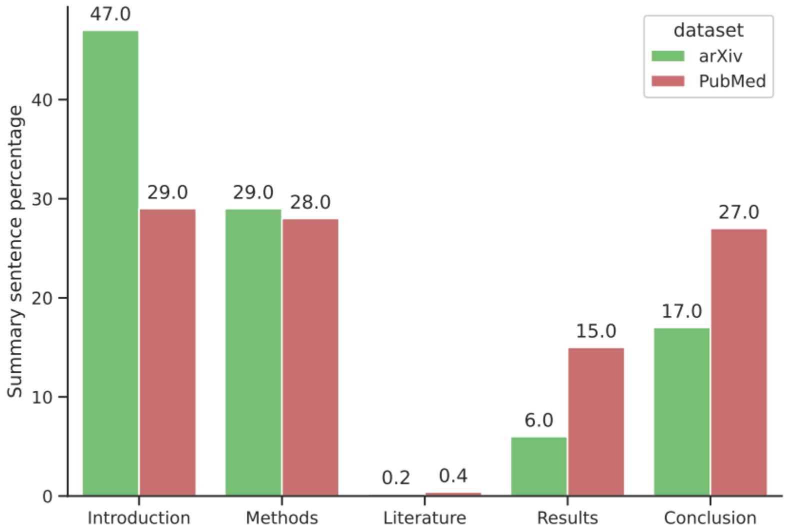
Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

Section Selection

We filter uninformative sections

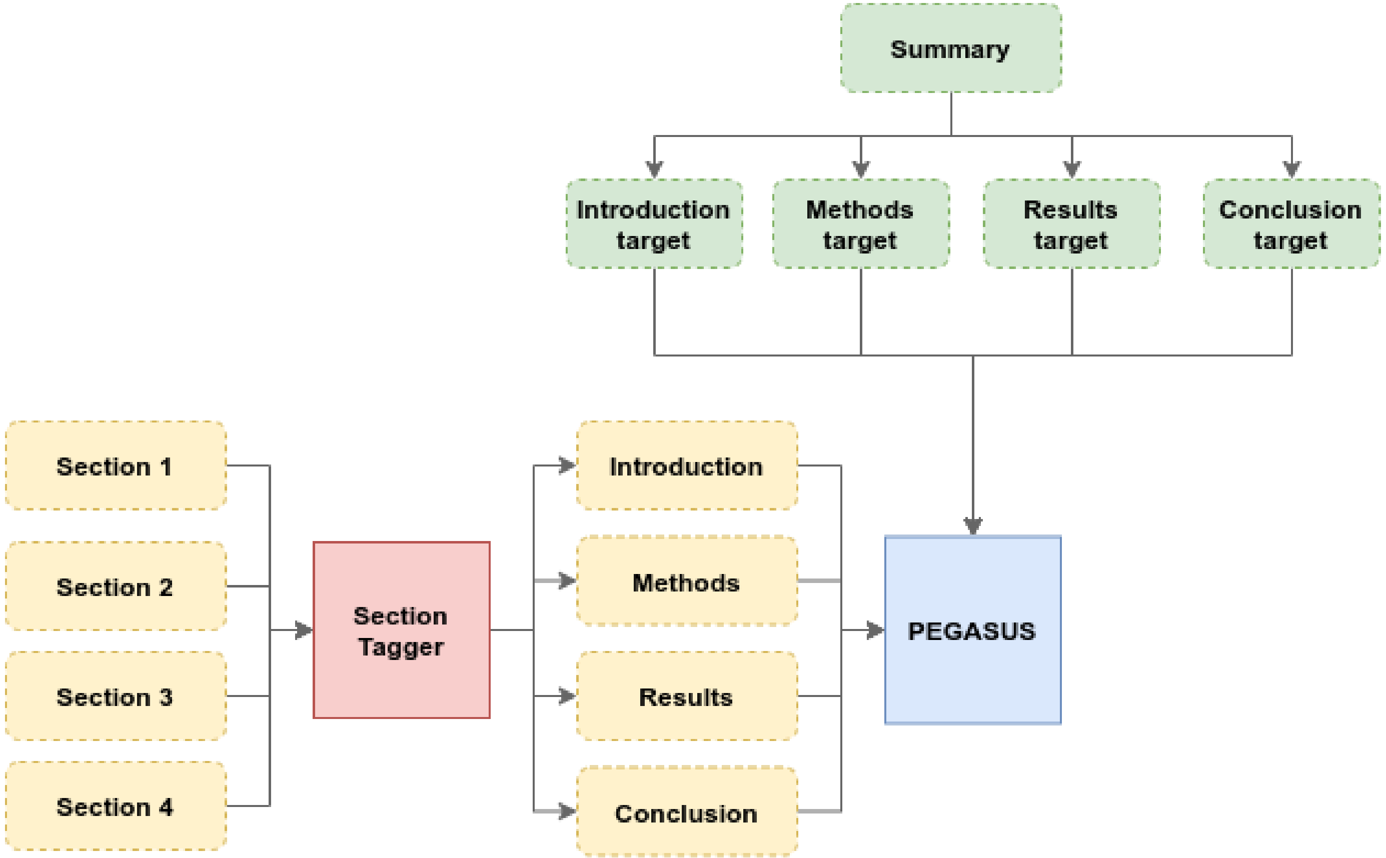
- E.g., front-end sections vs financial statements in financial reports
- E.g., introduction, conclusions vs related work, background in papers

Section	Keywords
Introduction	Introduction, case
~~Literature~~	~~Background, literature, related~~
Methods	Method(s), techniques, methodology
Results	Result(s), experimental, experiment(s)
Conclusions	Conclusion(s), concluding, discussion, limitations



Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

Summarizing Academic Papers



Gidiotis, A., & Tsoumakas, G. (2020). A Divide-and-Conquer Approach to the Summarization of Long Documents. IEEE/ACM Transactions on Audio Speech and Language Processing, 28, 3029–3040.

Results

arXiv

	R-1	R-2	R-L
PEGASUS	44.21	16.95	38.83
DANCER	45.01	17.60	40.56
BigBird	**46.63**	**19.02**	**41.77**

PubMed

	R-1	R-2	R-L
PEGASUS	45.97	20.15	41.34
DANCER	**46.34**	19.97	**42.42**
BigBird	46.32	**20.65**	42.33

⊖ Loss of dependencies between the different sections

⊕ No architectural change requirements, can do inference in parallel, can deal with large outputs too

Bayesian Active Summarization

The Problem

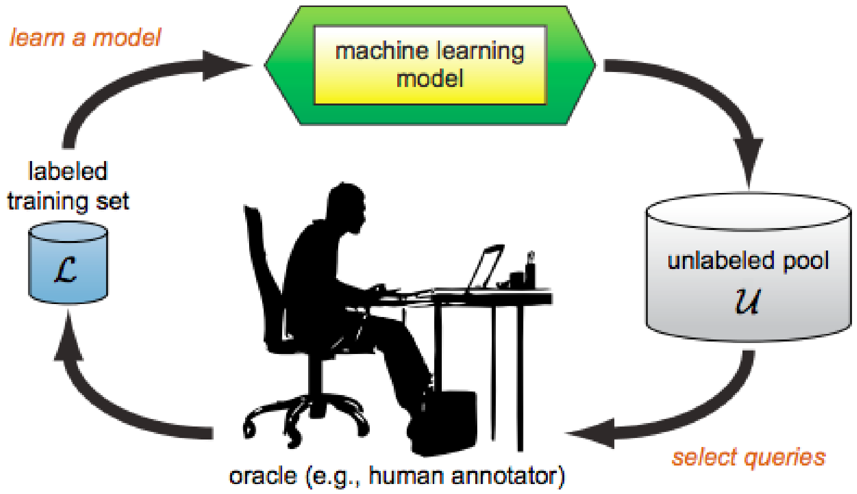
Deep learning models are data hungry

Collecting high quality training data is costly

- Especially if domain expertise is required, as in the financial, legal and health domains

Active learning can help make the most out of a finite budget

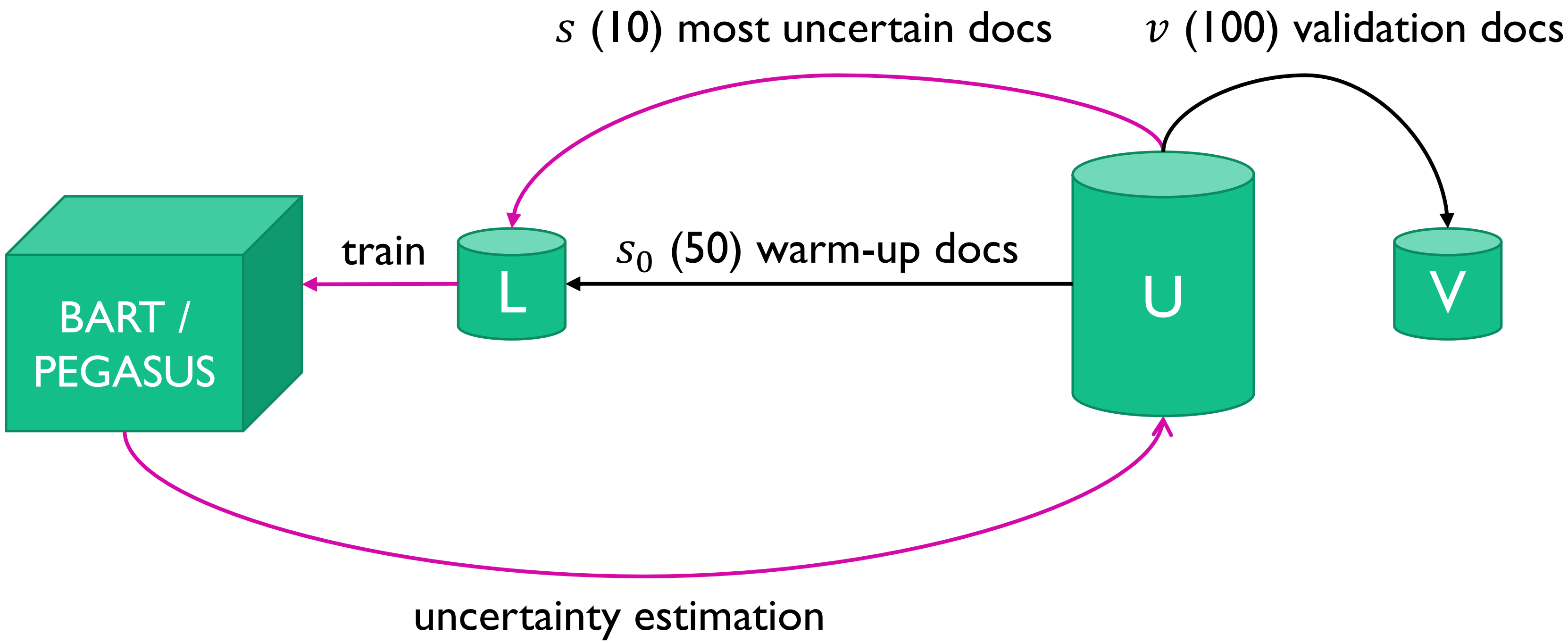
Almost no work on active summarization



The pool-based active learning cycle. Source: Settles, B. (2012). Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), 1–114.

