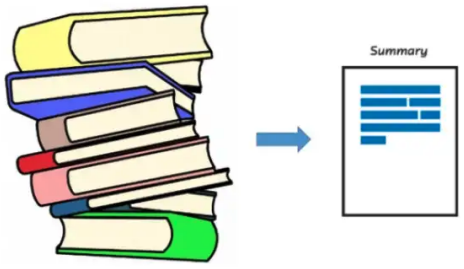
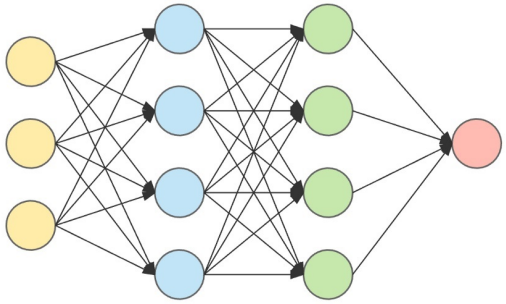
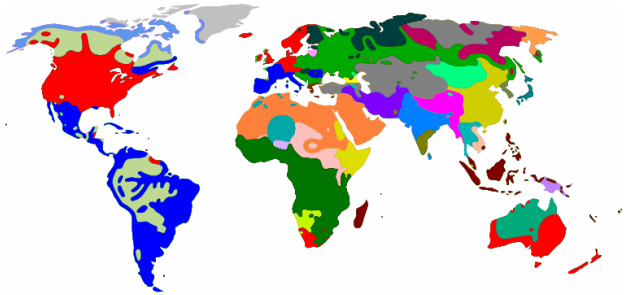
Summarization with Latent Structure, Context Factors and Quantitative Precision

Shay Cohen
ILCC, School of Informatics
University of Edinburgh

March 9, 2023

Stuff We Do in My Group...

 Text Generation	 Grammar Learning	 Rep. Learning
 Semantics	 Multilinguality	 Applications

Understand better the interface between *computation* and *language*

Large Language Models

Large volumes of data lead to an *ungrounded* understanding of text, divorced from *computation* - a fantastic achievement

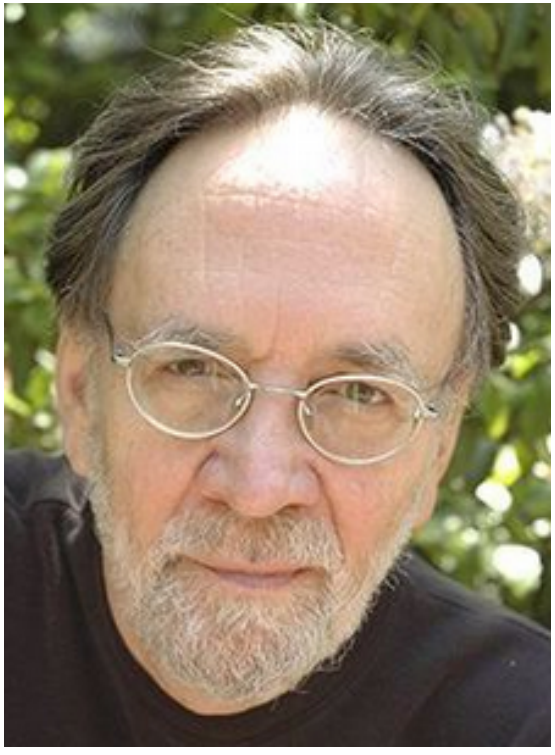
Large Language Models

George Miller
(1920–2012)

Humans can recognize sets of strings (formal languages) such as $DC^nD \cup CD^nC$ without grounding

Large Language Models

George Miller
(1920–2012)



Arthur Reber

They do not need to explicitly construct rule sets for that even where there is an underlying grammar

Octopuses and Language

Bender and Koller (2020)

Two Types of Text Understanding

Understanding, loosely, can refer to two different things:

- Attaching meaning (representations?) which is **divorced of computation** and direct denotation (large language models)
- Attaching meaning which is either executable and has **denotation** or is grounded in an environment (semantic parsing, agents)

Two Types of Text Understanding

Understanding, loosely, can refer to two different things:

- Attaching meaning (representations?) which is **divorced of computation** and direct denotation (large language models)
- Attaching meaning which is either executable and has **denotation** or is grounded in an environment (semantic parsing, agents)

Today: Understanding of the first kind, in the context of summarization

Machine Translation vs. Summarization

Feature	**Machine Translation**	**Summarization**
Text generation/transduction	✓	✓
Complex evaluation	✓	✓
Assistant Tools	✓	✓
Implicit background factors	✓	✓✓
Synthesis	✓	✓✓✓

Kung Fu Nuns in Nepal boost their health in the fight for women's rights

By Sandee LaMotte, CNN

Updated 1637 GMT (0037 HKT) December 21, 2019

Nepali Buddhist nuns practise kung fu at the Amitabha Drukpa Nunnery on the outskirts of Kathmandu.

(CNN) — Swords swirl around their bodies, coming perilously close to piercing flesh. Blades flashing in the morning sun, the young women twirl, cartwheel and then kick in unison, finishing their graceful movements in a centuries old kung fu fighting stance.

