Sci-Blogger: A Step Towards Automated Science Journalism

by

Raghuram Vadapalli, syed.b , Nishant Prabhu, Balaji Vasan Srinivasan, Vasudeva Varma

in

27th ACM International Conference on Information and Knowledge Management
(*CIKM-2018*)

Lingotto, Turin, Italy

Report No: IIIT/TR/2018/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
October 2018

Sci-Blogger: A Step Towards Automated Science Journalism

Raghuram Vadapalli
IIIT Hyderabad
raghuram.vadapalli@research.iiit.ac.in

Bakhtiyar Syed*
IIIT Hyderabad
syed.b@research.iiit.ac.in

Nishant Prabhu*
IIIT Hyderabad
nishant.prabhu@research.iiit.ac.in

Balaji Vasan Srinivasan
Adobe Research
balsrini@adobe.com

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

ABSTRACT

Science journalism is the art of conveying a detailed scientific research paper in a form that non-scientists can understand and appreciate while ensuring that its underlying information is conveyed accurately. It plays a crucial role in making scientific content suitable for consumption by the public at large. In this work, we introduce the problem of automating some parts of the science journalism workflow by automatically generating the 'title' of a blog version of a scientific paper. We have built a corpus of 87, 328 pairs of research papers and their corresponding blogs from two science news aggregators and have used it to build ***Sci**ence-**Blogger*** - a pipeline-based architecture consisting of a two-stage mechanism to generate the blog titles. Evaluation using standard metrics indicate viability of the proposed system.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**;

KEYWORDS

science journalism; content repurposing; automated blogging

ACM Reference Format:
Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018. Sci-Blogger: A Step Towards Automated Science Journalism. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269303

1 INTRODUCTION

Scientific research, new technologies, paradigm shifts and challenges to accepted scientific "truths" play a major role in key political, economic, cultural and social policy discussions and also in public dialogue. Approximately 2.5 million new scientific papers are published each year [5]. Some of these papers are breakthroughs and need to be brought to wider public knowledge. The aim of science journalism is to render the detailed, often jargon-laden information produced by scientists into a form that non-scientists can comprehend and appreciate, while still communicating the underlying information accurately. Science blog aggregator sites like *sciencedaily.com* try to precisely do this.

Table 1: A sample from the dataset

Blog title: Cancer-fighting nanorobots programmed to seek and destroy tumors **Paper title:** A DNA nanorobot functions as a cancer therapeutic in response to a molecular trigger in vivo **Abstract:** Nanoscale robots have potential as... . Using DNA origami we constructed an autonomous DNA robot... transport payloads and present them specifically in tumors...

While news-writing bots have captured headlines in the recent past[1,2], it is a non-trivial task to extend this framework for science journalism since this includes understanding scientific content and translating it into simpler language without losing its underlying information, thus spanning the domains of Natural Language Generation, Text Simplification and Text Adaptation. The lack of an appropriate training corpus in this space makes this task even more challenging. To address this, we have created a parallel corpus of scientific papers and their corresponding blogs with the aim of furthering the advances in automated science journalism.

While understanding the entire research paper and generating a complete blog is our larger vision - in this paper with *Sci-Blogger*, we present our initial exploration towards generating the title of the blog. The title of the blog conveys a lot about the research that is being talked about. For example, consider a blog titled *Cancer-fighting nanorobots programmed to seek and destroy tumors*. This conveys that the article is talking about certain advancement in nanotechnology to fight tumors. In *Sci-Blogger*, we use a combination of the title of the research paper and its abstract to generate the corresponding blog title. Samples of abstracts, paper titles and their corresponding blog titles are given in Table 1.

Sci-Blogger models the blog title generation task via a two-stage process: first, using an input mechanism to extract relevant information from the title and abstract of the research paper and then using the extracted information to generate the blog titles with the help of recent advances in sequence-to-sequence neural networks for natural language generation. The results and generated samples indicate the viability of the proposed model to produce semantically sound blog titles. Our contributions can be summed up as follows:

*The authors contributed equally.

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269303

[1] Washington Post's robot reporter has published 850 articles.
[2] New York Times is using bots to create more one-to-one experiences.

- A new parallel corpus of 87,328 pairs of blog titles along with their corresponding research paper titles and abstracts.
- Introducing *Sci-Blogger*, a pipeline-based architecture in order to facilitate the generation of blog titles, experimenting with various heuristic functions and recent sequence-to-sequence architectures and reporting various evaluation measures on the results.

