Abstractive Summarization

state of the art and the problem of factuality

Deep Learning Sessions Portugal

Diogo Pernes (diogo.pernes@priberam.pt)

February 2023

Outline

1. Background
2. The hallucination problem
3. Automatic evaluation
4. Our approach
5. Summarization with GPT3.5
6. Conclusion

Background

Summarization

The UN has said media restrictions and violence meant the environment was not conducive to free, credible elections.
Unrest started in April after President Pierre Nkurunziza said he would run for a third term - something protesters say is illegal.
The president says he is entitled to a third term because he was appointed for his first term, not elected.
The presidential election is scheduled for 15 July. East African leaders have called for a further two-week delay.
Africa news highlights: 7 July
The electoral commission spokesman told the BBC turnout for the parliamentary poll had been low in the districts of Bujumbura where there had been protests, but that in some provinces outside the capital it was as high as 98%.
The ruling party - the CNDD FDD - was ahead in every province of the country, Burundi's electoral commission announced.
They won 77 out of 100 elected seats in parliament, AFP news agency says."

The UN has stated that the current situation in Burundi is not conducive to free and credible elections.

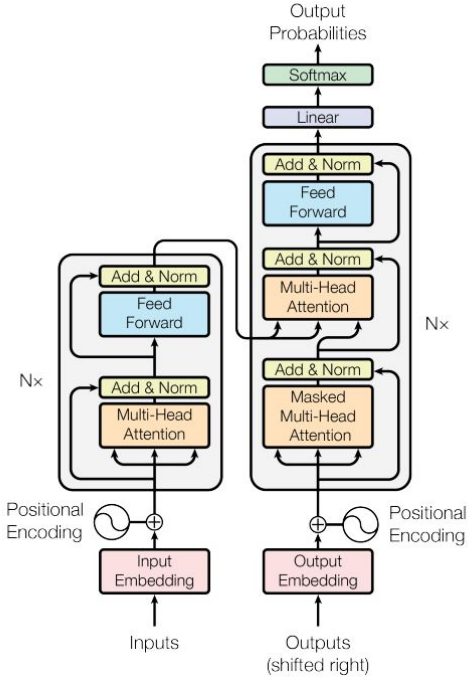
Extractive vs. abstractive

- Extractive summarization: select sentences from the source document and concatenate them to form the summary.
  - Problems: fluency, relevance

- Abstractive summarization: generate a new text that captures the essential ideas of the source document.
  - Problems: factual consistency, relevance

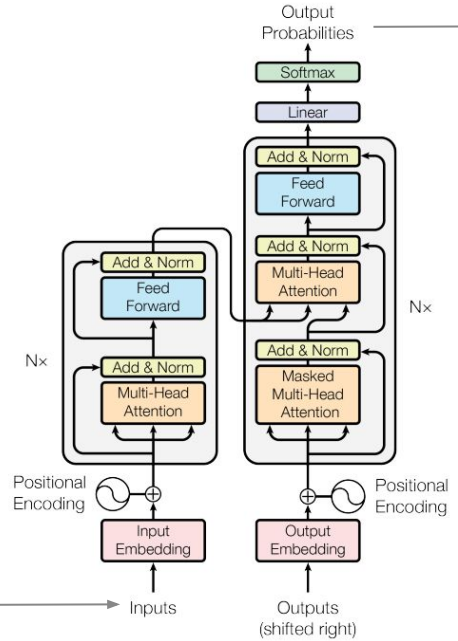
Typical abstractive summarizer

- Transformer-based, encoder-decoder seq2seq architecture (e.g. BART, T5, Pegasus, …).
- Self-supervised pre-training on a large corpus.
- Supervised fine-tuning on a summarization dataset.



Vaswani et al. 2017

Training a transformer for summarization



distribution $p_\theta(\mathrm{y} \mid x)$

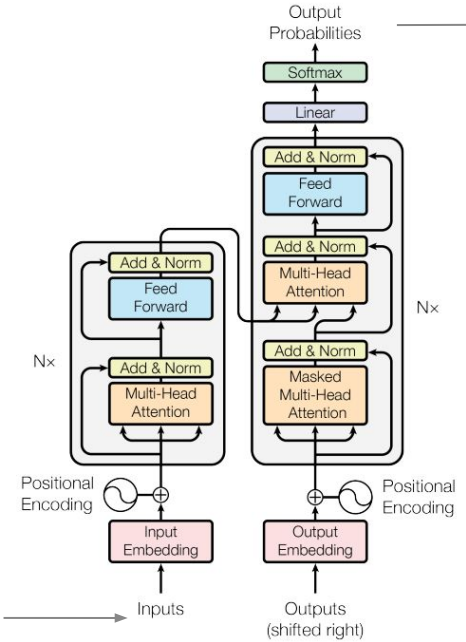
human-written summary y_{ref}

objective:

