Semi-supervised and Unsupervised Abstractive Summarization

Romain Paulus (rp@you.com)

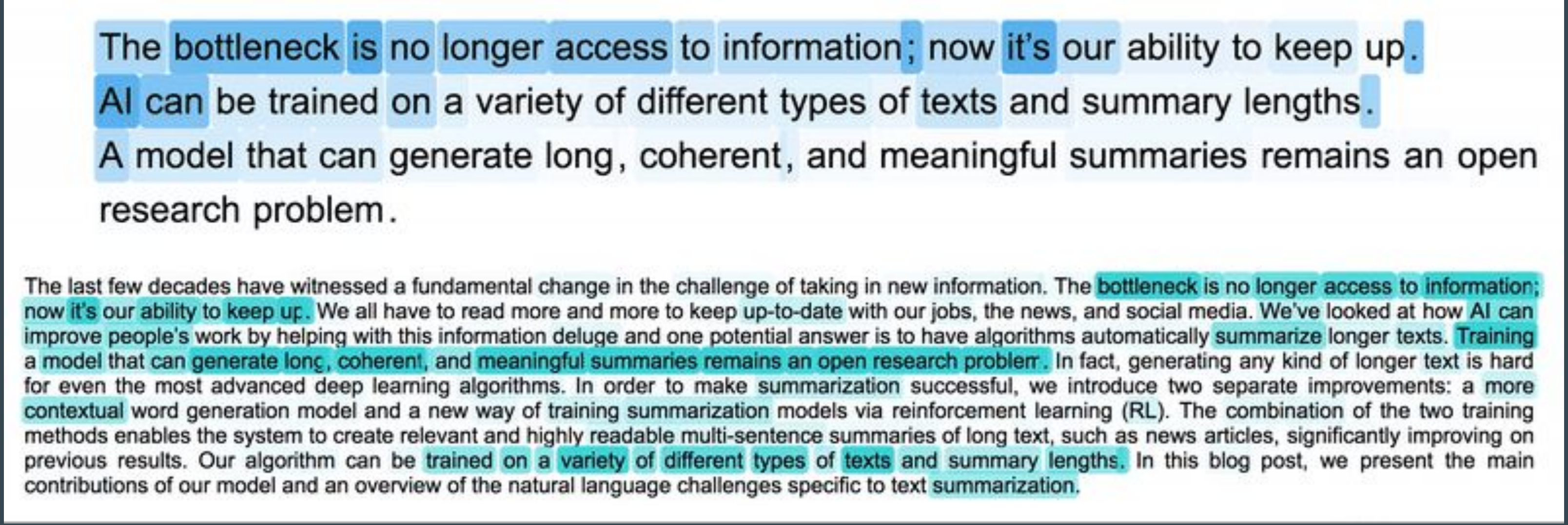
November 20th, 2020

About the speaker

- **Founding engineer @ you.com** (2020)

- Lead research scientist @ **Salesforce** (2016-2020)
- Founding engineer @ **MetaMind** (2014-2016)
- M.S. from **ISEP** (Paris, France) (2014)

What is abstractive summarization?



The bottleneck is no longer access to information; now it's our ability to keep up. AI can be trained on a variety of different types of texts and summary lengths. A model that can generate long, coherent, and meaningful summaries remains an open research problem.

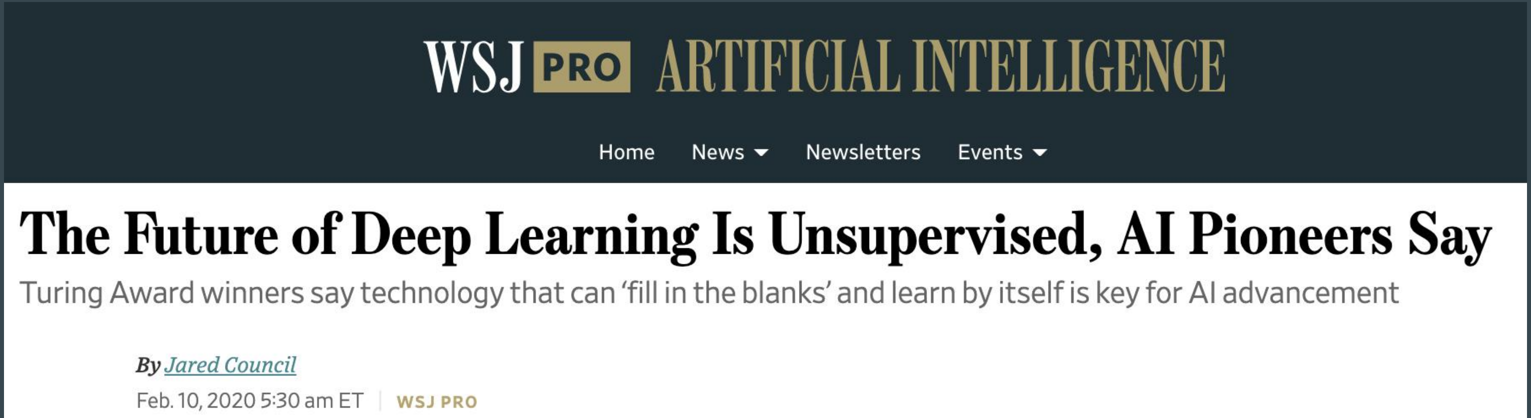
The last few decades have witnessed a fundamental change in the challenge of taking in new information. The bottleneck is no longer access to information; now it's our ability to keep up. We all have to read more and more to keep up-to-date with our jobs, the news, and social media. We've looked at how AI can improve people's work by helping with this information deluge and one potential answer is to have algorithms automatically summarize longer texts. Training a model that can generate long, coherent, and meaningful summaries remains an open research problem. In fact, generating any kind of longer text is hard for even the most advanced deep learning algorithms. In order to make summarization successful, we introduce two separate improvements: a more contextual word generation model and a new way of training summarization models via reinforcement learning (RL). The combination of the two training methods enables the system to create relevant and highly readable multi-sentence summaries of long text, such as news articles, significantly improving on previous results. Our algorithm can be trained on a variety of different types of texts and summary lengths. In this blog post, we present the main contributions of our model and an overview of the natural language challenges specific to text summarization.