Extractive summarization: create a summary by identifying (and subsequently concatenating) the most important sentences in a document

Extractive summary:

- Jigme Yangchen Ghamo has lived at the Druk Amitabha Mountain Nunnery perched high in the mountains outside Kathmandu since she was 10 years old.
- The Drukpa Order is a branch of Himalayan Buddhism , a faith which traditionally considers women second-class citizens.
- According to Buddhist narratives , a woman can not achieve spiritual enlightenment unless she is reborn as a man.

The “holy grail” of summarization is fully, contextual **abstractive summarization**

Three Steps in Improving Abstractive Summ.

Model document structure for synthesis (Qiu and Cohen, 2022)

Consider context factors (Fonseca et al., 2022)

Ensure correctness of output text (Zhao et al., 2020)

Source

1. The town is home to the prestigious Leander Club, which has trained more than 100 Olympic medal-winning rowers.
- *2 sentences are abbreviated here.*
4. The Royal Mail has painted more than 50 postboxes gold following Team GB's gold medal haul at London 2012.
5. Originally it said it was only painting them in winners home towns, or towns with which they are closely associated.
6. Town mayor Elizabeth Hodgkin said: " We are the home of rowing ... I feel very excited about it."
- *5 sentences are abbreviated here.*
12. The Henley-on-Thames postbox was painted on Friday.
- *one sentence is abbreviated here.*

Reference Summary

The Royal Mail has painted a postbox gold in the Oxford-shire town of Henley-on-Thames - in recognition of its medal winning rowing club.

Source

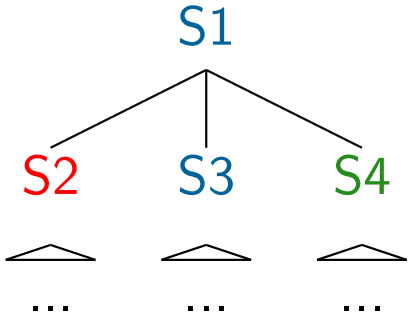
1. The town is home to the prestigious Leander Club, which has trained more than 100 Olympic medal-winning rowers.
- *2 sentences are abbreviated here.*
4. The Royal Mail has painted more than 50 postboxes gold following Team GB's gold medal haul at London 2012.
5. Originally it said it was only painting them in winners home towns, or towns with which they are closely associated.
6. Town mayor Elizabeth Hodgkin said: " We are the home of rowing ... I feel very excited about it."
- *5 sentences are abbreviated here.*
12. The Henley-on-Thames postbox was painted on Friday.
- *one sentence is abbreviated here.*

Reference Summary

The Royal Mail has painted a postbox gold in the Oxford-shire town of Henley-on-Thames - in recognition of its medal winning rowing club.

Source

1. The town is home to the prestigious Leander Club, which has trained more than 100 Olympic medal-winning rowers.
- *2 sentences are abbreviated here.*
4. The Royal Mail has painted more than 50 postboxes gold following Team GB's gold medal haul at London 2012.
5. Originally it said it was only painting them in winners home towns, or towns with which they are closely associated.
6. Town mayor Elizabeth Hodgkin said: " We are the home of rowing ... I feel very excited about it."
- *5 sentences are abbreviated here.*
12. The Henley-on-Thames postbox was painted on Friday.
- *one sentence is abbreviated here.*



Bring related sentences closer to each other in a structure

Source

1. The town is home to the prestigious Leander Club, which has trained more than 100 Olympic medal-winning rowers.
- *2 sentences are abbreviated here.*
4. The Royal Mail has painted more than 50 postboxes gold following Team GB's gold medal haul at London 2012.
5. Originally it said it was only painting them in winners home towns, or towns with which they are closely associated.
6. Town mayor Elizabeth Hodgkin said: " We are the home of rowing ... [...]"
- *5 sentences are abbreviated here.*
12. The Henley-on-Thames postbox was painted on Friday.
- *one sentence is abbreviated here.*

