Towards Summary Candidates Fusion (EMNLP 2022) - Mathieu Ravaut, Shafiq Joty, Nancy F. Chen

Presentation Author
Mathieu Ravaut

Nanyang Technological University

Table of content

SOTA Summarization

State-of-the-art in abstractive summarization relies on :

- A pre-trained encoder-decoder network like PEGASUS [13] or BART [4].
- There is large variance in quality among beam search candidates generated by these models.
- Second-stage methods provide noticeable extra gains :
  - **Re-rank** candidates with another model : SimCLS [6], SummaReranker [9].
  - Fine-tune the model with a **different loss** (in another fine-tuning round) like SeqCo [12], ConSum [10], BRIO [7].

Motivation

All the previous methods so far are bounded by the base model oracle.

To break the oracle barrier and allow the model to depart from first-stage summary candidates, we use a second-stage generation model which **fuses** first-stage candidates.

Fusing entire summary candidates has never been explored so far.

Model

How to design the model ?

- Concatenating the source document with all first-stage candidates is impossible (exceeds input length).
- We follow the **Fusion-In-Decoder** idea [2] :
  - Encode the source and each summary candidate separately (encoding is fast).
  - Concatenate the representations.
  - The decoder attends to a unique, long context vector of concatenated representations.
- Include candidate-level information :
  - **Special token** for candidate position.
  - **Classification head** to predict whether the candidate is the oracle or not (binary classification like SummaReranker [9]).

Architecture

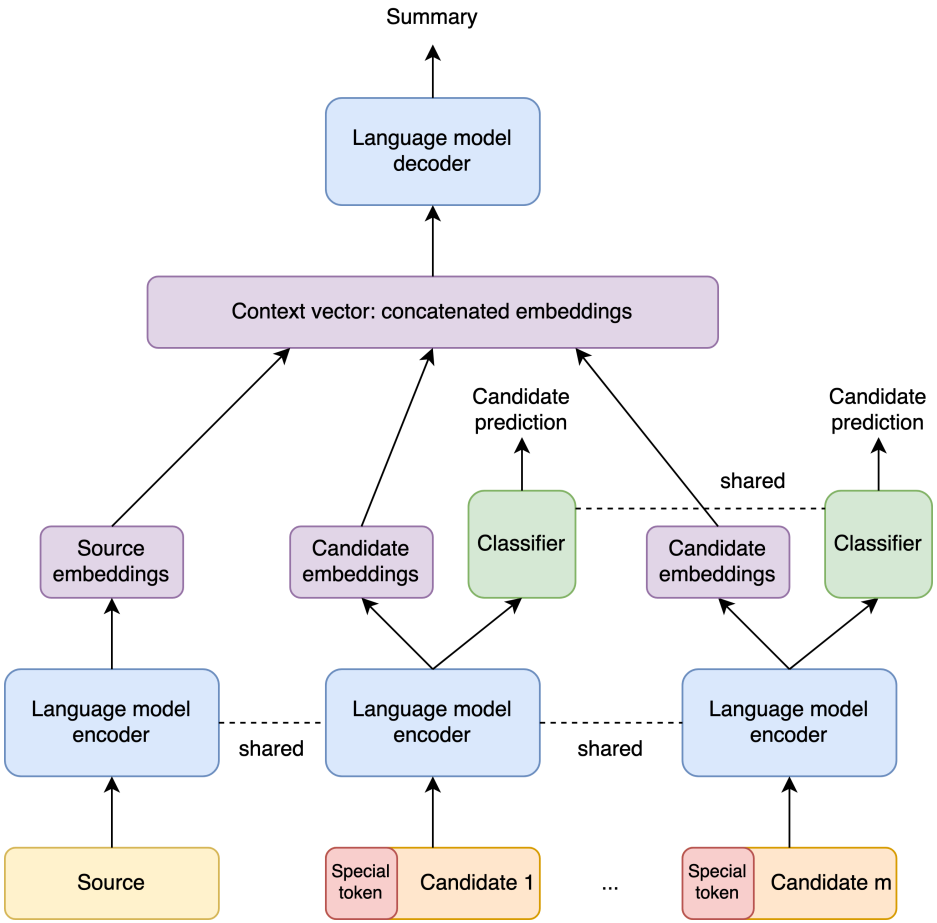

Figure – **SummaFusion model architecture.**

Architecture

In practice :

- We generate candidates with a PEGASUS-large [13] model.
- Candidates are generated through **diverse beam search** [11] to enhance diversity and push the oracle.
- We sample 15 candidates per data point.
- The language model encoder-decoder is a BART [4].
- Classifier conditions on the source and the candidate representations, and is a 2-layer MLP with ReLU.

