Transfer-Learning based on Extract, Paraphrase and Compress Models for Neural Abstractive Multi-Document Summarization

Yllias Chali and Elozino Egonmwan
University of Lethbridge
4401 University Drive
Lethbridge, Alberta, Canada
yllias.chali@uleth.ca and elozino.egonmwan@uleth.ca

Abstract

Recently, transfer-learning by unsupervised pre-training and fine-tuning has shown great success on a number of tasks. The paucity of data for multi-document summarization (MDS) in the news domain, especially makes this approach practical. However, while existing literature mostly formulate unsupervised learning objectives tailored for/around the summarization problem we find that MDS can benefit directly from models pre-trained on other downstream supervised tasks such as sentence extraction, paraphrase generation and sentence compression. We carry out experiments to demonstrate the impact of zero-shot transfer-learning from these downstream tasks on MDS. Since it is challenging to train end-to-end encoder-decoder models on MDS due to i) the sheer length of the input documents, and ii) the paucity of training data. We hope this paper encourages more work on these downstream tasks as a means to mitigating the challenges in neural abstractive MDS.

1 Introduction

Text summarization aims at presenting salient points of a text, concisely and fluently. In MDS the sources of text albeit multiple, convey a central idea or topic. For example, news article from different sources on a defined topic (Hong et al., 2014), questionnaires completed by various individuals (Luo and Litman, 2015; Luo et al., 2016) or varied reviews from different users on a certain product (Gerani et al., 2014). This paper addresses summarization of multiple news articles.

Despite the applications of MDS, not much neural-based approaches (Jin et al., 2020; Zhang et al., 2019; Liu et al., 2018; Lebanoff et al., 2018; Zhang et al., 2018a) exist in literature due to two main challenges – the lack of enormous parallel training data and the lengthy size of the input documents. The latter makes it especially challenging to encode and decode in an end-to-end fashion owing to memory constraints of the machine (Fabbri et al., 2019). To solve both problems, we propose transfer-learning (Dai et al., 2007) from pre-trained supervised models. First, we extract salient sentences by directly applying a pre-trained extractive summarization model. Next, we implement key abstraction techniques such as paraphrase generation (Gupta et al., 2018; Egonmwan and Chali, 2019a) and sentence compression (Filippova et al., 2015) on the extracted sentences using supervised pre-trained models to generate abstractive summaries. In contrast to existing works that require further adaptation (Zhang et al., 2019, 2018a; Lebanoff et al., 2018) of the pre-trained models, our transfer-learning method is direct and requires no MDS training data. Our main contributions are highlighted as follows: (1) We present a method for transfer-learning from transformer-based models pre-trained on downstream tasks, (2) We demonstrate the utility of downstream tasks, such as sentence extraction, paraphrase generation and sentence compression on MDS, and (3) Our method is simple and requires no MDS training data.

2 Methodology

Our method investigates how models tailored specifically for downstream tasks pre-trained on their dedicated labelled datasets can be directly beneficial for MDS. We investigate the utility of three (3) downstream tasks for this experiment – sentence extraction, paraphrase generation and sentence compression. Additionally, we compare our method against the performance of two (2) recent pre-trained language models – GPT2 and T5.

2.1 Extract, Paraphrase and Compress

This approach is motivated by the way humans generate summaries by highlighting salient points and re-writing in "own words". In fact, this concept

Proceedings of the 17th International Natural Language Generation Conference, pages 213–221
September 23–27, 2024.

is familiar in literature (Chen and Bansal, 2018; Gehrmann et al., 2018; Liu et al., 2018; Hsu et al., 2018). More-so, our method helps to address the challenges in training neural abstractive MDS models such as paucity of training data and the sheer length of input documents. We refer to this transfer-learning pipeline approach as EXPARCOM.