Gidiotis, A., & Tsoumakas, G. (2023). Bayesian active summarization. Computer Speech and Language, 83.
https://doi.org/10.1016/j.csl.2023.101553

BAS (Bayesian Active Summarization)



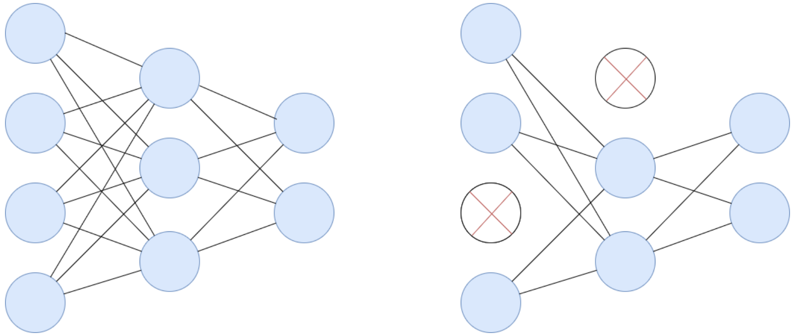
Annotation Budget b	Training Docs
(900)	(0)
(800)	(50)
(750)	(60)
(740)	(70)
(730)	…
…	(800)
(0)	

Gidiotis, A., & Tsoumakas, G. (2022). Should We Trust This Summary? Bayesian Abstractive Summarization to The Rescue. In Findings of the Association for Computational Linguistics: ACL 2022, pages 4119–4131, Dublin, Ireland.

Uncertainty Estimation

Monte Carlo Dropout (Gal & Ghahramani, 16)

- Train model with dropout
- Multiple stochastic inference passes with dropout turned on (different masks)



Michał Oleszak. Monte Carlo Dropout. https://bit.ly/3cKiPGL

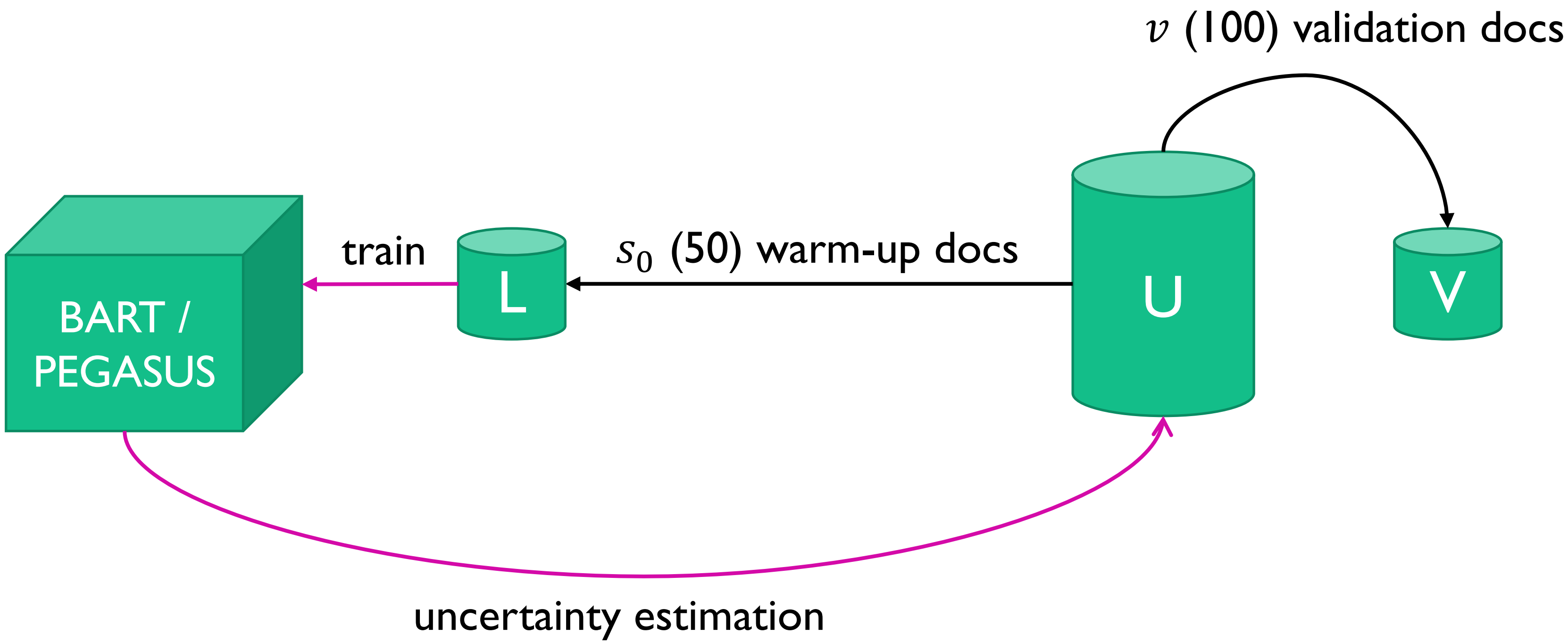
Following related work in machine translation (Xiao, Gomez & Gal, 20)

- Sample n (10) stochastic summaries for a given input
- Compute $\mathrm{BLEUVarN} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(1 - \mathrm{BLEU}\left(y_i, y_j\right)\right)^2$

Gidiotis, A., & Tsoumakas, G. (2023). Bayesian active summarization. Computer Speech and Language, 83. https://doi.org/10.1016/j.csl.2023.101553

Complexity Issue

Generating 10 summaries and computing their BLEUVarN for each document in U can be very costly for large |U|



v (100) validation docs

train

s_0 (50) warm-up docs

BART / PEGASUS

L

U

V

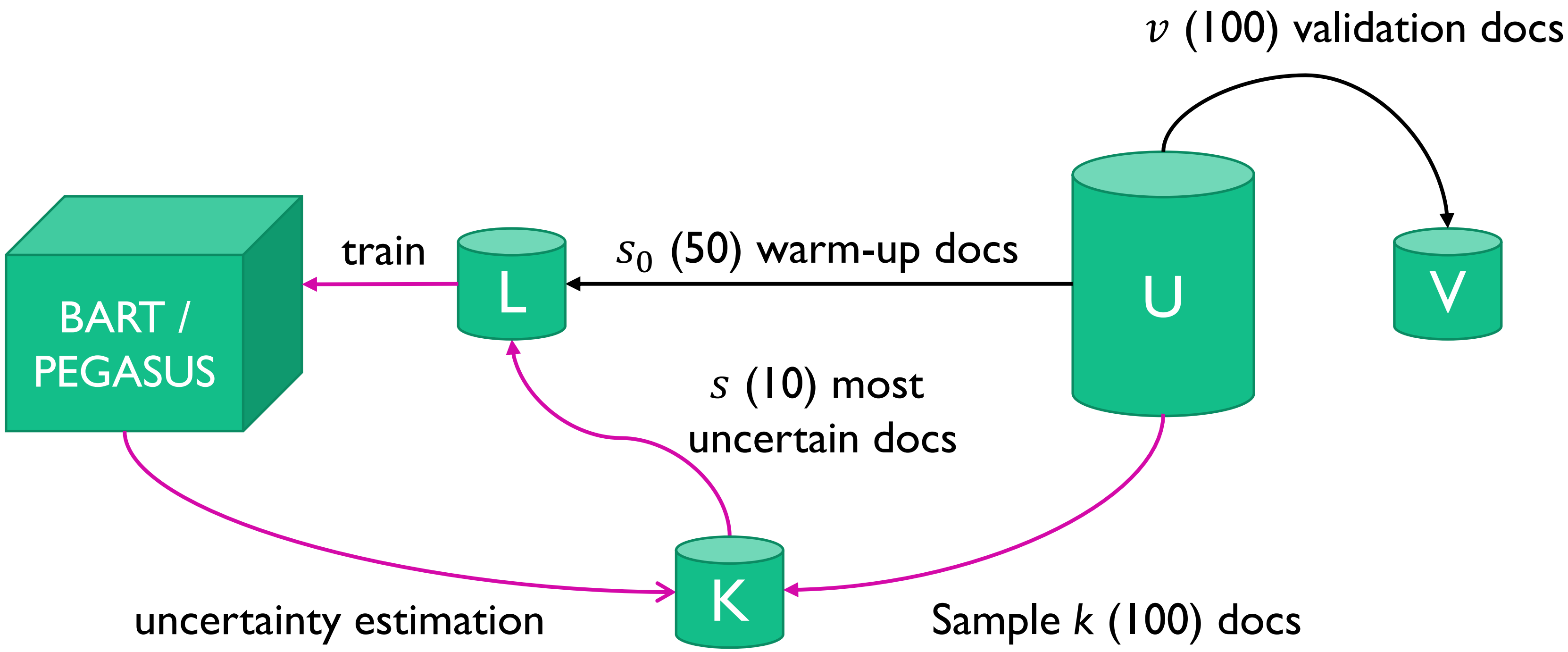
uncertainty estimation

Annotation Budget b	Training Docs
(900)	(0)
(800)	(50)
(750)	

Gidiotis, A., & Tsoumakas, G. (2023). Bayesian active summarization. Computer Speech and Language, 83. https://doi.org/10.1016/j.csl.2023.101553

Complexity Issue

Generating 10 summaries and computing their BLEUVarN for each document in U can be very costly for large |U|