Input Dropout

During training we use **input dropout** :

- Randomly shuts down the source or some of the candidates so that the model does not only rely on either.
- **Source dropout** : replace the source with a placeholder token with probability p_{src}.
- **Candidates dropout** : subsample with uniform probability $k \in \{2, \ldots, m\}$ summary candidates to keep, replacing the other $m - k$ by placeholder tokens.

Scope

Our experimental scope comprises :

- 3 summarization datasets with very high abstractiveness :
  - XSum [8].
  - Reddit-TIFU [3].
  - SAMSum [1].
- 2 BART backbone models :
  - SummaFusion-base with BART-base.
  - SummaFusion-large with BART-large.
- 2 training setups :
  - Full-shot.
  - Few-shot : 10-shot, 100-shot, and 1000-shot.

Full-shot results

Model	Stage	Candidates	XSum R-1	XSum R-2	XSum R-L	XSum Gain (%)	Reddit TIFU R-1	Reddit TIFU R-2	Reddit TIFU R-L	Reddit TIFU Gain (%)	SAMSum R-1	SAMSum R-2	SAMSum R-L	SAMSum Gain (%)
PEGASUS [13]	1	8	**47.21**	**24.56**	**39.25**	_	26.63	9.01	21.60	_	_	_	_	_
PEGASUS (ours)	1	15	46.78	23.77	38.70	_	25.67	8.07	20.97	_	50.31	26.20	42.06	_
PEGASUS (ours) - random	1	15	42.95	19.64	34.14	_	23.63	6.58	19.07	_	45.70	20.27	35.72	_
PEGASUS (ours) - *oracle*	1	15	*56.76*	*34.46*	*50.18*	*29.42*	*35.94*	*14.42*	*30.24*	*47.30*	*62.39*	*39.66*	*54.69*	*32.21*
PEGASUS (ours) + **SummaFusion-base**	2	15	46.16	23.55	38.53	-0.93	27.52	9.01	22.23	7.40	50.03	25.89	42.12	-0.43
PEGASUS (ours) + **SummaFusion-large**	2	15	47.08	24.05	38.82	0.63	**30.08**	**10.48**	**23.99**	17.98	**52.40**	**27.72**	**43.73**	4.48

Table – ROUGE **results** on the three datasets in full-shot. **Gain** is the relative gain over the mean of ROUGE-1, ROUGE-2 and ROUGE-L from our own PEGASUS baseline.

Strong relative improvements on Reddit-TIFU and SAMSum.

In the following, we stick to BART-large as backbone.

Few-shot results

Model	10-shot			100-shot		
	R-1	R-2	R-L	R-1	R-2	R-L
XSum						
PEGASUS [13]	19.39	3.45	14.02	39.07	16.44	31.27
PEGASUS (ours)	19.38	3.31	13.60	**39.90**	16.86	31.67
PSP [5]	_	_	_	32.50	10.83	25.03
SummaFusion-large	**30.41**	**9.92**	**22.93**	39.86	**17.01**	**31.68**
Reddit TIFU						
PEGASUS [13]	*15.36*	*2.91*	*10.76*	*16.64*	*4.09*	*12.92*
PEGASUS (ours)	18.39	3.34	13.23	22.82	5.85	17.88
SummaFusion-large	**20.79**	**4.77**	**14.58**	**26.09**	**7.51**	**20.22**
SAMSum						
PEGASUS (ours)	28.47	8.59	22.87	42.09	16.85	33.54
SummaFusion-large	**32.00**	**9.93**	**24.59**	**44.41**	**18.84**	**35.04**

Table – **Few-shot** ROUGE **results.**

SOTA in 10-shot and 100-shot summarization.

