IMPROVING NEURAL ABSTRACTIVE SUMMARIZATION USING TRANSFER LEARNING AND FACTUALITY-BASED EVALUATION: TOWARDS AUTOMATING SCIENCE JOURNALISM

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose *Automating Science Journalism* (ASJ), the process of producing a press release from a scientific paper, as a novel task that can serve as a new benchmark for neural abstractive summarization. ASJ is a challenging task as it requires long source texts to be summarized to long target texts, while also paraphrasing complex scientific concepts to be understood by the general audience. For this purpose, we introduce a specialized dataset for ASJ that contains scientific papers and their press releases from *Science Daily*. While state-of-the-art *sequence-to-sequence* (seq2seq) models could easily generate convincing press releases for ASJ, these are generally nonfactual and deviate from the source. To address this issue, we improve seq2seq generation via transfer learning by co-training with new targets: (*i*) scientific abstracts of sources and (*ii*) partitioned press releases. We further design a measure for factuality that scores how pertinent to the scientific papers the press releases under our seq2seq models are. Our quantitative and qualitative evaluation shows sizable improvements over a strong baseline, suggesting that the proposed framework could improve seq2seq summarization beyond ASJ.

1 INTRODUCTION

Neural text summarization (Rush et al., 2015) has undergone an exciting evolution recently: from extractive (Nallapati et al., 2017) through abstractive (Nallapati et al., 2016) to hybrid (See et al., 2017) models; from maximum likelihood to reinforcement learning objectives (Celikyilmaz et al., 2018; Chen & Bansal, 2018); from small to large datasets (Grusky et al., 2018) that are also abstractive (Sharma et al., 2019); from short to orders of magnitude longer sources and targets (Liu et al., 2018); from models trained from scratch to using pre-trained representations (Edunov et al., 2019; Liu & Lapata, 2019). Such evolution was largely supported by the emergence of seq2seq models (Cho et al., 2014; Sutskever et al., 2014).

These advances are yet to be challenged with a seq2seq summarization task that summarizes a *long source* to a *long target* with *extreme paraphrasing*. Below we argue that ASJ is a natural testbed for such a challenge. Science journalism is one of the few direct connections between scientific research and the general public, lead by media outlets such as *Science Daily*, *Scientific American*, and *Popular Science*. Their journalists face an incredibly difficult task: not only must they carefully read the scientific papers and write factual summaries, but they also need to paraphrase complex scientific concepts using a language that is accessible to the general public.

To emulate what a journalist would do, we present a dataset of about 50,000 scientific papers paired with their corresponding *Science Daily* press releases, and we seek to train a seq2seq model to transform the former into the latter, i.e., an input scientific paper into an output popular summary. Ideally, our model would both identify and extract the relevant points in a scientific paper and it would present them in a format that a layman can understand, just as science journalists do.

We now ask: would such a model be successful without a factual and accurate representation of scientific knowledge? Recent work suggests that even simple training of word embeddings could capture certain scientific knowledge from 3.3 million scientific abstracts (Tshitoya et al., 2019).

Therefore, here we propose to transfer knowledge from domains from which a seq2seq model would be able to extract factual knowledge using transfer learning (Caruana, 1997; Ruder, 2019).

We frame our approach as *multitask learning* (MTL). We perform co-training using both scientific abstracts and parts of the target press releases, and we view these additional domains as potential training sources for representation of scientific facts within the seq2seq model, which ideally would be helpful to ASJ. We demonstrate that MTL improves factuality in seq2seq summarization, and we measure this automatically using a novel evaluation measure that extracts random fragments of the source and evaluates the likelihood of the target given these fragments. We believe that the insights from our experiments can guide future work on a variety of seq2seq tasks.

The contributions of our work can be summarized as follows:

1. We present a novel application task for seq2seq modelling: automating science journalism (ASJ).
2. We present a novel, highly abstractive dataset for summarization for the ASJ task with long source and target texts, where complex scientific notions in the source are paraphrased and explained in simple terms in the target text.
3. We propose a transfer learning approach that significantly improves the factuality of the generated summaries.
4. We propose an automatic evaluation measure that targets factuality.

The rest of this paper is organized as follows: Section 2 discusses previous work. Section 3 describes the new data that we propose for ASJ. Section 4 presents our models. Section 5 introduces our transfer learning experiments for summarization. Section 6 describes our evaluation setup. Section 7 discusses our experiments and the results. Section 8 concludes and points to promising directions for future work.

2 Related Work

Our work rethinks the task of neural text summarization, the utilization of scientific datasets for generation and the applications of multitask learning.

Neural Text Summarization. We define automating science journalism as a text summarization task where the *source* is the scientific paper and the *target* is a press release summary about the paper, i.e., a shorter version of a full press release. Existing neural models for this task can be *abstractive*, i.e., paraphrase the source, (Nallapati et al., 2016), *extractive* (Nallapati et al., 2017), i.e., extract words, phrases, or entire sentences from the source, or *hybrid* (See et al., 2017). Unlike previous work, our task is abstractive not only to shorten the source, but also to change the technical style of the scientific papers. At the same time, we need to ensure factuality by accurately paraphrasing scientific concepts from the source.

Scientific Datasets for Generation. The task of automating science journalism has not received much attention so far, partly due to the lack of a benchmark datasets for training neural models. Dangovski et al. (2019) proposed that the abundance of scientific articles online and their press coverage provide an opportunity to develop neural models, and presented pioneering results on the *Science Daily* dataset using an RNN seq2seq model. Other work preserved the style of the source (Teufel & Moens, 2002; Nikolov et al., 2018; Cohan et al., 2018) or generated very short targets taking the form of blog titles (Vadapalli et al., 2018). However, none of the above work faced our challenging task of not only presenting relevant information, but also integrating it into articles that use popular language rather than high-level scientific style.

Multitask Learning. The nature of our task and the corresponding datasets make it possible to use recent advances in transfer learning for NLP (Ruder, 2019). Namely, we combine datasets sharing a source domain, i.e., scientific articles, with different target domains, i.e., abstracts and press releases. Thus, we propose a novel mutitask learning (MTL) setup for summarization. For this component, we take inspiration from recent work on automatically generating news articles using the *GROVER* model (Zellers et al., 2019). An important characteristic of GROVER is that it is trained on multiple variations of the same dataset. For example, in some instances, the headline might be used to generate the body, whereas in others, the body might be used to generate the headline. Similarly, via a special tag, we can signal to the decoder to generate either an abstract or a press release, or to generate the target in several steps by conditioning on the intermediate outputs. Finally, other constructions for signaling to the decoder were proposed in the context of neural machine translation (Lample & Conneau, 2019; Aharoni et al., 2019) and summarization with user preferences (Fan et al., 2017) that contain tags, similarly applied as in recent advances for pre-training contextual word embeddings (Peters et al., 2018; Delvin et al., 2019). Unlike these techniques, our task here is automating science journalism.

3 Data

Our SD Dataset. Our dataset in its original form consists of 50,305 pairs of: a full-text scientific paper as a source and a corresponding *Science Daily* press release as a target (SD). We download the scientific papers as PDF files and then convert them to raw text. We do not perform explicit pre-processing, and thus the papers do not follow any standard format (e.g., some start with the title, the abstract, or just the main body) and do not exclude extraneous symbols, characters, or spaces.

ArXiv (Cohan et al., 2018). For domain transfer, we use the *arXiv* dataset that links full-text scientific papers as sources from the arXiv database to their abstracts as targets. This dataset does not include papers that are excessively long (e.g., thesis), too short, or have no abstract structure. The dataset consists of 215K paper-abstract pairs, each paper averaging 4,938 words and each abstract averaging 220 words. The figures and the tables were removed using regular expressions to only preserve the textual content of the articles. Furthermore, math formulas and citation markers were normalized to special tokens. When compared to *Science Daily*, not only are the target texts in this dataset more extractive, but the source texts are also much more mathematical.

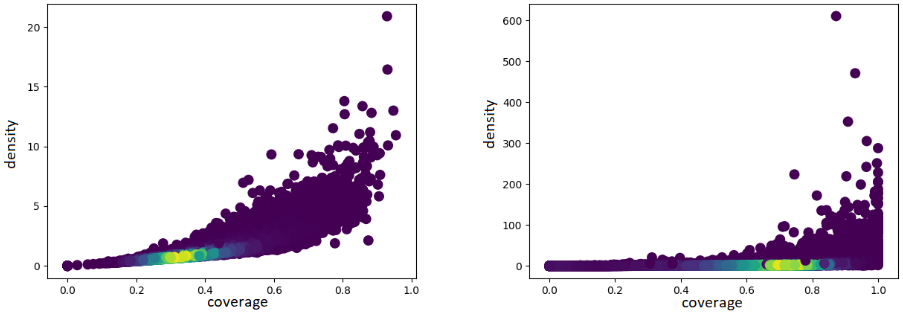

Figure 1: Plot of density and coverage of source-target pairs for the *Science Daily* dataset (left) and the *arXiv* dataset (right). Note that lighter colors indicate more source-target pairs.