“A Deep Reinforced Model for Abstractive Summarization” (Paulus et al. 2017)

Why supervised summarization isn’t enough

- Lack of human-written annotations for most domains
- Limited domain transfer capabilities
- “Ground-truth” is elusive

About unsupervised learning



The Future of Deep Learning Is Unsupervised, AI Pioneers Say

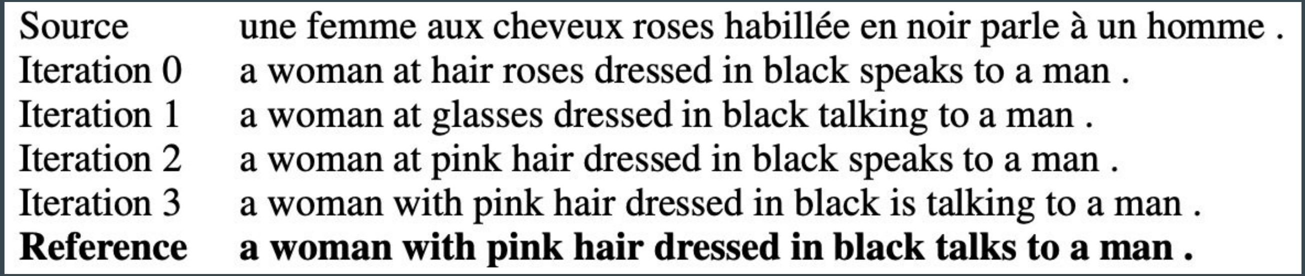
Turing Award winners say technology that can 'fill in the blanks' and learn by itself is key for AI advancement

By Jared Council

Feb. 10, 2020 5:30 am ET | WSJ PRO

Other applications of unsupervised learning in NLP

- Language models (i.e. **GPT-3, BERT**)
- Representation learning (i.e. auto-encoders)
- Unsupervised machine translation



Source	une femme aux cheveux roses habillée en noir parle à un homme .
Iteration 0	a woman at hair roses dressed in black speaks to a man .
Iteration 1	a woman at glasses dressed in black talking to a man .
Iteration 2	a woman at pink hair dressed in black speaks to a man .
Iteration 3	a woman with pink hair dressed in black is talking to a man .
Reference	**a woman with pink hair dressed in black talks to a man .**

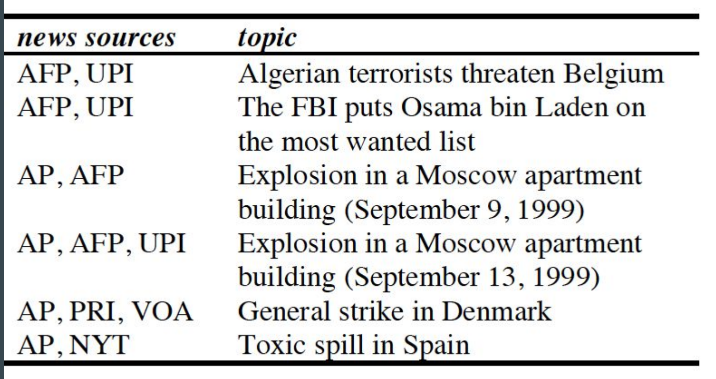
“Unsupervised Machine Translation Using Monolingual Corpora Only” (Lample et al. 2018)

Early attempts at unsupervised summarization

…based on TF-IDF

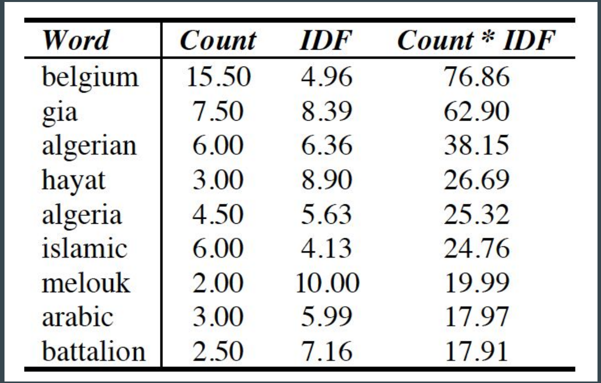
Early attempts at unsupervised summarization

Input documents (on a given topic)



news sources	*topic*
AFP, UPI	Algerian terrorists threaten Belgium
AFP, UPI	The FBI puts Osama bin Laden on the most wanted list
AP, AFP	Explosion in a Moscow apartment building (September 9, 1999)
AP, AFP, UPI	Explosion in a Moscow apartment building (September 13, 1999)
AP, PRI, VOA	General strike in Denmark
AP, NYT	Toxic spill in Spain

TF-IDF “centroid”



Word	*Count*	*IDF*	*Count * IDF*
belgium	15.50	4.96	76.86
gia	7.50	8.39	62.90
algerian	6.00	6.36	38.15
hayat	3.00	8.90	26.69
algeria	4.50	5.63	25.32
islamic	6.00	4.13	24.76
melouk	2.00	10.00	19.99
arabic	3.00	5.99	17.97
battalion	2.50	7.16	17.91

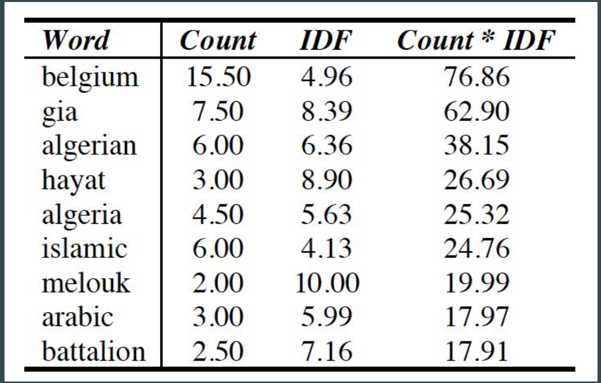
“Centroid-Based Summarization of Multiple Documents” (Radev et al. 2004)

Centroid-based summarization method

1. **One by one**, extract sentences that are **most similar to the centroid**

2. … but also **don’t pick sentences that are too similar** to the ones already picked
   (cross-sentence informational subsumption)

TF-IDF “centroid”



Word	*Count*	*IDF*	*Count * IDF*
belgium	15.50	4.96	76.86
gia	7.50	8.39	62.90
algerian	6.00	6.36	38.15
hayat	3.00	8.90	26.69
algeria	4.50	5.63	25.32
islamic	6.00	4.13	24.76
melouk	2.00	10.00	19.99
arabic	3.00	5.99	17.97
battalion	2.50	7.16	17.91

“Centroid-Based Summarization of Multiple Documents” (Radev et al. 2004)

Limitations

- **Extractive** summarization only
- M**ulti-document summarization** only
- Doesn’t use **word embeddings**

“Centroid-Based Summarization of Multiple Documents” (Radev et al. 2004)

Deep learning approaches

…for abstractive unsupervised summarization!

Common trick for deep abs. unsupervised summ.

