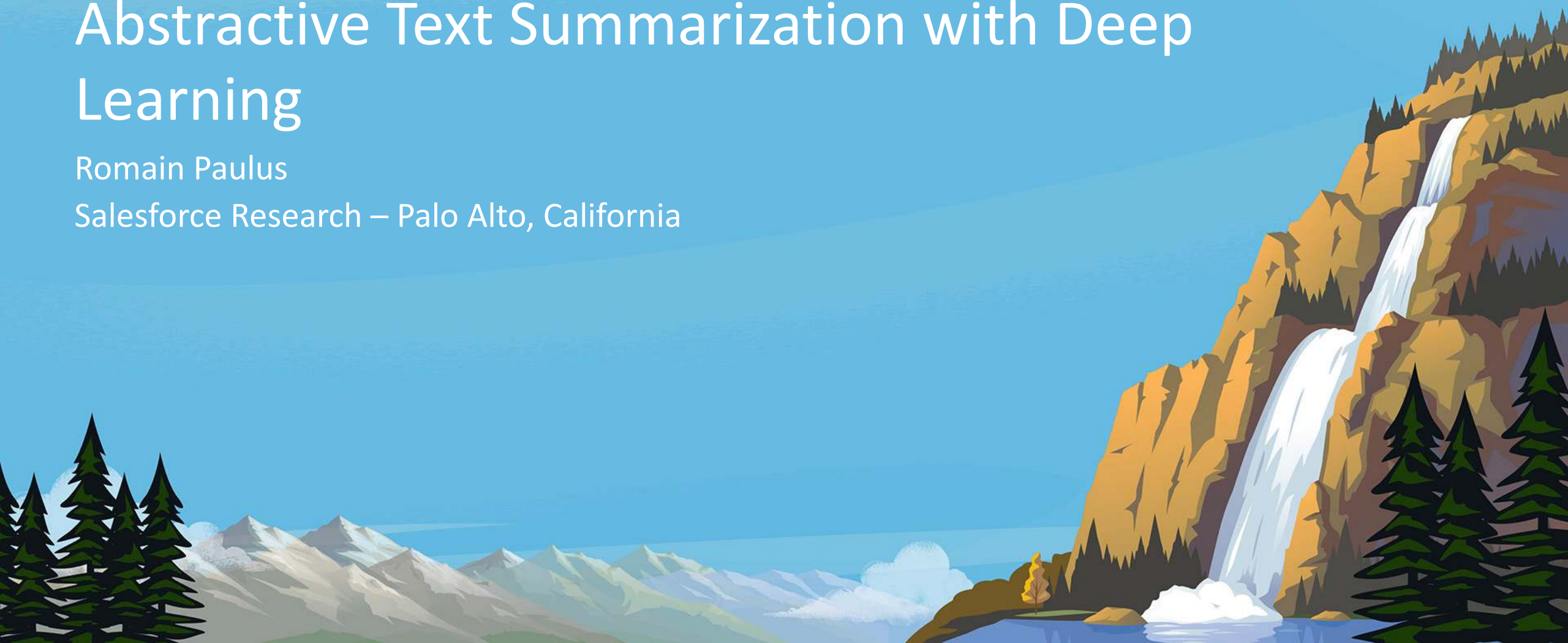
Abstractive Text Summarization with Deep Learning

Romain Paulus
Salesforce Research – Palo Alto, California

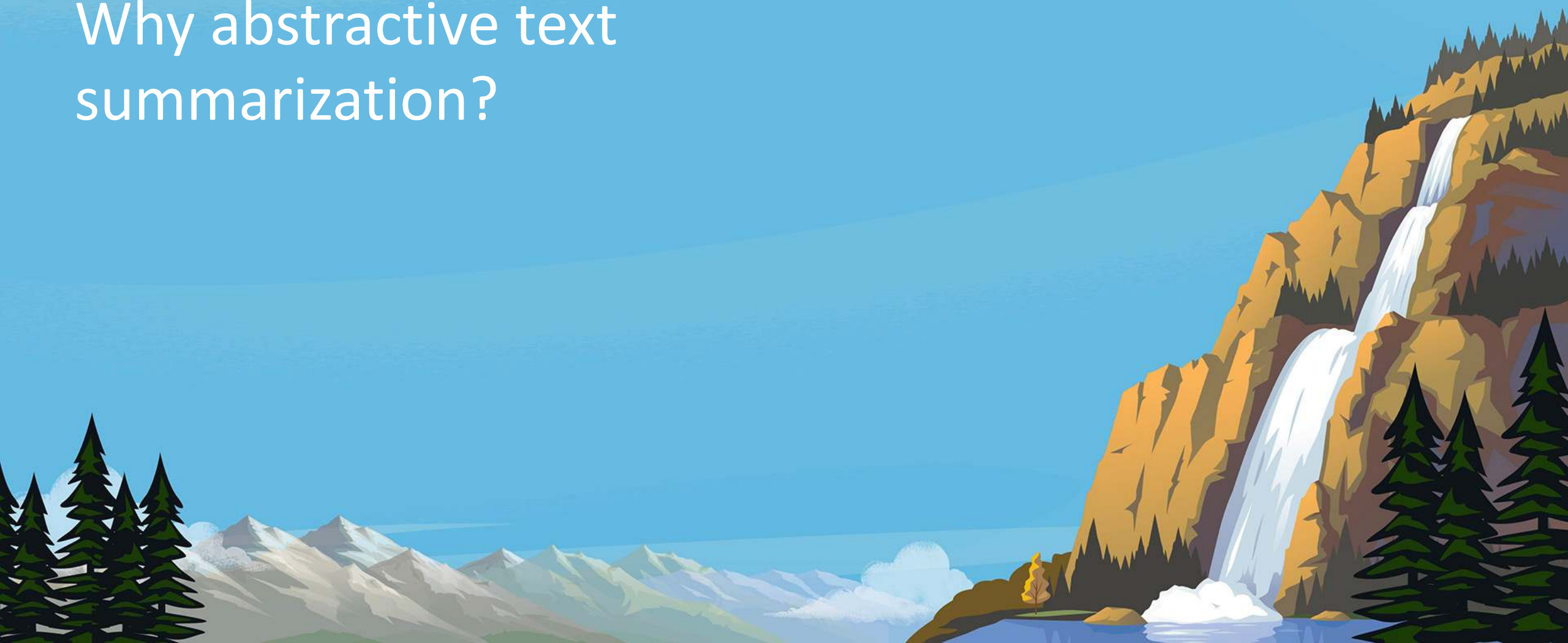


About the speaker

Romain Paulus

- Lead Research Scientist @ **Salesforce Research** (2016-present)
- Previously, founding engineer of **MetaMind** (2014-2016)
- M.Sc. From **ISEP Paris** (2014)

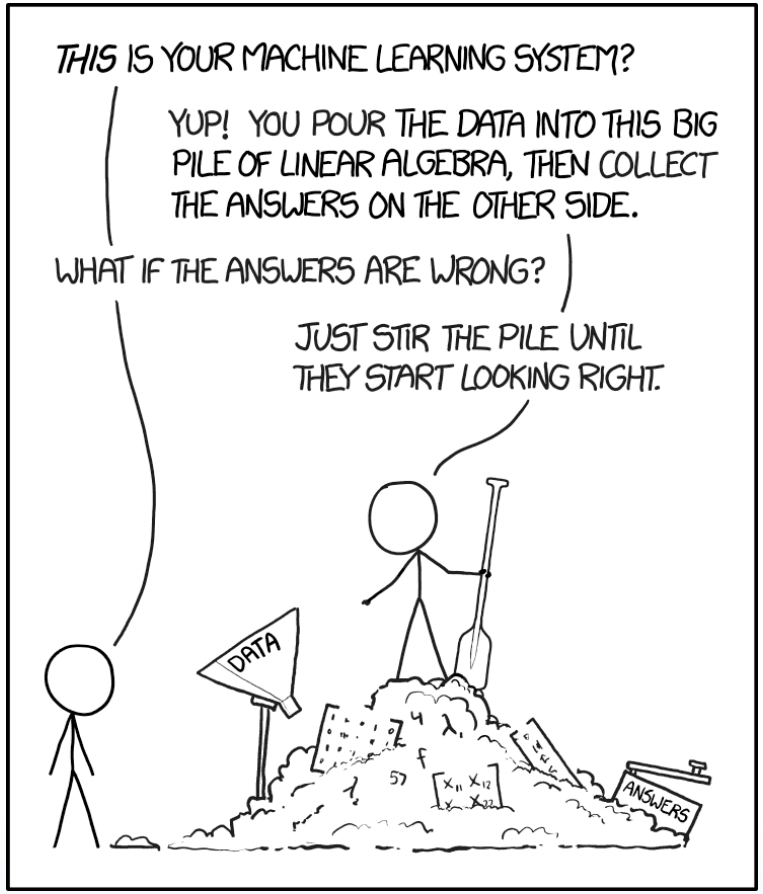
Why abstractive text summarization?