$$\max_\theta \sum_{i=1}^{n} \log p_\theta(\mathrm{y} = y_{\mathrm{ref}_i} \mid x_i)$$

document x

The output of the transformer



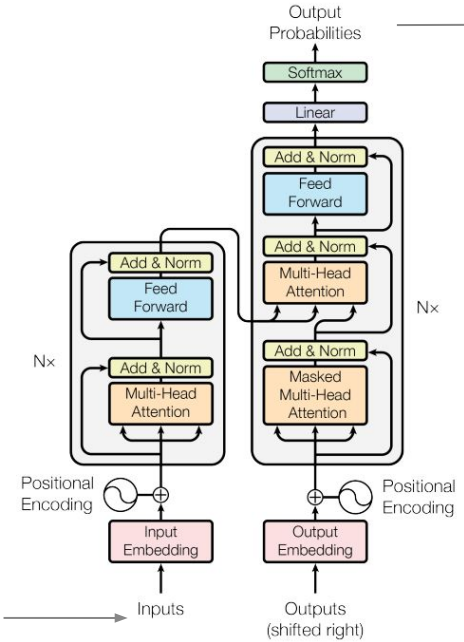
x

$p_\theta(\mathrm{y} \mid x)$

For a summary y with m tokens (words):

$$p_\theta(y \mid x) = p_\theta(y^{(1)} \mid x)p_\theta(y^{(2)} \mid x, y^{(1)}) \dots p_\theta(y^{(m)} \mid x, y^{(1)}, y^{(2)} \dots y^{(m-1)})$$

The output of the transformer



x

$p_\theta(\mathrm{y} \mid x)$

At timestep t, the output of the transformer is:

$p_\theta(\mathrm{y}^{(t)} \mid x, y^{<t})$ → vector with length equal to the vocabulary size

The transformer is an *autoregressive* model.

Summarizing a document

document x → Abstractive Summarizer → distribution $p_\theta(\mathrm{y} \mid x)$

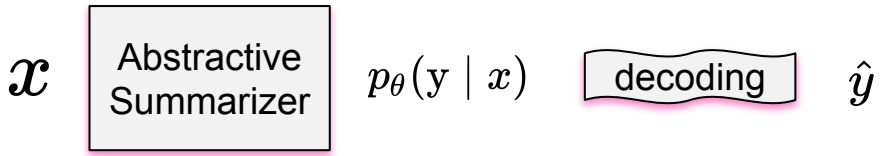
How to get from $p_\theta(\mathrm{y} \mid x)$ to an actual summary y ?

Summarizing a document

How to get from $p_\theta(\mathrm{y} \mid x)$ to an actual summary $\hat{y}$? Through a decoding algorithm!

- Deterministic approaches: look for $\hat{y}$ that maximizes $p_\theta(\mathrm{y} \mid x)$
  - Greedy decoding
  - Beam search
- Sampling-based approaches: randomly sample $\hat{y}$ according to $p_\theta(\mathrm{y} \mid x)$
  - Nucleus sampling
  - Top-k sampling
  - Typical decoding
  - …

Summarizing a document



The hallucination problem

Hallucination

"Hanamanthappa Koppad was tapped under 8m of snow at a height of nearly 6,000m along with nine other soldiers who all died. Their bodies have now been recovered. The **critically ill soldier** has been airlifted to a hospital in Delhi. (...) The army added that "he has been placed on a ventilator to protect his airway and lungs in view of his **comatose state**". (...)"

source: BBC

"**A soldier** who was trapped in an avalanche on the Siachen glacier in Indian-administered Kashmir last week **has been declared dead**, the army says."

summary: BART (fine-tuned on XSum)

Hallucination

- *Hallucinations* are factual inconsistencies (w.r.t. the source document) that occur in the generated summaries.

- Tackling hallucinations is a central problem in abstractive summarization.

Intrinsic hallucination

"**Zack Goldsmith,** who was the favourite for the Tory nomination, balloted his constituents earlier this year to seek permission to stand. At the very point of his entry into the **race for London mayor,** Goldsmith's decision revealed two big characteristics (...)"

Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.

adapted from Maynez et al., ACL 2020

Extrinsic hallucination

"**Zack Goldsmith,** who was the favourite for the Tory nomination, balloted his constituents earlier this year to seek permission to stand. At the very point of his entry into the **race for London mayor,** Goldsmith's decision revealed two big characteristics (...)"