Find structures in the input data

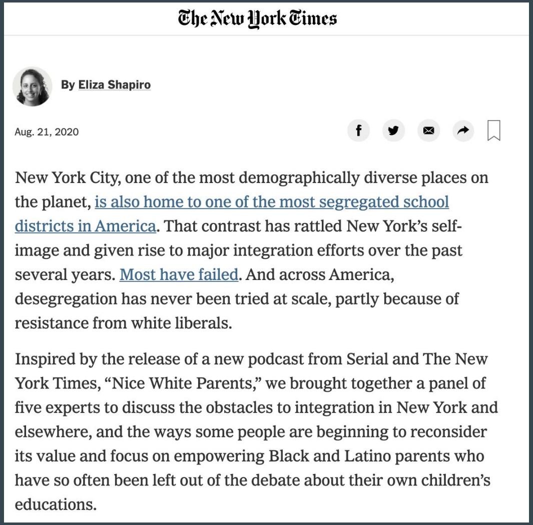
Common trick for deep abs. unsupervised summ.

Find structures in the input data

- Redundancy/implicit structures
- Domain-specific assumptions
- Use some information theory

Simple unsupervised (pre-)training

Input document



The New York Times

By Eliza Shapiro

Aug. 21, 2020

New York City, one of the most demographically diverse places on the planet, is also home to one of the most segregated school districts in America. That contrast has rattled New York's self-image and given rise to major integration efforts over the past several years. Most have failed. And across America, desegregation has never been tried at scale, partly because of resistance from white liberals.

Inspired by the release of a new podcast from Serial and The New York Times, "Nice White Parents," we brought together a panel of five experts to discuss the obstacles to integration in New York and elsewhere, and the ways some people are beginning to reconsider its value and focus on empowering Black and Latino parents who have so often been left out of the debate about their own children's educations.

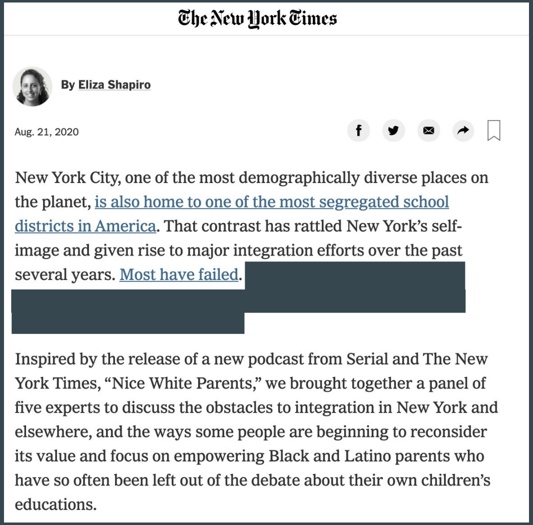
SUMMARIZATION MODEL →

Target summary

?

Take out a random* sentence from input document

Input document



The New York Times

By Eliza Shapiro

Aug. 21, 2020

New York City, one of the most demographically diverse places on the planet, is also home to one of the most segregated school districts in America. That contrast has rattled New York's self-image and given rise to major integration efforts over the past several years. Most have failed. [redacted]

Inspired by the release of a new podcast from Serial and The New York Times, "Nice White Parents," we brought together a panel of five experts to discuss the obstacles to integration in New York and elsewhere, and the ways some people are beginning to reconsider its value and focus on empowering Black and Latino parents who have so often been left out of the debate about their own children's educations.

SUMMARIZATION MODEL

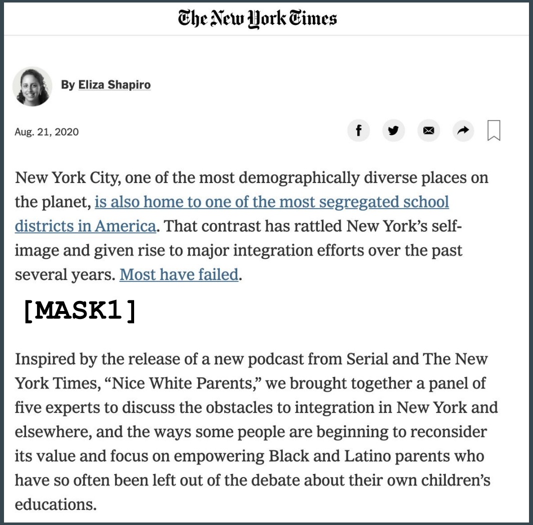
Target summary

And across America, desegregation has never been tried at scale, partly because of resistance from white liberals.

"PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization" (Zhang et al. 2019)

Replace with special mask token in input

Input document



The New York Times

By Eliza Shapiro

Aug. 21, 2020

New York City, one of the most demographically diverse places on the planet, is also home to one of the most segregated school districts in America. That contrast has rattled New York's self-image and given rise to major integration efforts over the past several years. Most have failed.

[MASK1]

Inspired by the release of a new podcast from Serial and The New York Times, "Nice White Parents," we brought together a panel of five experts to discuss the obstacles to integration in New York and elsewhere, and the ways some people are beginning to reconsider its value and focus on empowering Black and Latino parents who have so often been left out of the debate about their own children's educations.

SUMMARIZATION MODEL →

Target summary

And across America, desegregation has never been tried at scale, partly because of resistance from white liberals.

"PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization" (Zhang et al. 2019)

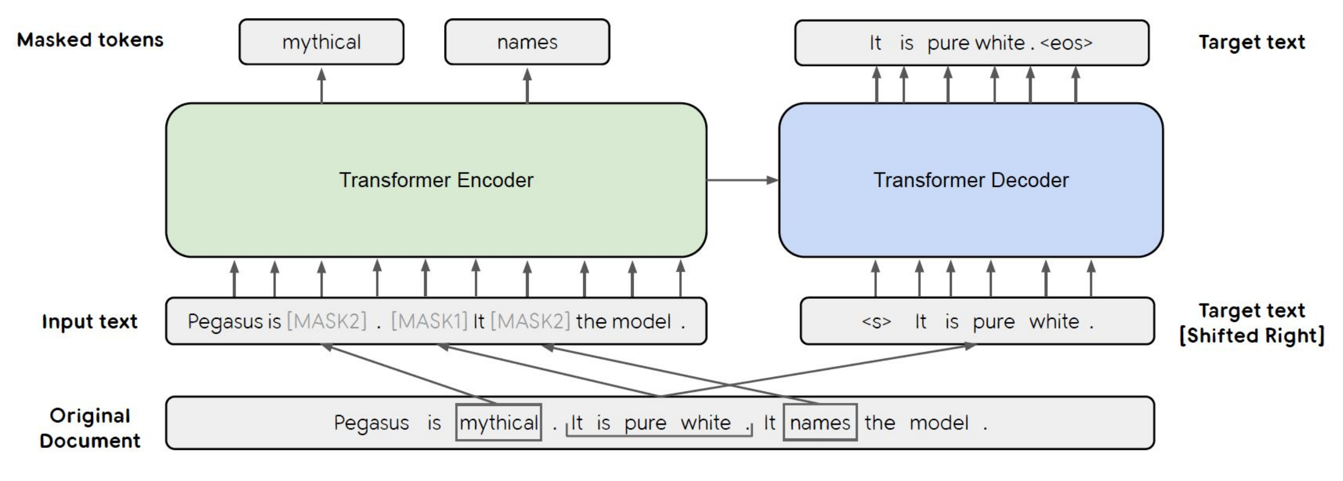
Full PEGASUS model

- **Gap-Sentence Generation (GSG)**
- Masked language model (MLM) (like BERT)

"PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization" (Zhang et al. 2019)

Full PEGASUS model

- **Gap-Sentence Generation (GSG)**
- Masked language model (MLM) (like BERT)



“PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization” (Zhang et al. 2019)

Applications of PEGASUS

- Applicable to **any data domain**
- Used as a **pre-training** self-supervised method, can still be fine-tuned

“PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization” (Zhang et al. 2019)

Can we refine these ideas for specific domains?