The age of Information Overload

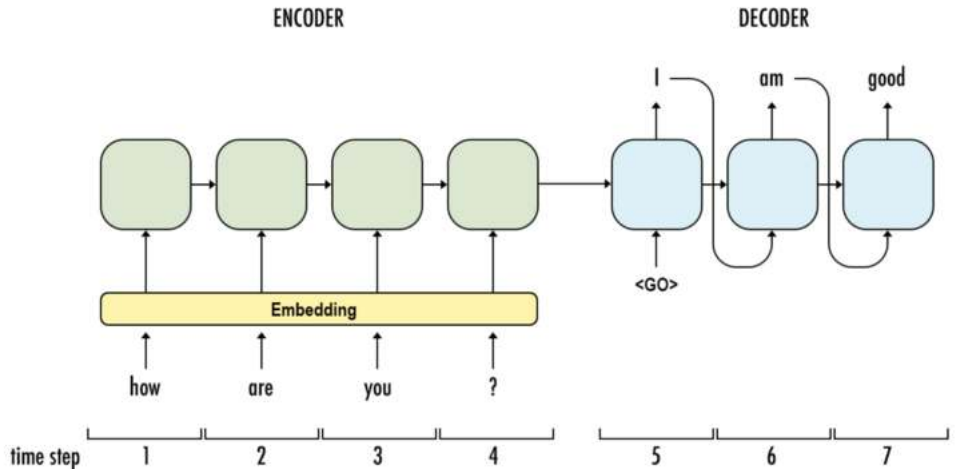
- Top priority: make sense of large amounts of textual data
- Can be analyzed for: sentiment, keywords, meaning…
- Text summarization is **more complex** and **high-level**

Deep learning to the rescue!

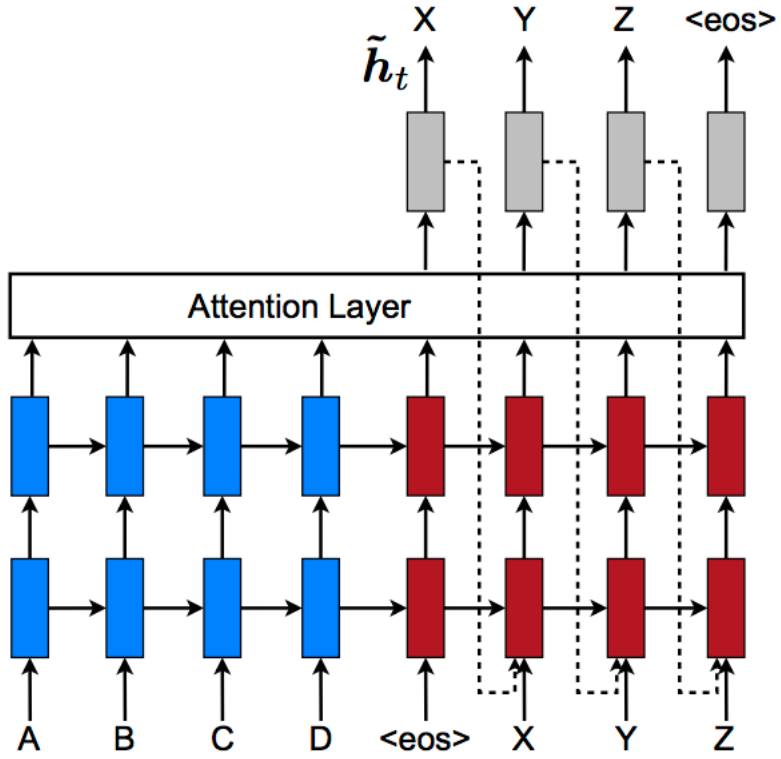


https://xkcd.com/1838/

Deep learning applied to Natural Language Processing



Sequence-to-sequence models



Attention mechanisms

"Sequence to Sequence Learning with Neural Networks" (Sutskever et al. 2014)
"Neural Machine Translation by Jointly Learning to Align and Translate" (Bahdanau et al. 2014)

Deep learning applied to Natural Language Processing (continued)

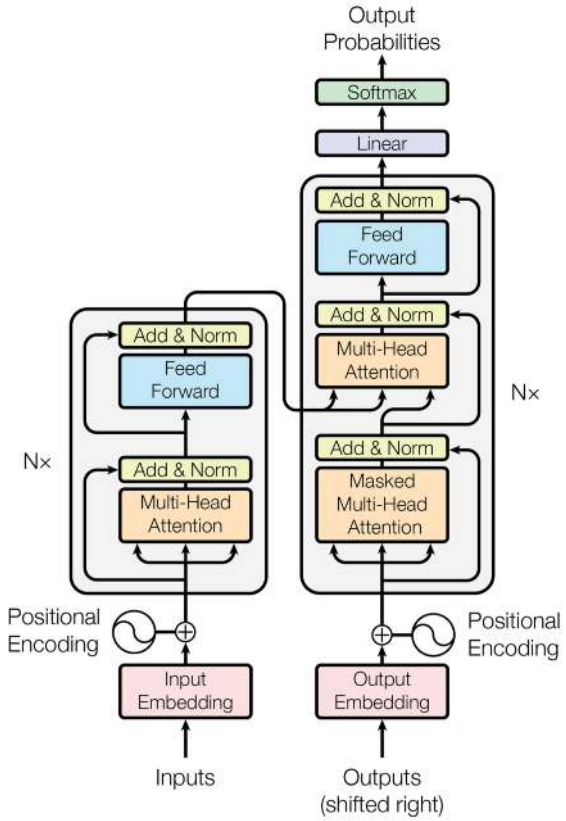
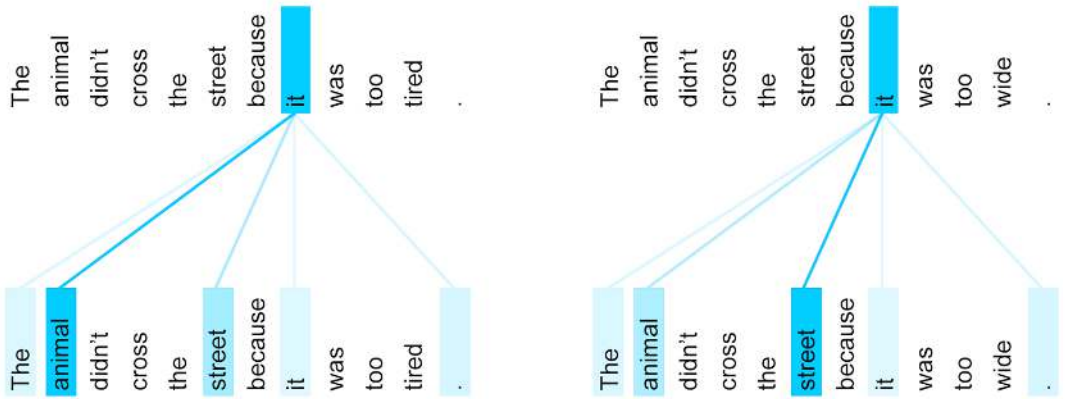


Figure 1: The Transformer - model architecture.