UKIP leader Nigel Goldsmith has been chosen to stand in the London mayoral election.

adapted from Maynez et al., ACL 2020

Automatic evaluation

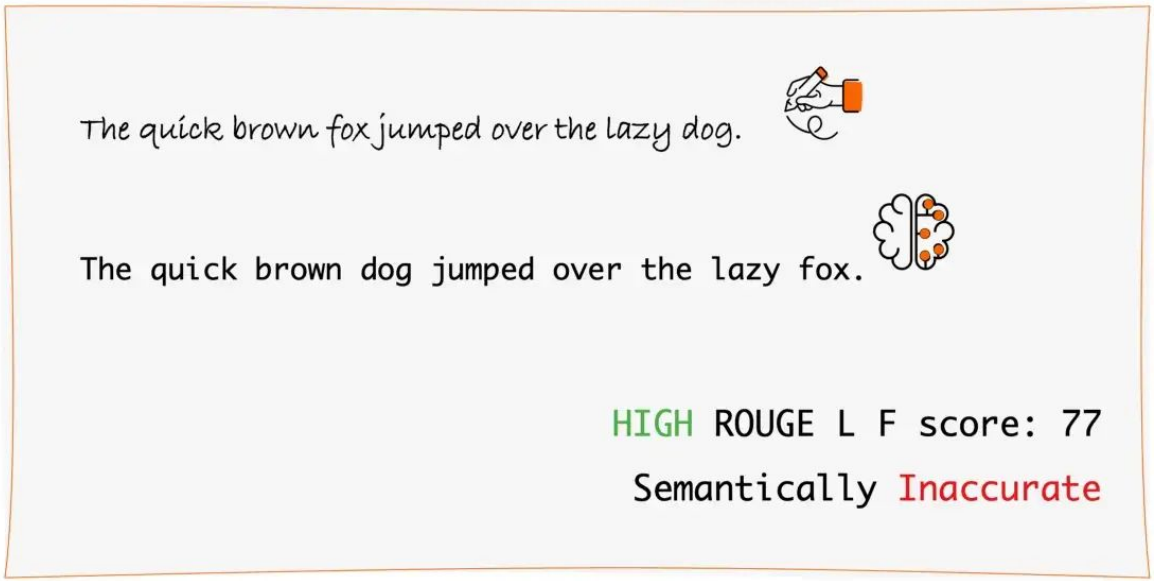
ROUGE scores

(Recall-Oriented Understudy for Gisting Evaluation)

ROUGE-*n* Precision = #{*n*-grams shared by $\hat{y}$ and y_{ref}} / #{*n*-grams in $\hat{y}$}

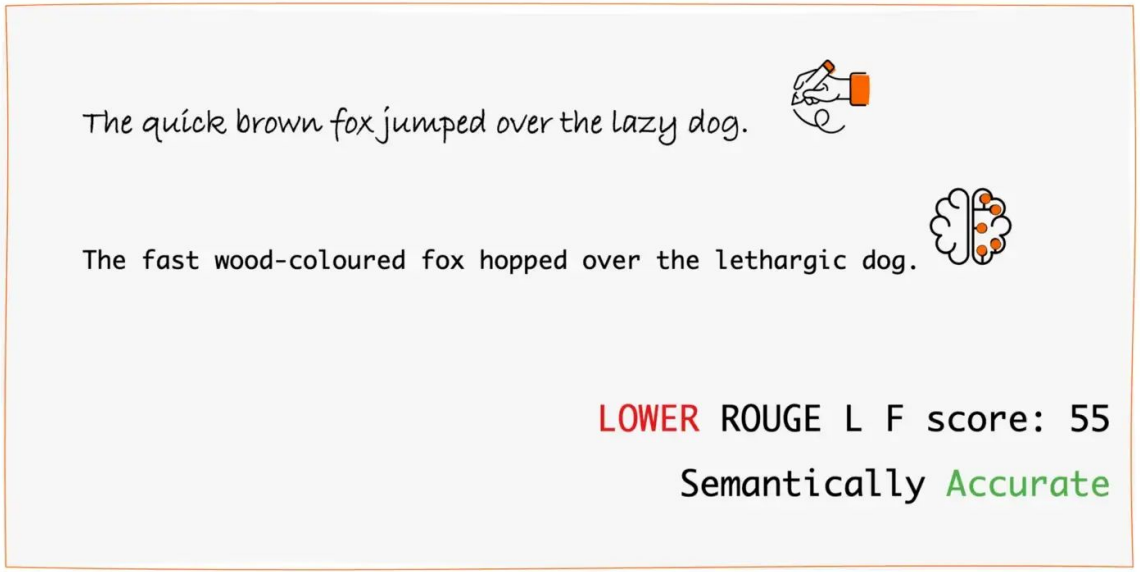
ROUGE-*n* Recall = #{*n*-grams shared by $\hat{y}$ and y_{ref}} / #{*n*-grams in y_{ref}}