2 RELATED WORK

Recent advances in sequence-to-sequence neural networks have benefited a lot of natural language generation tasks. Among these, two particular areas relevant to our problem are *neural headline generation* and *style transfer*.

In the space of **Neural Headline Generation**, Ayana et al. [2] explored the LSTM based seq2seq architectures with attention mechanism for headline generation. However, they generate headlines in the same domain and thus is not directly applicable to our problem. While directly using seq2seq architectures was helpful in this case as we will show later, cross domain headline generation requires the consideration of aspects such as style, readability, etc in the two domains.

Existing work in **non-parallel style transfer** assumes the unavailability of sufficient parallel data [4, 6, 14]. Shen et al. [14] first addressed the problem of style transfer on non-parallel data by trying to separate the content from the style of the article. They assume a shared latent content distribution across different text corpora, and propose a method that leverages refined alignment of latent representations to perform style transfer. While Shen et al. [14] demonstrated their results on sentiment transfer, this cannot be accepted as style transfer from a linguistic point of view.

Fu et al. [4] address the style-transfer problem by learning separate content and style representations using adversarial networks. They report results of their models on their custom Paper-News Title dataset. The samples reported in the paper either copy the entire source text or replace few words. Their evaluation criteria leaves a lot to be desired as they evaluate transfer strength using a classifier and content preservation using word embeddings. This is yet another drawback of a lack of parallel data.

A key assumption across all the non-parallel style transfer works is a significant overlap between the vocabulary of the source and target style. However, in the context of science journalism, the overlap in vocabulary between the source and the target is not significant which is the prime reason why the non-parallel style transfer methods cannot be directly extended to our problem.

3 PARALLEL CORPUS FOR SCIENCE JOURNALISM

We have built a corpus of parallel scientific papers and their corresponding blog articles from two science news aggregation sites: *sciencedaily.com* and *phys.org*. Both these sites report on significant advancements in several fields of science and are rich sources of parallel data. Most of the posts include a link to the original research paper. We began with 300,000 blog titles out of which only around 100,000 had links to original research papers from more than 1000 different research publication websites. We used manual rules to extract abstracts and titles from the most frequent websites which included *nature.com*, *pnas.org*. Our final dataset comprised of 87,328 (blog title, paper title, abstract) triples.

Out of 87,328 triples, 77,604 are obtained from *sciencedaily.com*, 9724 tuples are obtained from *phys.org*. Figure 1 provides statistics on the websites from which titles and abstracts have been extracted. Average length of blog titles is observed to be 9.55 words. Average length of research paper titles is 12.07 words. Average length of abstracts is 179.54 words. Average word overlap between blog titles and paper titles is 1.93 words. Average overlap between abstracts and blog titles is 3.64 words. The models must therefore learn which words in the target vocabulary correspond to the words in the source so that the generated output adheres to the target style.

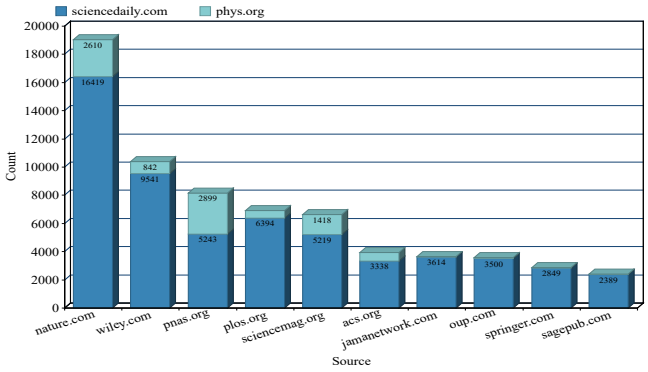

Figure 1: Abstract-title pairs from various sources

4 BLOG TITLE GENERATION

Figure 2 illustrates the proposed architecture. *Sci-Blogger* consists of a two-stage pipeline, which is described in detail as follows:

(1) We employ a heuristic-based function which takes the title and abstract of the research paper and extracts relevant information to feed it into the next step. This is done by experimenting with various heuristics as we will describe below.
(2) The output from the previous step is fed into a sequence-to-sequence neural generation model in order to generate the title of the blog post.

Formally, for **stage 1** - given our dataset, $\mathcal{T} = \{ (bt, pt, abs) \}$, where, bt is the blog title, pt is the paper title and abs is the abstract, we define a **heuristic function** $\mathcal{H}(pt, abs)$ which takes a paper title and abstract as parameters and outputs a sequence s. The various heuristic functions $\mathcal{H}$ we explored are outlined below:

$\mathcal{H}(pt, abs) = pt$: In this heuristic, we assume that the paper title will encapsulate sufficient information to generate the blog title.