Attention mechanisms

"Attention is all you need" (Vaswani et al. 2017)

Deep learning for NLP breakthroughs

- Question Answering
- AI beats humans on the Squad 2.0 QA problem!

Leaderboard

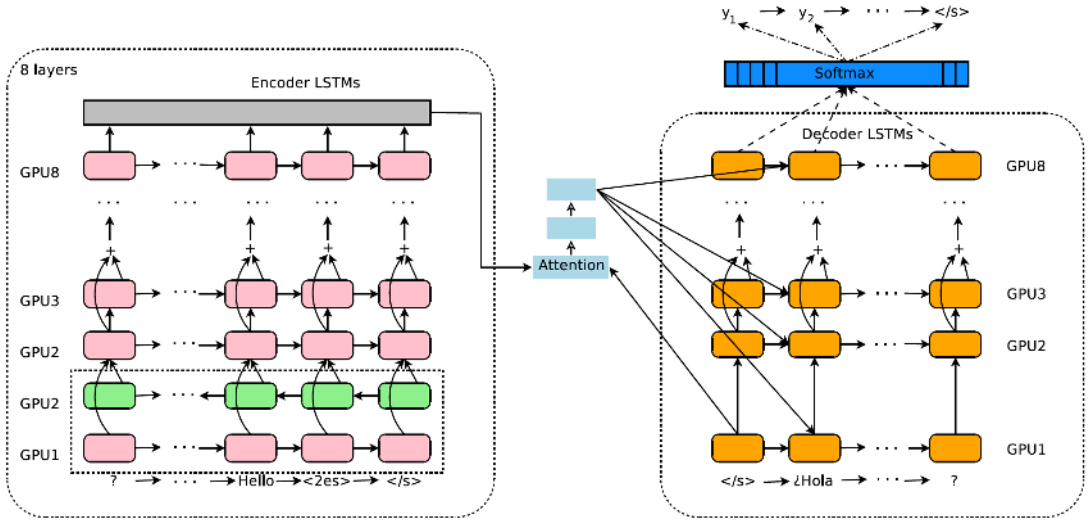
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4	XLNet + SG-Net Verifier++ (single model)	87.238	90.071

https://rajpurkar.github.io/SQuAD-explorer/

Deep learning for NLP breakthroughs (continued)

- Multi-lingual machine translation
- Used by Google Translate in production



"Google's Multilingual Neural Machine Translation System- Enabling Zero-Shot Translation" (Johnson et al. 2016)

Deep learning for NLP breakthroughs (continued)

- Unsupervised machine translation

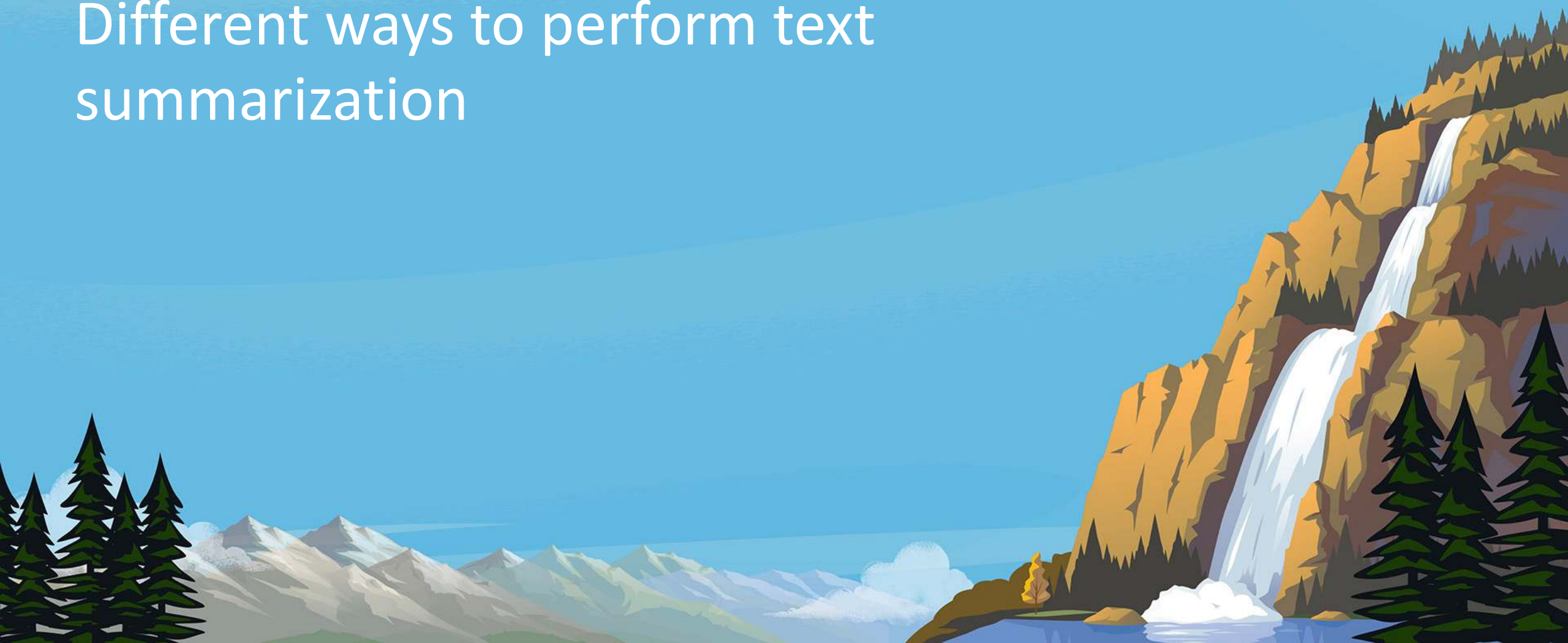
- Can some of these breakthrough techniques be used for text summarization as well?

Source	un homme est debout près d' une série de jeux vidéo dans un bar .
Iteration 0	a man is seated near a series of games video in a bar .
Iteration 1	a man is standing near a closeup of other games in a bar .
Iteration 2	a man is standing near a bunch of video video game in a bar .
Iteration 3	a man is standing near a bunch of video games in a bar .
Reference	**a man is standing by a group of video games in a bar .**
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 .**
Source	une photo d' une rue bondée en ville .
Iteration 0	a photo a street crowded in city .
Iteration 1	a picture of a street crowded in a city .
Iteration 2	a picture of a crowded city street .
Iteration 3	a picture of a crowded street in a city .
Reference	**a view of a crowded city street .**

Table 3: **Unsupervised translations.** Examples of translations on the French-English pair of the Multi30k-Task1 dataset. Iteration 0 corresponds to word-by-word translation. After 3 iterations, the model generates very good translations.

"Unsupervised Machine Translation Using Monolingual Corpora Only" (Lample et al. 2017)

Different ways to perform text summarization



Fixed systems

- Using no or little ML, but heuristics or methods like TF*IDF
- Sometimes with heuristics or expert knowledge too

- **Pros:** easy to implement
- **Cons:** not the most flexible or accurate. Don't benefit from more data