ROUGE scores



reproduced from Nina Hristozova's blogpost

ROUGE scores



reproduced from Nina Hristozova's blogpost

Entailment scores

(FactCC metric)

[CLS] x [SEP] $\hat{y}$ [SEP] → Transformer encoder → $s(x, \hat{y}) \in [0, 1]$

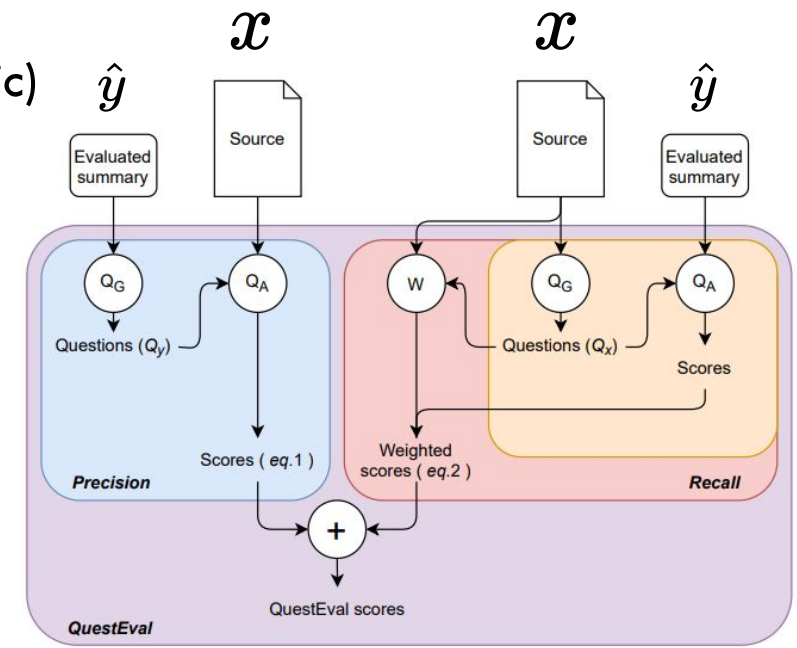
$s(x, \hat{y}) \approx 0$ if $\hat{y}$ is not consistent with x

$s(x, \hat{y}) \approx 1$ if $\hat{y}$ is consistent with x

Kryscinski et al., EMNLP 2020

Question answering scores

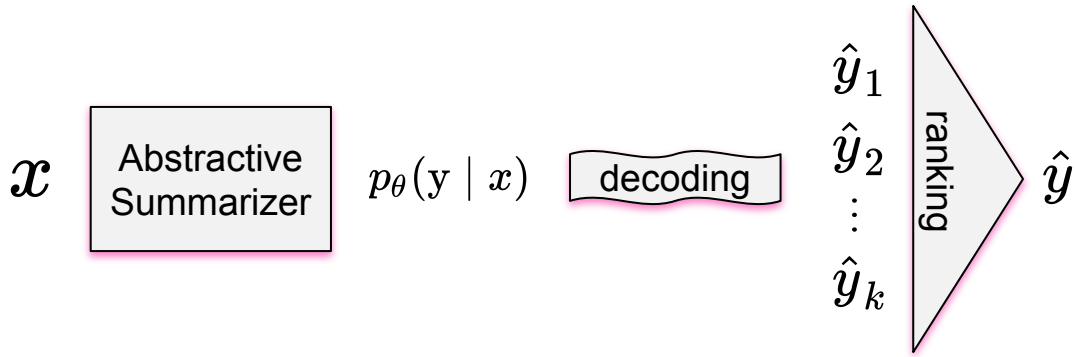
(QuestEval metric)



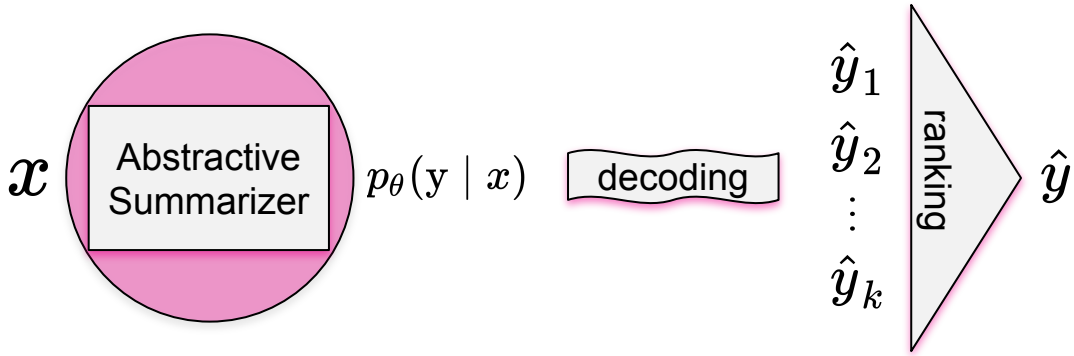
Scialom et al., EMNLP 2021

Improving summaries

Where to intervene?