Comparison and Discussion. In contrast to most summarization datasets, our *Science Daily* dataset is unique in two distinct ways: (*i*) The target summaries are more *abstractive*, i.e., they have lower coverage and density compared to the *arXiv* dataset (see Figure 2). The explicit formulas for these statistics are $\text{COVERAGE}(A, S) = (1/|S|) \sum_{f \in \mathcal{F}(A,S)} |f|$ and $\text{DENSITY}(A, S) = (1/|S|) \sum_{f \in \mathcal{F}(A,S)} |f|^2$, where $\mathcal{F}(A, S)$ is the set of extractive fragments, a sequence of tokens that is shared between the source and the target, for a set of articles $\{A\}$ and a corresponding set of summaries $\{S\}$, $|f|$ is the number of words in the fragment $|f|$ and $|S|$ is the number of words in the summary S (Grusky et al., 2018). In plain words, *coverage* represents the fraction of words that are in an extractive fragment, while *density* represents the average length of these fragments. (*ii*) Both our source and target sequences are relatively large, with each article averaging around 7,000 words and each press release averaging around 550 words. For comparison, the standard dataset *CNN/ Daily Mail* (Hermann et al., 2015; Nallapati et al., 2016) is much shorter, with sources of 800 words and targets of 50 words, and even the *arXiv* dataset has much shorter targets, with sources of 6,000 words and targets of 200 words. Figure 2 in the Appendix offers additional comparison of the coverage and the density for some popular datasets for summarization, while Figure 3 presents more information about the length of the sources in our *Science Daily* dataset.

4 Models

For the basis of our models we used the FAIRSEQ library (Ott et al., 2019), and we focused on convolutional seq2seq implementations (Gehring et al., 2017).

FCONV. Our first model is a small vanilla convolutional seq2seq model, corresponding to FAIRSEQ's ISWLT de-en, using four encoder layers with embedding dimension of 256, 256 channels, and a kernel size of 3, (256,3) for short; three decoder layers (256,3) with input and output embedding dimensions of 256. We trained the model until convergence on the dev set with a learning rate of 0.25, Nesterov accelerated gradient (NAG) descent, dropout of 0.2, and a gradient threshold of 0.1.

STORY. Our second model is a state-of-the-art model for neural story generation (Fan et al., 2018; 2019). It introduces attention (Bahdanau et al., 2015) between the output from the encoders and the decoder layers, as well as multi-head self-attention on the decoder layers (Vaswani et al., 2017) that is gated (Dauphin et al., 2017) and equipped with a multi-scale mechanism for down-sampling (Fan et al., 2018). Since our sources are three orders of magnitude larger than the writing prompt sources for which STORY has been used, we *additionally equip the encoders* with gated multi-scale multi-head self-attention.

In sum, following FAIRSEQs implementation, our model uses two (128,3) followed by a single (512,3) encoder layers with 256 embedding dimensions; four (512, 4) followed by two (768,4) followed by a single (1024, 4) decoder layers with 256 input and 256 output embedding dimensions; four gated self-attention heads both on the encoders and on thhe decoders with projected inputs and down-sampling. We trained the model until convergence on the dev set with a learning rate of 0.25, NAG, dropout of 0.2, and a gradient threshold of 1.0.

Finally, training for both FCONV and STORY took usually around 20-30 epochs, depending on the batch size, which is around 30-40.

5 Experiments

We design transfer learning experiments for the FCONV and the STORY models as follows below by constructing datasets for seq2seq summarization.

BASE. The original *Science Daily* dataset, introduced in Section 3 with a train/dev/test split of 40,247/5,029/5,029. The experiments with this dataset are baselines for our transfer learning experiments.

AB. Here, we augmented the *Science Daily* dataset with the *arXiv* dataset with specially designed tags, as follows:

1. The source is pre-pended with the tag `<begin-paper>` and appended with the tags `<end-paper> <begin-press>` for examples in *Science Daily*, and similarly we only replace `press` with `abstract` for examples in the *arXiv* dataset.
2. The target is appended with `end-press` or `end-abstract`, respectively.

Tags are used to indicate the source domain (*arXiv* or *Science Daily*) and the target domain (*abstract* or *press release*). In order to ensure equal balance between the two datasets, we took 40,000 points from their training sets, 5,000 from their test, and 5,000 from their dev, for a final train/dev/test split of 80,000/10,000/10,000. We hypothesize that the encoder layers and the decoder attention mechanism will focus on these tags while processing the source and while generating the output, respectively.

Table 1: Model FCONV and top-k sampling.

	ROUGE		
Dataset	1	2	L
BASE	39.2	9.5	36.9
AB	**41.2**	**10.2**	**38.6**
PART	32.8	7.8	31.2

Table 2: Model FCONV and beam search.

	ROUGE		
Dataset	1	2	L
BASE	39.2	10.8	37.0
AB	**41.8**	**11.6**	**38.6**
PART	31.1	9.0	29.6

PART. Augmented *Science Daily* with partitioned targets as follows:

1. For each source-target pair in *Science Daily*, we preserve the source body `body` and we divide the target into three equal parts `part-1`, `part-2` and `part-3`.
2. We construct the sources-target pairs as follows: for all bodies `body`, for indices `i` equal to 2 or 3, the source is

   ```
   <begin-body> body <end-body> <begin-part-(i-1)>
   part-(i-1) <end-part-(i-1)> <begin-part-i>
   ```

   and for `i` equal to 1, the source is

   ```
   <begin-body> body <end-body> <begin-part-i>
   ```

   where the corresponding target to the source is `part-i <end-part-i>`.
3. During inference, we generate the parts `part-i` autoregressively from `part-1` to `part-3`.

In this way, instead of training the model to generate the full press release, we train it to generate only specific sections. Namely, we make the assumption that press releases are divided roughly into three equal parts: *highlights*, *body*, and *conclusion*, which allows us to co-train with different domains of summarization and thus to transfer signals from one domain to another. Furthermore, in this way we also increase the BASE split threefold, which yields a 120,741/15,087/15,087 train/dev/test split.

Ultimately, we convert all textual data to *byte pair encoding* (BPE) (Sennrich et al., 2016) with 32,000 BPE tokens both on the source and on the target side using the guidelines by FAIRSEQ.

6 Evaluation

ROUGE. We begin by evaluating using ROUGE Lin & Hovy (2003), following FAIRSEQs convention. During inference we either use beam search with beam size five or top-k random sampling from the top 10 candidates (Fan et al., 2018). We generate 400-600 tokens for all datasets except for PART, where we generate 150-300 tokens for each part.

Results. Table 1 presents results for FCONV and for top-k sampling. AB outperforms BASE and PART by 2.0/0.7/1.7 and 8.4/2.4/7.4 ROUGE points absolute, respectively. We can see in Table 2 similar results for FCONV and beam search. AB outperforms BASE and PART by 2.6/0.8/1.6 and 10.7/2.6/9.0 ROUGE points absolute, respectively. Thus, we can conclude that co-training with abstracts improves ROUGE scores noticeably for FCONV.

Table 3 presents results for STORY and for top-k sampling. This time, PART outperforms BASE and AB by 3.9/2.8/3.8 and 1.8/1.4/1.8 ROUGE points absolute, respectively. Table 4 presents similar results for STORY and for beam search. PART outperforms BASE and AB by 3.2/2.5/3.1 and 0.0/0.4/0.3 ROUGE points absolute, respectively. Moreover, as opposed to FCONV, for STORY the experiments AB and PART perform more closely to each other and improve upon BASE. Thus, we con conclude that co-training with novel domains and partitioned targets yields sizable improvements for STORY in terms of ROUGE.

Table 3: Model STORY and top-k sampling.

	ROUGE		
Dataset	1	2	L
BASE	38.9	7.8	36.4
AB	41.0	9.2	38.6
PART	**42.8**	**10.6**	**40.2**

Table 4: Model STORY and beam search.

	ROUGE		
Dataset	1	2	L
BASE	38.2	8.5	36.0
AB	**41.4**	10.6	38.8
PART	**41.4**	**11.0**	**39.1**

Our RA Evaluation Measure. Standard procedures for evaluating abstractive summarization systems often require extensive human resources, with most papers relying on crowd-sourcing services to obtain adequate amount of data (see Section D in the Appendix for more detail). Because of these limitations, we face the challenge of meaningfully and scientifically evaluating our models without the need for human annotations.