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

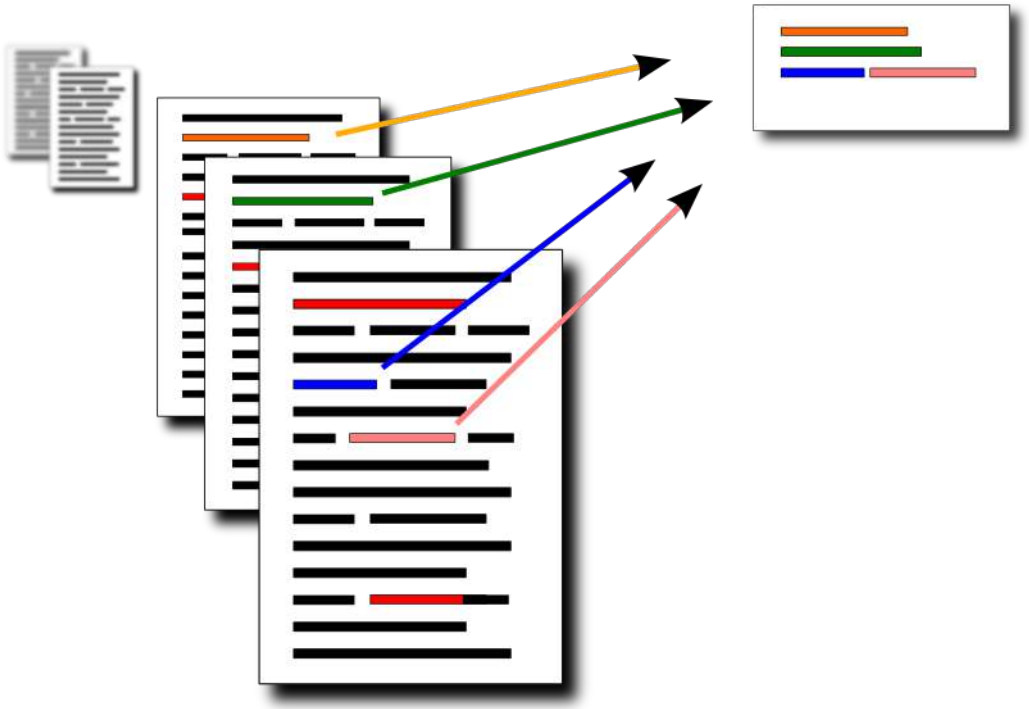
Table 2: Sample centroid produced by CIDR

"Centroid-based summarization of multiple documents" (Radev et al. 2004)

Extractive (or compressive) summarization

- “Highlighter” strategy: take full phrases or sentences from the document(s) to form a summary

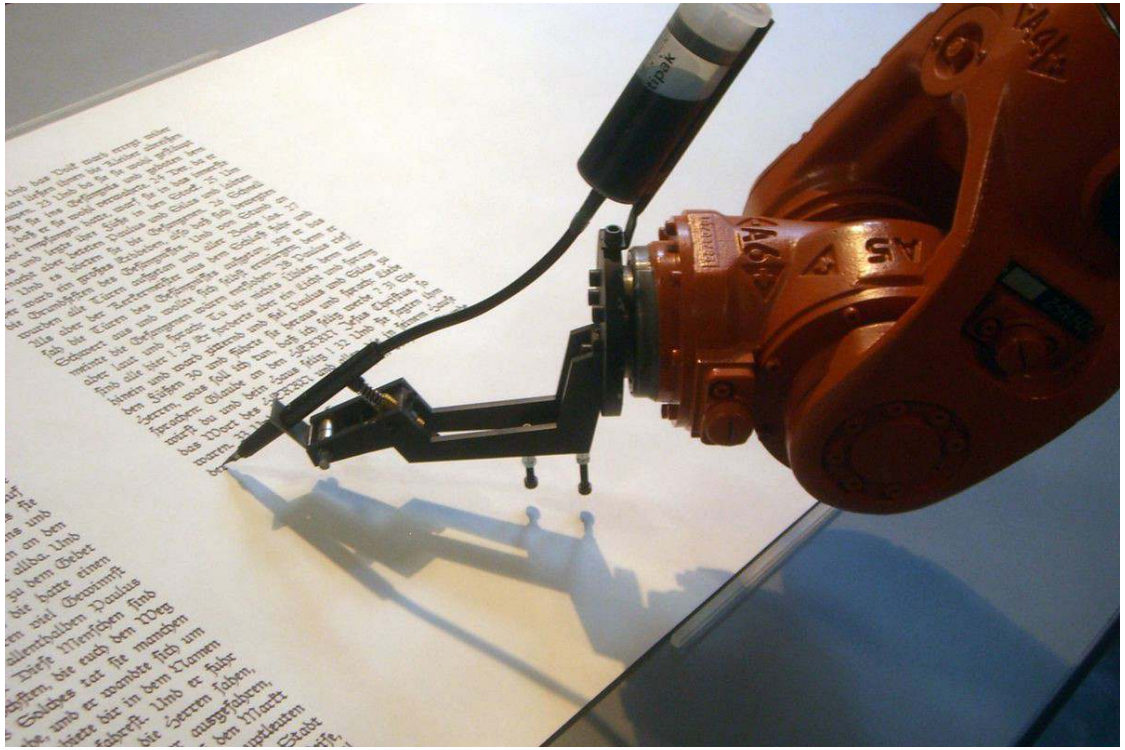
- **Pros:** robust, little room for big mistakes
- **Cons:** not flexible, can’t summarize like humans do



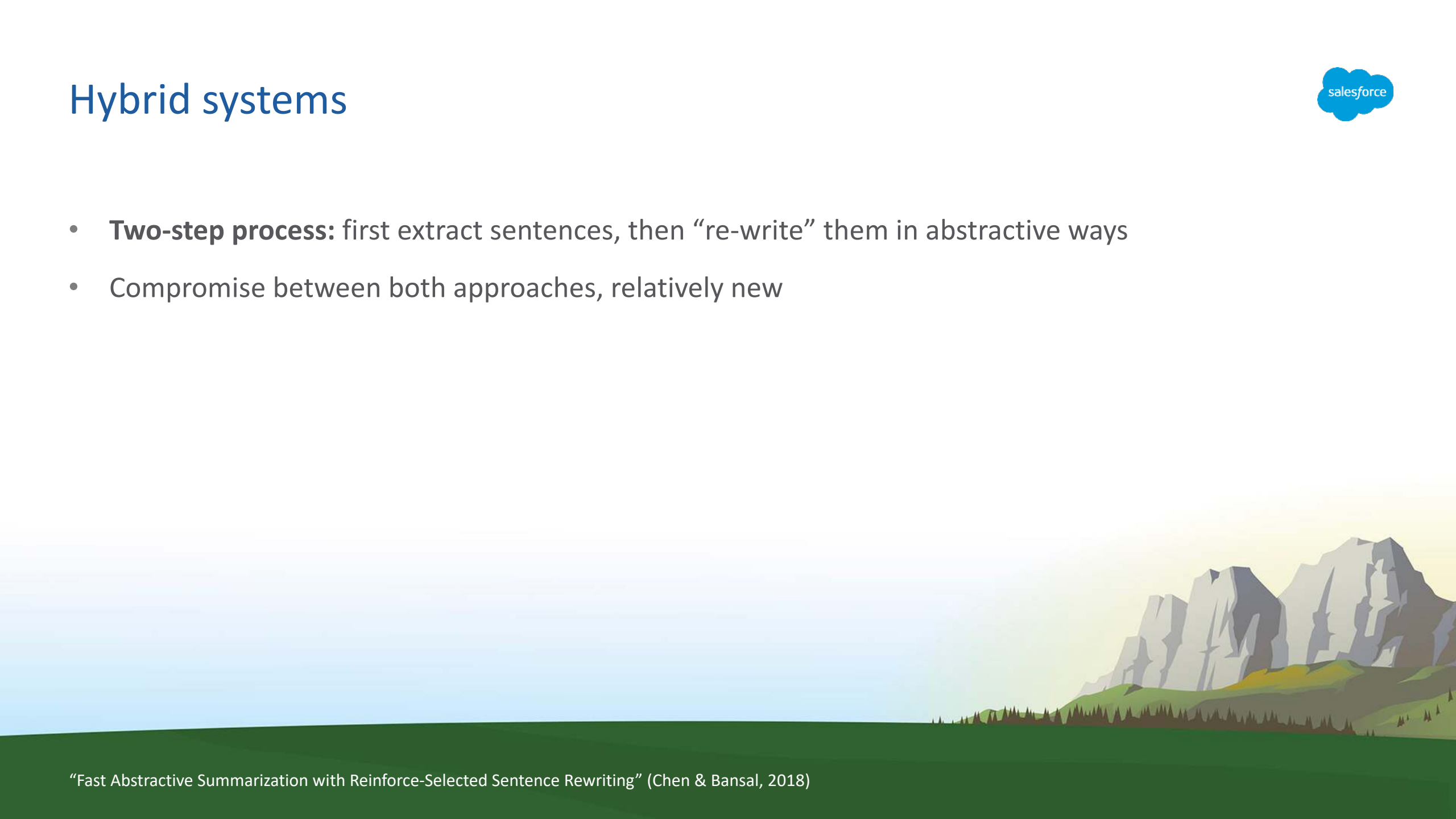
Abstractive summarization

- “pen” strategy: generate a new summary from scratch using any words (in theory)