Where to intervene?

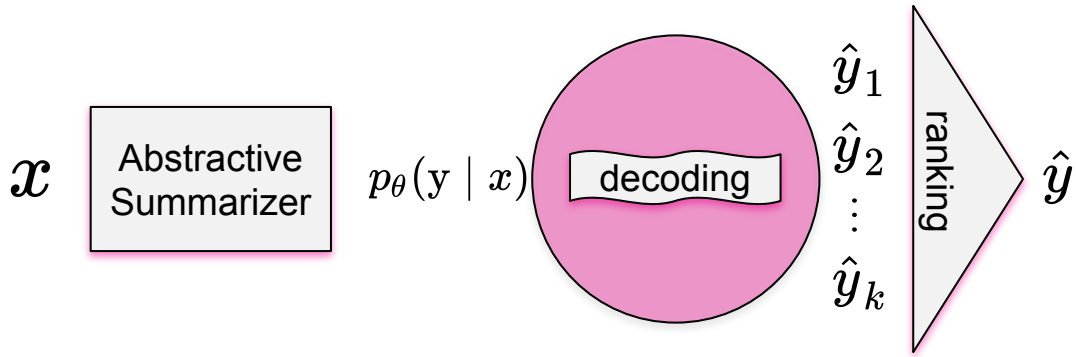


Optimize for a metric (e.g. BRIO, Liu et al., ACL 2022)

Synthetic hallucinations + contrastive learning (e.g. CLIFF, Cao and Wang, EMNLP 2021)

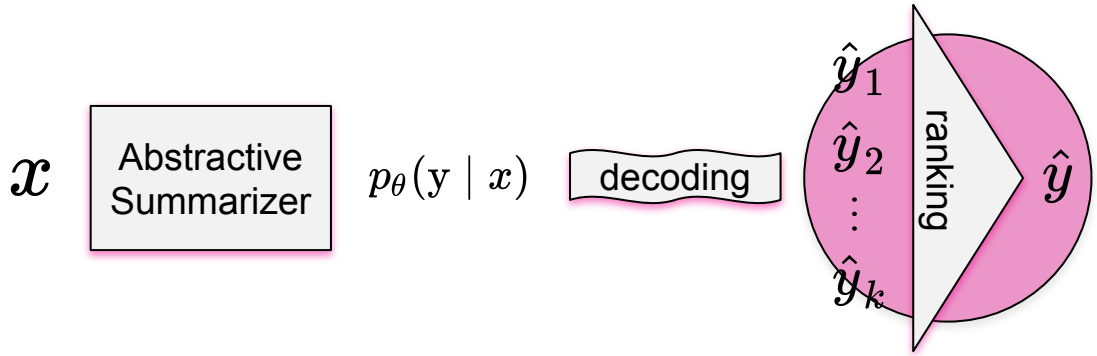
Reinforcement learning from human feedback (Stienon et al., NeurIPS 2020)

Where to intervene?



Avoid tokens whose probability is insensitive to the source document (CPMI, van der Poel et al., EMNLP 2022)

Where to intervene?



Distill metrics into a reference-free function (e.g. EBR, Pernes et al., GEM workshop at EMNLP 2022)

Our approach

Idea

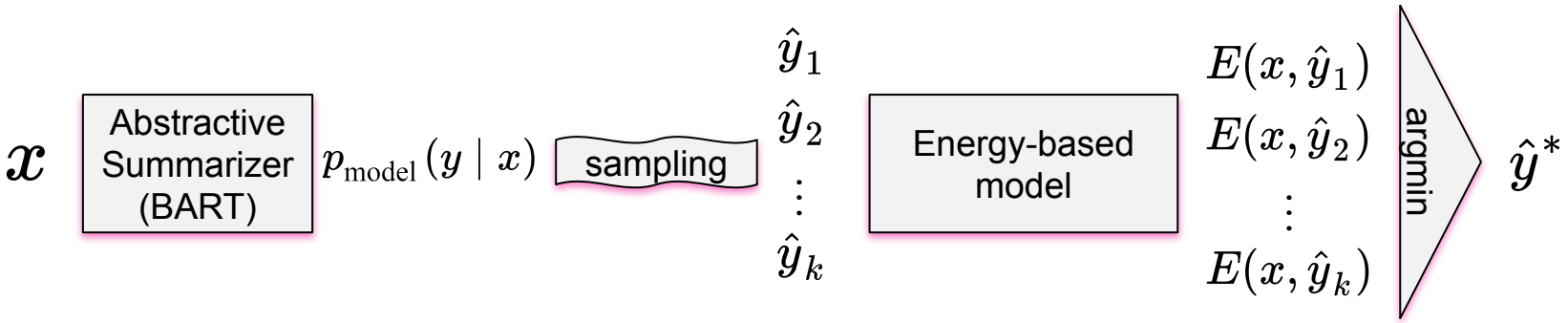
"Hanamanthappa Koppad was tapped under 8m of snow at a height of nearly 6,000m on the Siachen glacier along with nine other soldiers who all died. Their bodies have now been recovered. The **critically ill soldier** has been airlifted to a hospital in Delhi. (...) The army added that "he has been placed on a ventilator to protect his airway and lungs in view of his **comatose state**". (...)"