2.1.1 Sentence Extraction

First, we identify the most salient parts of the document, similar to text highlighting by humans. In tune with our transfer-learning focus, we use the pre-trained extractive summarization model of Zhong et al. (2020) – MATCHSUM [1] in zero-shot settings. The main idea behind MATCHSUM is that a good summary should be more semantically similar as a whole to the source document than the unqualified summaries (Zhong et al., 2020). Hence, the extractive summarization problem is formulated as one of semantic text matching between a set of candidate summaries and the document. The candidate summaries are obtained through a content selection module – BERTSUM (Liu and Lapata, 2019b), that pre-selects salient sentences. To obtain the candidates from these pre-selected sentences, Zhong et al. (2020) generates all combinations of sel sentences subject to the pre-selected sentences, and re-organize the order of sentences according to the original position in the document, arriving at a total of $\binom{n}{sel}$ candidate sets, where n is the number of pre-selected sentences and sel is the desired number of sentences to form the candidate summary. sel is subjectively chosen based on the statistics of the dataset (see section 3.1). A Siamese-BERT architecture is then constructed to match the document and each candidate summary. We refer readers to the literature on MATCHSUM by Zhong et al. (2020) for more details.

2.1.2 Sentence Paraphrasing

Research has shown gains in paraphrasing extracted document sentences as abstracts, either by training encoder-decoder models on extracted summarization sentences (Cao et al., 2018) or leveraging the abundance of data from machine-translation (Wieting and Gimpel, 2017; Mallinson et al., 2017) to back-translate the sentences. Inspired by such research and our transfer-learning goal, we utilize the pre-trained paraphrase generation model of Krishna et al. (2020) – STRAP[2]. STRAP (**S**tyle **Tra**nsfer via **P**araphrasing) generates diverse paraphrases by fine-tuning GPT2 (Radford et al., 2019) language model on paraphrase data. Because this is a single sentence-level model, we split the extracted output from section 2.1.1 into single sentences with document markers per sentence [3].

2.1.3 Sentence Compression

Xu and Durrett (2019); Desai et al. (2020) demonstrated that sentence extraction with compression improves the conciseness of summaries. This experiment has mostly been implemented for single document summarization (SDS) by training the sentence compression model to map a sentence selected by the extractive model to a sentence in the summary (Zhang et al., 2018b). Moreover, the gains of sentence compression for summarization would be more evident in MDS due to the lengthy nature of the source documents. In line, with our transfer-learning objective we use the pre-trained sentence compression model of Malireddy et al. (2020) – SCAR. SCAR is an unsupervised autoencoder-based model for deletion-based sentence compression primarily composed of two (2) encoder-decoder pairs – a compressor and a reconstructor. The compressor masks the input, and the reconstructor tries to regenerate it (Malireddy et al., 2020). In EXCOMPAR, the input to this pre-trained compression model are the sentence paraphrases from section 2.1.2.

2.1.4 Ablation Studies

To investigate the impact of each of these pre-trained models (2.1.2 - 2.1.3) on MDS, we conduct ablation test. Given the extractive summaries, we apply paraphrase generation only (EXPAR), sentence compression only (EXCOM), paraphrase+compression (EXPARCOM) and compression+paraphrase (EXCOMPAR).

3 Experiments

3.1 Datasets

DUC 2004 (Paul and James, 2004): This is a test corpus provided by NIST for Task 2 – Multi-document summarization. It contains 50 document

[1] https://github.com/maszhongming/MatchSum

[2] https://github.com/martiansideofthemoon/style-transfer-paraphrase

[3] this way, we know what sentences initially belonged to what documents, similar to the approach in Fabbri et al. (2019); Lebanoff et al. (2018); Liu et al. (2018).

clusters, with 10 documents per cluster. The documents contain about 4,600 words spanning 173.15 sentences on an average while the summaries consist of about 110 words and 5 sentences.

MULTINEWS (Fabbri et al., 2019): This dataset contains about 2 – 10 documents per document cluster. The documents contain about 2,100 words spanning 82.73 sentences on an average while the summaries consist of about 264 words and 10 sentences[4].

Table 1: Statistics of the MDS dataset test samples.

	MULTINEWS	DUC04
Avg. #words/psg.	2100	4600
Avg. #words/summ.	264	173

3.2 Baselines

We implement two (2) additional baselines for comparison.

3.2.1 Fine-tuning GPT2 LM for MDS

Lack of coherency/fluency is a challenge in text summarization (Christensen et al., 2013). Since LMs like GPT2 are great at generating syntactically coherent text (Radford et al., 2019) we attempt to leverage this ability in generating coherent summaries for MDS. Besides, similar to LM, the task of text summarization can be expressed in a probabilistic framework as – $p(summary|document)$, that is, learning the conditional distribution of a summary given some document(s).