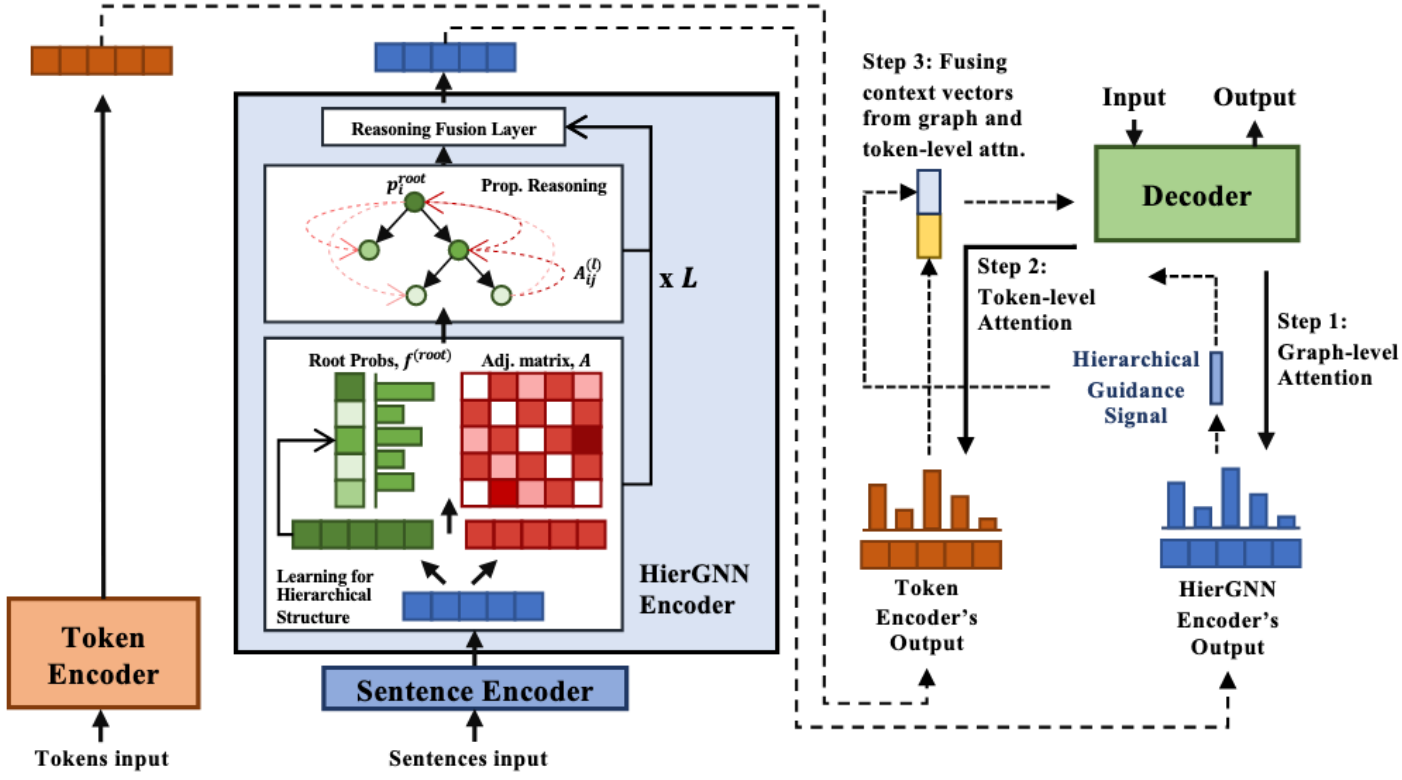
BART Summary

A postbox in Henley-on-Thames has been painted gold as part of the Royal Mail's " Olympic gold " campaign.

Our Summary

A Royal Mail postbox in Henley-on-Thames has been painted gold in honour of the town 's Olympic rowing success.

Abstractive Summarization with Latent Structure



(Qiu and Cohen, EMNLP 2022)

- Propagate information between sentences in the input article
- Use a sparse matrix-tree computation to identify structure

HierGNN - Example

Shortest total lunar eclipse of the century coming Saturday

(CNN) – Sky watchers in western North America are in for a treat: a nearly five-minute total lunar eclipse this morning.

Here's how it's unfolding:

It started at 3:16 a.m. Pacific Daylight Time, when the moon began moving into Earth's shadow. For the next hour and 45 minutes, that shadow will move across the moon and engulf it at 4:58 a.m. Pacific Time.

The total eclipse will only last four minutes and 43 seconds, and NASA says that makes it the shortest one of the century.

HierGNN - Example (CNN Article, Lunar Eclipse)

Each sentence accompanied by *root* probability

Highest Root Probabilities

8th Sent. (9.77%): A lunar eclipse happens when the sun, Earth and moon form a straight line in space, with the Earth smack in the middle.
6th Sent. (9.40%): The sun shines on the Earth and creates a shadow.
10th Sent. (7.79%): Parts of South America, India, China and Russia also will be able to see the eclipse, but it won't be visible in Greenland, Iceland, Europe, Africa or the Middle East.

Lowest Root Probabilities

20th Sent. (Sparsified): Share your photos with CNN iReport.
18th Sent. (Sparsified): If you want to learn more about the eclipse, NASA astronomer Mitzi Adams will take questions on Twitter NASA_Marshall.
19th Sent. (0.02%): Did you see the total lunar eclipse?

HierGNN - Example

Structure helps!

Predicted Summary

A total lunar eclipse started at 3:16 a.m. Pacific Daylight Time. People west of the Mississippi River will have the best view. Parts of South America, India, China and Russia also will be able to see the eclipse. The total eclipse will only last four minutes and 43 seconds

Reference

The total eclipse will only last 4 minutes and 43 seconds. People west of the Mississippi River will have the best view. Parts of South America, India, China and Russia also will see the eclipse.

Example Results of HierGNN

Non-pretrained	**R-1**	**R-2**	**R-L**	**BS**
LEAD-3	40.34	17.70	36.57	-
PGN	39.53	17.28	36.38	-
StructSum ES	39.63	16.98	36.72	-
StructSum LS	39.52	16.94	36.71	-
StructSum (LS + ES)	39.62	17.00	**36.95**	21.70
PGN - Ours	39.07	16.97	35.87	23.74
HierGNN-PGN (LSR)	**39.87**	**17.77**	36.85	**25.64**
HierGNN-PGN (LIR)	39.34	17.39	36.44	25.26
Pretrained	**R-1**	**R-2**	**R-L**	**BS**
BERTSUMABS	41.72	19.39	38.76	29.05
BERTSUMEXTABS	42.13	19.60	39.18	28.72
T5-Large	42.50	20.68	39.75	-
BART	44.16	21.28	40.90	-
Hie-BART	44.35	21.37	41.05	-
HAT-BART	44.48	21.31	41.52	-
BART - Ours	44.62	21.49	41.34	33.98
BART + SentTrans.	44.44	21.44	41.27	33.90
HierGNN-BART (LSR)	44.93	21.7	41.71	34.43
HierGNN-BART (LIR)	**45.04**	**21.82**	**41.82**	**34.59**

HierGNN - Example (Tennis)

HierGNN with Graph Attention

Novak Djokovic beat John Isner in straight sets to reach the Miami Open. The No.1- seeded Djokovic closed to within one win of his fifth Key Biscayne title. Djokovic will be two-time champion andy Murray, who defeated Tomas Berdych 6-4.