Results

Overview of results :

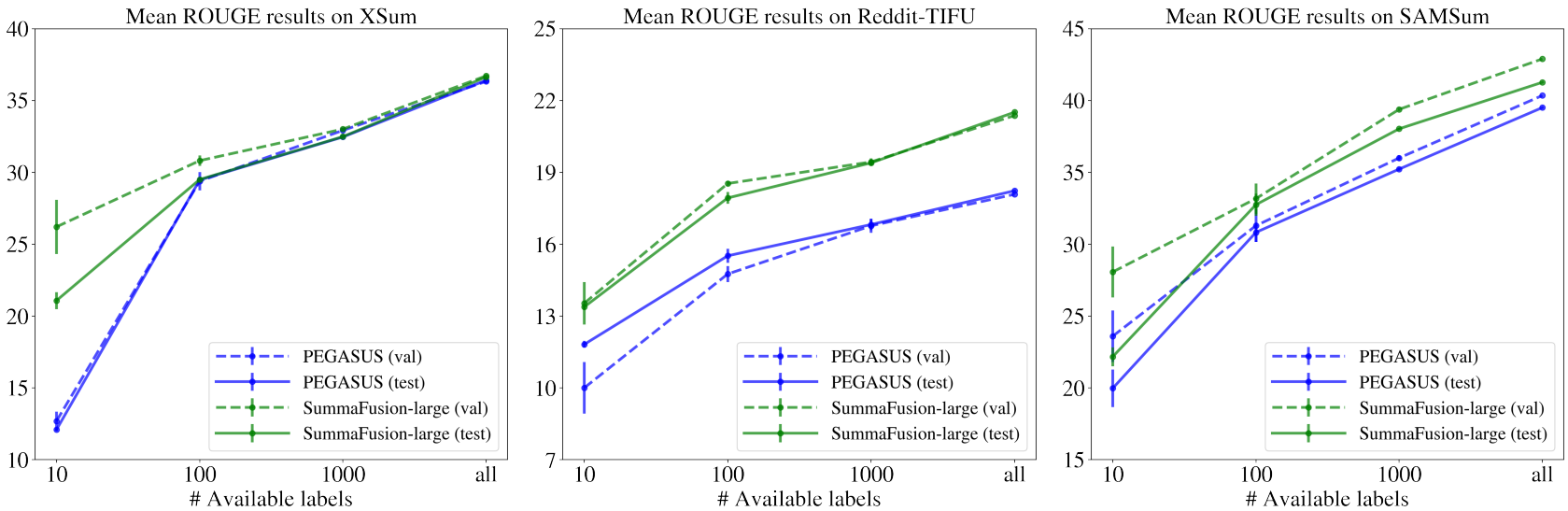

Figure – ROUGE **results** on the three datasets.

Architecture

Example summary on Reddit-TIFU :

Source document :
this happen yesterday afternoon. i been trying to dual boot mac os and windows on my wife's macbook pro. it's a late 2011 model so support from apple is almost nonexistent which is great when they wanted to charge me chat with them. i convinced them to a free chat and learned that apparently my hardware is to out dated to have boot camp make a bootable usb. boot camp assistant on this macbook only does cd iso img.
...
3 : delete the wrong hard dive part and corrupt the hard drive and have to re format the whole computer and lose every file that was saved on her computer since 2011. i got yelled at for a good hour. i knew it was my fault but at the same time... how in the world have you not backed your things in 4 years !

Summary candidates (PEGASUS with diverse beam search) :
1 : got yelled at for 4 years.
2 : i tried to back up my wife's files on her macbook pro and ended up deleting 4 years worth of data.
3 : tried to back up my wife's files on her macbook pro and ended up deleting 4 years worth of data.
4 : got yelled at by a bunch of people because i backed up my wife's files.
5 i was trying to back up my wife's files on her macbook pro when i accidentally backed it up.
...
14 : tried to back up my wife's files and ended up deleting her entire computer.
15 : fired because i backed up my wife's data without realizing it.

SummaFusion summary :
tried to *dual boot* my wife's macbook pro *with boot camp assistant* and ended up deleting 4 years worth of data.

Ground truth summary :
tried to install windows on macbook and ended up erasing everything without backing up and losing 4 years of my wife's work.