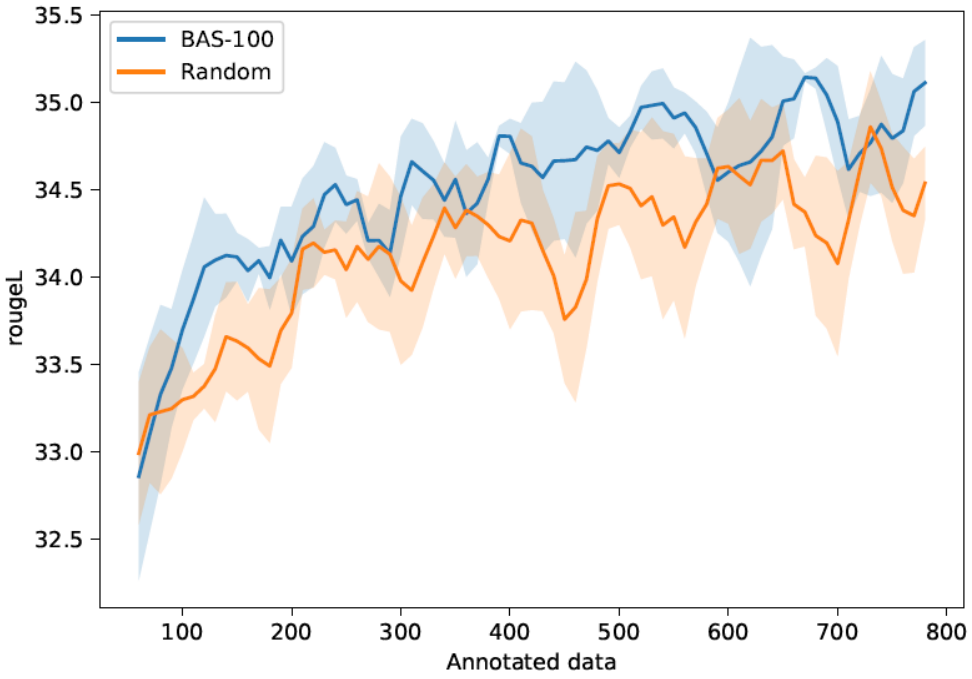


Annotation Budget b	Training Docs
(900)	(0)
(800)	(50)
(750)	(60)
(740)	(70)
(730)	…
…	(800)
(0)	

Gidiotis, A., & Tsoumakas, G. (2023). Bayesian active summarization. Computer Speech and Language, 83.
https://doi.org/10.1016/j.csl.2023.101553

Results on XSum

	R-1	R-2	R-L
PEGASUS pre-trained	17.84	2.65	12.71
b=150 Random	42.06	19.14	33.77
b=150 BAS-100	42.39	19.45	34.20
b=150 BAS-200	**42.55**	**19.59**	**34.31**
b=800 Random	43.25	20.07	35.02
b=800 BAS-100	**43.40**	**20.32**	**35.26**
b=800 BAS-200	43.38	20.24	35.11
PEGASUS full	44.90	23.33	37.74



Controlling the Output's Topic

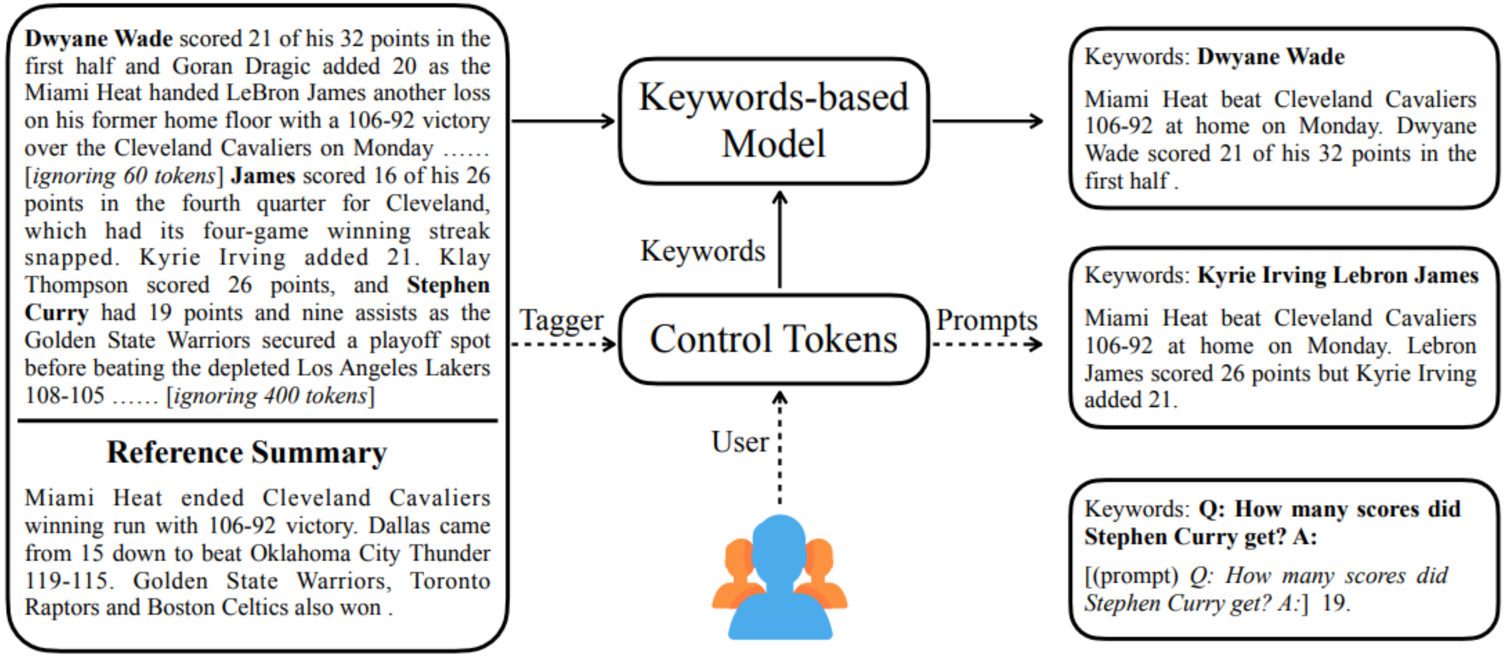
Controllable Summarization

Named entities

Length

Style

Topic



Dwyane Wade scored 21 of his 32 points in the first half and Goran Dragic added 20 as the Miami Heat handed LeBron James another loss on his former home floor with a 106-92 victory over the Cleveland Cavaliers on Monday ...... [*ignoring 60 tokens*] **James** scored 16 of his 26 points in the fourth quarter for Cleveland, which had its four-game winning streak snapped. Kyrie Irving added 21. Klay Thompson scored 26 points, and **Stephen Curry** had 19 points and nine assists as the Golden State Warriors secured a playoff spot before beating the depleted Los Angeles Lakers 108-105 ...... [*ignoring 400 tokens*]

Reference Summary

Miami Heat ended Cleveland Cavaliers winning run with 106-92 victory. Dallas came from 15 down to beat Oklahoma City Thunder 119-115. Golden State Warriors, Toronto Raptors and Boston Celtics also won .

Keywords: **Dwyane Wade**

Miami Heat beat Cleveland Cavaliers 106-92 at home on Monday. Dwyane Wade scored 21 of his 32 points in the first half .

Keywords: **Kyrie Irving Lebron James**

Miami Heat beat Cleveland Cavaliers 106-92 at home on Monday. Lebron James scored 26 points but Kyrie Irving added 21.

Keywords: **Q: How many scores did Stephen Curry get? A:**

[(prompt) *Q: How many scores did Stephen Curry get? A:*] 19.

CTRLsum: Towards Generic Controllable Text Summarization, Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, Caiming Xiong, arXiv 2020

Topic Control

One-hot encoded topic vectors concatenated to the embedding of each token

- Krishna & Srinivasan. Generating Topic-Oriented Summaries Using Neural Attention, NAACL 2018

Incorporate topical information into the attention mechanism of encoder-decoder RNNs

- Bahrainian, Zerveas, Crestani, and Eickhoff. 2021. CATS: Customizable Abstractive Topic-based Summarization. ACM Trans. Inf. Syst. 40, 1, Article 5 (January 2022), 24 pages

Proposed for RNNs

Require architectural changes to models

Not clear how to apply to Transformers

Evaluation is based on traditional ROUGE metrics

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods. In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

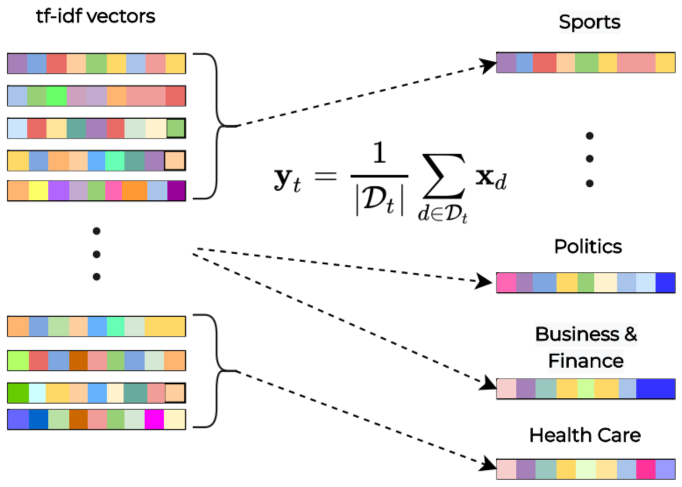
Topic Representation

We assume a set of topics T, and a set of documents D_t for each topic $t \in T$

For simplicity, we use a tf-idf representation, $\boldsymbol{x}_d$, for each document $d \in D_t$, of each topic $t \in T$, with IDF computed across $\bigcup_{t \in T} D_t$

The representation for a topic is computed as the average representation of its documents

- $\boldsymbol{y}_t = \frac{1}{|D_t|} \sum_{d \in D_t} \boldsymbol{x}_d$


tf-idf vectors
Sports
$\mathbf{y}_t = \frac{1}{|\mathcal{D}_t|} \sum_{d \in \mathcal{D}_t} \mathbf{x}_d$
Politics
Business & Finance
Health Care

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods. In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

Topic-Aware Evaluation

Given

- The tf-idf topic representation $\boldsymbol{y}_t = \frac{1}{|D_t|}\sum_{d \in D_t} \boldsymbol{x}_d$
- A representation of the summary $\boldsymbol{y}_s$ using the same tf-idf model

Summarization Topic Affinity Score (STAS)

- $\mathrm{STAS}(\boldsymbol{y}_s, \boldsymbol{y}_t) = \frac{\cos(\boldsymbol{y}_s, \boldsymbol{y}_t)}{\max_{z \in T}\{\cos(\boldsymbol{y}_s, \boldsymbol{y}_z)\}}$

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods. In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

Topic Control for Transformers

Topic embeddings (inspired from Krish. & Srin.)

- Trainable topic embeddings that are summed with the token embeddings and positional encodings

Prepending (inspired from CTRLsum)

- Add the gold/desired topic at the beginning of the input during training/inference

Tagging

- Tag with a special token the words of the topic representation with the top N tf-idf scores

$$\boldsymbol{z}_i = WE(w_i) + PE(i) + TE$$

Politics From Michael Jordan to LeBron James, how the NBA became a powerful political organization. Four decades ago, back when the NBA televised its championship games at midnight …

From Michael Jordan to LeBron James, how the NBA became a powerful **[TAG]**political **[TAG]**organization. Four decades ago, back when the NBA televised its championship games at midnight …

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods. In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

Topic Control for Transformers

Topic embeddings (inspired from Krish. & Srin.)