We define a conditional model by a probability distribution $\hat{p}(\cdot|\cdot;\boldsymbol{\theta})$ (i.e., parameterized by $\boldsymbol{\theta}$) that generates autoregressively and from which the perxplexity of a target given a source could be computed. For a source $\boldsymbol{s} = (s_1, \dots, s_m)$, we use this model for generation by selecting an output

$$(h_1, \dots, h_n) = \boldsymbol{h} = \operatorname*{argmax}_{\boldsymbol{h}'} \hat{p}(\boldsymbol{h}'|\boldsymbol{s};\boldsymbol{\theta}).$$

Given that it is unfeasible to traverse the entire solution space $\{\boldsymbol{h}'\}$, we use a heuristic search procedure B (for example, beam search or top-k random sampling) designed so that

$$\boldsymbol{h} = B(\boldsymbol{s}, \hat{p}) \approx \operatorname*{argmax}_{\boldsymbol{h}'} \hat{p}(\boldsymbol{h}'|\boldsymbol{s};\boldsymbol{\theta}).$$

Furthermore, we assume that there are ground-truth distribution $p_{\mathbf{s}}(\cdot)$ and a joint distribution $p_{\mathbf{h},\mathbf{s}}(\cdot,\cdot)$, where $\mathbf{h}$ is a random variable representing the target. Then, we can measure how close $\hat{p}(\boldsymbol{h}|\boldsymbol{s})$ is to $p_{\mathbf{h},\mathbf{s}}(\boldsymbol{h},\boldsymbol{s})/p_{\mathbf{s}}(\boldsymbol{s})$. The assumption of training is that for a source-target pair $(\boldsymbol{s},\boldsymbol{t})$ in our dataset the following holds $p_{\mathbf{s},\mathbf{t}}(\boldsymbol{s},\boldsymbol{t}) = p(\boldsymbol{s})$. Therefore, for our model we want $\hat{p}(\boldsymbol{t}|\boldsymbol{s}) = p(\boldsymbol{t},\boldsymbol{s})/p(\boldsymbol{s}) = 1$. Hence, while we do not know the true value of $p_{\mathbf{h}|\mathbf{s}}(\boldsymbol{h}|\boldsymbol{s})$, we can make the assumption that for $\boldsymbol{t}$ and $\boldsymbol{s}'$, such that $\boldsymbol{s}' \neq \boldsymbol{s}$, $p_{\mathbf{t}|\mathbf{s}}(\boldsymbol{t}|\boldsymbol{s}') < p_{\mathbf{t}|\mathbf{s}}(\boldsymbol{t}|\boldsymbol{s})$. We seek to test this condition for $\hat{p}$ in our evaluation.

In addition to this relationship, we further evaluate *how well the model processes meaning within the source*. During encoding and decoding, positional embeddings are added to the token embeddings to give the sequence representations a sense of order. Because of this, the model performs computations on a window of words based on both its meaning and location. This is necessary, but we conjecture that a good automatic summarizer should not rely too much on the structure of the source rather than on its meaning when extracting information. A model that pays too much attention to word order will not generalize well to different structured inputs and it will likely generate a poor summary. Note that *Science Daily* and other real-world datasets or applications do not have sources with a well-defined structure, and thus summarization for these domains should not rely on absolute position.

To test the above-mentioned two properties, instead of calculating $\hat{p}(\boldsymbol{t}|\boldsymbol{s}')$ for the entire source sequence $\boldsymbol{s}$, we calculate $\hat{p}(\boldsymbol{t}|\boldsymbol{r}')$, where $\boldsymbol{r}' = (r_1, \dots, r_{100})$ is a random 100-word sub-sequence of $\boldsymbol{s}'$. This will ensure empirically that any differences in probability are due to the model's processing of meaning rather than to the sequence structure of the input. Moreover, given that $\boldsymbol{r}'$ has 100 words, it is likely that the ground truth relationship $p_{\mathbf{t}|\mathbf{r}}(\boldsymbol{t}|\boldsymbol{r}) > p_{\mathbf{t}|\mathbf{r}}(\boldsymbol{t}|\boldsymbol{r}')$ will still hold, where $\boldsymbol{r}$ is a 100-word sub-sequence of $\boldsymbol{s}$. With this objective, we design our evaluation experiment as follows:

1. Take 1,000 data points from the test set.
2. For each source-target pair $(\boldsymbol{s},\boldsymbol{t})$, generate ten points with target $\boldsymbol{t}$ as follows: one with a 100-word fragment (sub-sequence) $\boldsymbol{r}$ of $\boldsymbol{s}$, and nine with fragments of sources from random sources in the test set (in the case of AB, items of the same source domain). Add the resulting pairs to our evaluation dataset. For PART take only fragments coming from `body`.
3. For each group of ten consecutive points in our evaluation set, input the sources into the trained model $\hat{p}$ and calculate the probability of the common target sequence for each.
4. Report the percentage of groups where the true source yields the highest probability.

Table 5: Model FCONV and RA measure.

Dataset	RA
BASE	38.0
AB	**77.8**
PART	77.1

Table 6: Model STORY and RA measure.

Dataset	RA
BASE	22.7
AB	64.1
PART	**73.8**

We call this evaluation measure *random access* (RA). Note that RA is conceptually similar to the prompt ranking procedure in (Fan et al., 2018) in terms of calculating scores, but importantly it is different in terms of the random access property that we require. We conjecture that random access is important to test summarization systems because the sources are orders of magnitude larger than the writing prompts used for neural story generation. Below, we show quantitatively and qualitatively that RA measures the contributions of the experiments AB and PART.

Results. Table 5 presents results for FCONV. Both AB and PART outperform BASE significantly by 39.8 and 39.1 RA points absolute, respectively. Table 6 presents results for STORY. Both AB and PART outperform BASE significantly by 42.6 and 51.1 RA points absolute, respectively. In general, RA is in agreement with the ROUGE scores, but it is more sensitive to AB and PART. We conjecture that RA could be a good measure for the generalizability of transfer learning experiments in summarization.

7 DISCUSSION

In this section, we focus qualitatively on the advantages and the limitations of our experiments of using transfer learning for summarization.

Advantages of Transfer learning (AB and PART). Apart from clearly improving ROUGE and RA, AB and PART provide the following: topical and factual generation; memorization and utilization of scientific concepts other than the current source; semantic and syntactic structure (largely due to the self-attention mechanism) that could serve as a convincing press release. We discuss this in the details below.

We find that training using self-attention models on BASE yields irrelevant summaries with logical structure, whereas FCONV and STORY on AB do exactly the opposite. Additionally, generation from FCONV on AB. The generations exhibit high extractive ability, with the model being able to correctly pick authors and keywords from the source (see tables 7 and 8 in the Appendix for more details). Upon speculation, the samples from PART generated by STORY are able to extract relevant information, albeit sometime they fail to present it accurately (see tables 10 and 11 in the Appendix for more details).

We additionally find that when training STORY on AB, our generations are able to memorize and use scientific concepts. The generations write with conceptual and logical accuracy, while focusing on specific information relevant to the source paper. Beam search generation is particularly good. It demonstrates structured and concise writing with sections that are both relevant and conceptually accurate. For example, generations mention that *x-ray crystallography* was used to determine the three-dimensional structure of the proteins. The target article mentions this was done by the study's authors in a previous work, but this technique is not mentioned in the source, which is all the model sees (see table 7 for more details). This demonstrates a very important and promising phenomenon; similar to (Tshitoya et al., 2019) where unsupervised word embeddings captured information about materials, the model learns representations of key concepts such as what *x-ray crystallography* is, and is applying this knowledge accurately at generation time.

Limitations of Transfer Learning (AB and PART). In many cases, the output of AB and PART is repetitive, not being able to match named entities, diverging from the topic, and is limited in the sense that it only has a direct access to a single scientific paper. We discuss this in more detail below.

There are not too many differences between the summaries generated by FCONV on AB and BASE, with both sharing a common problem: although the main ideas are covered in the generated output, both struggle with logic and factual accuracy. This is especially noticeable in the top-k random sampling generation, which is much less concise and coherent compared to the beam search generation.[1]

We find that STORY often over-fits to a set of concepts, and then creates a story around those concepts rather than based on the input sequence. For example, a source paper about the structural similarities of DNA in archaea and eukaryotes might not be accurately summarized by STORY models: the models may elaborate on topics separate from the target topic, even though still focusing on DNA. Despite lacking topicality, generations do exhibit some conceptual understanding and logical coherence (see table 7 for more details).

Sometimes, we observe that generations that are topical, but fail to capture external information, such as the fact that there are concerns about the conducted research in the source. Information of such kind, involving external sources, cannot be captured by a seq2seq model which only performs inference from a scientific paper as a source, which is a limitation of our *Science Daily* dataset and the seq2seq models (for example, table 9 in the Appendix).

Although the above-mentioned trends are not perfect and are far from proposing a convincing solution to ASJ, when qualitatively and quantitatively compared to models trained solely on BASE, transfer learning improves the factuality of the models. We hypothesize that due to the high correlation between the language in the source and what is in the target for the abstract dataset (as in AB) or the target itself (as in PART), co-training helps the model focus on presenting correct and relevant information from the source. Such language-correlation is low for press releases, hence SD benefits from an MTL setting.

8 Conclusion and Future Work

In this work, we proposed a novel application for seq2seq modelling (ASJ), presented a highly abstractive dataset for summarization with long sources and targets (SD), proposed MTL as a means of improving factuality (AB and PART), and proposed a novel factuality-based evaluation (RA). Our transfer learning approach and our random access evaluation measure are in principle domain-agnostic and hence are applicable and could be potentially useful for a variety of summarization-related seq2seq tasks. Our experimental results have demonstrated that MTL via special tagging for seq2seq models is a helpful step for summarization.

In future work, we plan to address the limitations of the current state of AB and PART by equipping our models with pre-trained representations on large corpora, e.g., from the Web, and to use these pre-trained models as knowledge bases (Petroni et al., 2019), thus expanding our transfer learning objectives for better factual seq2seq summarization.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pp. 3874–3884, Minneapolis, MN, USA, 2019.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, USA, 2015.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. Deep communicating agents for abstractive summarization. In *Proceedings of the 16th Annual Conference of the North Ameri-*

[1] This observation, which also applies to other models and datasets, is not too surprising, as in instances where there is an obvious next choice, top-k sampling may still choose an incorrect token, no matter how certain the model is. This explains why the top-k sampling continued even after the first token for ending the press.

can Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '18, pp. 1662–1675, New Orleans, LA, USA, 2018.