Training Details We transform the *{document, summary}* pairs into a contiguous sequence of texts suitable for the GPT2 LM model by appending each summary to its source document article along with a delimiter (Khandelwal et al., 2019; Radford et al., 2018). Similar to Radford et al. (2019), we use Top-k random sampling (Fan et al., 2018) with k=2 to reduce repetition and encourage abstractiveness. We use a batch size of 10[5]. We observe that fine-tuning the GPT2 model tends to exhibit a tendency referred to as *catastrophic forgetting* (Kirkpatrick et al., 2017) leading to overfitting (Chen et al., 2019). Hence, similar to Khandelwal et al. (2019) we train 3000 randomly chosen with token length less than 1024 for 5 epochs with 32 `gradient_accumulation_steps` and a learning rate of 5e-5.

3.2.2 Zero-shot transfer of T5 model to MDS

Raffel et al. (2020) proposed a unified framework – **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer (T5) that converts text-based language problems into a text-to-text format. The model was pre-trained on an enormous English text corpora and fine-tuned on a variety of downstream tasks, including abstractive SDS. We investigate the zero-shot ability of this model by directly applying it on MDS data.

3.3 Evaluation

We measure the performance of our models by automatic evaluation using ROUGE[6] metric (Lin, 2004). Additionally, we also perform human evaluation to confirm the performance of our three (3) top models by ROUGE. We design the following Amazon MTurk experiment: we randomly select 50 samples (Luo et al., 2019) from the DUC 2004 and MULTINEWS and ask the human testers (3 per sample) to rank between outputs. We presented the testers[7] with the reference summary and our system's summary, X, of each model. The testers were required to scale (1 – 5, with 5 being of superior quality to 1) the system's output on *informativeness* (how well does it cover the information in the reference summary?), *fluency* (how well does the information in the systems summary flow?) and *non-redundancy* (how well are information not being repeated?). Results are presented in Table 2 and 3.

3.4 Results Analysis

From Table 2, we notice an increase in ROUGE points from model **ex** to EXPARCOM. On average, EXPARCOM had a performance gain of 2.9% and 3.7% for DUC 2004 and MULTINEWS respectively over EX, with an average of about 7.61%, 21.55% and 70.84% of the gain coming from the compression, paraphrase and compression+paraphrase (and paraphrase+compression) modules respectively, across both datasets. We observe higher gain/performance in MULTINEWS corpus because one of the pre-trained models – BERTSUM, used for sentence extraction was fine-tuned on MULTINEWS. The percentage contribution of each of these modules to EXPARCOM, proves that

[4]based on these statistics, we choose $sel = 6$ and $sel = 9$ for DUC04 and MULTINEWS respectively in MATCHSUM .

[5]due to memory constraints of our machine

[6]https://github.com/andersjo/pyrouge/tree/master/tools/ROUGE-1.5.5

[7]We selected testers who were located in US or Canada, have Mechanical Turk Masters qualification and had HIT approval rate greater than or equal to 95%.

Table 2: Average **ROUGE-F1** (%) scores (with 95% confidence interval) of various MDS models on the **DUC04** and **MULTINEWS** test sets. The first section reports published models while the second section reports our's.

DUC 2004	R-1	R-2	R-SU4	MULTINEWS	R-1	R-2	R-SU4
(Lebanoff et al., 2018)	36.42	9.36	**13.23**	(Jin et al., 2020)	46.00	16.81	20.09
(Zhang et al., 2018a)	36.70	7.83	12.40	(Zhang et al., 2019)	**47.52**	**18.72**	**24.91**
(Fabbri et al., 2019)	35.78	8.90	11.43	(Fabbri et al., 2019)	43.47	14.89	17.41
EX	36.52	9.27	11.85	EX	46.20	16.51	19.43
EXCOM	36.70	9.39	11.87	EXCOM	46.02	16.53	19.47
EXPAR	36.77	9.48	11.85	EXPAR	46.25	16.55	19.50
EXCOMPAR	36.89	**9.79**	11.94	EXCOMPAR	46.61	16.78	20.15
EXPARCOM	**37.08**	9.59	12.34	EXPARCOM	47.15	16.93	20.86
GPT2	24.71	3.66	6.30	GPT2	27.60	5.49	10.22
T5	27.21	4.84	6.61	T5	30.01	7.16	12.38

Table 3: Human Evaluation scores of our top 3 models based on Informativeness, Fluency and Non-Redundancy against some existing models.