- Trainable topic embeddings that are summed with the token embeddings and positional encodings

Prepending (inspired from CTRLsum)

- Add the gold/desired topic at the beginning of the input during training/inference

Tagging

- Tag with a special token the words of the topic representation with the top N tf-idf scores

$$\boldsymbol{z}_i = WE(w_i) + PE(i) + TE$$

Sports From Michael Jordan to LeBron James, how the NBA became a powerful political organization. Four decades ago, back when the NBA televised its championship games at midnight …

From Michael Jordan to LeBron James, how the **[TAG]**NBA became a powerful political organization. Four decades ago, back when the **[TAG]**NBA televised its **[TAG]**championship **[TAG]**games at midnight …

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods. In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

Topic-Oriented Summarization Data

Not one but two very familiar faces will be ranged against Andy Murray on the support benches as he revisits one of the most highly charged matches of his career. Britain struck oil in the Falklands yesterday, a discovery likely to escalate already heightened tensions with Argentina over the ownership of the islands. Tomas Berdych is his opponent in the semi-final of the Miami Open, the man Murray met — and eventually beat — at the same stage of the Australian Open in January. After nine months of exploratory drilling, a group of British companies found oil and gas in a remote field north of the islands. …

Energy & Environment: British companies found oil and gas in a remote field north of the islands. Comes days after minister warned of 'very live threat' from Argentina.

Sports: British No 1 faces Tomas Berdych in the Miami Open semi-finals. Former coach Dani Vallverdu and now fitness trainer Jez Green left Andy Murray's team to join up with the Czech. Murray defeated Berdych in a controversial Australian Open semi-final.

Krishna & Srinivasan. Generating Topic-Oriented Summaries Using Neural Attention, NAACL 2018

Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods.
In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

Human Evaluation of STAS

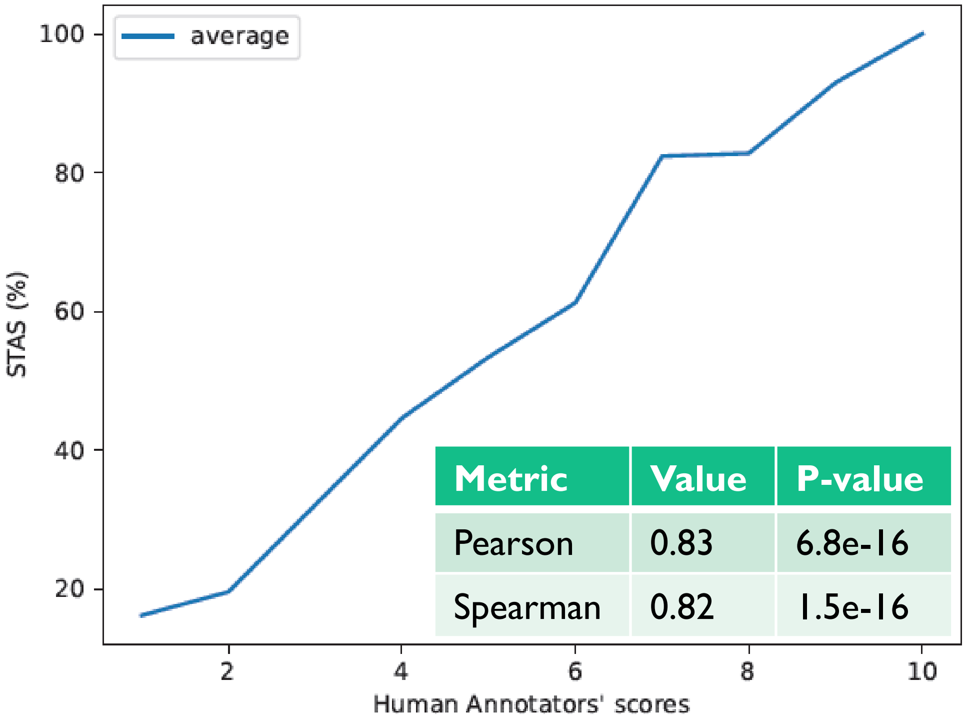
62 volunteers

- Graduate and undergraduate students

How relevant is this summary to this topic in a scale from 1 to 10?

- Randomly show them one of 10 summaries
- Randomly show them the correct or a different topic

Compute STAS for summary and topic



Metric	Value	P-value
Pearson	0.83	6.8e-16
Spearman	0.82	1.5e-16

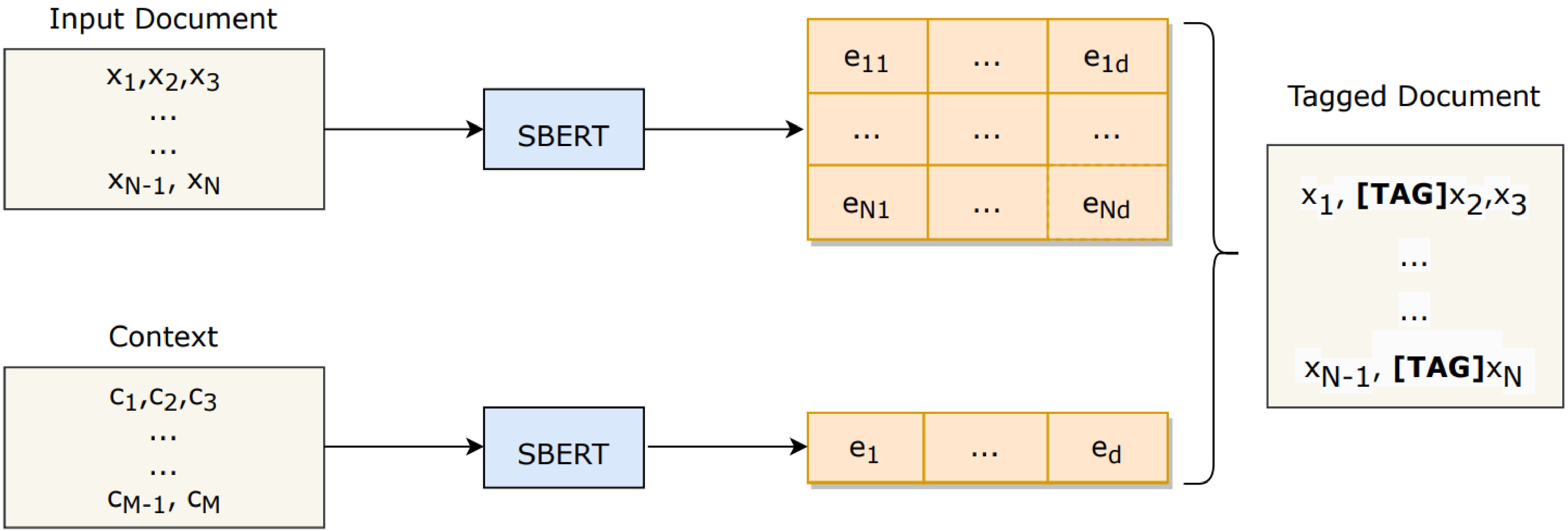
Passali, T., & Tsoumakas, G. (2024) Topic-Controllable Summarization: Topic-Aware Evaluation and Transformer Methods.
In Proceedings of LREC-COLING 2024, pages 16282–16292, Torino, Italia

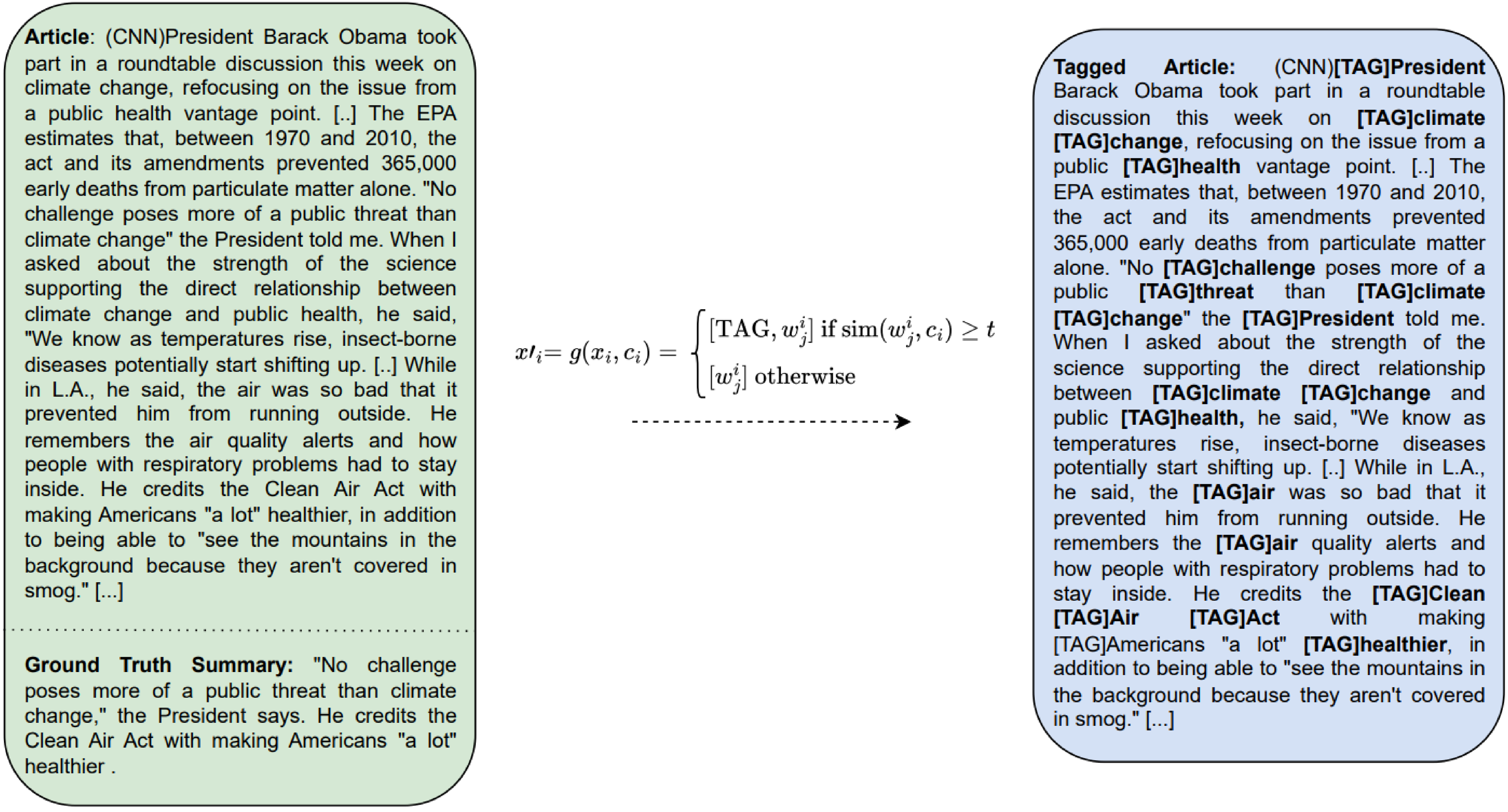
Evaluation of Methods

Model	Method	R-1	R-2	R-L	STAS (%)	Time (s)
BART	-	30.46	11.92	20.57	51.86	
BART	TAG	39.30	18.06	36.67	68.42	39
BART	EMB	40.15	18.53	37.41	68.50	303
BART	PRE	41.58	19.55	38.74	71.90	31
BART	PRE+TAG	**41.66**	**19.57**	**38.83**	**72.36**	40

Passali, T., & Tsoumakas, G. (2024) Controllable Abstractive Summarization with Arbitrary Textual Context
Under review (Natural Language Engineering)

Towards Arbitrary Textual Context





Article: (CNN)President Barack Obama took part in a roundtable discussion this week on climate change, refocusing on the issue from a public health vantage point. [..] The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early deaths from particulate matter alone. "No challenge poses more of a public threat than climate change" the President told me. When I asked about the strength of the science supporting the direct relationship between climate change and public health, he said, "We know as temperatures rise, insect-borne diseases potentially start shifting up. [..] While in L.A., he said, the air was so bad that it prevented him from running outside. He remembers the air quality alerts and how people with respiratory problems had to stay inside. He credits the Clean Air Act with making Americans "a lot" healthier, in addition to being able to "see the mountains in the background because they aren't covered in smog." [...]