- **Pros:** can potentially write concise, highly abstractive and complex summaries
- **Cons:** hard to train, potential for more mistakes



Hybrid systems



- **Two-step process:** first extract sentences, then “re-write” them in abstractive ways
- Compromise between both approaches, relatively new

“Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting” (Chen & Bansal, 2018)

Abstractive Summarization Systems

Issues with previous abstractive summarization models

- **High amounts of repetition**, especially with RNNs for sentence generation

- Works well for short summaries (1 sentence or less)
- **Doesn’t scale to longer summaries** or longer documents

Issues with previous abstractive summarization models (continued)

- **Low levels of real abstraction**
- Summarization models copy words in the summary more often than humans do

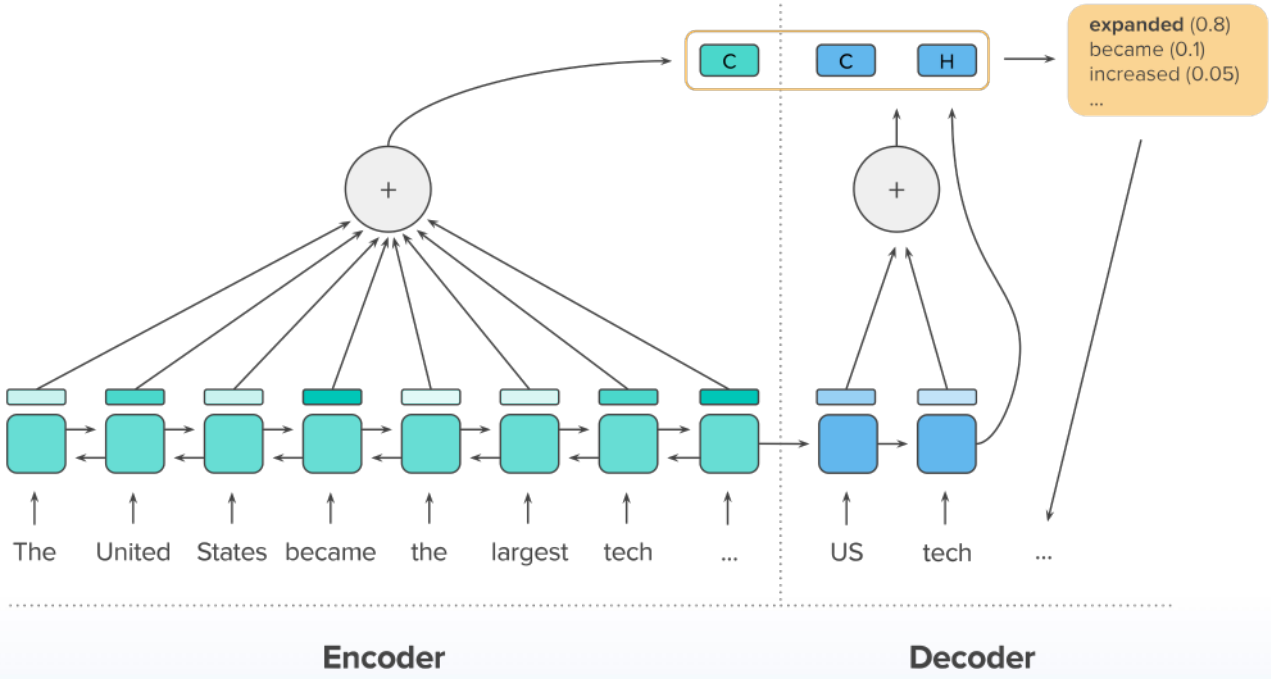
Issues with previous abstractive summarization models (continued)

- **Low ROUGE scores**
- == overlap between generated summary and human-written summary for the same document

"ROUGE: A Package for Automatic Evaluation of Summaries" (Lin 2004)

A Deep Reinforced model for Abstractive Summarization

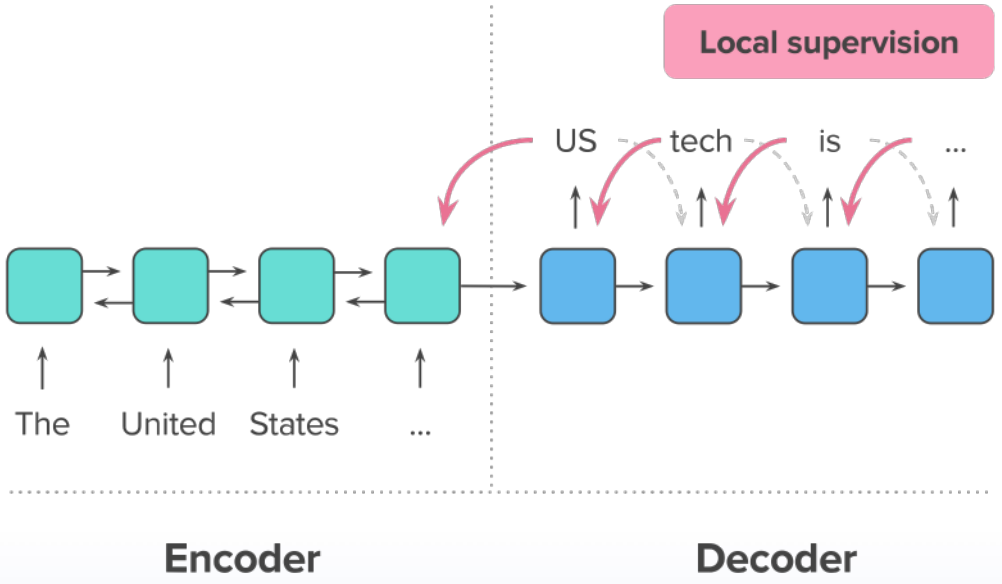
- Encoder-decoder model for summary generation
- Uses **temporal attention** on the input sequence, and **self-attention** on the output



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

Limitations of supervised learning for sequence generation

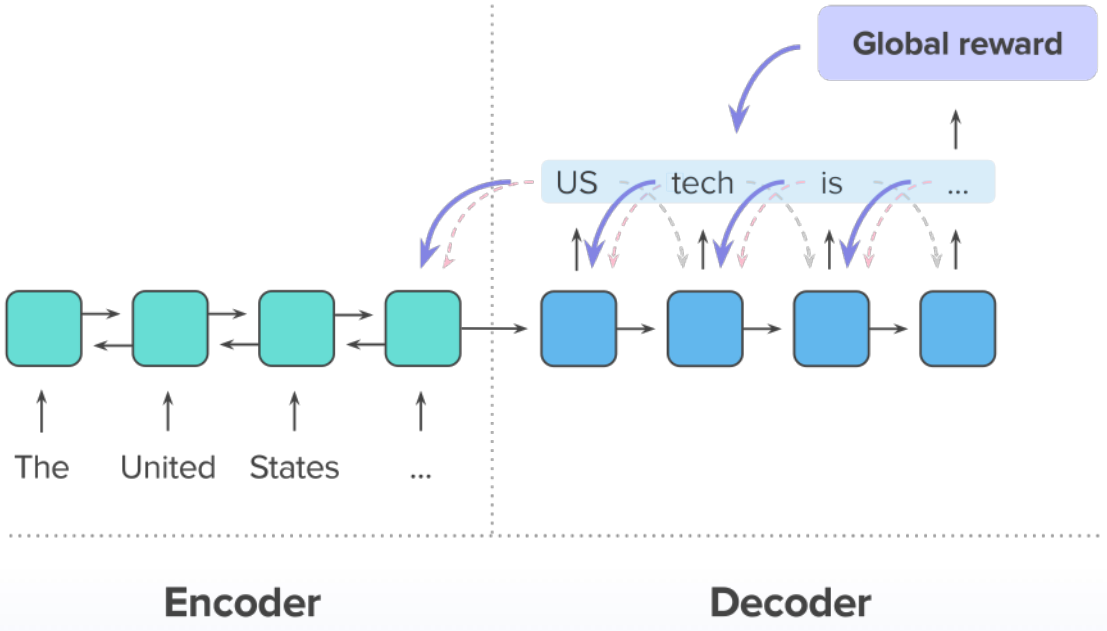
- Only local reward during training (= **teacher forcing** algorithm)
- Issues with **long-term coherence** and **repeated phrases** in the output