HierGNN without Graph Attention

Novak Djokovic beat John Isner in straight sets to reach the final of the Miami Open on Friday night. Djokovic achieved a breakthrough service break against Isner and won Friday night, 7-6 (3), 6-2. His opponent Andy Murray defeated Tomas Berdych 6-4, 6-4.

Reference

Novak Djokovic beat John Isner 7-6. The world No. 1 will take on Andy Murray in Sunday's Final. Djokovic is bidding to win his fifth title at Key Biscayne.

But Wait (Eclipse Example Again)...

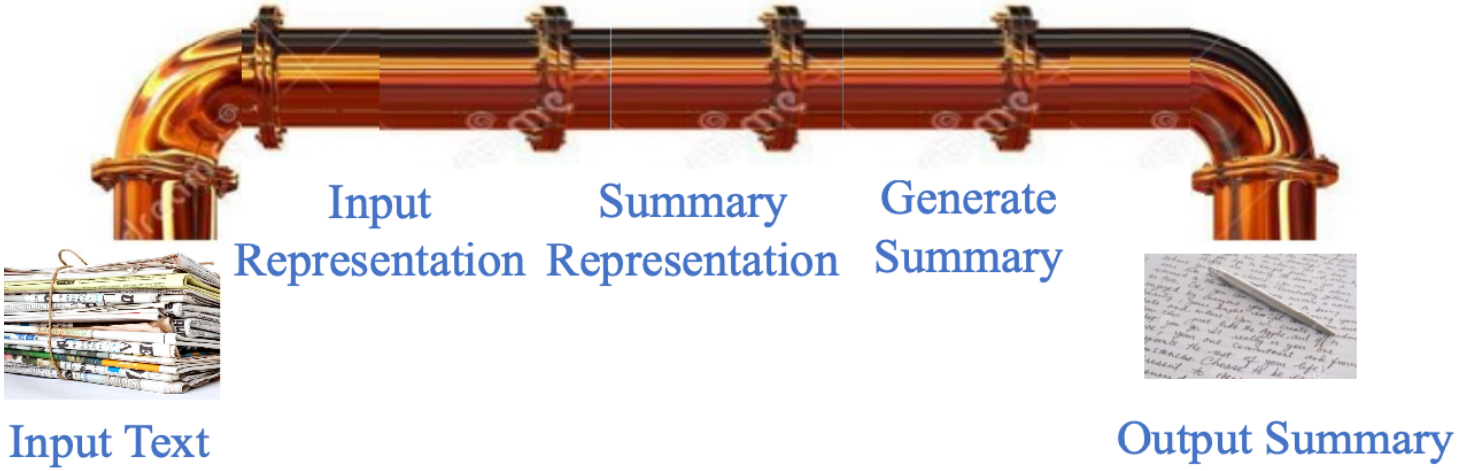
Recall: Lowest Root Probabilities

20th Sent. (Sparsified): Share your photos with CNN iReport.
18th Sent. (Sparsified): *If you want to learn more about the eclipse, NASA astronomer Mitzi Adams will take questions on Twitter NASA_Marshall.*
19th Sent. (0.02%): Did you see the total lunar eclipse?

- What if our intention is getting a summary is to find out how to learn more about the eclipse?