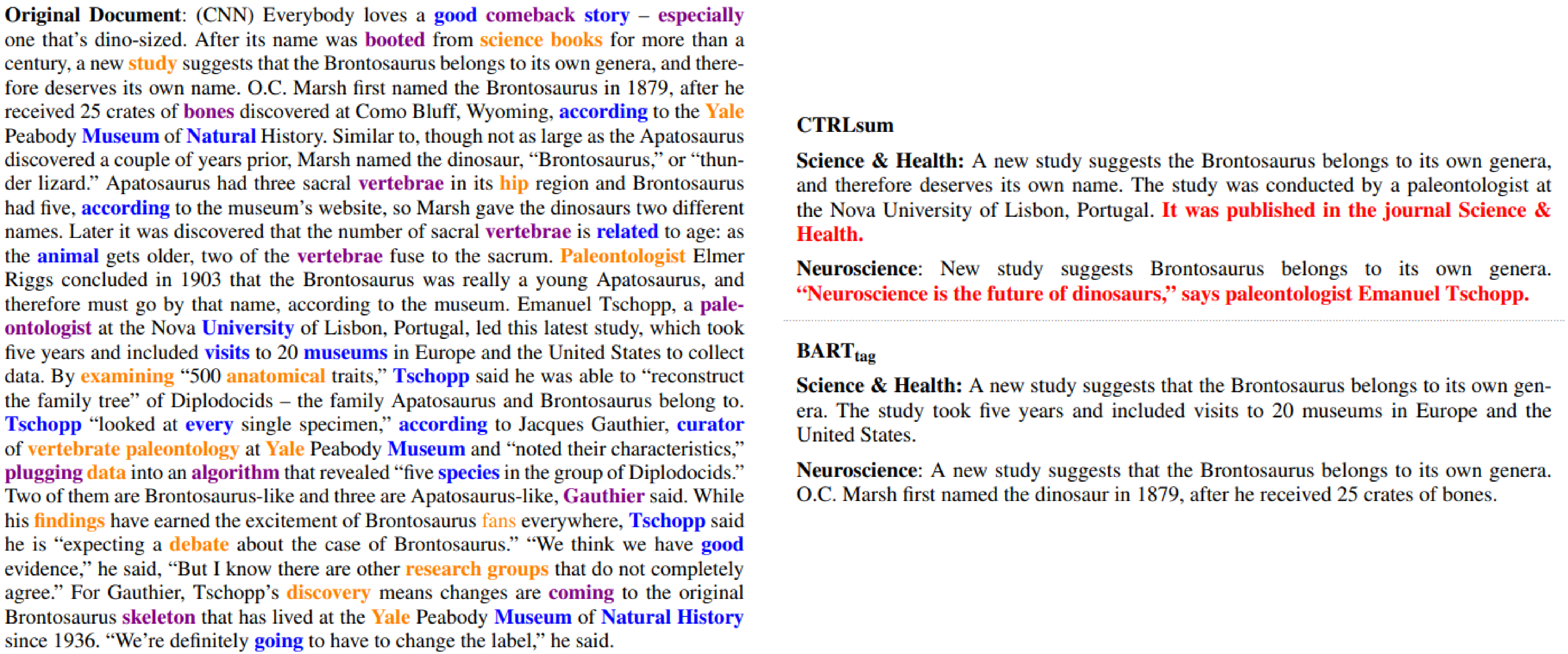
Ground Truth Summary: "No challenge poses more of a public threat than climate change," the President says. He credits the Clean Air Act with making Americans "a lot" healthier .

$$x\prime_i = g(x_i, c_i) = \begin{cases} [\mathrm{TAG}, w_j^i] \text{ if } \mathrm{sim}(w_j^i, c_i) \geq t \\ [w_j^i] \text{ otherwise} \end{cases}$$

Tagged Article: (CNN)**[TAG]President** Barack Obama took part in a roundtable discussion this week on **[TAG]climate [TAG]change**, refocusing on the issue from a public **[TAG]health** vantage point. [..] The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early deaths from particulate matter alone. "No **[TAG]challenge** poses more of a public **[TAG]threat** than **[TAG]climate [TAG]change**" the **[TAG]President** told me. When I asked about the strength of the science supporting the direct relationship between **[TAG]climate [TAG]change** and public **[TAG]health,** he said, "We know as temperatures rise, insect-borne diseases potentially start shifting up. [..] While in L.A., he said, the **[TAG]air** was so bad that it prevented him from running outside. He remembers the **[TAG]air** quality alerts and how people with respiratory problems had to stay inside. He credits the **[TAG]Clean [TAG]Air [TAG]Act** with making [TAG]Americans "a lot" **[TAG]healthier**, in addition to being able to "see the mountains in the background because they aren't covered in smog." [...]

Passali, T., & Tsoumakas, G. (2024) Controllable Abstractive Summarization with Arbitrary Textual Context
Under review (Natural Language Engineering)

Need for Hallucination Aware Evaluation



Original Document: (CNN) Everybody loves a **good** **comeback** **story** – **especially** one that's dino-sized. After its name was **booted** from **science** **books** for more than a century, a new **study** suggests that the Brontosaurus belongs to its own genera, and therefore deserves its own name. O.C. Marsh first named the Brontosaurus in 1879, after he received 25 crates of **bones** discovered at Como Bluff, Wyoming, **according** to the **Yale** Peabody **Museum** of **Natural** History. Similar to, though not as large as the Apatosaurus discovered a couple of years prior, Marsh named the dinosaur, "Brontosaurus," or "thunder lizard." Apatosaurus had three sacral **vertebrae** in its **hip** region and Brontosaurus had five, **according** to the museum's website, so Marsh gave the dinosaurs two different names. Later it was discovered that the number of sacral **vertebrae** is **related** to age: as the **animal** gets older, two of the **vertebrae** fuse to the sacrum. **Paleontologist** Elmer Riggs concluded in 1903 that the Brontosaurus was really a young Apatosaurus, and therefore must go by that name, according to the museum. Emanuel Tschopp, a **paleontologist** at the Nova **University** of Lisbon, Portugal, led this latest study, which took five years and included **visits** to 20 **museums** in Europe and the United States to collect data. By **examining** "500 **anatomical** traits," **Tschopp** said he was able to "reconstruct the family tree" of Diplodocids – the family Apatosaurus and Brontosaurus belong to. **Tschopp** "looked at **every** single specimen," **according** to Jacques Gauthier, **curator** of **vertebrate** **paleontology** at **Yale** Peabody **Museum** and "noted their characteristics," **plugging** **data** into an **algorithm** that revealed "five **species** in the group of Diplodocids." Two of them are Brontosaurus-like and three are Apatosaurus-like, **Gauthier** said. While his **findings** have earned the excitement of Brontosaurus fans everywhere, **Tschopp** said he is "expecting a **debate** about the case of Brontosaurus." "We think we have **good** evidence," he said, "But I know there are other **research** **groups** that do not completely agree." For Gauthier, Tschopp's **discovery** means changes are **coming** to the original Brontosaurus **skeleton** that has lived at the **Yale** Peabody **Museum** of **Natural** **History** since 1936. "We're definitely **going** to have to change the label," he said.

CTRLsum

Science & Health: A new study suggests the Brontosaurus belongs to its own genera, and therefore deserves its own name. The study was conducted by a paleontologist at the Nova University of Lisbon, Portugal. **It was published in the journal Science & Health.**

Neuroscience: New study suggests Brontosaurus belongs to its own genera. **"Neuroscience is the future of dinosaurs," says paleontologist Emanuel Tschopp.**