… especially **news articles**?

What’s so special about news datasets?

… lead bias!

What’s so special about news datasets?

… lead bias!

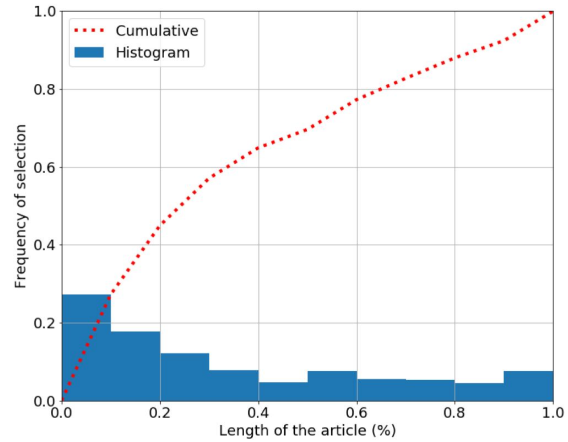


Figure 1: The distribution of important sentences over the length of the article according to human annotators (blue) and its cumulative distribution (red).

(On CNN/Daily Mail dataset)

“Neural Text Summarization: A Critical Evaluation” (Kryściński et al. 2019)

What’s so special about news datasets?

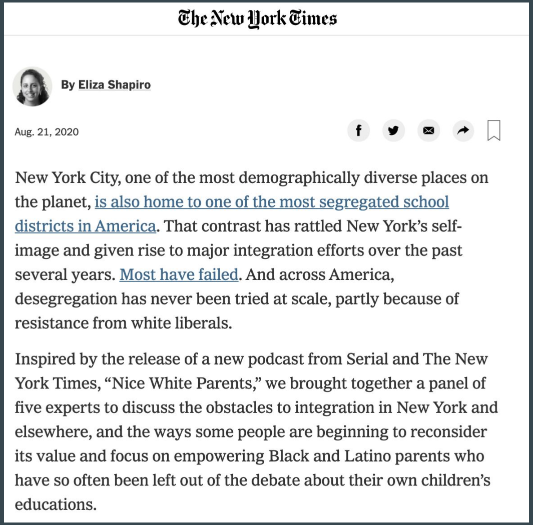
… lead bias!

Using the first 3 sentences in a **news article** as a summary is still a hard baseline to beat for abstractive summarization models

“Get To The Point: Summarization with Pointer-Generator Networks” (See et al. 2017)

Pre-training abstractive summarization with lead bias

Input document



The New York Times

By Eliza Shapiro

Aug. 21, 2020

New York City, one of the most demographically diverse places on the planet, is also home to one of the most segregated school districts in America. That contrast has rattled New York's self-image and given rise to major integration efforts over the past several years. Most have failed. And across America, desegregation has never been tried at scale, partly because of resistance from white liberals.

Inspired by the release of a new podcast from Serial and The New York Times, "Nice White Parents," we brought together a panel of five experts to discuss the obstacles to integration in New York and elsewhere, and the ways some people are beginning to reconsider its value and focus on empowering Black and Latino parents who have so often been left out of the debate about their own children's educations.

SUMMARIZATION MODEL

→

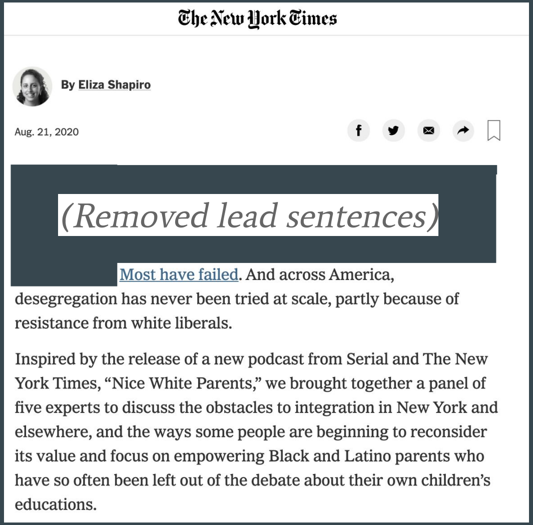
Target summary

?

"TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising" (Yang et al. 2020)

Pre-training abstractive summarization with lead bias

Input document



The New York Times

By Eliza Shapiro

Aug. 21, 2020

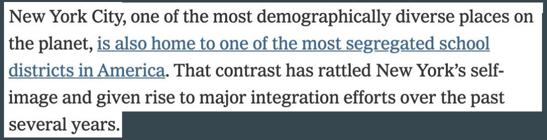
(Removed lead sentences)

Most have failed. And across America, desegregation has never been tried at scale, partly because of resistance from white liberals.

Inspired by the release of a new podcast from Serial and The New York Times, “Nice White Parents,” we brought together a panel of five experts to discuss the obstacles to integration in New York and elsewhere, and the ways some people are beginning to reconsider its value and focus on empowering Black and Latino parents who have so often been left out of the debate about their own children’s educations.

SUMMARIZATION MODEL

Target summary

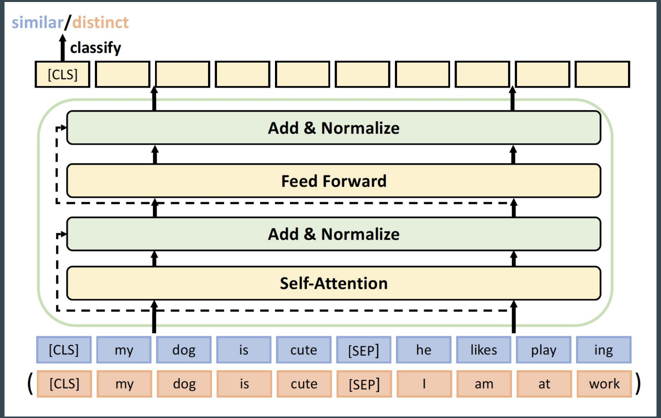


New York City, one of the most demographically diverse places on the planet, is also home to one of the most segregated school districts in America. That contrast has rattled New York’s self-image and given rise to major integration efforts over the past several years.

“TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising” (Yang et al. 2020)

In addition to this pre-training:

- Theme modeling
  - **Additional loss function** making the summ. output **more similar to the input domain**
  - Implemented as a simple **discriminator classifier:**



similar/distinct
classify
[CLS]
Add & Normalize
Feed Forward
Add & Normalize
Self-Attention
[CLS] my dog is cute [SEP] he likes play ing
([CLS] my dog is cute [SEP] I am at work)

Summary sentence "Theme" sentence

"TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising" (Yang et al. 2020)

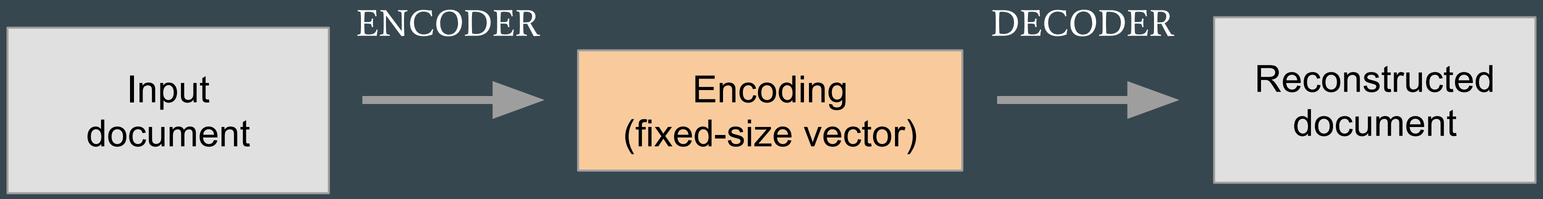
In addition to this pre-training:

- **Denoising autoencoder**
  - Also used in unsupervised machine translation
  - help with **reconstruction** from an imperfect input (shuffled words, etc)

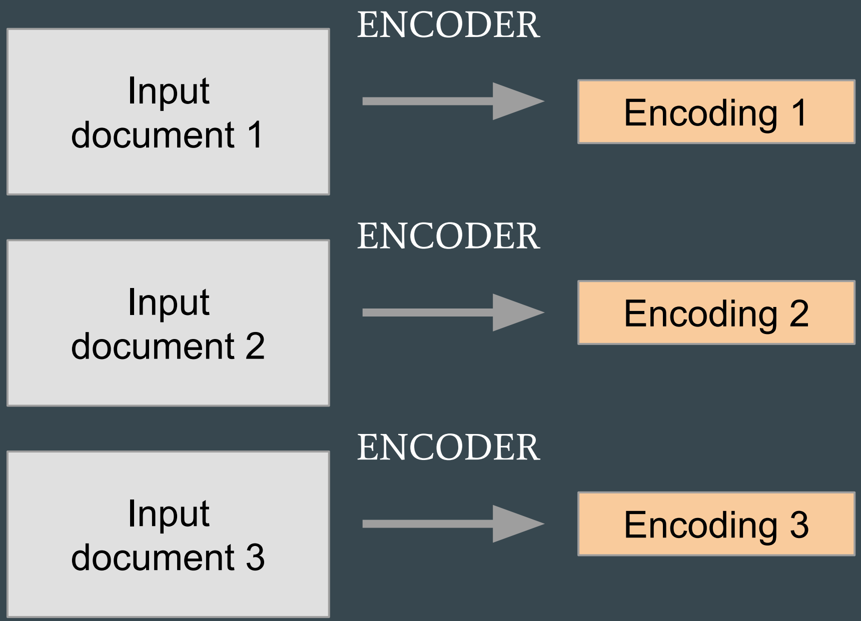
“TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising” (Yang et al. 2020)

Why autoencoders are great

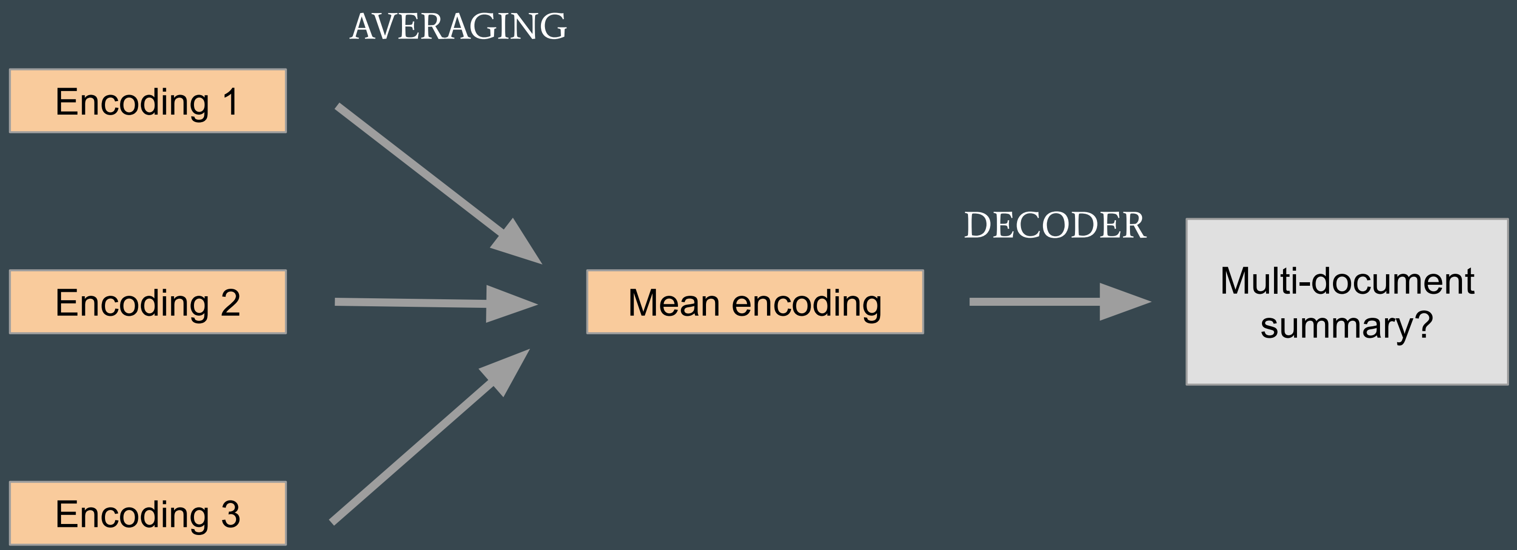
- **Self-supervised** training
- Compressed representation, could reveal some signal