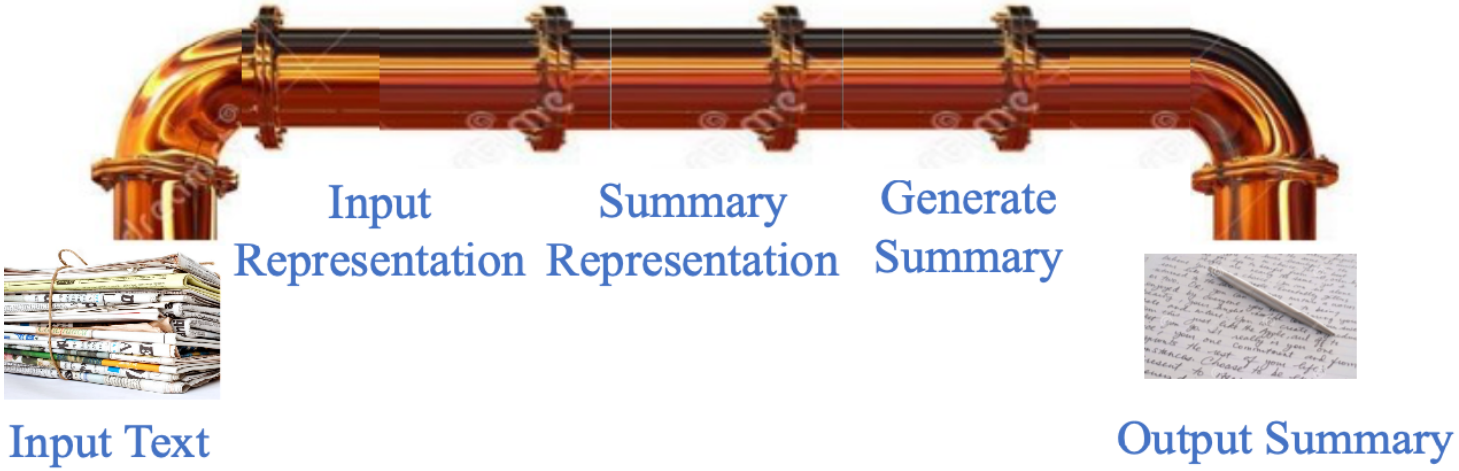
General Formulation of Summarization

Interpret, Transform, Generate



General Formulation of Summarization

Interpret, Transform, Generate

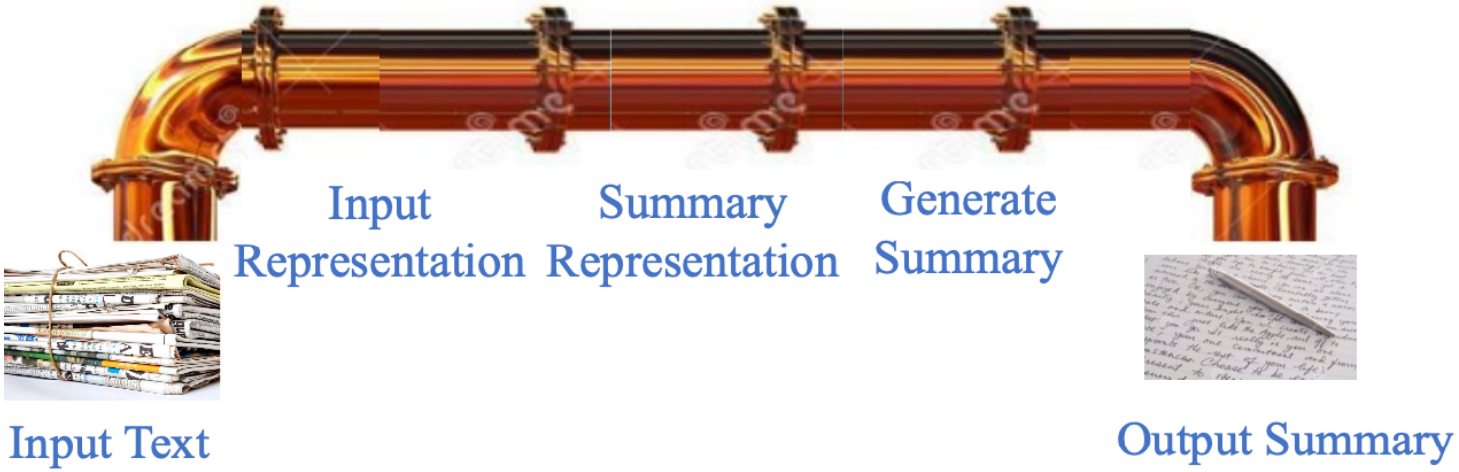


But ...

- Stiennon et al. 2020 report summary quality agreement between 65% and 80% between researchers
- "... after six weeks, subjects selected on average only 55% of the sentences themselves selected" (Teufel and Moens 2002)
- Task of summarization is underconstrained and "too ambiguous to be solved by end-to-end models." (Kryscinski et al. 2019)