Abstractiveness

Surprisingly, SummaFusion summaries are ***not*** more abstractive :

Dataset	**Model**	***Source-abstractiveness***			***Candidates-abstractiveness***		
		1-grams	**2-grams**	**3-grams**	**1-grams**	**2-grams**	**3-grams**
XSum	Ground truth	33.79	83.29	95.51	_	_	_
	PEGASUS	27.38	76.79	91.53	_	_	_
	PEGASUS - *oracle*	*28.53*	*78.52*	*93.06*	_	_	_
	SummaFusion-large	27.18	75.71	90.94	5.46	16.62	25.78
Reddit-TIFU	Ground truth	28.77	77.43	92.48	_	_	_
	PEGASUS	12.96	57.20	78.72	_	_	_
	PEGASUS - *oracle*	*14.19*	*60.92*	*82.86*	_	_	_
	SummaFusion-large	10.26	48.85	69.62	21.23	46.31	59.88
SAMSum	Ground truth	34.13	79.31	90.51	_	_	_
	PEGASUS	23.23	65.95	79.74	_	_	_
	PEGASUS - *oracle*	*25.32*	*69.82*	*83.48*	_	_	_
	SummaFusion-large	22.43	63.11	76.8	3.35	11.37	18.87

Table – **Abstractiveness**. Novel n-grams on all datasets.

Model Ablation

Each model component contributes to the performance :

Model	R-1	R-2	R-L	New source 2-grams	New candidates 2-grams
SummaFusion-large	**29.95**	**10.32**	**23.86**	49.03	46.04
- candidates classification	29.16	10.09	23.34	48.46	42.12
- input dropout	29.23	10.31	23.35	46.58	46.06
- position token	29.09	10.23	23.29	45.98	44.59
Concat-baseline	27.26	8.74	21.93	64.50	13.27
PEGASUS	25.67	8.07	20.97	57.20	_

Table – **Ablation study** on Reddit-TIFU.

Human Evaluation

Summary	Overall preference	Reasons: More informative	Reasons: More fluent or grammatical	Reasons: More factually consistent
		XSum		
PEGASUS	15.33 (6.11)	6.67 (2.52)	3.00 (5.20)	8.67 (3.79)
SummaFusion-large	**24.33** (7.57)	**11.67** (2.31)	**8.33** (8.74)	**9.00** (2.65)
Tie	10.33 (8.08)	_	_	_
		Reddit-TIFU		
PEGASUS	9.67 (1.53)	5.33 (0.71)	1.00 (1.41)	5.00 (0.71)
SummaFusion-large	**30.67** (3.79)	**24.33** (6.36)	**6.00** (2.83)	**16.67** (10.61)
Tie	9.67 (4.93)	_	_	_
		SAMSum		
PEGASUS	14.67 (1.53)	9.33 (1.15)	0.33 (0.58)	8.33 (2.52)
SummaFusion-large	**26.00** (4.36)	**18.67** (1.53)	**3.33** (1.15)	**12.00** (10.82)
Tie	9.33 (2.89)	_	_	_

Table – **Human evaluation** on all datasets. We show mean counts over three humans rating 50 data points in each dataset, with standard deviation in parenthesis.

Breaking the oracle barrier

SummaFusion can surpass the 1st-stage oracle :

Number of candidates (m)	**Available supervision** 10-shot	100-shot	1000-shot	all data
$m = 5$	45.78%	30.51%	30.08%	37.04%
$m = 10$	27.33%	21.34%	21.44%	28.66%
$m = 15$	17.50%	17.57%	16.72%	24.87%

Table – **SummaFusion surpassing the first-stage oracle** counts (as percentages) on Reddit-TIFU.

Especially with less candidates and less supervision.

Fine-grained analysis

Breaking data points across 4 features :