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

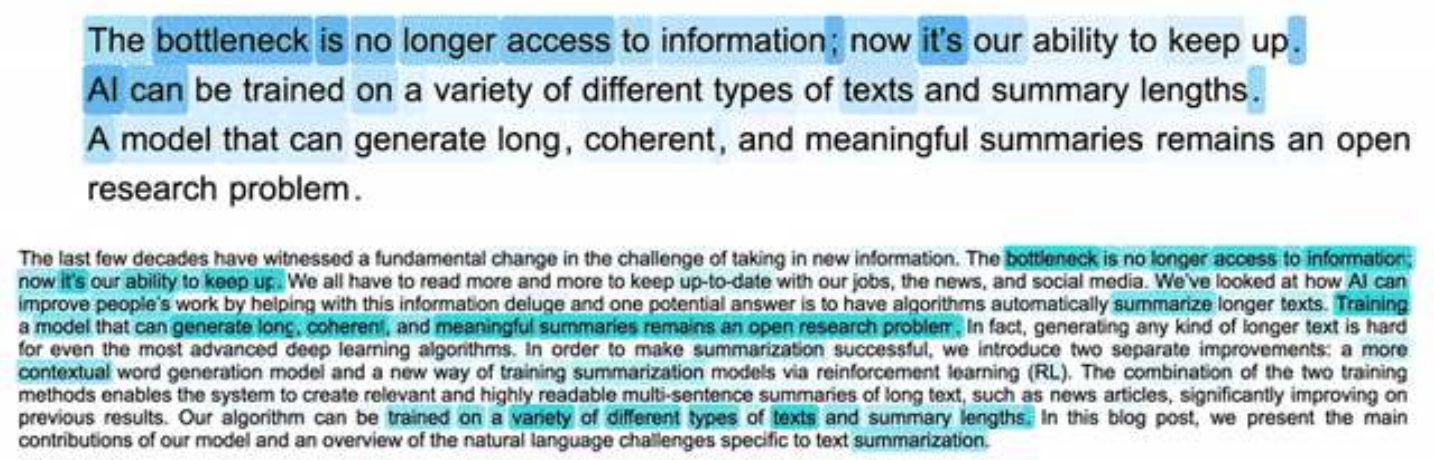
Reinforcement learning (RL) for abstractive summarization

- Add a **global reward** (ROUGE score) to validate the summary as a whole



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

Abstractive summarization results

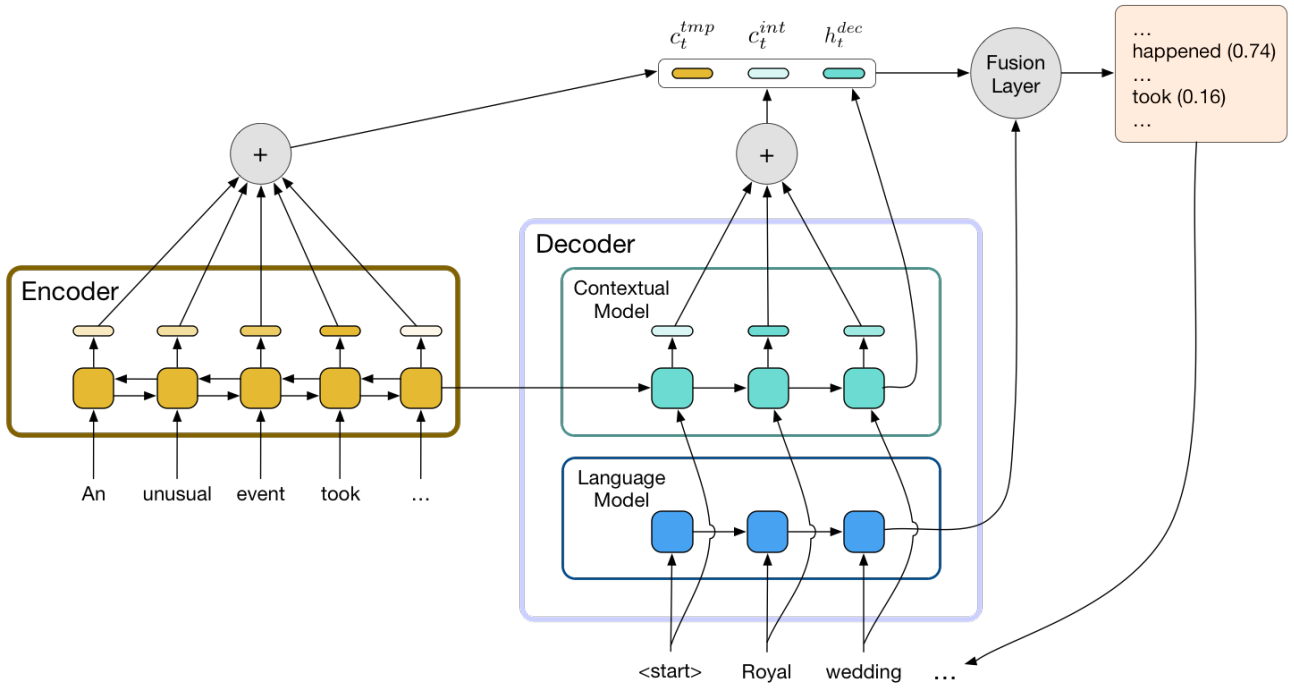


"A Deep Reinforced Model for Abstractive Summarization" (Paulus et al. 2017)

Limitations

- This model is still **far less abstractive** than humans
- Does word copying because it is the “safest” way to obtain high ROUGE scores
- How can we make the model more abstractive while maintaining relevant summaries?

Improving abstraction in text summarization

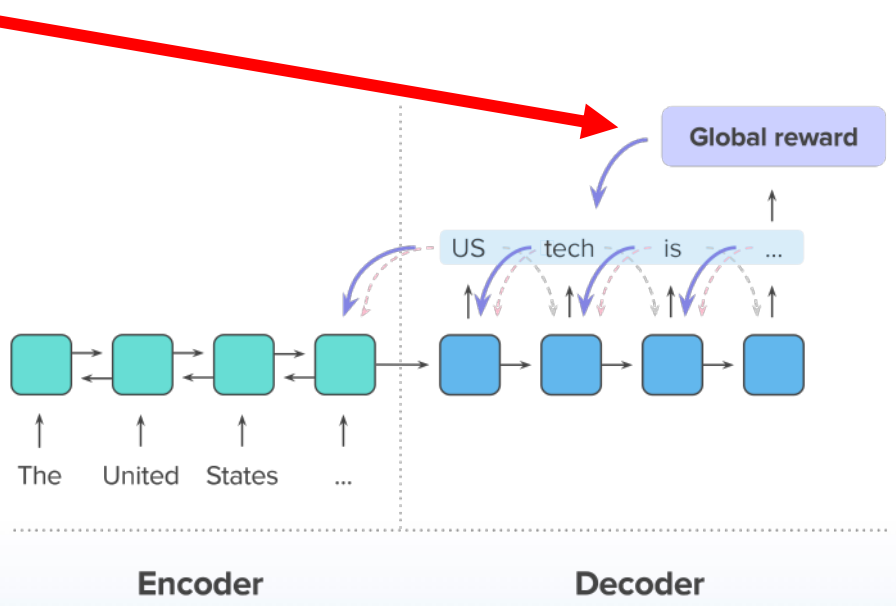


"Improving Abstraction in Text Summarization" (Kryściński et al. 2018)