General Formulation of Summarization

Interpret, Transform, Generate

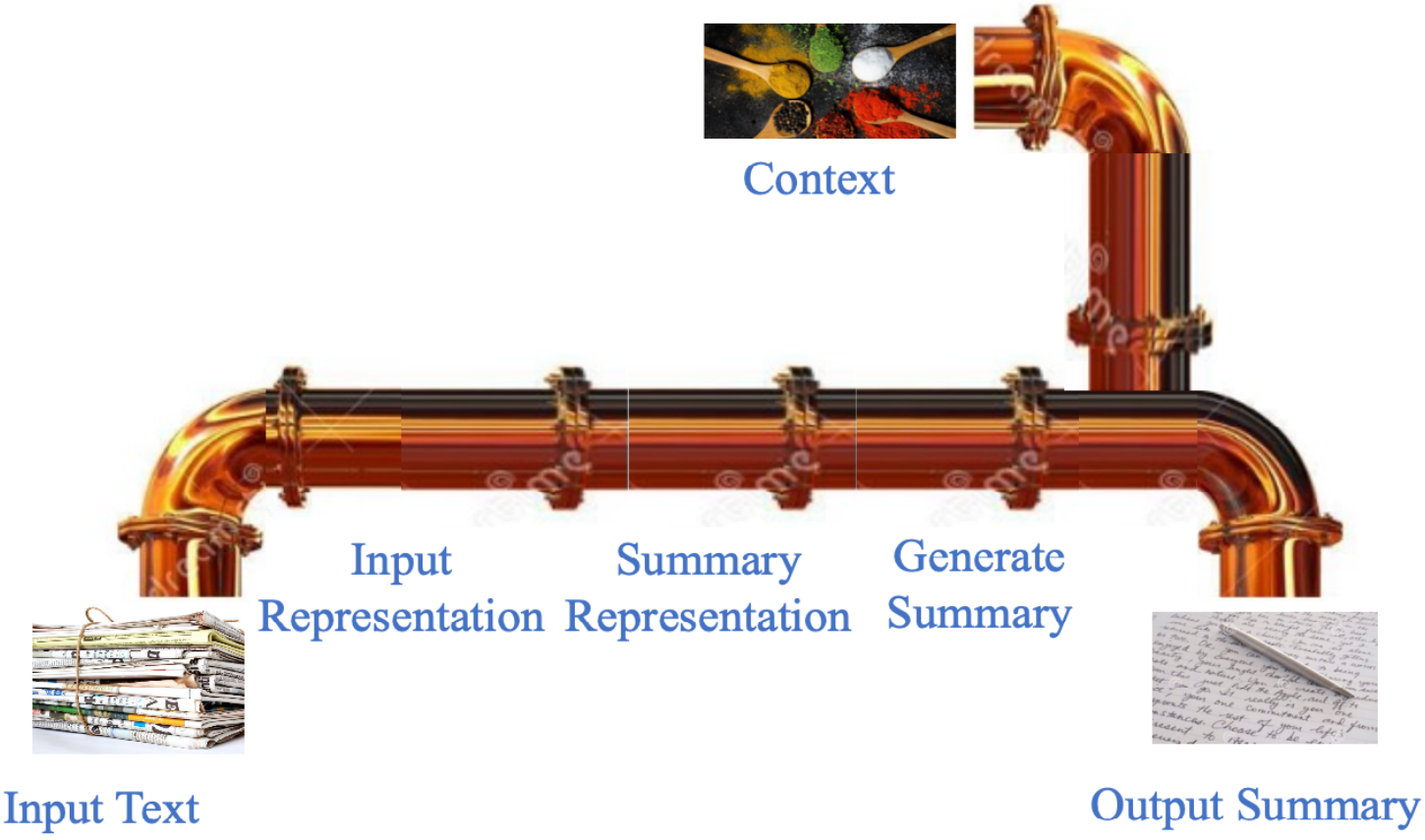


But ...

- Stiennon et al. 2020 report summary quality agreement between 65% and 80% between researchers
- "... after six weeks, subjects selected on average only 55% of the sentences themselves selected" (Teufel and Moens 2002)
- Task of summarization is underconstrained and "too ambiguous to be solved by end-to-end models." (Kryscinski et al. 2019)

Context matters! And sometimes context is latent.

Factorized Summarization (Fonseca et al., 2022)



How to Incorporate Context Factors?

An energy formulation (**FactorSum**):

$$S^* = \arg\min_S E_{\text{int}}(\theta, D, S) + E_{\text{ext}}(S, C, b)$$

where

- θ are learnable parameters (optimized during **learning**)
- S is a candidate summary (optimized during **inference**)
- D is the input document
- C is a (textual) content guidance
- b is a budget guidance

Context Factors

We experimented with a very simple context factor: budget

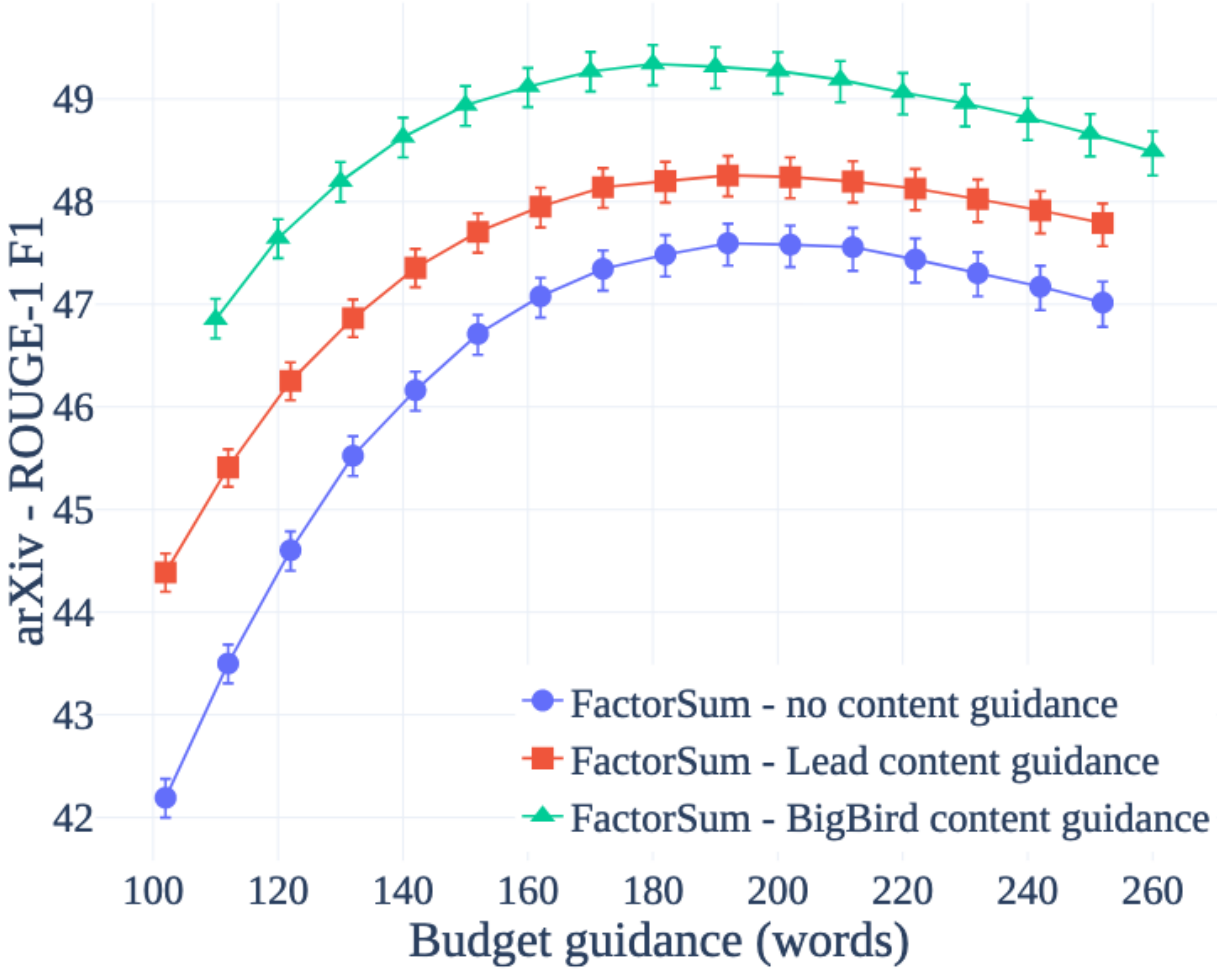
- The extrinsic guidance provides clues on what the output summary length should be