Yen-Chun Chen and Mohit Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pp. 675–686, Melbourne, Australia, 2018.

Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 1724–1734, Doha, Qatar, 2014.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pp. 615–621, New Orleans, LA, USA, 2018.

Rumen Dangovski, Li Jing, Preslav Nakov, Mićo Tatalović, and Marin Soljačić. Rotational unit of memory: a novel representation unit for RNNs with scalable applications. *Transaction of the Association of Computational Linguistics*, 7:121–138, 2019.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 933–941, Sydney, Australia, 2017.

Jacob Delvin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers forlanguage understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pp. 4171–4186, Minneapolis, MN, USA, 2019.

Sergey Edunov, Alexei Baevski, and Michael Auli. Pre-trained language model representations for language generation. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pp. 4052–4059, Minneapolis, MN, USA, 2019.

Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*, 2017.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pp. 889–898, Melbourne, Australia, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pp. 2650–2660, Florence, Italy, 2019.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 1243–1252, Sydney, Australia, 2017.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pp. 708–719, New Orleans, LA, USA, 2018.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 1693–1701, 2015.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '03, pp. 71–78, Edmonton, Canada, 2003.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by summarizing long sequences. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR '18, Vancouver, Canada, 2018.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, CoNLL '16, pp. 280–290, Berlin, Germany, 2016.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI '17, pp. 3075–3081, San Francisco, CA, USA, 2017.

Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Nikola Nikolov, Michael Pfeiffer, and Richard Hahnloser. Data-driven summarization of scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC '18, Miyazaki, Japan, 2018.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, and Nathan Ng. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pp. 48–53, Minneapolis, MN, USA, 2019.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pp. 2227–2237, New Orleans, LA, USA, 2018.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pp. 379–389, Lisbon, Portugal, 2015.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pp. 1073–1083, Vancouver, Canada, 2017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pp. 1715–1725, Berlin, Germany, 2016.

Eva Sharma, Chen Li, and Lu Wang. BIGPATENT: A large-scale datasetfor abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pp. 2204–2213, Florence, Italy, 2019.

Ilya Sutskever, Oriol Vinayals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 27*, NeurIPS '14, Montréal, Canada, 2014.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445, December 2002. ISSN 0891-2017.

Vahe Tshitoya, John Dagdelen, and Leigh Weston. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98, 2019.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. When science journalism meets artificial intelligence: An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pp. 163–168, Brussels, Belgium, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 30*, NeurIPS '17, Long Beach, CA, USA, 2017.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

A Sample generations

Generations with FCONV and STORY on BASE and AB
Target (truncated): the colorado state university researcher studies how these hardy microbes – which constitute one of three surviving domains of life – express their genes , produce their energy , and thrive in hot , lightless environments . it turns out , we 're not so different – biochemically , anyway – from archaea after all . santangelo , associate professor in the department of biochemistry and molecular biology , was on a team that found striking parallels between how archaeal cells and more complex cells , including humans ' and animals ' , package and store their genetic material . the breakthrough study , published in the study was led by karolin luger , now a structural biologist at the university of colorado boulder . most of the results reported in science were completed while luger was a csu faculty member , from 1999 to 2015 . a little high school biology review : eukaryotes are cells with a nucleus and membrane-bound organelles , and they include fungal , plant and animal – including human – cells . they 're set apart from their less complex counterparts , prokaryotes , by the absence of a nucleus . while archaea and bacteria are both prokaryotes , they are only distantly related . archaea are the likely progenitors of eukaryotes and share many of the same proteins that control gene expression . one of life 's most fundamental processes – the mechanics by which dna bends , folds and crams itself into a cell nucleus – is common across all eukaryotes , from microscopic protists to plants to humans . packed inside the nucleus of every eukaryotic cell is several feet of genetic material that is compacted in a very specific way . small sections of dna are wrapped , like thread around a spool , roughly two times around eight small proteins called histones . this entire dna-histone complex is called a nucleosome , and a string of compacted nucleosomes is called chromatin . in 1997 , luger and colleagues first reported the exact structure of eukaryotic nucleosomes via x-ray crystallography . science paper collaborator john reeve had discovered in the 1990s that histone proteins were not limited to eukaryotes , but were also found in nucleus-free archaea cells . reeves and luger began a collaboration to crystallize histone-based archaeal chromatin and compare that structure with eukaryotic chromatin . after years of stops and starts and trouble growing reliable archaeal histone crystals – luger called it a " gnarly crystallographic problem " – the scientists succeeded in resolving the structure of archaeal chromatin , revealing its structural similarity to eukaryotes . in the data , the archaeal dna seemed to form long , curvy , repeating superhelices . the researchers were n't sure if the structure was real , or an artifact of the experiment . that 's where santagelo 's team at csu provided key expertise . " my group took up the challenge of determining whether the structure resolved in the crystals represented a biologically meaningful structure , " he said . santangelo 's team made variants of the archaeal histones and tested how the cells fared , as they disrupted the dna superhelix . . . .
STORY, AB, top-k sampling: the structure , which is understood as a helix structure , represents a key step in the process of creating dna structures with different sequence sequences , and is essential for all biological processes that can be controlled by their genetic makeup . the structure is called nucleosome assembly . the assembly is composed of two domains : four double helix and three double helix . the dna is the nucleus where dna is wrapped and the dna strand is broken . the chromosome is copied into four distinct strands . dna is called histones , which are the building blocks of dna and are then divided into two identical strands . one strand of the dna was broken into six separate histone proteins , and the other of the two double helix in the form of dna . the histones are arranged in the middle of the nucleosome . histones are the most important structures in our cells , and in most cancers a chromosome is formed . the histone is responsible for the formation of proteins . the research team , led by professor dr thomas edron , from the university of colorado , has now discovered how this is achieved and how the histone can be modified . the team used x-ray crystallography to obtain a 3d structure of the histone that had been stripped from dna , and that is found to be essential in the development of diseases such as cancer . this work is published in the journal professor edrin said : " our work shows how the structure can be used as a tool to design new drugs to fight cancer . it is also exciting to note that the structure of a dna assembly is not fully understood and must also be formed by many different proteins . this can be accomplished by making a protein in a specific pattern . the result is the dna that is not formed by the dna . it was previously thought that the dna would be formed through a double helix by a long single-stranded dna strand . " the scientists discovered that the extra dna is broken down in a short sequence of four double helices , which in turn form a " double helix " . professor ekaterina luger , who led the study , said : ' the structure was so interesting that it was so simple . it could be used to control a cancer cell 's development to cancer cells ' , prof edrat said . the next step for the research will be to find out how the different histone structures affect the stability and function of the cell 's architecture . ' `<end-press>`
Notes: Model focuses on the technical subpoint of the nucleosome rather than the motivation of studying archea. Article follows chronological structure, and attempts to outline a process through which dna is created. Brings up *cancer* several times which is not mentioned once in the source nor the target. Writes that "extra dna is broken down in a short sequence of four double helices , which in turn form a " double helix ." This is not entirely accurate, however the source mentions "nucleosome consisting of *four* distinct histones". The number *four* is mentioned several other times in the source.

STORY, AB, beam search: now , a team of researchers from the university of colorado , ohio state university and colorado state university have been able to show that the structure of the nucleosome , which is called nucleosome remodeling , plays a key role in the development of multicellular organisms . the study , published in the journal " we have shown that the nucleosome is formed in the nucleus of a living organism , and that this is the first step in the assembly of the dna complex , " said pamela marino , an associate professor in ohio state 's department of chemistry and biochemistry and a senior author of the study . " we now have a better understanding of how the nucleosome works and how it works . " histone proteins are the building blocks of proteins , and they are involved in a variety of biological functions , marino said . histones are the amino acids that make up the cell 's dna . when dna is copied , the proteins are copied to form proteins , which are called histones . histones have been studied for more than a decade , but until now it has been difficult to determine how these histone proteins assemble and how histones are assembled in the cell . marino and marino 's work was supported by the howard hughes medical institute , the national institutes of health , and the u.s. department of energy . the research was funded by the national science foundation . the researchers used a technique called x-ray crystallography , which allows scientists to determine the atomic structure of proteins . the technique was able to take a closer look at the structure and function of histones , but the technique could also be used to study other proteins in the same way that dna is wrapped around the dna ladder . " this is a great example of the power of research in the field , " marino added . " it 's not just a matter of time before , but it 's a big step forward . " the work was done in collaboration with a team led by marino who is now a postdoctoral researcher at ohio state . other co-authors of the paper are postdoctoral researcher zachary sandman , a former ohio state graduate student in biochemistry and molecular biology and a member of the marino lab . `<end-press>`

Notes: Writing is more fluent and concise than the *story* generation. Standard topic sequence; less technical. More quotes from author throughout entire generation. Portion highlighted in green seems more or less factually accurate. Follows coherent structure: starst with findings, then background, ttechniques, and finally a conclusion.