Improving abstraction in text summarization

- Hybrid training with an **abstraction/novelty reward** in addition to ROUGE-L score reward

$$R(y) = \lambda^{\text{rou}} R^{\text{rou}}(y^{\text{sam}}) + \lambda^{\text{nov}} R^{\text{nov}}(y^{\text{sam}})$$



"Improving Abstraction in Text Summarization" (Kryściński et al. 2018)

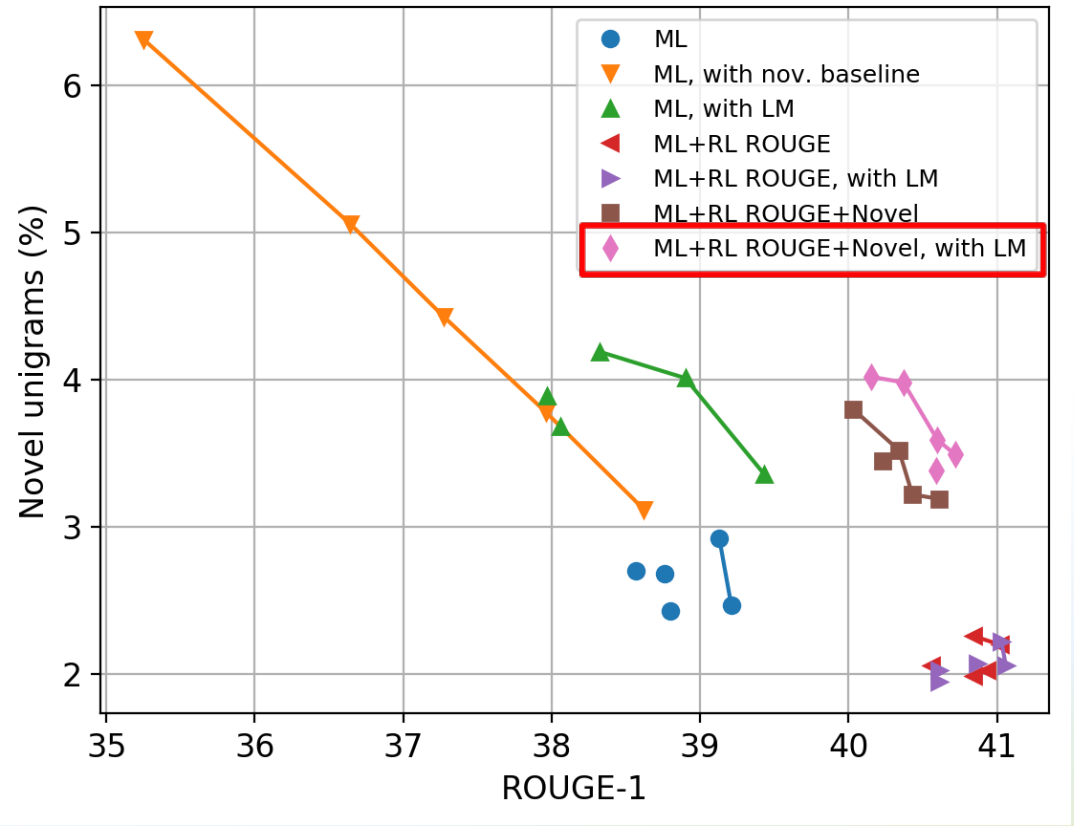
Improving abstraction results

Yellow = sequences of 3+ words **copied from the article**

Article
(cnn) to allay possible concerns, boston prosecutors released video friday of the shooting of a police officer last month that resulted in the killing of the gunman. the officer wounded, john moynihan, is white. angelo west, the gunman shot to death by officers, was black. after the shooting, community leaders in the predominantly african-american neighborhood of (...)
Human-written summary
boston police officer john moynihan is released from the hospital. video shows that the man later shot dead by police in boston opened fire first. moynihan was shot in the face during a traffic stop.
Generated summary (See et al., 2017)
boston prosecutors released video friday of the shooting of a police officer last month. the gunman shot to death by officers , was black . one said the officers were forced to return fire. he was placed in a medically induced coma at a boston hospital.
Generated summary (Liu et al., 2018)
boston prosecutors released video of the shooting of a police officer last month . the shooting occurred in the wake of the boston marathon bombing. the video shows west sprang out and fired a shot with a pistol at officer's face.
Our summary (ML+RL ROUGE+Novel, with LM)
new: boston police release video of shooting of officer , john moynihan. new: angelo west had several prior gun convictions , police say. boston police officer john moynihan, 34, survived with a bullet wound . he was in a medically induced coma at a boston hospital , a police officer says.

"Improving Abstraction in Text Summarization" (Kryściński et al. 2018)

Abstraction/ROUGE tradeoff in summarization



"Improving Abstraction in Text Summarization" (Kryściński et al. 2018)

Remaining challenges and future directions

The field has been stagnating over the past year

- No matter which models, researchers **hit a performance “wall” far from human levels**
- Especially true when looking at **ROUGE scores for long summaries**

- Should we take a step back and **re-evaluate the current research field?**

The difficulty of evaluating summarization models

What makes it more challenging than other AI tasks:

- **Large number** and variety of **“good” answers**
- Summaries have to be **judged as a whole**
  - Small word differences can flip the meaning entirely, shouldn’t be ignored
- Human evaluation is **multi-dimensional**:
  - Fluency
  - Truthfulness (factual consistency)
  - Coherence
  - Relevance

Evaluation methods in summarization

- Increased scrutiny of evaluation methods (like ROUGE) over the years, as summarization problems became more complex
- Yet, ROUGE remained popular, and no clear consensus on what else should be used for evaluation, despite alternatives
  - BLEU
  - METEOR

Re-evaluating the summarization research field

- In-depth look at different sides of summarization research:

- **Evaluation metrics** issues
- **Datasets** issues
- **Models** issues

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

Evaluation metric issues

- **Weak correlation** between **ROUGE** and **human judgments** confirmed
  - Humans judged for: factual consistency, relevance, fluency, and coherence
  - Weaker correlation for abstractive models than extractive ones

- **Insufficient evaluation protocol**
  - Randomly selected summary outputs from SOTA summarization models
  - **30%** of them had **factual consistency issues**

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

Datasets issues

- **Layout bias**
- For news articles, the “inverted pyramid” structure means that more important facts are often written in the beginning

- 60% of the important information is found in the first 3rd or the article (in CNN/Daily Mail)
- Can this be scaled to non-biased data (books, legal docs)?

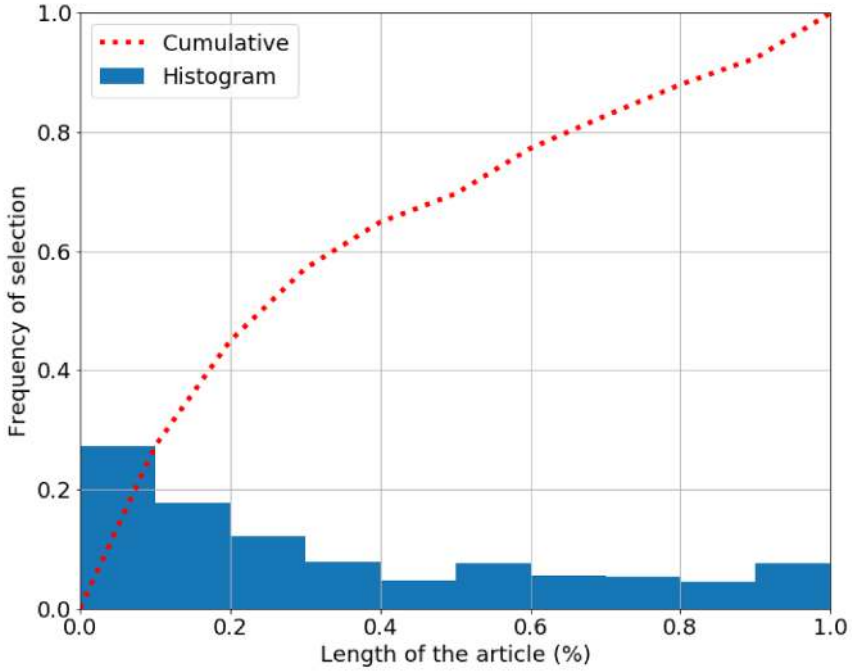


Figure 1: The distribution of important sentences over the length of the article according to human annotators (blue) and its cumulative distribution (red).