Models	Informativeness	Fluency	Non-Redundancy
EXPAR	3.31	3.10	3.28
EXCOMPAR	3.59	3.22	3.38
EXPARCOM	**3.61**	**3.33**	**3.43**
PG-MMR (Lebanoff et al., 2018)	3.52	3.24	3.42
(Zhang et al., 2019)	2.19	2.03	1.88

while only paraphrase generation or sentence compression applied over extracted sentences improves performance, a decoupled pipe-lined application of both paraphrase and compression yields better improvements. Table 2 shows that our models are competitive with existing abstractive MDS models. On the quality of the summaries generated, we observed that although GPT2 generated fluent summaries, they mostly contained hallucinations. We deduce that the GPT2 model is not fully capable of using a substantial part of the source (especially for long input documents) but rather behaves like a general domain LM. The T5 model is able to generate faithful summaries, but however starts to suffer from repetition and lack of fluency at some point. The EXPARCOM model displayed abstractive quality as some novel words were introduced while being concise. Tables 2 and 3 show that paraphrase generation and sentence compression improve the quality of summaries, giving credence to the utility of transfer-learning/combination of these specific tasks for MDS. From Table 2, we observe an average increase of about 0.2 R-1 points on top each previous output when each of paraphrase and/or compression module is added.

4 Related Work

Our work is related to paradigms such as Extract-and-Compress (Desai et al., 2020; Xu and Durrett, 2019; Mendes et al., 2019); Extract-and-Paraphrase (Egonmwan and Chali, 2019b; Chen and Bansal, 2018; Hsu et al., 2018). However it differs significantly from these models in the following ways: i) it requires zero training on data for the task it is being applied – MDS ii) it requires no architectural changes or augmentations to the pre-trained models iii) it consists of three (3) pipe-lined downstream tasks instead of two (2) in comparison to existing work. The simplicity of the pipeline enables various instantiations. Despite, its simplicity it is competitive with current state-of-the-art MDS models – PEGASUS (Zhang et al., 2019) and MGSUMABS (Jin et al., 2020) which are specifically tailored for abstractive text summarization with lots of parameters. Zhang et al. (2019) proposed a large transformer-based encoder-decoder model pre-trained on a massive text corpora with new self-supervised objective and fine-tuned on a variety of summarization datasets. Jin et al. (2020) proposed a multi-granularity interaction network that encodes semantic representations for documents, sen-

tences, and words for MDS. Lebanoff et al. (2018) and Zhang et al. (2018a) adapt the neural model trained on abstractive SDS for MDS. Our methods utilize transfer-learning from tasks/models not specifically engineered for abstractive summarization yet yields impressive results.

Existing MDS methods are mostly extractive. These extractive methods are majorly modelled as graph operations with peculiarities on edge weight assignment. Yasunaga et al. (2017) recently proposed a Graph Convolutional Neural (GCN) network with sentence embeddings obtained from RNNS as input node features.

Abstractive MDS on the other hand, has met with limited research due to data limitations. Liu and Lapata (2019a) proposed a neural model which is capable of encoding multiple input documents hierarchically. Liu et al. (2018) handled MDS in two stages – extract and abstract. Abstraction was performed by a decoder-only sequence transduction model. Our approach is much similar to Lebanoff et al. (2018) and Zhang et al. (2018a) that adapt the neural model trained on SDS for MDS by fine-tuning on the MDS dataset. We use the SDS model as-is in the extractive stage, making no changes to the encoder or decoder. Additionally, different from their methods, we incorporate other downstream tasks like paraphrasing and sentence compression.

5 Conclusion

We demonstrated the utility of sentence extraction, paraphrase generation and sentence compression for MDS. We show that these tasks need not be pre-trained on abstractive summarization corpora or with abstractive summarization learning objectives. We hope this paper serves as a test bed for experiments in MDS driven by transfer-learning and encourages similar approach to related tasks and for problems with limited training data.

6 Acknowledgments

The research reported in this paper was conducted at the University of Lethbridge and supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada discovery grant and the NSERC Alliance - Alberta Innovates Advance Program grant. We thank also the anonymous reviewers for their comments.