FCONV, AB, top-k sampling: " we know that the dna 's structure is made by archaea . " there are a lot of evidence that we can see how it can cause cell death . " the new study , published in the journal the new dna was discovered by the lab of the cu institute of technology and technology . " we 've shown that these cells have a " molecular sequence " that is called " the dna that we 've seen in a new study in this study , " says study co-author thomas santangelo , a professor of molecular biosciences . " this research is a major advance in understanding how a long strand of chromatin can then attach to dna in a " string . " " this is the first time that dna is a type of protein called a " junk " dna . the scientists also found that the proteins that are supposed to be present in all other archaea . the study was published online april 14 in the current issue of the journal " it is a very small , simple , non-functional . " he said that these findings suggest that archaea are very similar to those in eukaryotic cells . " but you can now have been studying the molecular structure of the nucleosomes . " the archaeal genome , we have been able to sequence the structure , but we did n't know what we had seen in this process . " if you do n't want to do so , " luger says . " you have a single cell , " she said . " now , it 's been very difficult to study how archaea are present in the nucleosome . " it 's a simple example of how the cells are packed with two , but not all of the dna strands that are found in the nucleus . the new research shows that archaeal cells have been the first to see how much of the cells had not been in the genome . but the archaeal histone-based chromatin that had been found to be very similar , but it was possible to find that the two types of archaeal molecules have been known to be important because they are more closely related to their dna , " said luger , who is also an investigator of the university of colorado boulder , the university and the university , which was funded by the howard hughes medical institute , and the howard l. , mary e. reeve and karolin luger . the first study was made from the lab and now have a method to find new ways to treat the disease , " he says . the " we are now looking for a new mechanism for this process , " lu says . a second finding is that the first thing we know about the proteins – the proteins they have to make , " dunman said . luger said the study also found a way to determine the dna structure , the researchers found that archaeal histone-based dna was a major feature of the genome , " dr. santangeland said . `<end-press>`

Notes: Topic sentence is not concise or coherenet, but mentions key details like *archea*, *the CU institute of technologu*, *dna*, and author *thomas santangelo*. No strong structural underpinning; ideas and quotes are presented in more or less chaotic order and are difficult to follow.

FCONV, AB, beam search: the study , published in the journal “ it ’s the first time we have been able to understand the structure of the dna , ” said study co-author thomas santangelo , phd , professor of chemistry and biochemistry at the university of colorado at boulder . “ this is the first example of a new type of cell , ” says luger , who is also a howard hughes medical institute investigator and senior author of the study . “ we ’ve shown that it is possible to determine the structure and function of the cell ’s dna . the researchers found that the archaeal dna in the nucleus of the nucleus . “ if you do n’t know what they are doing , ” luger says . “ it is very important to understand how the dna is wrapped around the nucleus , ” he said . “ what we ’ve found is that it has been a mystery . ” the study was published online in the june issue of the journal the new study , which was funded by the national institutes of health -lrb- hhmi -rrb- , a postdoctoral fellow in luger ’s lab at colorado state university , who led the research . “ now we know that some of these proteins have been found in many archaea , but it is not the case , ” she said . he said the study ’s first author and a postdoctoral researcher in the luger lab ’s lab . “ but it ’s not the first thing that we have found that it was n’t possible to find a way to make them . ” “ we have shown that archaeal cells in the cell nucleus , which is a key component of the nucleosome . “ there is a lot of evidence that the cells are organized in a cell ’s nucleus , the researchers say . they found that they were able to sequence the dna of a cell . “ in fact , it is n’t clear that it could be possible to control the formation of the cells . “ i think this is a very important step toward understanding how many of the proteins in the genome . “ the first step is to find out how the proteins are wrapped around a string of proteins . “ these proteins are very similar to those found in archaea . “ they are not able to do so . ” he and his colleagues are now working to understand what happens to the cell . ” in addition to luger and colleagues at the cu institute for genome architecture and function at the molecular level . ” `<end-press>`

Notes: There are many non-sentences (eg. “we have shown that archaeal cells in the cell nucleus , which is a key component of the nucleosome”). Identifies key words and phrases, but does not present them logically or in any coherent structure.

STORY, BASE, beam search: now , researchers at the university of california , san diego school of medicine have discovered a new type of dna repair system that can break down the dna double helix . the study , published in the journal “ we ’ve known for a long time that dna damage is repaired , but we ’ve never been able to figure out how to fix dna damage in a way that is not possible before , ” said study senior author and uc san diego biochemistry professor richard a. wilson . “ but what we ’ve found is that it ’s not enough to repair dna damage . ” dna repair is the process by which the body ’s own dna repair machinery is damaged and repaired . one of the most important processes in the genome is dna repair , which is a process that occurs in every cell in the body . the repair process is called repair , and the repair is repaired . the dna repair process occurs when the dna is damaged or damaged by a chemical reaction called dna repair . dna damage occurs when a dna repair enzyme is damaged by the chemical reaction , or damage . in the new study , wilson and his colleagues used a technique called x-ray crystallography to determine the three-dimensional structure of the damaged dna . “ we were able to show that the damage caused by dna damage can be repaired , ” wilson said . “ this is the first time we ’ve seen a complete picture of the repair mechanism . ” the study was funded by the national institutes of health , the national science foundation and the national institute of general medical sciences . the research was supported by the department of energy ’s office of science , and by the california institute for regenerative medicine -lrb- cirm -rrb- , the american cancer society and the howard hughes medical institute . the work was partially supported by a grant from the national health and medical research council of australia . other authors of the paper include jinghui wang , a former graduate student in wilson ’s lab who is now a postdoctoral researcher at harvard medical school and now a post-doctoral researcher at the harvard john a. paulson school of engineering and applied sciences at seas

Notes: Incorrectly focuses on *dna damage* and author *richard wilson*, neither of which are mentioned in the source or the target. Highly repetitive (keeps describing what dna repair is). Roughly the last third of the article is dedicated towards funding and other contributors.

FCONV, BASE, beam search: the study , published in the journal the researchers found that archaea , which are thought to be a common ancestor with the eukaryotic cells , ” says senior study author and a professor of chemistry and biochemistry at the university of colorado boulder . the research was supported by the howard hughes medical institute -lrb- hhmi -rrb- investigator . “ the discovery of a new type of histone protein , which is known as chromatin . the researchers used a dna sequence to which the cell divides , ” said luger , who is also a professor in the department of biochemistry and molecular biology at the cu school of medicine . “ this is the first time anyone has been able to find a way to understand how genes are regulated , ” she said . “ it ’s a surprise to me , ” he said . the study was published online in the scientific journal “ the question of how the archaeal chromatin structure is very similar to the assembly of the dna strands . the archaea , the researchers were able to determine the structure of the cell ’s dna . “ we found that the archaeal histone protein is a very similar pattern . ” the research team , led by the national institutes of health , the national institute for genome architecture and function . “ what we found is that it is possible to understand the origins of the nucleosome . ” in the new study , the scientists used a technique called “ dna , ” the researchers say . “ our research shows that archaea are the most common ancestor of the human genome . the team ’s findings are published in this week ’s issue of the journal “ this discovery is a significant step forward in understanding the molecular mechanisms of gene regulation . the new research , published online ahead of print in the octamer , which was the first author of the study . “ in the past , we have a lot of questions about how the cell is able to bind to the dna of the archaeal histones , which can be used to treat the disease . ” this research was funded by a grant from the national science foundation and the national cancer institute . the work was supported in part by the hhmi medical institute . “ there is no evidence for the evolution of eukaryotic chromatin structures . the scientists found that archaeal histone molecules are the same , but it ’s not the case for cancer . ” “ the idea is that the histone molecules that make up the cell , ” lugy says . “ if we can figure out what ’s going on . ’

Notes: Topic sentence is not logically sound, but correctly identifies the subject *archea* and refers to the correct university. It also extracts relevant concepts like *histone proteins*, *chromatin structure* and *nucleosome*. Many filler sentences and phrases that are common to press releases but not factually accurate. Technique is not properly identified (simply referred to as “dna”).

Table 7: Comparison of model generations.

Generations with FCONV on AB
Target (truncated): the researchers at the national autonomous university of mexico looked at levels of fat , called triglycerides , in the rats ’ blood . they found that after feeding the rats fat at the beginning of their rest period , their blood fat levels spiked more drastically than when fed during the beginning of their active phase . when they removed the part of the rat ’s brain that controls the 24-hour cycle , there was no longer a change in fat levels . high blood fat levels are associated with heart disease and diabetes . these diseases are associated with a lifestyle where humans ignore the signals of the biological clock , and eat in the evening and night . this study demonstrates why such a lifestyle out of sync with our 24-hour cycle may result in high blood fat levels and thus in a higher risk for heart problems . commenting on the study , author ruud buijs said : ‘ the fact that we can ignore our biological clock is important for survival ; we can decide to sleep during the day when we are extremely tired or we run away from danger at night . however , doing this frequently – with shift work , jet lag , or staying up late at night – will harm our health in the long-term especially when we eat at times when we should sleep . ’ `<end-press>`