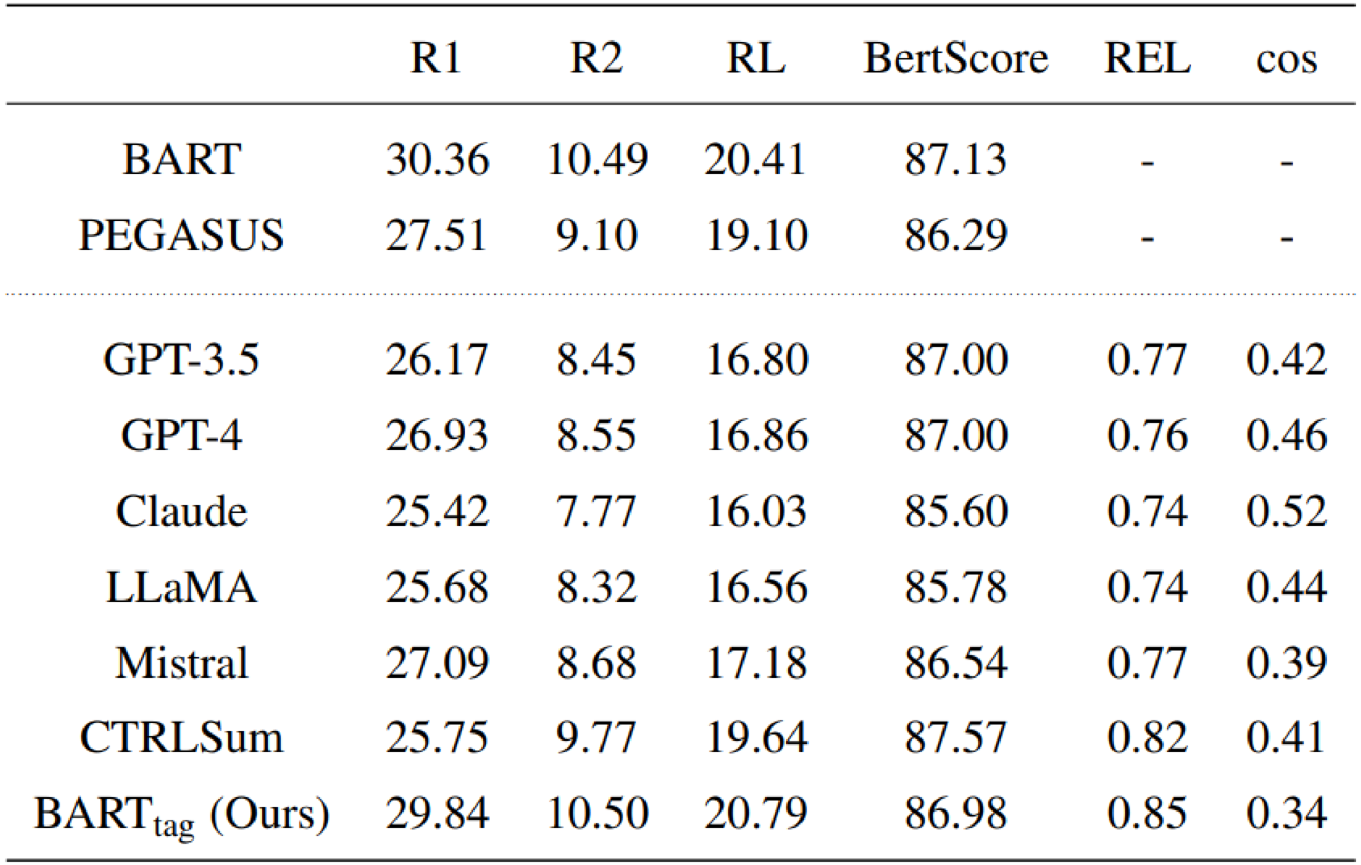
BART$_{tag}$

Science & Health: A new study suggests that the Brontosaurus belongs to its own genera. The study took five years and included visits to 20 museums in Europe and the United States.

Neuroscience: A new study suggests that the Brontosaurus belongs to its own genera. O.C. Marsh first named the dinosaur in 1879, after he received 25 crates of bones.

Passali, T., & Tsoumakas, G. (2024) Controllable Abstractive Summarization with Arbitrary Textual Context
Under review (Natural Language Engineering)

Results on the MacDoc dataset



	R1	R2	RL	BertScore	REL	cos
BART	30.36	10.49	20.41	87.13	-	-
PEGASUS	27.51	9.10	19.10	86.29	-	-
GPT-3.5	26.17	8.45	16.80	87.00	0.77	0.42
GPT-4	26.93	8.55	16.86	87.00	0.76	0.46
Claude	25.42	7.77	16.03	85.60	0.74	0.52
LLaMA	25.68	8.32	16.56	85.78	0.74	0.44
Mistral	27.09	8.68	17.18	86.54	0.77	0.39
CTRLSum	25.75	9.77	19.64	87.57	0.82	0.41
BART$_{tag}$ (Ours)	29.84	10.50	20.79	86.98	0.85	0.34

REL

- Given a generated summary S, we extract the sentence from the summary that is closest to the requested topic
- Then, REL is computed as the maximum of all the similarities between the selected sentence representation and each of the sentence representations of the original document

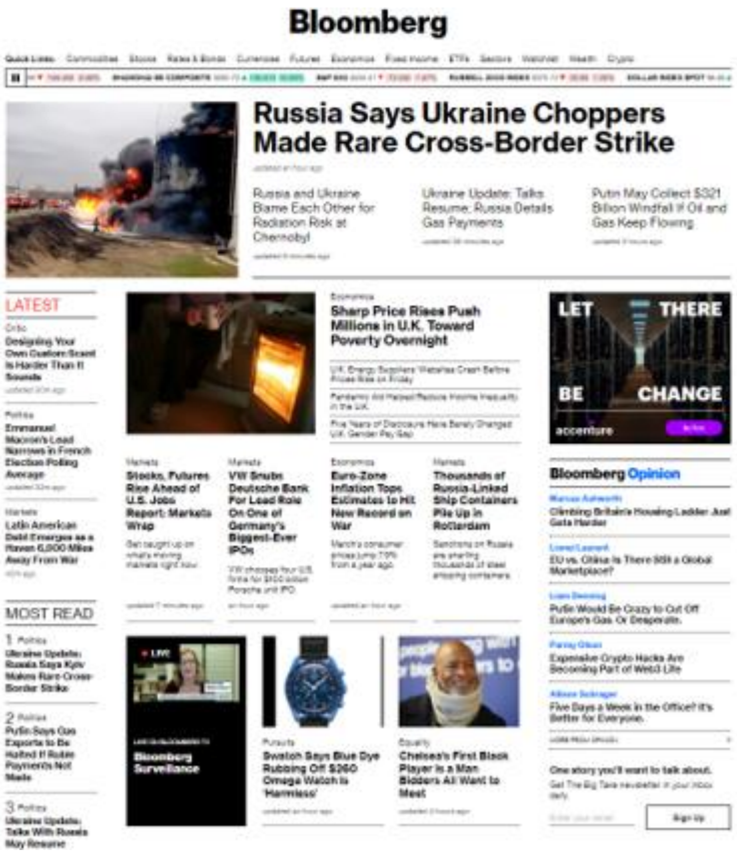
Healthcare and Finance Apps

Passali, Gidiotis, Chatzikyriakidis, Tsoumakas (2021) Towards Human-Centered Summarization: A Case Study on Financial News, Proc. EACL Workshop on Bridging HCI and NLP

Financial Summarization

Data collection from Bloomberg's Market and Financial News API

PEGASUS model pre-trained on C4 and HugeNews, fine-tuned on the XSum news dataset, and further fine-tuned on our financial data set



Bloomberg

Russia Says Ukraine Choppers Made Rare Cross-Border Strike

Russia and Ukraine Blame Each Other for Radiation Risk at Chernobyl

Ukraine Update: Talks Resume; Russia Details Gas Payments

Putin May Collect $321 Billion Windfall If Oil and Gas Keep Flowing

LATEST

Designing Your Own Custom Scent Is Harder Than It Sounds

Emmanuel Macron's Lead Narrows in French Election Polling Average

Latin American Debt Emerges as a Haven 6,000 Miles Away From War

Sharp Price Rises Push Millions in U.K. Toward Poverty Overnight

Stocks, Futures Rise Ahead of U.S. Jobs Report: Markets Wrap

VW Snubs Deutsche Bank For Lead Role On One of Germany's Biggest-Ever IPOs

Euro-Zone Inflation Tops Estimates to Hit New Record on War

Thousands of Russia-Linked Ship Containers Pile Up in Rotterdam

MOST READ

1 Ukraine Update: Russia Says Kyiv Makes Rare Cross-Border Strike

2 Putin Says Gas Exports to Be Halted If Ruble Payments Not Made

3 Ukraine Update: Talks With Russia May Resume

Bloomberg Surveillance

Swatch Says Blue Dye Rubbing Off $260 Omega Watch Is 'Harmless'

Chelsea's First Black Player Is a Man Bidders All Want to Meet

Bloomberg Opinion

Climbing Britain's Housing Ladder Just Gets Harder

EU vs. China: Is There Still a Global Marketplace?

Putin Would Be Crazy to Cut Off Europe's Gas. Or Desperate.

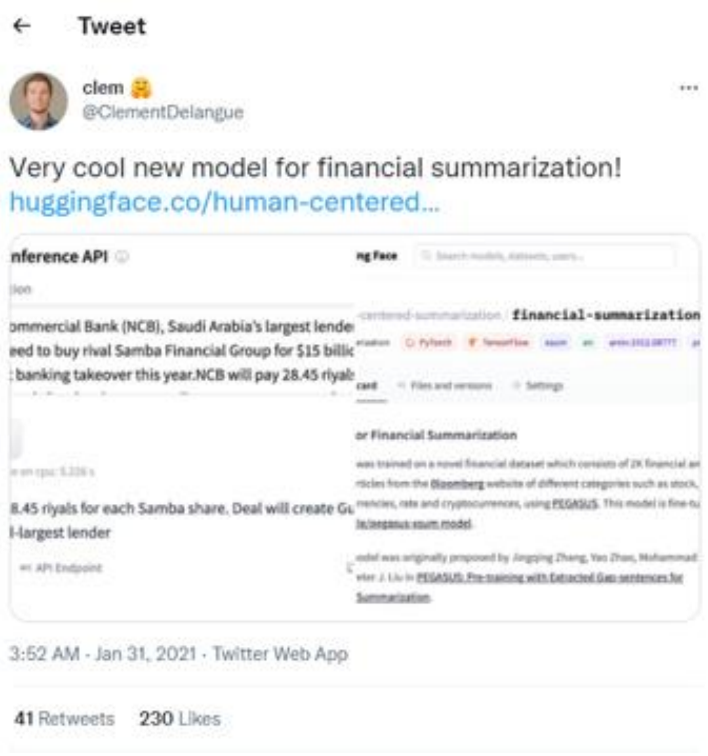
Expensive Crypto Hacks Are Becoming Part of Web3 Life

Five Days a Week in the Office? It's Better for Everyone.

One story you'll want to talk about.

Sign Up

Financial Summarization



← Tweet

clem 🤗
@ClementDelangue

Very cool new model for financial summarization! huggingface.co/human-centered...

nference API

ommercial Bank (NCB), Saudi Arabia's largest lende
eed to buy rival Samba Financial Group for $15 billic
: banking takeover this year.NCB will pay 28.45 riyal:

8.45 riyals for each Samba share. Deal will create Gu
l-largest lender

API Endpoint

-centered-summarization / financial-summarization

PyTorch · TensorFlow · xsum · en · arxiv:1912.08777

card · Files and versions · Settings

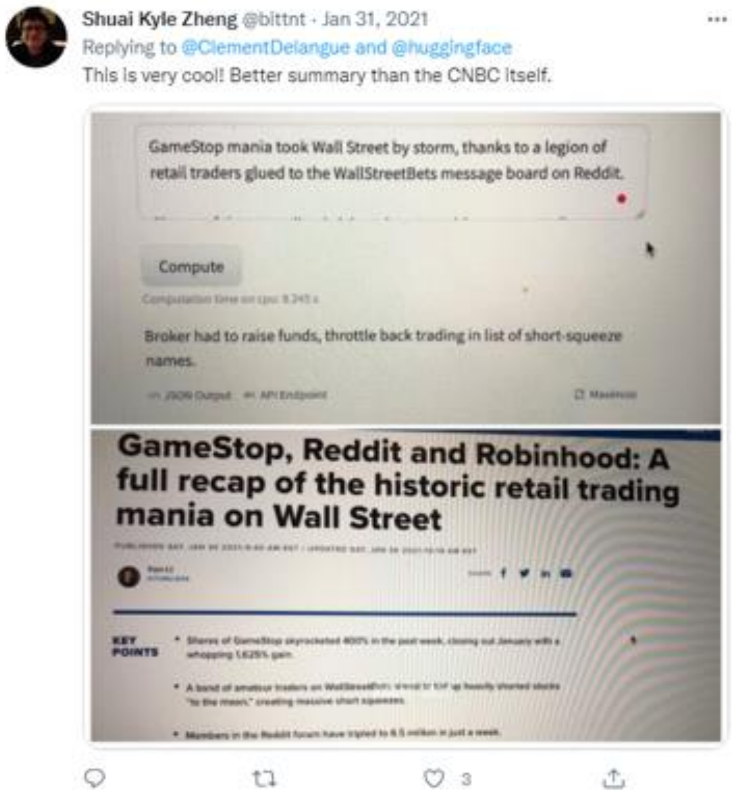
or Financial Summarization

was trained on a novel financial dataset which consists of 2K financial ar ticles from the Bloomberg website of different categories such as stock, rencies, rate and cryptocurrences, using PEGASUS. This model is fine-tu le/pegasus-xsum model.

odel was originally proposed by Jingqing Zhang, Yao Zhao, Mohammad eter J. Liu in PEGASUS: Pre-training with Extracted Gap-sentences for Summarization.