References

Yash Atri, Arun Iyer, Tanmoy Chakraborty, and Vikram Goyal. 2023. Promoting topic coherence and inter-document consorts in multi-document summarization via simplicial complex and sheaf graph. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2154–2166, Singapore. Association for Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *Advances in Neural Information Processing Systems*, pages 1908–1918.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

Janara Christensen, Stephen Soderland, Oren Etzioni, et al. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200.

Shrey Desai, Jiacheng Xu, and Greg Durrett. 2020. Compressive Summarization with Plausibility and Salience Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6259–6274.

Elozino Egonmwan and Yllias Chali. 2019a. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255.

Elozino Egonmwan and Yllias Chali. 2019b. Transformer-based model for single documents neural summarization. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 70–79.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *arXiv preprint arXiv:1905.08836*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yang Liu and Mirella Lapata. 2019b. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like her: Human reading inspired extractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3024–3034.

Wencan Luo and Diane Litman. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

Wencan Luo, Fei Liu, Zitao Liu, and Diane Litman. 2016. Automatic summarization of student course feedback. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 80–85.

Chanakya Malireddy, Tirth Maniar, and Manish Shrivastava. 2020. SCAR: Sentence Compression using Autoencoders for Reconstruction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 88–94.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference*

of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 881–893.

Alfonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André FT Martins, and Shay B Cohen. 2019. Jointly Extracting and Compressing Documents with Summary State Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3955–3966.

Over Paul and Yen James. 2004. An introduction to duc-2004. In *Proceedings of the 4th Document Understanding Conference (DUC 2004)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

John Wieting and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Jiacheng Xu and Greg Durrett. 2019. Neural Extractive Text Summarization with Syntactic Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3283–3294.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 2018a. Adapting neural single-document summarization model for abstractive multi-document summarization: A pilot study. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 381–390.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018b. Neural Latent Extractive Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Examples

Source document (truncated): speaking at a conference in sweden's third-largest city of malmö , home to a large immigrant population , the dalai lama – who won the nobel peace prize in 1989 – said europe was " morally responsible " for helping " a refugee really facing danger against their life " . " receive them , help them , educate them . . . but ultimately they should develop their own country , " said the 83-year-old tibetan who fled the capital lhasa in fear of his life after china poured troops into the region to crush an uprising . " i think europe belongs to the europeans , " he said , adding they should make clear to refugees that " they ultimately should rebuild their own country " . the dalai lama at the conference in malmö [...]
EX-PAR-COM summary: speaking at a conference in the city of malmö, home to a lot of immigrants, the dalai lama – who won the nobel peace prize – said europe was "morally responsible" for assisting "a refugee really facing danger against their life" . " receive them , help them , educate them . . . but finally they should develop their countries, " said the 83-year-old tibetan . " i think europe belongs to the europeans , " he said , adding refugees should know that " they ultimately should rebuild their own country ".
Reference summary: addressing a conference in malmö , sweden , home of many immigrants , the dalai lama said europe was " morally responsible " for helping refugees who are in danger — but that ultimately those refugees should return to their homelands . " receive them , help them , educate them ... but ultimately they should develop their own country , " he said , per the local . " i think europe belongs to the europeans , " and that refugees " ultimately should rebuild their own country , " he added . the 83-year-old buddhist spiritual leader and nobel peace prize winner fled tibet as chinese communist troops took over the area in 1959 , and settled in india , where he was granted asylum , the daily caller notes .

Figure 1: An example of a summary generated by **EXPARCOM** model. Red colored texts shows novel words generated.

Source document (truncated): the second ex-wife of peter cook has issued a written apology to his first wife — supermodel christie brinkley — for ignoring her warning about the architect' s philandering . "christie and i have talked recently and i have privately apologized to her, but, given the public nature of their divorce and custody battle, i feel a public apology is also appropriate and deserved , " suzanne shaw wrote in a letter obtained by the post thursday [...] christie ' s six-year marriage to cook imploded in 2007 after she learned of his affair with 18-year-old office clerk diana bianchi [...]
GPT2 summary: the second ex-wife of peter and jennifer davis, who was also a lawyer. The couple divorced in 2005 after the divorce court ruled that she had been cheating on him with another man for years before they separated last year.
Reference summary (truncated): – christie brinkley' s marriage to peter cook ended in 2008 after he had an affair with an 18-year-old ; he went on to wed suzanne shaw [..] shaw and cook filed for divorce this year ... after cook allegedly cheated on shaw . now shaw has apologized to brinkley for bad-mouthing her years ago [...]