→ "**A soldier** who was trapped in an avalanche **on the Siachen glacier is in a comatose state**, the army says." 😃

→ "**A soldier** who was trapped in an avalanche **is in a comatose state**, the army says." 🙂

→ "**A soldier** who was trapped in an avalanche on the Siachen glacier **has been declared dead**, the army says." 😜

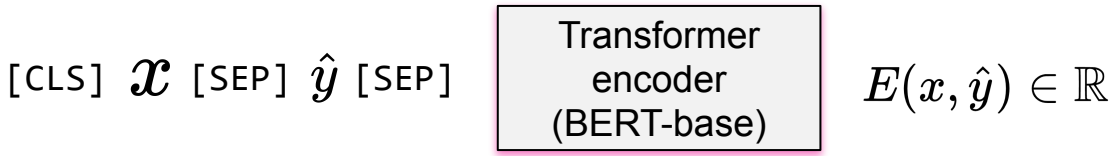
Idea



For a chosen gold-metric ϕ, E should be such that:

$$\phi(x, y_{\text{ref}}, \hat{y}_i) > \phi(x, y_{\text{ref}}, \hat{y}_j) \Rightarrow E(x, \hat{y}_i) < E(x, \hat{y}_j)$$

The EBR



$p_{\text{true}}(y \mid x) \propto \exp(-E(x, y))$

the metric ϕ works as a proxy for $p_{\text{true}}(y \mid x)$ (unknown)

Training the EBR

ListMLE ranking loss:

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k : i < j \Rightarrow \phi(x, y_{\text{ref}}, \hat{y}_i) \geq \phi(x, y_{\text{ref}}, \hat{y}_j)$$

(candidates sorted in decreasing order of quality)

$$\mathcal{L}_\phi(\theta) = -\mathbb{E}_{(x, y_{\text{ref}}, \hat{\mathbf{y}}) \sim \mathcal{D}} \log \prod_{i=1}^{k} \frac{\exp(-E(x, \hat{y}_i; \theta)/\tau)}{\sum_{j=i}^{k} \exp(-E(x, \hat{y}_j; \theta)/\tau)}$$

Training the EBR

ListMLE ranking loss:

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k : i < j \Rightarrow \phi(x, y_{\text{ref}}, \hat{y}_i) \geq \phi(x, y_{\text{ref}}, \hat{y}_j)$$

(candidates sorted in decreasing order of quality)

$$\mathcal{L}_\phi(\theta) = -\mathbb{E}_{(x, y_{\text{ref}}, \hat{\mathbf{y}}) \sim \mathcal{D}} \log \prod_{i=1}^{k} \frac{\exp(-E(x, \hat{y}_i; \theta)/\tau)}{\sum_{j=i}^{k} \exp(-E(x, \hat{y}_j; \theta)/\tau)}$$

prob. that $\hat{y}_i$ ranks first among $\{\hat{y}_i, \hat{y}_{i+1}, \ldots, \hat{y}_k\}$

Automatic evaluation

	CNN/DailyMail							XSum						
	R1	R2	RL	QE	Cons	Rel	FCC	R1	R2	RL	QE	Cons	Rel	FCC
BART	43.64	20.75	40.52	43.28	95.01	61.75	55.68	42.67	19.42	34.48	28.27	83.18	52.23	26.28
BRIO	**47.97***	**24.06***	**44.86***	43.49	89.61	60.75	33.05	–	–	–	–	–	–	–
CLIFF	43.86	20.88	40.63	43.28	94.68	60.38	55.85	44.50	**21.41**	36.41	29.34	82.57	51.92	24.86
DAE	–	–	–	–	–	–	–	37.61	14.19	28.84	29.20	79.45	51.05	19.46
FASum	40.40	17.68	37.26	42.87	94.30	57.91	51.20	30.22	9.97	23.69	24.35	75.45	39.42	26.96
SummaReranker	45.07	21.73	41.87	43.61	95.07	62.49	54.50	**44.93**	21.40	**36.76**	28.76	83.00	52.75	26.27
EBR [RL]	44.90	21.58	41.75	43.60	95.01	62.16	54.95	43.63	20.28	35.78	28.55	84.47	52.92	**27.21**
EBR [QE]	44.07	21.13	40.94	**44.27***	95.71	62.48	59.23	42.94	19.42	34.62	**29.89**	83.34	52.50	26.34
EBR [Rel]	44.04	20.98	40.85	43.78	95.93	**63.40**	60.28	43.39	19.75	35.03	28.60	85.49	**54.80**	26.28
EBR [Cons+Rel]	43.88	20.87	40.69	43.79	**96.15**	63.32	**61.67***	43.28	19.72	34.92	28.66	**86.03***	54.74	27.12