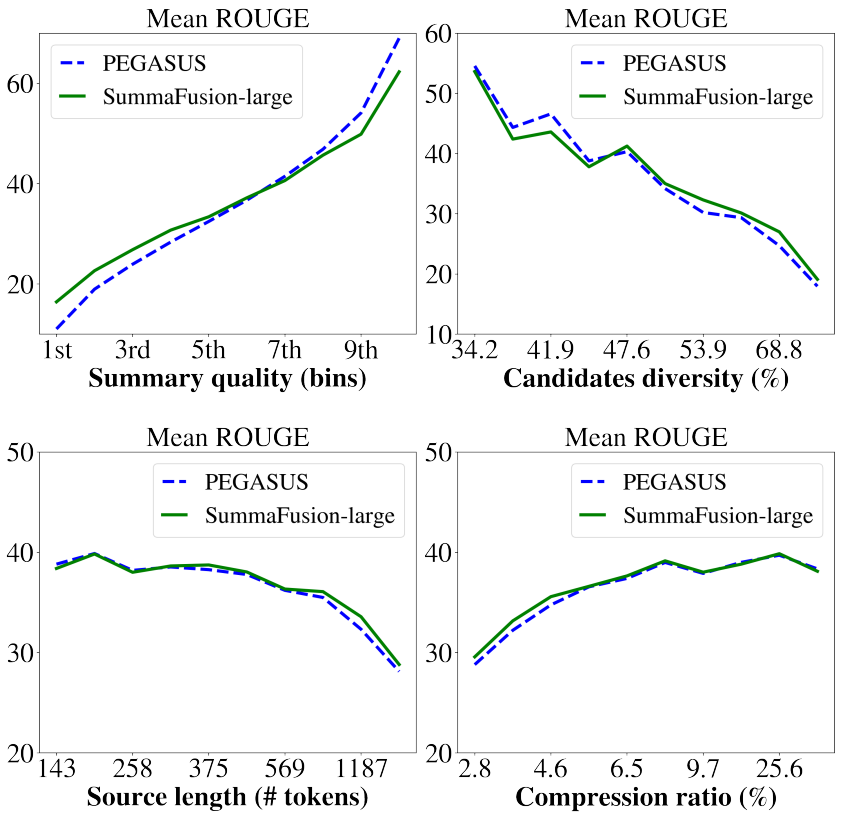

Figure – **Fine-grained analysis** on XSum.

Fine-grained analysis

Breaking data points across 4 features :

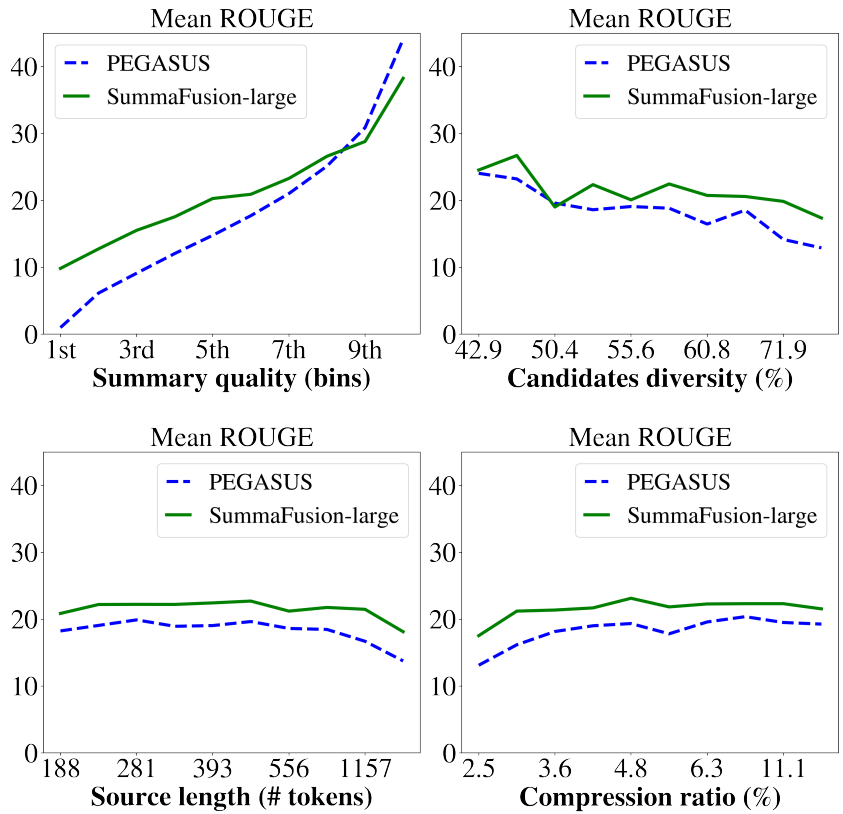

Figure – **Fine-grained analysis** on Reddit-TIFU.