Supercharging autoencoders for multi-doc summarization

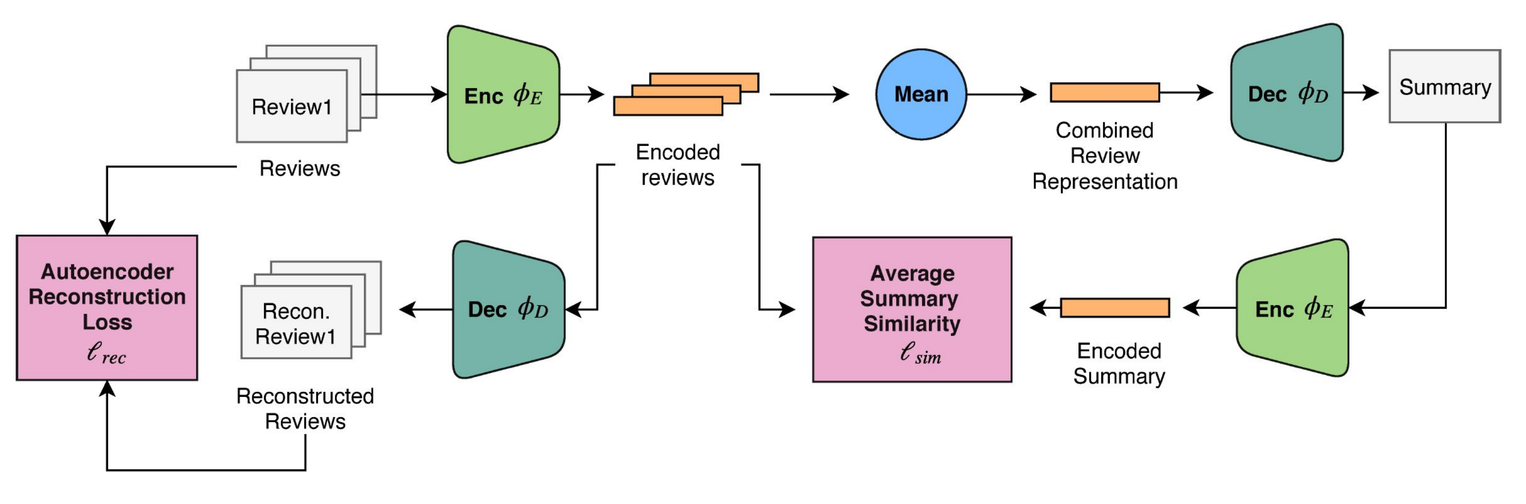


Supercharging autoencoders for multi-doc summarization



“MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization” (Chu et al. 2019)

Multi-document unsupervised abstractive summarization



“MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization” (Chu et al. 2019)

How to evaluate multi-doc summ. without references?

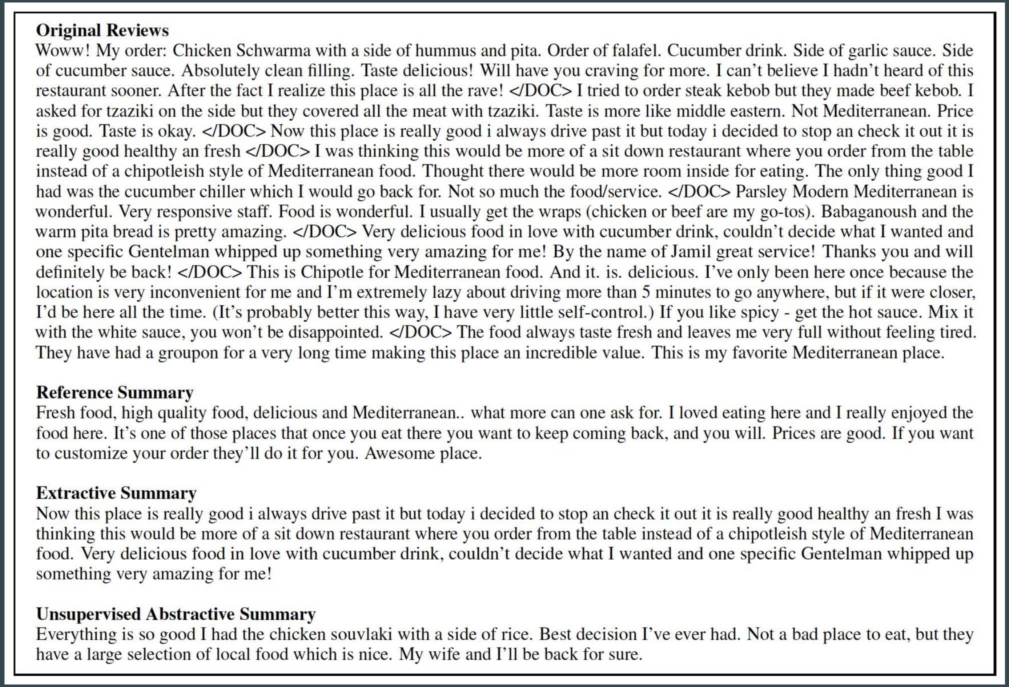
- no ground truth ==> no ROUGE score

Instead:

- Sentiment accuracy (for product reviews)
- Word overlap
- Negative log-likelihood
- Human evaluation of quality

“MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization” (Chu et al. 2019)

Output examples



Original Reviews

Woww! My order: Chicken Schwarma with a side of hummus and pita. Order of falafel. Cucumber drink. Side of garlic sauce. Side of cucumber sauce. Absolutely clean filling. Taste delicious! Will have you craving for more. I can't believe I hadn't heard of this restaurant sooner. After the fact I realize this place is all the rave! </DOC> I tried to order steak kebob but they made beef kebob. I asked for tzaziki on the side but they covered all the meat with tzaziki. Taste is more like middle eastern. Not Mediterranean. Price is good. Taste is okay. </DOC> Now this place is really good i always drive past it but today i decided to stop an check it out it is really good healthy an fresh </DOC> I was thinking this would be more of a sit down restaurant where you order from the table instead of a chipotleish style of Mediterranean food. Thought there would be more room inside for eating. The only thing good I had was the cucumber chiller which I would go back for. Not so much the food/service. </DOC> Parsley Modern Mediterranean is wonderful. Very responsive staff. Food is wonderful. I usually get the wraps (chicken or beef are my go-tos). Babaganoush and the warm pita bread is pretty amazing. </DOC> Very delicious food in love with cucumber drink, couldn't decide what I wanted and one specific Gentelman whipped up something very amazing for me! By the name of Jamil great service! Thanks you and will definitely be back! </DOC> This is Chipotle for Mediterranean food. And it. is. delicious. I've only been here once because the location is very inconvenient for me and I'm extremely lazy about driving more than 5 minutes to go anywhere, but if it were closer, I'd be here all the time. (It's probably better this way, I have very little self-control.) If you like spicy - get the hot sauce. Mix it with the white sauce, you won't be disappointed. </DOC> The food always taste fresh and leaves me very full without feeling tired. They have had a groupon for a very long time making this place an incredible value. This is my favorite Mediterranean place.