Human evaluation

	CNN/DailyMail			XSum		
	FC	R	F	FC	R	F
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 4: Proportion of times (in %) that each model was considered the best for the human judges in each pairwise comparison according to each criteria (FC: factual consistency, R: relevance, F: fluency). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons.

Human evaluation

	CNN/DailyMail			XSum		
	FC	R	F	FC	R	F
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 4: Proportion of times (in %) that each model was considered the best for the human judges in each pairwise comparison according to each criteria (FC: factual consistency, R: relevance, F: fluency). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons.

Human evaluation

	CNN/DailyMail			XSum		
	FC	R	F	FC	R	F
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 4: Proportion of times (in %) that each model was considered the best for the human judges in each pairwise comparison according to each criteria (FC: factual consistency, R: relevance, F: fluency). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons.

Human evaluation

	CNN/DailyMail			XSum		
	FC	R	F	FC	R	F
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 4: Proportion of times (in %) that each model was considered the best for the human judges in each pairwise comparison according to each criteria (FC: factual consistency, R: relevance, F: fluency). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons.

Human evaluation

	CNN/DailyMail			XSum		
	FC	R	F	FC	R	F
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 4: Proportion of times (in %) that each model was considered the best for the human judges in each pairwise comparison according to each criteria (FC: factual consistency, R: relevance, F: fluency). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons.

Summary examples

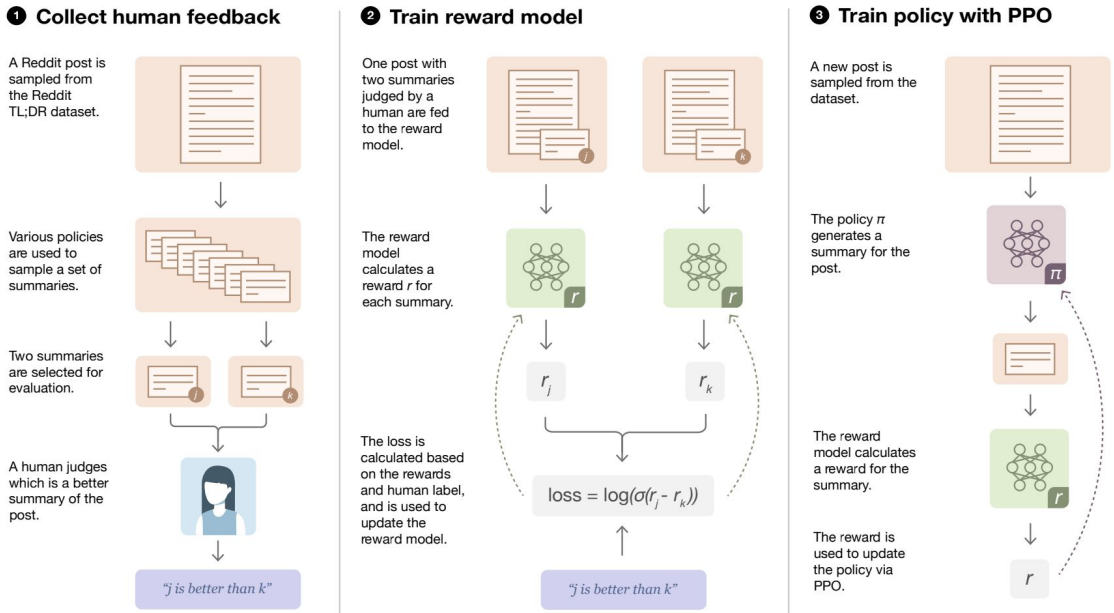
	Text	Cons ↑	E ↓
Source (CNN/DM)	Kell Brook has finally landed the Battle of Britain he craved, but will take on Frankie Gavin rather than bitter rival Amir Khan. Just sixty four days after the first defence of his IBF belt against Jo Jo Dan, Brook will return to action on a packed pay-per-view show on May 30 at the O2 in London. The welterweight bout has been added to a card that includes world title challenges for Kevin Mitchell and Lee Selby while Anthony Joshua faces his toughest test to date against Kevin Johnson. Kell Brook poses outside London's O2 Arena where he will fight Frankie Gavin on May 30. Brook posing on the train as he headed to London for the announcement of his fight. Brook (left) was back in action as he beat Jo Jo Dan for the IBF World Welterweight title in Sheffield last month. Brook poses with Gavin inside the O2 arena after announcing their world title fight. Brook had been desperate to face Khan at Wembley in June but his compatriot ruled out a fight until at least later in the year. (...)		
BART	Kell Brook will fight Frankie Gavin at the O2 in London on May 30. The welterweight bout has been added to a card that includes world title challenges for Kevin Mitchell and Lee Selby. Anthony Joshua faces his toughest test to date against Kevin Johnson. Click here for more boxing news.	88.6%	0.09
EBR	Kell Brook will fight Frankie Gavin on May 30 at the O2 in London. The welterweight bout has been added to a card that includes world title challenges for Kevin Mitchell and Lee Selby. Anthony Joshua faces his toughest test to date against Kevin Johnson. Brook had been desperate to face Amir Khan at Wembley in June.	**97.3%**	**−3.67**