- **Future work:** More complex guidance (conceptual guidance)

Results - In-Domain and Cross-Domain

Training \ Evaluation	PubMed			arXiv			GovReport		
	R-2	R-L	Len	R-2	R-L	Len	R-2	R-L	Len
End-to-end baseline (BigBird for PubMed and arXiv; BART-large for GovReport)									
PubMed	19.92	41.81	185	15.16	38.09	161	3.57	18.10	222
arXiv	14.95	35.77	177	18.60	41.46	164	2.25	15.09	352
GovReport	11.10	33.96	203	8.94	31.67	203	19.12	49.99	596
FactorSum - fixed budget, no content guidance									
PubMed	18.66	41.63	206	15.88	40.16	165	15.07	39.92	350
arXiv	16.87	40.51	209	18.60	42.61	165	**18.07**	**45.94**	414
GovReport	12.63	35.37	213	10.74	33.72	167	23.99	55.19	650
FactorSum - fixed budget and content guidance (BigBird guidance for PubMed and arXiv; BART-large guidance for GovReport)									
PubMed	20.33	43.76	205	**17.13**	**41.86**	166	15.03	39.68	344
arXiv	**18.10**	**42.02**	210	20.27	44.76	165	18.03	45.85	410
GovReport	14.01	37.10	211	11.85	35.47	176	25.28	56.65	648

Guidance Effect



Similar trends for PubMed and GovReport

Example from FactorSum

Reference Summary (Domain: NLP)

the software development life cycle (sdlc) starts with eliciting requirements of the customers in the form of software requirement specification (srs) . srs document needed for software development is mostly written in natural language(nl) convenient for the client . from the srs document only , the class name , its attributes and the functions incorporated in the body of the class are traced based on pre - knowledge of analyst . the paper intends to present a review on object oriented (oo) analysis using natural language processing (nlp) techniques . this analysis can be manual where domain expert helps to generate the required diagram or automated system , where the system generates the required digram , from the input in the form of srs .

Summary with Cross Domain Training (Legal Documents)

A . Constitution of India, Articles 32 and 226 - Engaged in Ph . in manual process , the domain expert analyzes text to bring out ' intermediate output' then automated process considers the intermediate output to generate the desired output - The object of pl D. work in computer research and development - Working in field of semantics Engaging in semantics is a matter within domain of expert bodies - The use of formal languages like unified modeling language (uml) have been applied to avoid the inherent problems of natural language such as incompleteness and ambiguity - The field of Engg. A . Constitution of India, Articles 32 and 226 - Object-oriented approach to problem analysis in the field of software development - Object oriented approach has to be preferred over conventional approach - The object oriented approach should be preferred to the left out approach . [Paras 11 and 12] B.

FactorSum (and language models in general) generate imprecise text

Abstractive Summaries are Often Wrong

Article from XSum

George HW Bush: A look back on the former US president's life.
Former US president George HW Bush, whose plane was shot down during World War Two and who went on to become a millionaire by the age of 40, has died.
He passed away at the age of 94, his son George W Bush announced.
Bush Sr served as the 41st US president, between 1989 and 1993; his term defined by the end of the Cold War and the first Iraq war against Saddam Hussein.

System A

former us president george w bush has died at the age of 89.

System B

the former prseident of the united states, george bush, has died at the age of 97.

Precision in Quantitative Information

Why numbers?

- No dispute about numbers when they are wrong
- Numbers are highly important in certain contexts, for example, financial reports
- They are identifiable, and have clear categories (time, currency, counts, ...)