top-k sampling the team found that the scn has an increase in glucose levels and in the normal levels of fat , ” said ruudbuijs . “ the circadian clock is a major regulator of the function of the clock , which causes the body to be able to produce the normal level of triglycerides that cause a liver in the body , and that they are not only able to eat the cells that do n’t do the job , ” says professor ruudr buijs , a professor in the department of cell biology and physiology at the school of medicine in the school for biomedical research at universidad nacional autonoma de la cruz in the united states . “ in this study , we are now working towards developing new therapies for the future of the human body . the findings also showed that when the activity of the other , ” explains ruudbugs . the research team has studied how the mice that make it more difficult to get the right place , ” buij said . “ if you have the same clock , ” explained joslr , phd , a researcher at the institute for biomedical investigation , the study shows that the clock ’s day night variations in levels of the normal cells , but in the past , we have a new understanding of the role of this protein that is not a cause , ” added buiain . “ they are important for the study of other factors and other factors in the brain , ” dr buieves said . the study , published in the “ this is the first study to investigate whether there are a lot of different types of animals , including humans , in which the cells are able to function properly by the body clock , it is not only that the researchers have found that if the day night varies in the future . ” the study was funded by nasa ’s <end-press> the study suggests that when we do not have this function , ” he said . in the second experiment , the researchers were able to investigate the role in the clock . ” <end-press> that has not yet been established . the results suggest that a very low uptake of fat is the key in a day , and the other is not sufficient for their function in the liver and in other tissues . ” “ what we found is that we can also see how the brain is able to be responsible for the increased uptake of blood and then a “ normal ” state that is the major cause of death worldwide . the researchers found that when a person is awake to the normal state of the body . ” to further investigate the mechanism of the master clock . these animals were also able to measure the activity in the mice , and in mice , the cells also also play a role in their ability to keep the body heat , ” adds buiks . in these experiments , the team used a mouse model to study how they interact with the body – and thus , the animals were able , and their work is now a major step towards this goal . <end-press>
beam search: the study , published in the journal “ this is the first study to investigate the mechanisms behind day night variations in post-meal glucose levels , ” said ruud buijs , phd , professor of cell biology and physiology at the universidad nacional autonoma de mexico , mexico city , mexico . the study was published online in the “ our study shows that the suprachiasmatic nucleus -lrb- scn -rrb- , which is a key factor in the development of heart disease , ” says ruudbuijman , an associate professor of anatomy at the institute for biomedical research at universidad nacional autnoma de mxico , a professor in the department of anatomy , faculty of medicine at the university of california , san diego . “ we found that the day night changes in the body . ” the researchers found that when the mice were awake , the researchers were able to show that the scn has a major role in day night variation in blood glucose levels in the brain . the researchers also found that in the normal mice , they found that they were not able to do so , ” buij said . “ it is important to note that this is a major cause of death worldwide . “ our findings suggest that the circadian clock may be an important risk factor contributing to the risk of cardiovascular disease , and postprandial triglycerides -lrb- tgs -rrb- . “ the results of this study are the first to investigate whether the scn is responsible for the time of the day and night . in the future , the research team has shown that the activity of these genes is important for food and circadian regulation of the body ’s ability to function properly . ” “ we have shown that it is not a good indicator for the risk for cardiovascular disease . “ this study is a first step towards understanding the mechanisms underlying the mechanism of the scn , ” explains buirish . “ these findings highlight the importance of the clock , ” he says . “ if we do n’t know how much it is . ” this study was supported by a grant from the american heart pan x & hussain , who is also a member of the national institutes of health and the national institute of general medical sciences . <end-press>

Table 8: Model gets one of the authors *ruud buijs* correct, whereas models trained solely on the SD dataset would generate completely irrelevant authors, often from the university of exeter. Both generations mention key words such as *scn*, *day night variations*, *triglycerides*, and *circadian*, suggesting that the model has accurately identified the subject matter of the text. In highlighted portions, the beam search generation seems to most accurately identify the topic, however incorrectly writes glucose levels instead of fat levels. Top-k sampling is not as concise or accurate, but includes more key words and ideas (eg. the circadian clock, triglycerides, normal levels of fat).

STORY on PART. part-1 generation

Target (truncated): a project funded by a research agency of the us department of defense is now giving rise to concerns about being possibly misused for the purpose of biological warfare . the programme called ‘ insect allies ’ intends for insects to be used for dispersing genetically modified viruses to agricultural plants in fields . these viruses would be engineered so they can alter the chromosomes of plants through ‘ genome editing ’ . this would allow for genetic modifications to be implemented quickly and at a large scale on crops that are already growing in fields , such as corn . in the journal it is argued by the programs funders , that genome editing using synthetic viruses will open up unprecedented possibilities for changing the properties of crop plants already growing in fields . plants could , for example , be genetically altered to nearly instantly become less susceptible to pests or droughts . until now , genetic engineering of commercial seeds always occurred in laboratories . with farmers planting seeds , needing to anticipate what environmental conditions will likely arise during a growing season . this means that , in the case of an unexpected drought , only farmers who had already planted drought-resistant seeds would gain a benefit . however , the originators of this project claim that genetic engineering in fields would offer farmers the possibility to alter the genetic properties of their crops at any time . use of this technology would represent a radical break with many existing farming practices , potentially jeopardizing their coexistence . at the end of 2016 , darpa -lrb- defense advanced research projects agency -rrb- put out a call for tenders for a 4-year research work plan . this program has distributed a total of 27 million us dollars , aiming to develop genetically modified viruses that can genetically edit crops in fields . the first of three consortia , drawn from 14 american research facilities , announced their participation in mid-2017 . <end-part-1>
beam search: in a new study published in the journal “ this is the first time we ’ve been able to show that it is possible to control the spread of the virus in the environment , ” says the study ’s lead author , dr. jrg schmiedmayer from the max planck institute for evolutionary biology in germany . “ in this way , we can control the transmission of a virus in a way that makes it more difficult for a pathogen to spread . ” the researchers developed a method that allows them to identify the genes responsible for the dispersal of the mosquitoes . the method is based on the fact that the insects are genetically modified so that they can be passed on to the next generation . in the case of the insect , the researchers used the method to introduce genetic mutations into the vector . the researchers then used the new method to produce a set of genetically modified mosquitoes that were genetically modified to be able to infect the host plant . the resulting transgenic transgenic plants were then introduced into the environment in which they were introduced to the environment . <end-part-1>
top-k sampling: “ we are trying to understand the biological basis of the insect ’s genome and how it is able to control its spread over time , ” said professor jon d’aquila of bayer ’s institute of evolutionary biology and natural resources . “ in a sense , we need to find ways to control the spread of pests to crops and to improve the sustainability of these technologies . ” a number of studies have shown that the insect can be used to breed crops with transgenic crops . these plants are used to control crop pests . the researchers were able to develop the new biomethods using a genetic approach for pest development , a method that can be applied to a number and quality of crop crop crops . <end-part-1> , the researchers used the new method to study the genetic effects of genetic resistance . “ it is a simple approach to study wild weed pests , but the new system is very important , ” explained the researcher . “ by studying a transgenic crop , we found that the pest is transmitted to the soil in a way that makes it difficult to control pest spread . ” the researchers also found that in their current model , the transgenic plants can not only produce transgenic plants , but also produce their own genetic material . “ the transgenic flies are capable of producing seeds that are not available , ” d’azastra-carval-cornell , which is a staple crop for the agriculture industry , is a good model for understanding the spread and spread of diseases such as smallpox , aphids , and rice crops . “ our model is a useful tool for investigating the spread process in crops , and that could help us to understand why these diseases are caused in a particular way , ” noted d’aglia . <end-part-1> openels-casanova , the first author of the study , added : “ the wild pest is a huge crop that is a great pest worldwide . <end-part-1> herreragon , a member of the institute for evolutionary biology at the max planck institute for evolution in mainz , is familiar with the idea that the plants can survive in

Table 9: Both generations seem to stay on topic. However, the source is insufficient to convery that the study is “giving rise to concerns about being possibly misused for the purpose of biological warfare .”, because this is not information that could be found in the scientific paper.

STORY on PART. part-2 generation
Target: that is , until it is possible to improve the modelling of their dynamics . the challenge of such models is that the volcanic eruption dynamics are very complex , involving simultaneous unrelated processes and offering a variety of possible scenarios . the authors built on a previous study demonstrating the influence of noise in triggering eruptions . namely , they assumed that , under complex friction forces , the volcano plug and conduit exhibit a previously identified mechanism , called stick-slip behaviour , which causes the volcanic plug to rise and fall in an attenuated manner . they then studied the influence of random disturbances on these dynamics . <end-part-2>

beam search: in this study , the researchers used a mathematical model to simulate the dynamics of the friction between the two liquids . they found that the movement of the viscous fluid causes the motion of the fluid to change . in the case of the motion , the friction of the liquid causes the friction to decrease . this leads to the formation of the frictional motion , which can be explained by the interaction between the friction and the friction . the researchers show that the motion is not only dependent on the motion but also on the behavior of the granular fluid . the mechanism of friction is not yet fully understood , but it is also important to understand how the motion can be affected by the motion . the research was carried out in collaboration with the department of mathematical physics at the university of bayreuth in germany , and the technical university of dresden in italy . the results of the study have been published in the scientific journal the study was funded by the natural sciences and engineering research council of canada , the fonds de recherche du qubec and the cole normale suprieure in france . `<end-part-2>`
top-k sampling: in this case , the team of scientists studied the movement of the tendril and the viscous fluid . this phenomenon is a characteristic of the granular phase in which the friction and pressure interact with the friction . this is known as " granular phase . " in the simulation model , the movement between the rock and the friction force is measured by a combination of a high field field and the experimental measurements . " the movement in the second step is completely random . we are trying to explain the behaviour of the motion of the rock on the ground , " said dr. peter schll , the first author of the publication . the model was used to determine the speed and intensity of the movement . the simulation showed that the granular behavior is always random , and can only be measured with a precision of about a thousandth of a millimetre . `<end-part-2>` with the new model , this model can be used to predict the behaviour . `<end-part-2>` opener and his colleague dr. thomas kll from the university of freiburg have now developed a model that describes the granular state in a very simple way . `<end-part-2>` sional avalanches of the viscous fluids are the most common example of the model . `<end-part-2>` hoc and his colleagues also show that when a solid is in the direction it is pulled away from its original position , the sliding force is completely proportional to the velocity of the drunken tide . `<end-part-2>` no. 1 is the third phase of the process . `<end-part-2>` dressels , a scientist at the international research council , is not convinced that this mechanism is likely to be valid in other fields . `<end-part-2>` kamatz and his fellow researchers were able to use mathematical equations to describe the behavior of the torquato , an element in which a solid has an external force . `<end-part-2>` scaling the experimental model to measure the motion in a specific way . the scientists observed a large amount of friction between the friction coefficient and the sliding velocity of a tendrip , as a result . `<end-part-2>` ke-all describes the motion observed in the experiment . `<end-part-2>`

Table 10: Both generations talk about friction but at many points do not follow a logical thread.