Summary examples

	Text	Cons ↑	E ↓
Source (XSum)	Many Sephardic Jews were killed, forced to convert to Christianity or leave at the end of the 15th Century. Parliament paved the way for a change in citizenship laws two years ago, but the move needed Cabinet approval. From now on, descendants of Sephardic Jews who can prove a strong link to Portugal can apply for a passport. Proof can be brought, the government says, through a combination of surname, language spoken in the family or evidence of direct descent. Thousands of Sephardic Jews were forced off the Iberian peninsula, first from Spain and then from Portugal. Some of those who fled to other parts of Europe or to America continued to speak a form of Portuguese in their new communities. The Portuguese government acknowledges that Jews lived in the region long before the Portuguese kingdom was founded in the 12th Century. (...)		
BART	Portugal has approved a law that will allow descendants of Jews who fled the country to become citizens.	86.8%	1.48
EBR	The Portuguese government has approved a law that will allow descendants of Jews who fled to Portugal to become citizens.	**93.1%**	**1.15**

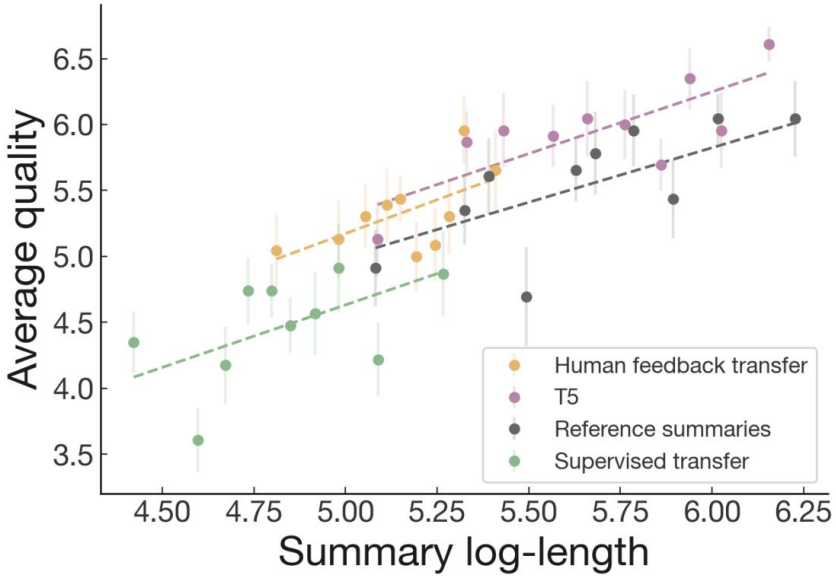
Summarization with GPT3.5

Human feedback is the key



Stienon et al., NeurIPS 2020

Reddit TL;DR → CNN/DailyMail



Stienon et al., NeurIPS 2020

Limitations

- 6.7B parameters (vs. 406M in BART and 770M in T5);
- RL training is slow (320 GPU-days);
- Expensive data collection process (“thousands of labeler hours”);
- No systematic evaluation of hallucinations (yet).

Conclusion

Take-home messages

- **Have a good summarization metric?** You can **use our approach** to generate better abstractive summaries!

- However, **current metrics struggle at detecting hallucinations** and therefore models trained with these metrics suffer from the same problem.

- **Human feedback helps,** but it is still **extremely costly.**

Public resources

- Paper: https://arxiv.org/abs/2210.15553
- Code and trained models: https://github.com/Priberam/SummEBR

Thank you

We're hiring!

priberam.com/careers