3:52 AM · Jan 31, 2021 · Twitter Web App

41 Retweets 230 Likes



Shuai Kyle Zheng @bittnt · Jan 31, 2021
Replying to @ClementDelangue and @huggingface
This is very cool! Better summary than the CNBC itself.

GameStop mania took Wall Street by storm, thanks to a legion of retail traders glued to the WallStreetBets message board on Reddit.

Compute

Computation time on cpu: 8.345 s

Broker had to raise funds, throttle back trading in list of short-squeeze names.

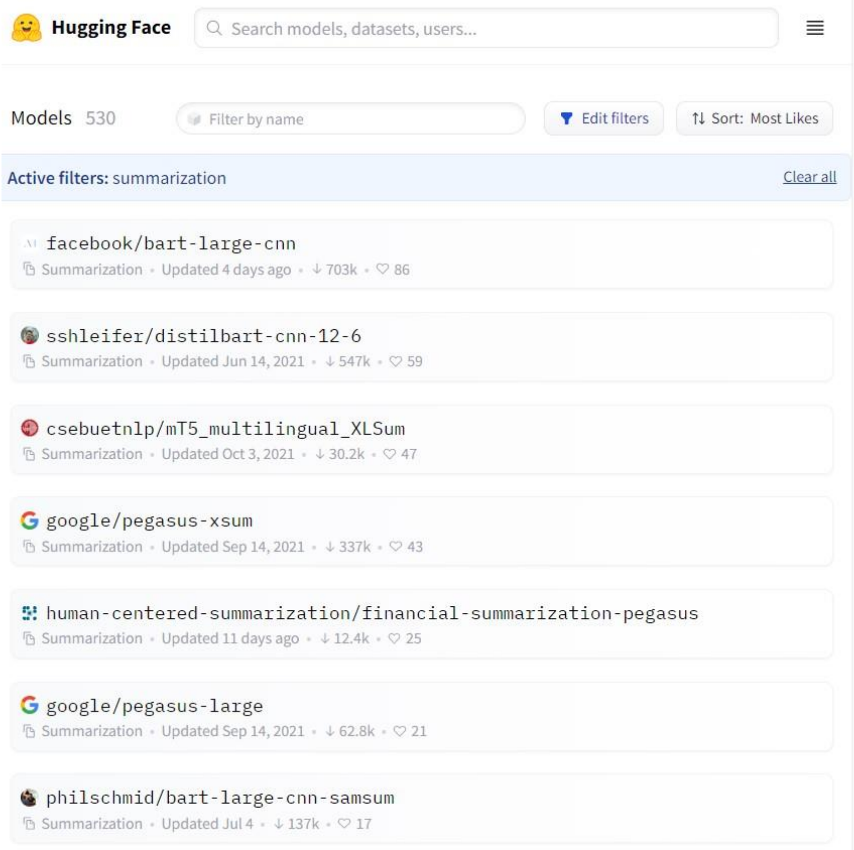
JSON Output · API Endpoint · Maximize

GameStop, Reddit and Robinhood: A full recap of the historic retail trading mania on Wall Street

KEY POINTS

- Shares of GameStop skyrocketed 400% in the past week, closing out January with a whopping 1,625% gain.
- A band of amateur traders on WallStreetBets vowed to lift up heavily shorted stocks "to the moon," creating massive short squeezes.
- Members in the Reddit forum have tripled to 8.5 million in just a week.

3



Hugging Face · Search models, datasets, users...

Models 530 · Filter by name · Edit filters · Sort: Most Likes

Active filters: summarization · Clear all

facebook/bart-large-cnn
Summarization · Updated 4 days ago · ↓ 703k · ♡ 86

sshleifer/distilbart-cnn-12-6
Summarization · Updated Jun 14, 2021 · ↓ 547k · ♡ 59

csebuetnlp/mT5_multilingual_XLSum
Summarization · Updated Oct 3, 2021 · ↓ 30.2k · ♡ 47

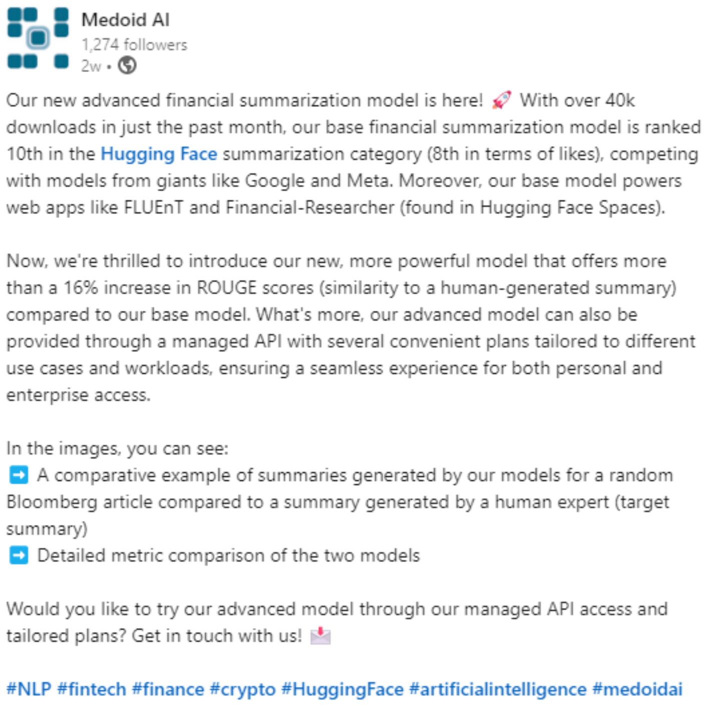
google/pegasus-xsum
Summarization · Updated Sep 14, 2021 · ↓ 337k · ♡ 43

human-centered-summarization/financial-summarization-pegasus
Summarization · Updated 11 days ago · ↓ 12.4k · ♡ 25

google/pegasus-large
Summarization · Updated Sep 14, 2021 · ↓ 62.8k · ♡ 21

philschmid/bart-large-cnn-samsum
Summarization · Updated Jul 4 · ↓ 137k · ♡ 17

Financial Summarization



Medoid AI
1,274 followers
2w •

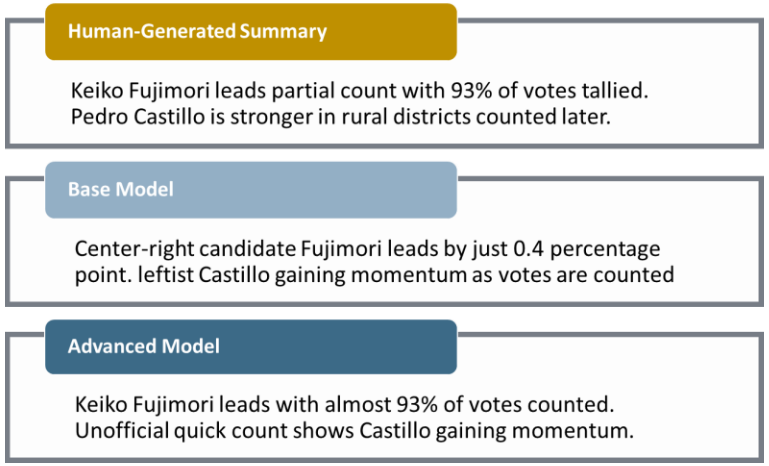
Our new advanced financial summarization model is here! 🚀 With over 40k downloads in just the past month, our base financial summarization model is ranked 10th in the Hugging Face summarization category (8th in terms of likes), competing with models from giants like Google and Meta. Moreover, our base model powers web apps like FLUEnT and Financial-Researcher (found in Hugging Face Spaces).

Now, we're thrilled to introduce our new, more powerful model that offers more than a 16% increase in ROUGE scores (similarity to a human-generated summary) compared to our base model. What's more, our advanced model can also be provided through a managed API with several convenient plans tailored to different use cases and workloads, ensuring a seamless experience for both personal and enterprise access.

In the images, you can see:
➡️ A comparative example of summaries generated by our models for a random Bloomberg article compared to a summary generated by a human expert (target summary)
➡️ Detailed metric comparison of the two models

Would you like to try our advanced model through our managed API access and tailored plans? Get in touch with us! 📩

#NLP #fintech #finance #crypto #HuggingFace #artificialintelligence #medoidai



Human-Generated Summary

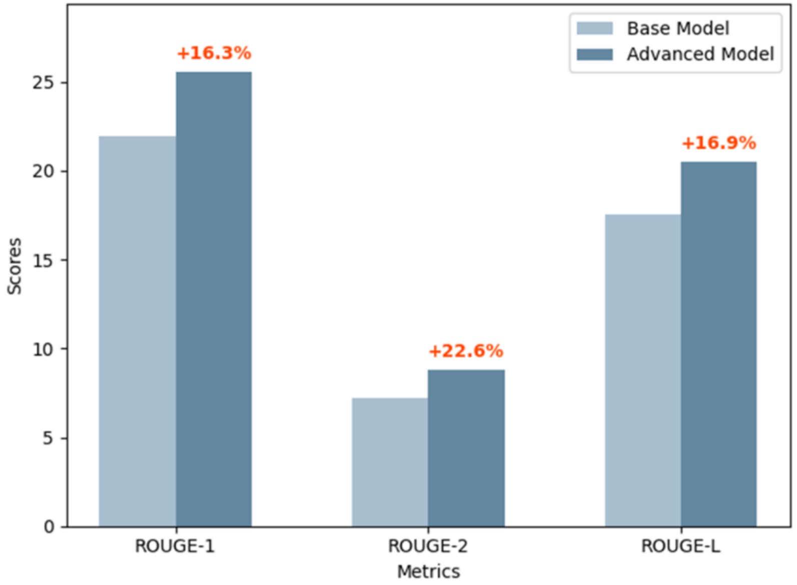
Keiko Fujimori leads partial count with 93% of votes tallied.
Pedro Castillo is stronger in rural districts counted later.