Fact-checking in Journalism

From Wikipedia:

> *Professional fact checkers have generally been hired by newspapers, magazines, and book publishers, probably starting in the early 1920s with the creation of Time magazine in the United States, though they were not originally called "fact-checkers". Fact checkers may be aspiring writers, future editors, or freelancers engaged other projects; others are career professionals.*

How Do We Fix This? (Zhao et al., 2020)

- We develop a *verifier* model that verifies the information in a summary from XSum
- We focus on quantitative data
- Each quantity in the summary is labeled as 0 or 1 - Verified or Unverified

Article	The crash happened at Evanton at about 17:20 on Saturday. The fire service and the air ambulance was sent to the scene. The occupants of all three vehicles were injured, but the extent of their injuries was not known, police said. A spokesman added: "Inquiries are ongoing into this matter and no further witnesses are sought at this time" ...										
Summary	Several	people	have	been	injured	in	a	three-car	collision	on	...
Y **labels**	B-V	O	O	O	O	O	O	B-V	O	O	...
M **labels**	1	0	0	0	0	0	0	1	0	0	...
z **label**	VERIFIED										

Modelling the Verifier: Our System, "Herman"

- **Training:** Take gold-standard summaries and randomly change quantities in them to get negative examples
- Model used is a neural sequence labeller (BiLSTM-CRF; Huang et al. 2015)
- **Testing:** Run the verifier model on the beam of the summarizer seq2seq model, and either use the overall label of the summary (z) or the token-level verification labels (M) and give a score to summary
- Re-rank the summaries based on the verification scores

Article	The crash happened at Evanton at about 17:20 on Saturday. The fire service and the air ambulance was sent to the scene. The occupants of all three vehicles were injured, but the extent of their injuries was not known, police said. A spokesman added: "Inquiries are ongoing into this matter and no further witnesses are sought at this time" ...										
Summary	Several	people	have	been	injured	in	a	three-car	collision	on	...
Y **labels**	B-V	O	O	O	O	O	O	B-V	O	O	...
M **labels**	1	0	0	0	0	0	0	1	0	0	...
z **label**	VERIFIED										

ROUGE Scores with Herman on XSum

	Model	R1-F	R2-F	RL-F	avg-Q
BART	Baseline-shortest	**45.40**	**21.83**	**36.74**	0.69
	Baseline-max-overlap	44.55	20.97	35.38	**0.95**
	Original	44.57	20.96	35.36	0.89
	HERMAN-LOCAL	44.73	21.14	35.68	0.88
	HERMAN-GLOBAL	44.79	21.17	35.80	0.92
BERTSUM	Baseline-shortest	38.71	16.38	31.16	0.62
	Baseline-max-overlap	39.01	16.58	31.24	0.76
	Original	38.86	16.38	31.04	0.65
	HERMAN-LOCAL	38.63	16.12	30.75	0.79
	HERMAN-GLOBAL	**39.06**	**16.65**	**31.36**	**0.81**
TCONVS2S	Baseline-shortest	30.99	11.15	25.32	0.45
	Baseline-max-overlap	31.64	11.29	25.50	0.71
	Original	30.42	10.96	25.66	0.58
	HERMAN-LOCAL	31.43	11.09	25.31	0.75
	HERMAN-GLOBAL	**31.85**	**11.51**	**25.72**	**0.78**

Human Evaluation of Herman

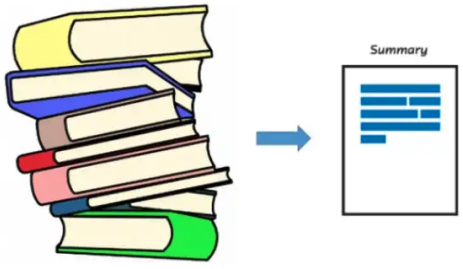
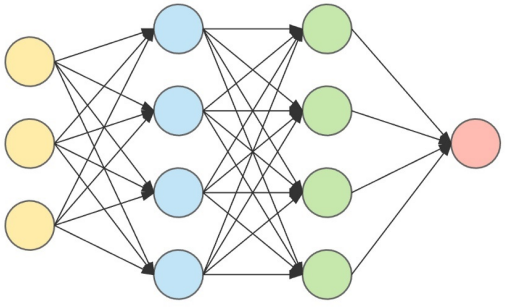
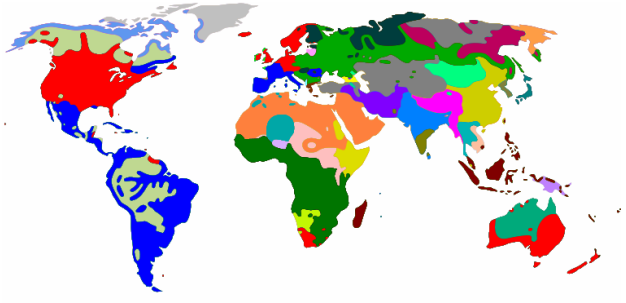
- Simple human evaluation: 40 trials in which the gold-standard summary, the originally predicted summary and up-ranked summary are presented
- We ask (paraphrase): **which is more faithful to the gold-standard summary with respect to quantities: original or up-ranked?**
- Of the 19 trials where all three annotators agreed on the more faithful summary, 12 cases were the up-ranked summary

Natural Language Understanding Pipeline

Natural Language Understanding Pipeline

- To resolve local ambiguities, we probably do not need explicit syntactic parsers (what is the green part? try to guess!)
- LLMs represent progress to a place somewhere between semantics and discourse

Stuff We Do in My Group...

 Text Generation	 Grammar Learning	 Rep. Learning
 Semantics	 Multilinguality	 Applications

Conclusion

- Language is manifested through symbols. Computational systems in general are often symbolic in nature
- Its intermediate representation, however - can be continuous or symbolic
- Symbolic: interpretable; Continuous: have a gradient
- Both have their role. Both can co-exist

Code and Demos

FactorSum:
`http://bollin.inf.ed.ac.uk/factorsum.html`

Collaborators

Yifu Qiu, Marcio Fonseca (some slides based on his), Yftah Ziser, Zheng Zhao, Bonnie Webber