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

Datasets issues (continued)

- **Under-constrained summarization task**
- Problem: Humans don't always agree on how to write a good summary
  - This makes it hard for AI to learn
- Solution: add constraints to the task by **asking specific questions** about the document to summarize

Article

The glowing blue letters that once lit the Bronx from above Yankee stadium failed to find a buyer at an auction at Sotheby's on Wednesday. While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, the only person who raised a paddle - for $260,000 - was a Sotheby's employee trying to jump start the bidding. The current owner of the signage is Yankee hall-of-famer Reggie Jackson, who purchased the 10-feet-tall letters for an undisclosed amount after the stadium saw its final game in 2008. No love: 13 letters that hung over Yankee stadium were estimated to bring in anywhere from $300,000 to $600,000, but received no bids at a Sotheby's auction Wednesday. The 68-year-old Yankee said he wanted 'a new generation to own and enjoy this icon of the Yankees and of New York City.', The letters had beamed from atop Yankee stadium near grand [illegible] joined the team. (...)

Summary Questions

When was the auction at Sotheby's?
Who is the owner of the signage?
When had the letters been installed on the stadium?

Unconstrained Summary A

There was not a single buyer at the auction at Sotheby's on Wednesday for the glowing blue letters that once lit the Bronx's Yankee Stadium. Not a single non-employee raised their paddle to bid. Jackson, the owner of the letters, was surprised by the lack of results. The venue is also auctioning off other items like Mets memorabilia.

Constrained Summary B

Glowing letters that had been hanging above the Yankee stadium from 1976 to 2008 were placed for auction at Sotheby's on Wednesday, but were not sold, The current owner of the sign is Reggie Jackson, a Yankee hall-of-famer.

Unconstrained Summary B

The once iconic and attractive pack of 13 letters that was placed at the Yankee stadium in 1976 and later removed in 2008 was unexpectedly not favorably considered at the Sotheby's auction when the 68 year old owner of the letters attempted to transfer its ownership to a member the younger populace. Thus, when the minimum estimate of $300,000 was not met, a further attempt was made by a former player of the Yankees to personally visit the new owner as an

Constrained Summary A

An auction for the lights from Yankee Stadium failed to produce any bids on Wednesday at Sotheby's. The lights, currently owned by former Yankees player Reggie Jackson, lit the stadium from 1976 until 2008.

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

Models issues

- Different models give **more similar outputs to each other** than they are similar to the ground truth

- Training data contains **easy to pick up patterns** that all models overfit to…
- **Or,** information in the **training signal too weak** to connect the source content with the reference summaries

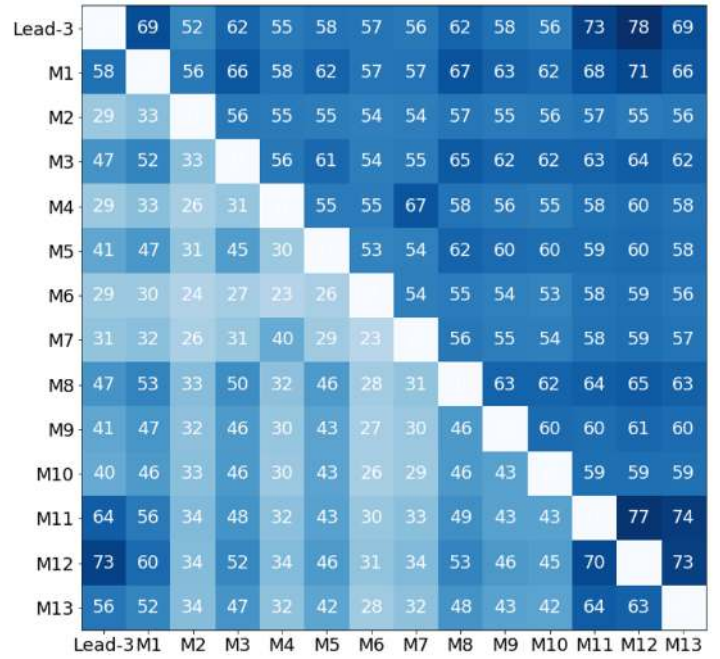


Figure 2: Pairwise similarities between model outputs computed using ROUGE. Above diagonal: Unigram overlap (ROUGE-1). Below diagonal: 4-gram overlap (ROUGE-4). Model order (M-) follows Table 6.

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

Critical evaluation: conclusion

"We hope that this critique provides the summarization community with practical insights for future research directions that include the **construction of datasets**, models **less fit to a particular domain bias,** and **evaluation that goes beyond current metrics** to capture the most important features of summarization."

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

Final challenges: from research to production

Reliably summarize facts

- Even bigger issue in the “fake news” era
- Promising direction: using textual entailment
  - “**Multi-Reward Reinforced Summarization with Saliency and Entailment**” (Pansuru et al 2018)
  - “**Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization**” (Li et al 2018)

Summarize long technical documents

- Books, legal, financial documents, etc

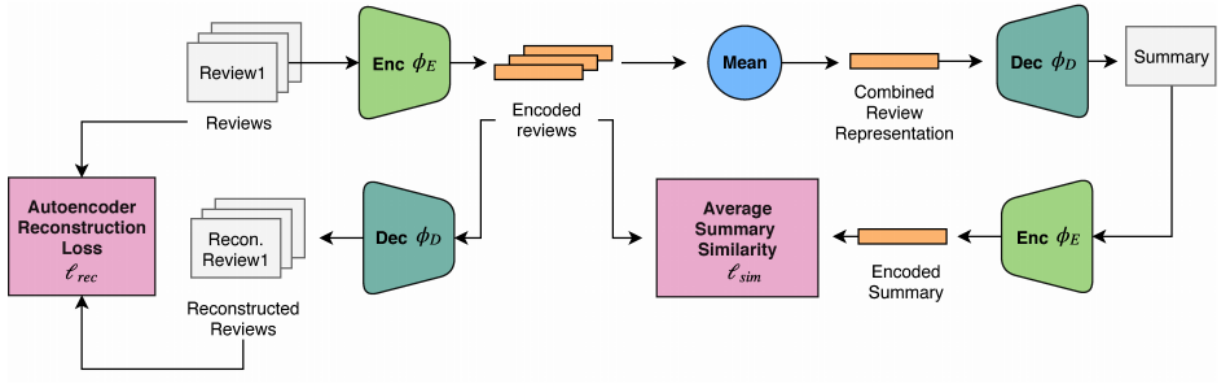
- First obstacle: **lack of large, good, consistent, public datasets**
- Second obstacle: **Way bigger scale** than current summarization problems (news articles)

- I’ve been personally asked to evaluate the feasibility of summarization for financial reports

Domain transfer/unsupervised learning

- **Problem:** summarization models trained on one domain rarely work well on another
- Lack of good labeled datasets make unsupervised learning an appealing, but difficult, solution

- **Promising direction:** use auto-encoders for unsupervised multi-document abstractive summarization



"MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization" (Chu et al 2018)

Personalized summaries

- Fine-tune the summary based on the reader’s profile, what they know, what they are looking for, etc

- Lack of good public data makes this currently hard
- But eventually there will be a large demand for this level of summarization
- Combines well with search results, AI personal assistants, etc

Conclusion


THANK YOU
We’re hiring full-time & interns!
https://einstein.ai/careers