Base Model

Center-right candidate Fujimori leads by just 0.4 percentage point. leftist Castillo gaining momentum as votes are counted

Advanced Model

Keiko Fujimori leads with almost 93% of votes counted.
Unofficial quick count shows Castillo gaining momentum.



Model	R-1	R-2	R-L
ChatGPT* zero shot	15.90	3.49	14.38
Medoid AI Base	21.98	7.20	17.56
Medoid AI Advanced	25.56	8.83	20.52

* Prompt: *Summarize the text below in two sentences*

“A global survey by 3M that found 88% of people think scientists should speak in easy-to-understand language”

Clinical Studies,
Scientific Publications

Plain Language Summary (PLS)

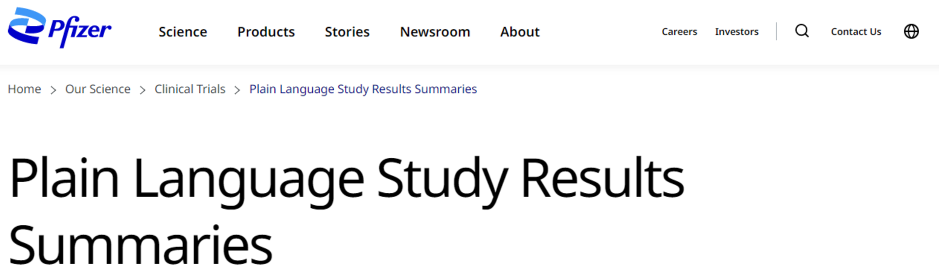
EU Regulation No 536/2014
US Public Health Service Act 2007

Patients

General Audience

Giannouris, Myridis, Passali, Tsoumakas (2024) Plain Language Summarization of Clinical Trials
In Proceedings of DeTermIt workshop @ LREC-COLING 2024, pages 60–67, Torino, Italia.

Lay Summarization of Clinical Trials



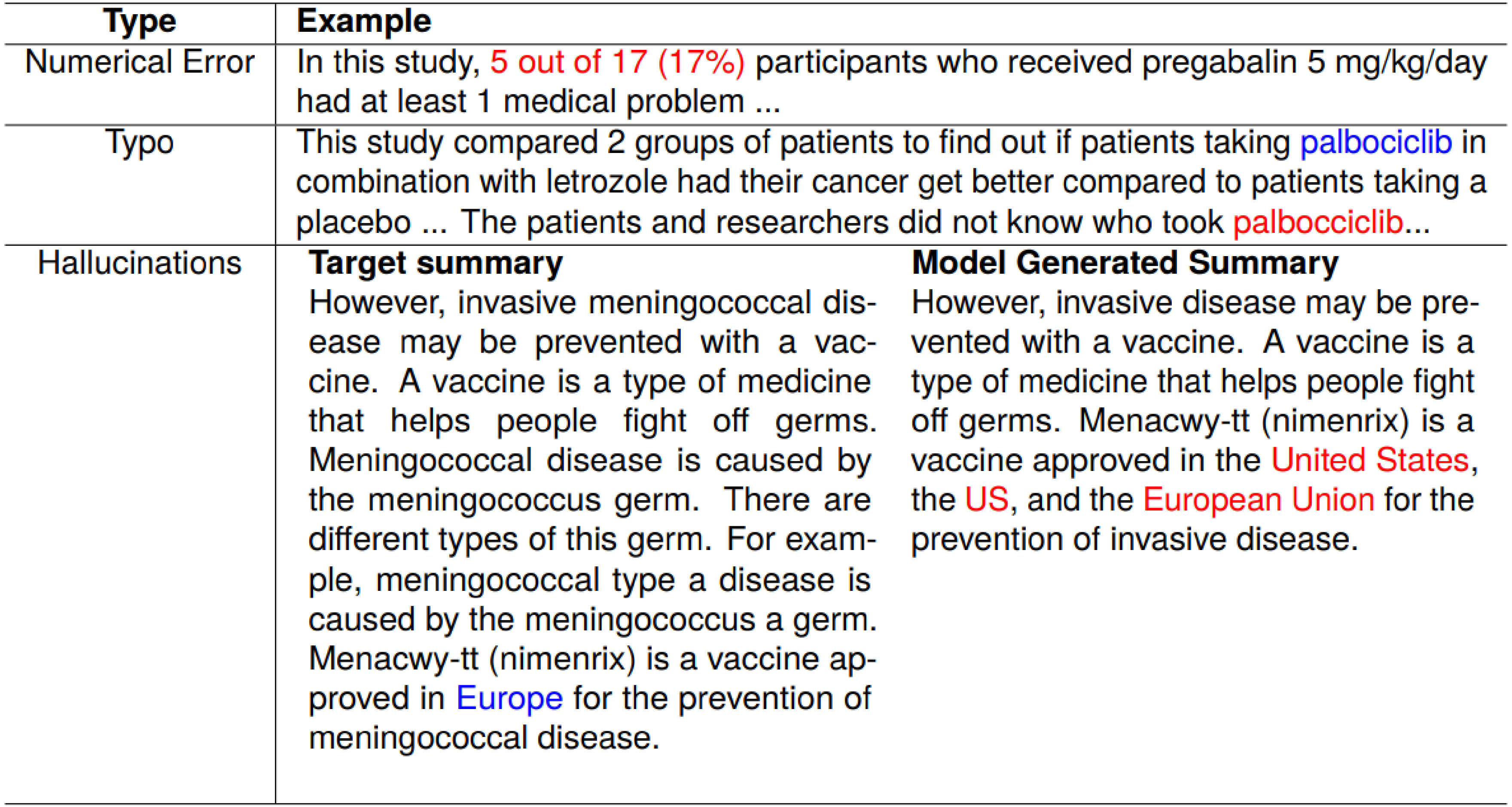
Home > Our Science > Clinical Trials > Plain Language Study Results Summaries

Plain Language Study Results Summaries

		Length		Clinical Trials			Results		
Id	Question	Source	Target	Train	Val	Test	R1	R2	RL
Q1	Why was this study done?	641	321	78	13	18	53,31	26,98	33,26
Q2	What happened during the study?	146	559	74	13	18	47,54	19,07	25,89
Q3	What were the results of the study?	-	-	-	-	-	-	-	-
Q4	What medical problems did patients have during the study?	663	421	103	13	18	77,49	68,98	73,09
Q5	Were there any serious medical problems?	663	131	107	13	18	55,47	38,44	45,48

Giannouris, Myridis, Passali, Tsoumakas (2024) Plain Language Summarization of Clinical Trials
In Proceedings of DeTermIt workshop @ LREC-COLING 2024, pages 60–67, Torino, Italia.

Lay Summarization of Clinical Trials



Type	Example	
Numerical Error	In this study, 5 out of 17 (17%) participants who received pregabalin 5 mg/kg/day had at least 1 medical problem ...	
Typo	This study compared 2 groups of patients to find out if patients taking palbociclib in combination with letrozole had their cancer get better compared to patients taking a placebo ... The patients and researchers did not know who took palbocciclib...	
Hallucinations	**Target summary** However, invasive meningococcal disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Meningococcal disease is caused by the meningococcus germ. There are different types of this germ. For example, meningococcal type a disease is caused by the meningococcus a germ. Menacwy-tt (nimenrix) is a vaccine approved in Europe for the prevention of meningococcal disease.	**Model Generated Summary** However, invasive disease may be prevented with a vaccine. A vaccine is a type of medicine that helps people fight off germs. Menacwy-tt (nimenrix) is a vaccine approved in the United States, the US, and the European Union for the prevention of invasive disease.

Stefanou, Passali, Tsoumakas (2024) AUTH at BioLaySumm 2024: Bringing Scientific Content to Kids.
In Proceedings of 23rd Workshop on Biomedical Natural Language Processing, pages 380–389, Bangkok, Thailand.

Lay Summarization for Kids

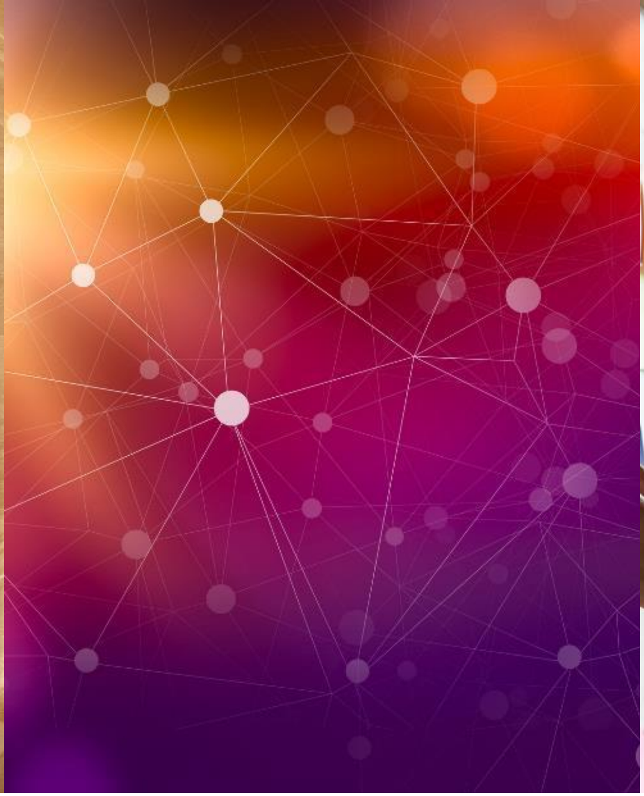
7th out of 57 participants in the BioLaySumm 2024 shared task

- Abstractive summarization of biomedical publications in lay terms

Our approach

- BioBART-v2 model fine-tuned using abstracts from eLife, PLOS
- Some training samples had high complexity summaries
- New SKJ dataset with content from the Science Journal for Kids
- Added synthetic summaries using GPT4 in a few-shot fashion, including summaries from the SJK dataset in the prompt
- Improved readability of lay summaries

SCIENCE JOURNAL FOR KIDS



Summary

Neural abstractive summarization

Interesting research challenges (long text, uncertainty, control)

Applications in two important domains (finance, healthcare)

Team

Alexios Gidiotis

PhD Student

Tatiana Passali

PhD Student

Stathis Chatzikyriakidis

Polydoros Giannouris

Loukritia Stefanou

Thodoris Myridis

ARISTOTLE UNIVERSITY OF THESSALONIKI

Medoid AI

Thank You

Grigorios Tsoumakas

greg@csd.auth.gr

greg@medoid.ai