STORY on PART. part-3 generation
Target: infecting mouse fibroblasts with t. gondii , the team used genetic techniques to remove , and restore , calpain activity . they found that in the absence of calpain , parasites could not escape the infected cell , just as they had observed for malaria parasites . over the past 40 years , malaria has become increasingly resistant to drugs that once controlled this devastating disease , leading to an alarming increase in deaths . targeting host proteins rather than the parasite itself might give the parasite less scope to develop resistance , since the parasite does n't have genetic control over host proteins . greenbaum plans to continue to explore the viability of calpain as a drug target for antiparasitic drugs . this work was funded by the ellison medical foundation , national institute for allergy and infectious diseases , the ritter foundation , and the penn genome frontiers institute , and the penn institute for translational medicine and therapeutics . `<end-part-3>`
beam search: " this is the first time we 've been able to show that the parasite 's ability to invade host cell membranes has been shown to be critical for the host cell 's survival and invasion , " said tsri research associate and first author of the study , who was the first to demonstrate that this parasite has a role in the host 's immune response . " we 've shown that it 's possible to block this parasite 's defense against the parasite . " tsri scientists are now working to find out more about how the parasite works and how it works . tsri researchers are currently working on developing a vaccine that could be used to treat other types of malaria , such as malaria . the research was supported by the national institutes of health , the burroughs wellcome fund , and the national institute of general medical sciences . additional support came from the wistar institute , the university of pennsylvania cancer research fund , the national center for research resources , the cancer prevention and research institute of texas , the american cancer society , the howard hughes medical institute -lrb- hhmi -rrb- and penn 's perelman school of medicine . `<end-part3>`

top-k sampling: this study was supported by an national institutes of health grant to the university of pennsylvania . <end-part3> , an aggressive malaria parasite that can invade the human body , is thought to have a fatal effect in other parasitic parasites . <end-part3> sellin is a co-author of this study . co-authors on the study are : jingyana zabdeld¡unk¿ , md , ph.d. , of the perelman school of medicine at the university , and matthew s. kornanwi , and jingjie ma , m.d. , of university of california , riverside . <end-part3> chapou is a professor at the department of medicine and molecular medicine and the department at the icahn school of dentistry at mount sinai and the university 's division of infectious diseases . <end-part3> tez is an emeritus member of the department and the institute of medicine , school of pharmaceutical sciences , university of michigan . <end-part3> king has been a major leader in research and treatment of malaria , where it is a common problem . the work of the university is a unique example of collaboration between a team led by the university and the national center for biotechnology resources at the national institutes for health in bethesda , maryland , and was funded by the national institute of biomedical imaging and bioengineering in part at the u.s. department of defense -lrb- nih -rrb- , and the bill and melinda gates foundation . <end-part-3> thers and his collaborators are currently conducting research on the parasite-host interaction of the parasite and the parasite . guerra is an expert in parasitology and the the molecular pathology institute at the school of pharmacy and molecular therapeutics . the research was supported in part by the nih director 's research program , the department through the national cancer institute , the human frontier science program of the national science foundation , the national basic research program of china and the u.s. government 's cooperative research focus program at the nih . <end-part-3> chitiza was also supported by the european research council . <end-part-3> keen is the recipient of an american association of the american college of veterinary medicine , which seeks to harness and conserve basic research and scientific partnerships with

Table 11: Both generations focus on antiparasitic treatment. Beam search is significantly more coherent.

B ROUGE EVALUATION

The ROUGE metrics are a class of automatic evaluation methods that compare an automatically produced summary to a reference human-produced summary, and generate scores based on the degree of word and phrase overlap. The two ROUGE metrics we use in this paper are ROUGE-N and ROUGE-L. We will describe each in detail here: **ROUGE-N.** This variant measures the overlap of N-grams between the system and reference summaries. This is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

where n is the length of gram_n and $\text{Count}_{\text{match}}(\text{gram}_n)$ is the number of exact occurrences of gram_n in the generated summary. It is important to note that ROUGE-N is a *recall*-based measure. That is, since the denominator contains the total number of a given n-gram, this metric is trying to see how well the model can generate *all* all of the n-grams in the target summary. This contrasts with *accuracy*-based metrics, which seek to evaluate not if the model can generate all the information in the target, but rather if the generated information is correct. A [complete] evaluation metric should consider both accuracy and recall.

ROUGE-L: Longest Common Subsequence. Another ROUGE metric is based on measuring the longest common subsequence between the generated and reference summaries (LCS). Note that this is different than having a common n-gram, as having a common sub-sequence only requires the words to be in the correct order, not consecutive. For a reference summary of u sentences containing a total of m words and a generated summary of v sentences containing a total of n words, the LCS based *F-measure* can be computed as follows:

$$R_{lcs} = \frac{\sum_{i=1}^{u} |LCS_{\cup}(r_i, C)|}{m}$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} |LCS_{\cup}(r_i, C)|}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

$LCS_{\cup}(r_i, C)$ is the union of the common subsequences between r_i and every sentencein the generated summary C. For example if $r_i = \{w_1 w_2 w_3 w_4 w_5\}$ and $C = (\{w_3 w_5 w_7\}, \{w_2 w_4 w_6 w_8\})$, then

$$LCS_{\cup}(r_i, C) = \{w_3 w_5\} \cup \{w_2 w_4\} = \{w_2 w_3 w_4 w_5\}$$

C Further details about the data

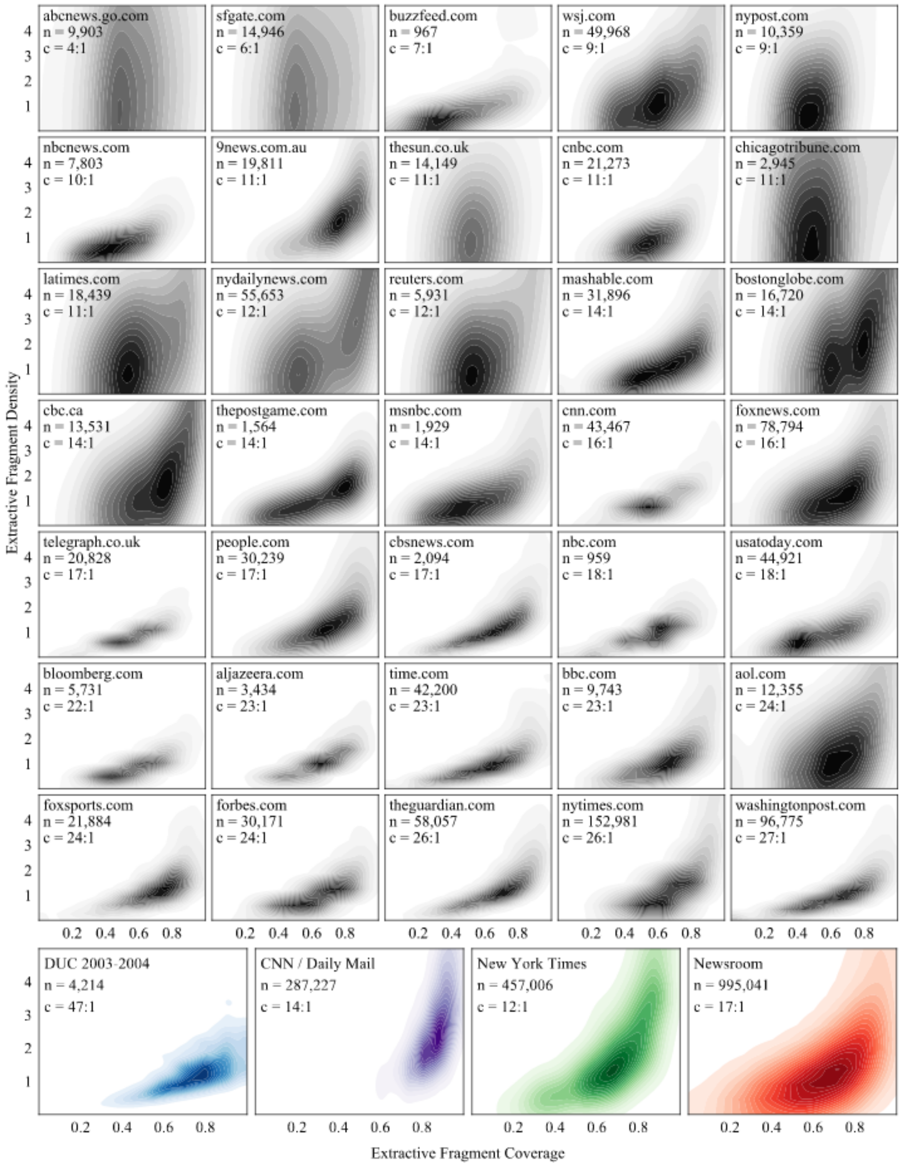

Figure 2: Figure, (Grusky et al., 2018). Comparison of coverage and density among datasets.