Figure 2: An example of a summary generated by fine-tuning **GPT2**. Red colored texts shows hallucinated content in the summary, with no corresponding context in the source document.

Source document (truncated): a man has doused himself with gasoline and set himself on fire in st. peter 's square in vatican city . the vatican said thursday the 51-year-old man was treated at a nearby hospital and transferred to a larger one , with serious burns to the upper part of his body . a vatican statement says a jesuit priest initially tried to smother the flames with a jacket . he was later joined by two police officers who used a fire extinguisher to put out the blaze [..] the man had a piece of paper with the phone number of his daughter . [...]
T5 summary: the 51-year-old man was treated at a nearby hospital and transferred to a larger one, with serious burns to the upper part of his body . a vatican statement says a jesuit priest initially tried to smother the flames with a jacket . the man was later joined by two police officers who used a fire extinguisher to put out the blaze . police found a piece of paper with his daughter's phone number written on it at the scene '. .. ... the a. the . at around . local . and . with . severe burns . to the lower part of . his . they a large , the , to . of his a the s e a priest's 'the man . in . it a police . after
Reference summary (truncated): – a man is in " grave " condition after he doused himself in gasoline and set himself on fire in st. peter ' s square today around 8 : 30am local time , ansa reports . the 51-year-old man was treated at a nearby hospital , then transferred to a larger one , with serious burns to his upper body , the ap notes . a jesuit priest was first on the scene and threw his jacket on the man before two police officers arrived to extinguish the flames ; both were treated for smoke inhalation and injuries to their hands . the motive for the man ' s act isn ' t clear , though a piece of paper with his daughter ' s phone number on it was found nearby .

Figure 3: An example of a summary generated by the **T5** model. Red colored texts shows content with repetition and grammatical errors.

B Human Evaluation Screenshot

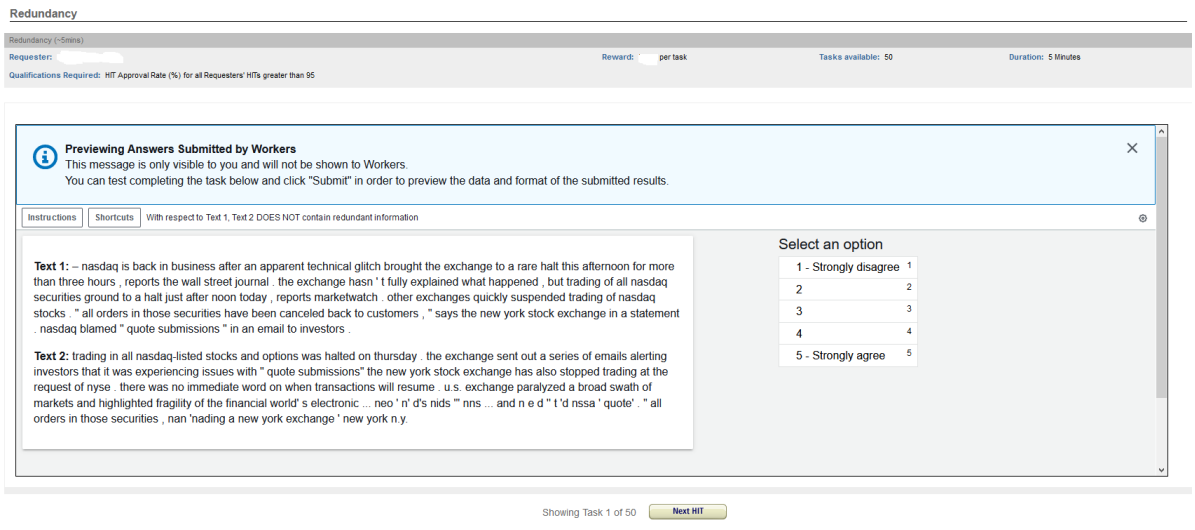

Figure 4: Screenshot of our human evaluation task on Amazon MTurk