Fine-grained analysis

Breaking data points across 4 features :

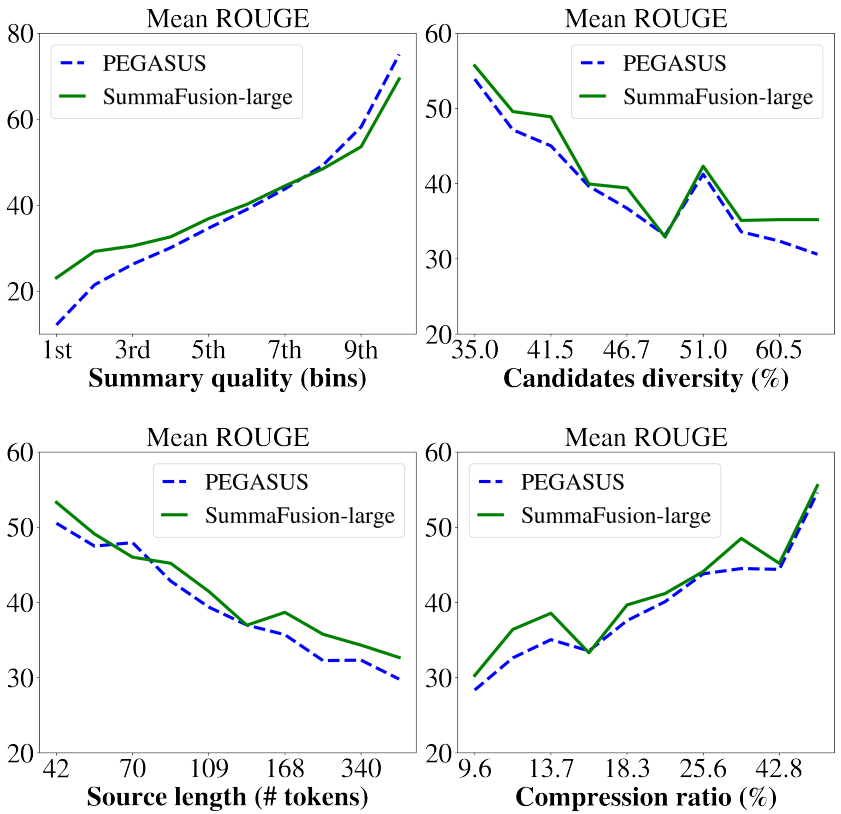

Figure – **Fine-grained analysis** on SAMSum.

Conclusion

We introduced **SummaFusion** :

- New paradigm in second-stage summarization : **fusion** of summary candidates.
- Encodes PEGASUS diverse beam search summaries as well as the source independently with another BART, concatenates representation from the source and each candidate into a unique long context vector.
- Improves a lot on *complicated* data points :
  - Few-shot setups.
  - Low-quality base candidates.
  - More diverse candidates.
  - More compressing data points.
- Humans prefer SummaFusion summaries over PEGASUS.

[1] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus : A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv :1911.12237*, 2019.

[2] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv :2007.01282*, 2020.

[3] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv :1811.00783*, 2018.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*, 2019.

[5] Xiaochen Liu, Yu Bai, Jiawei Li, Yinan Hu, and Yang Gao. Psp : Pre-trained soft prompts for few-shot abstractive summarization. *arXiv preprint arXiv :2204.04413*, 2022.

[6] Yixin Liu and Pengfei Liu. Simcls : A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv :2106.01890*, 2021.

[7] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio : Bringing order to abstractive summarization. *arXiv preprint arXiv :2203.16804*, 2022.

[8] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary ! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv :1808.08745*, 2018.

[9] Mathieu Ravaut, Shafiq Joty, and Nancy F Chen. Summareranker : A multi-task mixture-of-experts re-ranking

framework for abstractive summarization. *arXiv preprint arXiv :2203.06569*, 2022.

[10] Shichao Sun and Wenjie Li. Alleviating exposure bias via contrastive learning for abstractive text summarization. *arXiv preprint arXiv :2108.11846*, 2021.

[11] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search : Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv :1610.02424*, 2016.

[12] Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11556–11565, 2022.

[13] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus : Pre-training with extracted gap-sentences for

abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.