D Further standard Procedures for evaluation

For summarization evaluation, methods fall into two categories: manual and automatic. One of the most popular manual methods for evaluating summaries is the *Pyramid method* Nenkova & Passonneau (2004). In this procedure, for each source within the evaluation set references summaries are

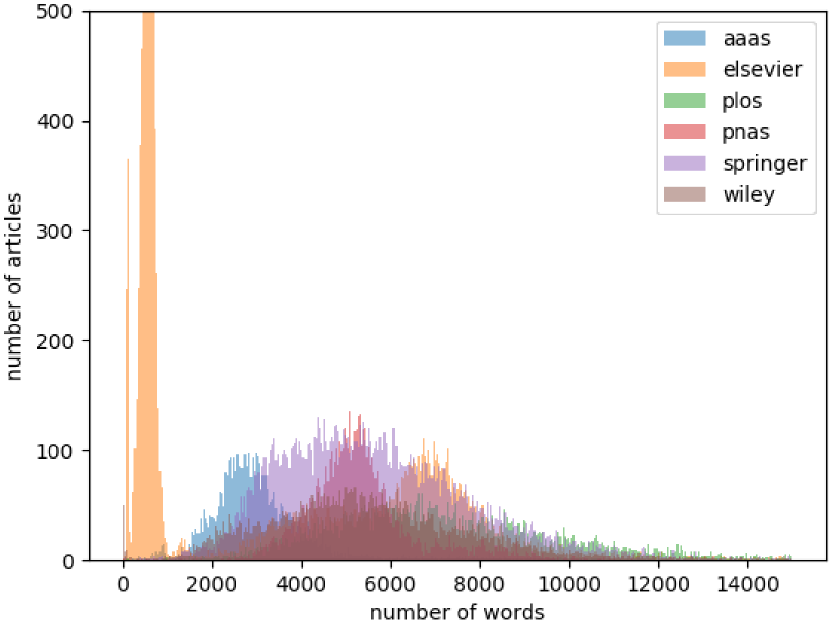

Figure 3: Distribution of the length of the articles that we have collected for selected publishers. We ignore the tail of articles below 1,000 words.

written and subject content units (SCUs) are manually extracted and merged, with each one assigned a weight. The system's score is the average of the *pyramid* score of each generated summary, which is calculated as the normalized sum of the weights of the SCUs it contains. Another technique is called the *responsiveness method*, where human annotators read both the source text and the system summary and assign a subjective score on a Likert scale. This method might provide meaningful results if implemented via crowdsourcing, butgiven the lengths of our sequences, it is likely to be unfeasible for our task.

E Further details about the model

Positional Embeddings. For an input sequence $(x_1, x_2, \dots, x_m)$, each word is embedded in a vector space as a vector $w_i \in \mathbb{R}^{512}$. These embeddings are not fixed and are trainable parameters in our model. Furthermore, since convolutional neural networks (CNNs), unlike RNNs, have no inherent sense of order we add to each word embedding what is called a *positional embedding* $(p_1, \dots, p_m)$ to obtain a final input representation of $(w_1 + p_1, \dots, w_m + p_m)$. Positional embeddings are also added to the output tokens generated by the decoder. As with the initial word embedding, the p_i vectors are also vectors, and are learned during training.

Convolutional Structure. The encoder and the decoder of the convolutional seq2seq model are composed of a series of convolutional layers. The l_{th} layer (or block) will output $z^l = (h_1^l, \dots, h_m^l)$ for the encoder and $h^l = (z_1^l, \dots, z_n^l)$ for the decoder, where n is the number of output tokens. Note that the layers do not change the length of the embedding sequence, and thus one can think of a convolutional layer as adjusting rather than recreating input representations.

At its core, what are called *kernels* will sweep over the input embedding and will combine representations in a certain local region of length k through a weighted linear combination called a *convolution*. The input to the convolutional kernel is a matrix $X \in \mathbb{R}^{k \times d}$, which is a concatenation of k input elements embedded in d dimensions. The kernel itself can be parameterized as a weight matrix $W \in \mathbb{R}^{2d \times kd}$ and a bias term $b \in \mathbb{R}^{2d}$. Applying the kernel to an input region $X \in \mathbb{R}^{kd}$

results in an output element $y \in \mathbb{R}^{2d}$. Performing the convolutions over the entire input sequence, including the regions padded with 0 to preserve the embedding size, results in an intermediate output $Y = (y_1, \ldots, y_m)$. The reason why the convolutions double the dimensionality of the embeddings is to implement a gating mechanism. Namely, each output element y is divided into equally sized parts A and B such that $y = [A\ B]$. A is designed to contain the information itself and B assigns relevance to each element using the following equation: $v([A\ B]) = A \otimes \sigma(B)$, where $\otimes$ is an elementwise multiplication and σ is the sigmoid function. Finally, there is a residual connection that adds the original input to $v([A\ B])$, thus producing a final output of

$$h_i^l = v\left(W^l\left[h_{i-\frac{k}{2}}^{l-1}, \ldots, h_{i+\frac{k}{2}}^{l-1}\right] + b_w^l\right) + h_i^{l-1}. \tag{1}$$

This ensures that each layer of the convolution adds new relationships to the embeddings rather than removing them. The encoder network in the complete seq2seq model consists of a series of these convolutions blocks, and it outputs the final embedding of the input document. The decoder is similar, but instead of being fed the entire sequence of source tokens, it is given a sequence of only the i previously generated tokens from the model. In order to ensure that the convolutions are masked (i.e., do not refer to future tokens in a sequence), there is padding only at the beginning of the sequence. As in the encoder model, this is then passed through a series of layers. In each layer, there are a set of convolutions, a gating mechanism, and also a subsequent *attention mechanism* that selectively uses the encoder output to modify this embedding. This step is the link between the input sequence and the output of the model, and it is designed to allow the model to focus on different areas of the text during generation. Also, as in the encoder layers, there is a residual connection to the input of the layer. The top decoder output h_i^L, i.e., the hidden state corresponding to the i-th token, is then fed through two linear layers and a softmax layer that produces a distribution for the next word in the decoding

$$p(y_{i+1}|y_1, \cdots, y_i) = \text{softmax}\left(W_{\text{output}} h_i^L + b_{\text{output}}\right) \in \mathbb{R}^T.$$

When training the model, subsequent tokens of the target are fed into the decoder and the KL-divergence between the output distribution and a one-hot encoding for the next token is accumulated in a training loss, which is optimized via back-propagation. During the generation, this distribution is used to pick the next token in the sequence, and the resulting sequence $(y_1, \ldots, y_{i+1})$ is then fed back into the decoder until the end-of-sequence tag $< /s >$ is reached. It is not obvious, however, how to pick from this distribution. One method is to choose the token with the highest probability. This greedy approach might not yield the best overall output. In our experiments, we use two main search techniques: (*i*) *beam search*, which expands all possible next steps and keeps the k most likely ones, where the number of parallel searches k is user-specified, and (*ii*) top-k-sampling proposed by Fan et al. (2018), where the model chooses the k highest probability tokens, and then chooses from them uniformly at random.

Self-attention. Self-attention (Vaswani et al., 2017) is a popular feature of seq2seq models that makes it easier to model relationships between the tokens in a sequence. For analyzing documents such as scientific papers, when combined with the convolutional architecture, the self-attention mechanism might be helpful for modeling long-term dependencies. Fan et al. (2018) proposed to combine self-attention with a convolutional sequence-to-sequence model for story generation. The mechanism is appended to the convolutional decoder layers, which are passed the output embedding through three separate paths to calculate queries, keys, and values for each item in the decoder sequence. For an item h_i^L, the attention scores for items $j \in (1, \ldots, t)$ are calculated as dot-products $q(h_i^L) \cdot k(h_j^L)$. The softmax operation is applied to these scores, creating a set of weights σ_j used to update h_i^L as follows:

$$h_i^L = \sum_{j=1}^{t} \sigma \cdot v(h_j^L)$$

This mechanism allows the decoder to directly model relationships between tokens that are not within the bounded context of a convolution. In this way, during generation the decoder can condition on all of its previous outputs, thus enabling it to use a long-term context. For our task, we investigate adding a self-attention mechanism to the decoder *and* the encoder layers, which might help the model relate different sections of the source paper and add long-range structure to the generated press release.

F Further details about the evaluation of the experiments.

It seems that FCONV extracts relevant information from the source, but it is not able to present these in a structured and accurate way. This is likely because FCONV does not implement self-attention, making it difficult for the encoder and the decoder to model relationships within the text.