Reference Summary

Fresh food, high quality food, delicious and Mediterranean.. what more can one ask for. I loved eating here and I really enjoyed the food here. It's one of those places that once you eat there you want to keep coming back, and you will. Prices are good. If you want to customize your order they'll do it for you. Awesome place.

Extractive Summary

Now this place is really good i always drive past it but today i decided to stop an check it out it is really good healthy an fresh I was thinking this would be more of a sit down restaurant where you order from the table instead of a chipotleish style of Mediterranean food. Very delicious food in love with cucumber drink, couldn't decide what I wanted and one specific Gentelman whipped up something very amazing for me!

Unsupervised Abstractive Summary

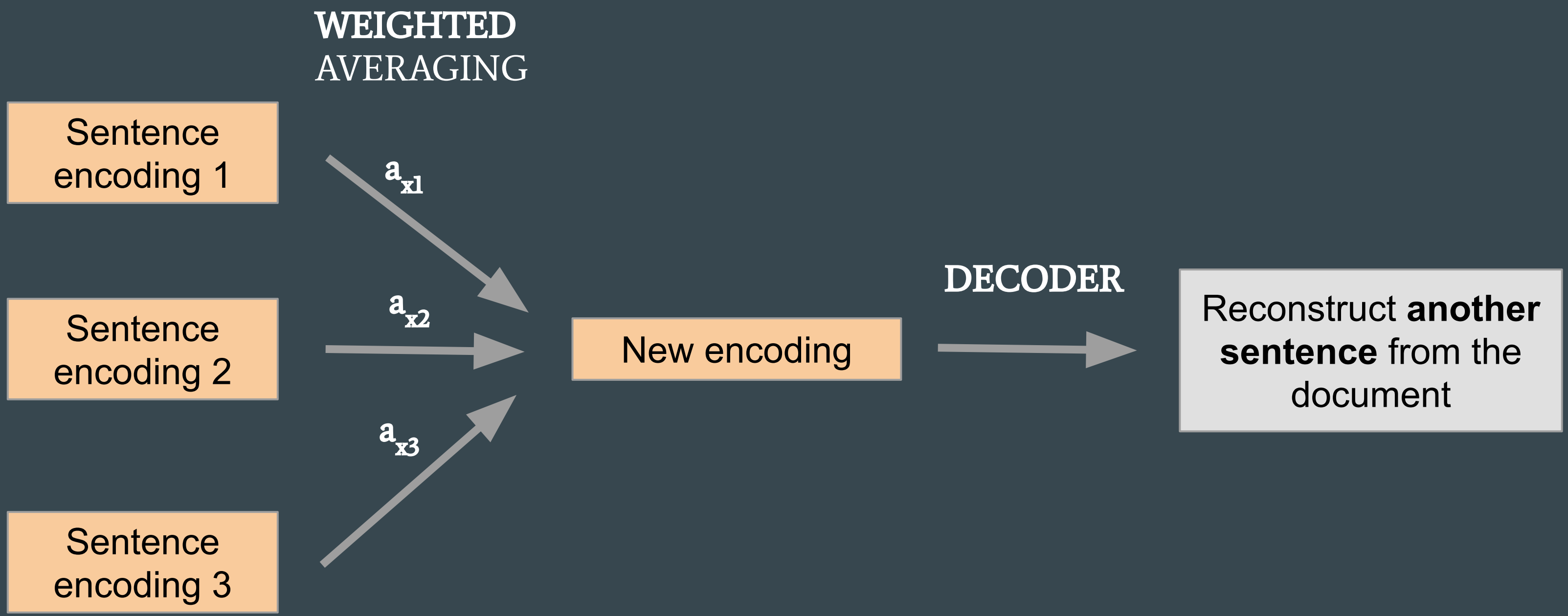
Everything is so good I had the chicken souvlaki with a side of rice. Best decision I've ever had. Not a bad place to eat, but they have a large selection of local food which is nice. My wife and I'll be back for sure.

"MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization" (Chu et al. 2019)

Can we use that for single-doc summarization?

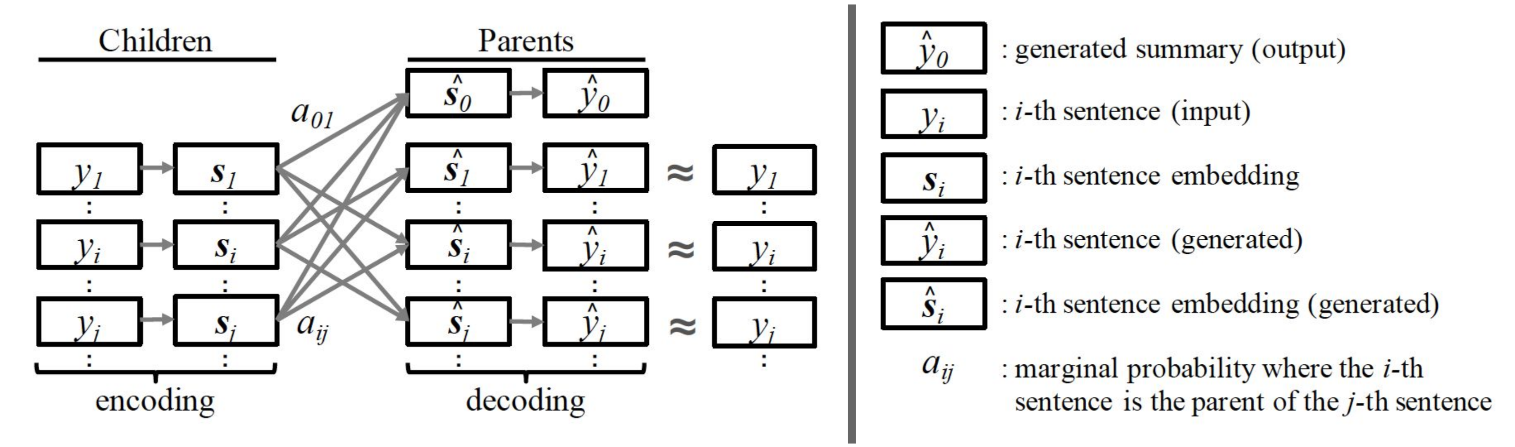
- Idea: train an autoencoder on **individual sentences** of an input document

Autoencoders for single-doc summ.



"Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking" (Isonuma et al. 2019)

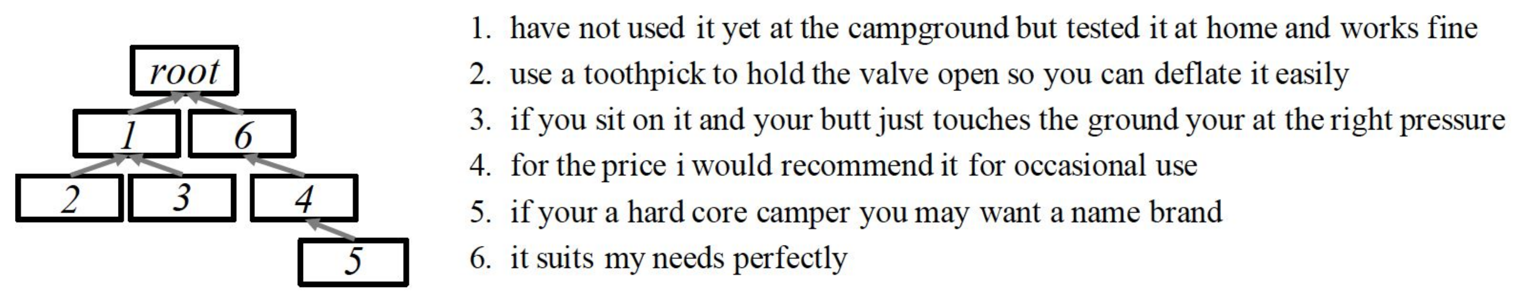
Full multi-sentence autoencoder model



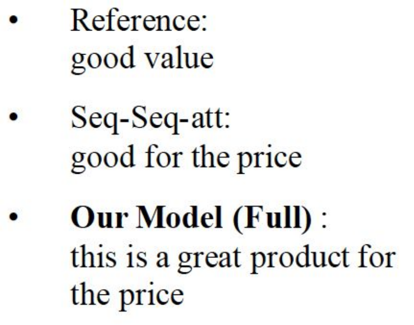
- $\hat{y}_0$: generated summary (output)
- y_i : i-th sentence (input)
- $\mathbf{s}_i$: i-th sentence embedding
- $\hat{y}_i$: i-th sentence (generated)
- $\hat{\mathbf{s}}_i$: i-th sentence embedding (generated)
- a_{ij} : marginal probability where the i-th sentence is the parent of the j-th sentence

“Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking” (Isonuma et al. 2019)

Generated discourse trees and summary



1. have not used it yet at the campground but tested it at home and works fine
2. use a toothpick to hold the valve open so you can deflate it easily
3. if you sit on it and your butt just touches the ground your at the right pressure
4. for the price i would recommend it for occasional use
5. if your a hard core camper you may want a name brand
6. it suits my needs perfectly



- Reference:
  good value
- Seq-Seq-att:
  good for the price
- **Our Model (Full)** :
  this is a great product for the price

“Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking” (Isonuma et al. 2019)

A new breakthrough?

> “In particular, for relatively long reviews, **our model achieved competitive or better performance** compared to supervised models.”

“Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking” (Isonuma et al. 2019)

BottleSum

- Use the **Information Bottleneck (IB)** (Tishby et al., 1999) principle
- Given two pieces of text, **how much of text 1 useful for predicting text 2**?
- Keep **significant details of text 1** only. More elegant than auto-encoders

“Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle” (West et al. 2019)

Formal notion of information relevance

- **S**: input (**sentence 1**)
- **Y**: external relevance variable (= **sentence 2**)
- **S~**: summary (generated)
- **I:** mutual information function (= conditional language model)
- **beta**: positive coefficient

Objective, **minimize**:

$$I(\tilde{S}; S) - \beta I(\tilde{S}; Y)$$

Pruning term

Relevance term

“Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle” (West et al. 2019)

Two ways of applying this method/objective

- **Extractive** (using a **pre-trained LM**)
  - Ensures that S~ is derived from S, even if we minimize I(S, S~)
  - Iteratively delete words or phrases in S to lower the pruning term (= increase **p(S~)**) and keep a high relevance term (p(Y|S~))

"Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle" (West et al. 2019)

Two ways of applying this method/objective

- **Extractive** (using a pre-trained LM)
  - Ensures that S~ is derived from S, even if we minimize I(S, S~)
  - Iteratively delete words of phrases in S to lower the pruning term (= increase **p(S~)**) and keep a high relevance term (p(Y|S~))

Example:

S: “Hong Kong, a bustling metropolis with a population over 7 million, was once under British Rule.”
Y: “The city returned to Chinese control in 1997.”
S~ (output): “**Hong Kong**, ~~a bustling metropolis with a population over 7 million~~, **was once under British Rule.**”

“Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle” (West et al. 2019)

Two ways of applying this method/objective

- **Abstractive (self-supervised)**
  - Use the previous **extractive** model to **generate summaries**
  - **Use these summaries as ground-truth** of an abstractive summarization model (in practice, **fine-tuning GPT-2**)
  - Profit!

“Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle” (West et al. 2019)

Conclusion

- Very promising methods for unsupervised abstractive summarization
- Rely on: Structures, domain knowledge, and information theory

- Next step: **better evaluation methods**

Thank you!

Get in touch: rp@you.com
We’re hiring! su-sea.github.io/jobs/

(order by which I’m gonna talk about the papers)

- PEGASUS (quite influential because it’s so simple for pre-training)
    - Large-scale self-supervised pre-training, following the steps of BERT-stuff (but different transformer model)
- TED (lead bias + more?)
- MeanSum (one of the 1st breakthroughs?)
    - Auto-encoder based
- BottleSum
- Discourse tree (Isonuma et al. 2020) (actually acyclic weighted graph?)
    - some similarities with PEGASUS in taking one of the sentences as target, but bigger) (also ideas from BottleSum too?)
    - Also auto-encoder based, kinda like MeanSum, but at the sentence-level
- Unsupervised Opinion Summarization? (hierarchical VAE this time. Bonus one?)

Overview

- Who am I, what I worked on
- Some of the issues I ran into while I worked on abs summ
- Why this is an important problem bigger than just for me, but for AI in general (give some numbers on scale of summ ground truths, and also qualitative issues w/ multiple ground truths)
- How unsupervised training worked on some related tasks like Translation (give cool examples from FAIR, and TODO see if there's anything new since then?)
- Different approaches:
  - Early approaches: (Radev et al., 2004; Mihalcea and Tarau (TextRank?"), 2004)
  - Pre-training (PEGASUS, TED?)
  - Fully unsupervised (BottleSum,)
  - Data-dependent approaches (TED with lead bias, MeanSum with multi-doc)
  - Other semi-supervision (GASP!, human-in-the-loop openAI)
- TODO: read papers with a red tag 🔴 (on-topic papers) and orange tag □ (slightly related) in my `papers` folder