$\mathcal{H}(pt, abs) = \mathcal{RP}(abs)$: Here, we define $\mathcal{RP}(abs)$ as the most representative sentence in *abs*. We used the sum of TF-IDF [1] values of words in a sentence as representativeness of a sentence and follow a similar procedure like the previous approach.

$\mathcal{H}(pt, abs) = \mathcal{RPD}(abs)$: Let $n\mathcal{RD}(abs)$ and $n\mathcal{RP}(abs)$ be the normalized readability and representativeness of a sentence respectively, where normalization is performed across all sentences. We define $\mathcal{RPD}(abs) = n\mathcal{RD}(abs) \times n\mathcal{RP}(abs)$.

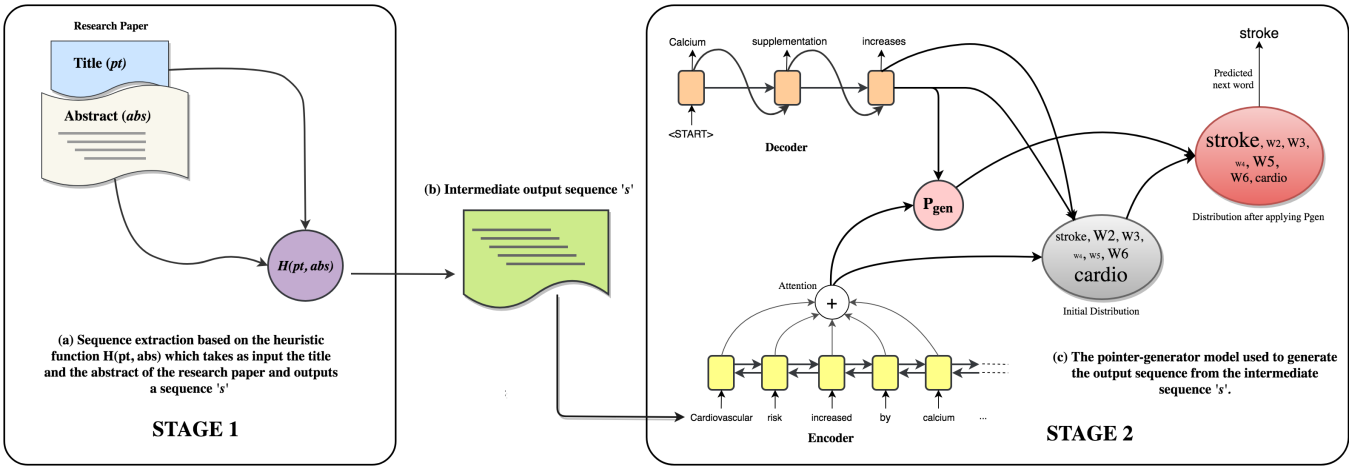

Figure 2: Model pipeline architecture for *Sci-Blogger*, describing the two stages in which we model blog title generation

We also experimented with different combinations of the above heuristics: $\mathcal{H}(pt, abs) = pt \mid \mathcal{RD}(abs)$, $\mathcal{H}(pt, abs) = pt \mid \mathcal{RP}(abs)$, $\mathcal{H}(pt, abs) = pt \mid \mathcal{RPD}(abs)$ and $\mathcal{H}(pt, abs) = pt \mid abs$; where $\mid$ implies concatenation of the associated heuristics.

In **stage 2** - we leverage a competent sequence-to-sequence (seq2seq) architecture [12] for generating the blog titles using the intermediate output sequence from stage 1. Sequence-to-sequence networks have been successfully applied to summarization and neural machine translation [10] tasks where an '**attention**' is defined over the input sequence to allow the network to focus on specific parts of the input text to generate the text. Such a framework has also been successfully extended for headline generation [9]. One of the recent advancements in this direction is the **pointer-generator** [12] framework where the model extends over the attention-based frameworks to compute a probability P_{gen} to decide whether the next word in sequence should be copied from the source or generated from the rest of the vocabulary. Such a framework aids in copying factual information from the source, and we hypothesize that this will be useful when generating blog titles. We thus use this pointer-generator model as our sequence-to-sequence framework for the 2nd stage. Formally, the sequence $\boldsymbol{s}$ obtained from the first stage is trained via the model which generates $\boldsymbol{bt'}$ as output with a loss function $\mathcal{L}(\boldsymbol{bt}, \boldsymbol{bt'})$, given by sum of cross entropy loss at all time-steps:

$$\mathcal{L}(\boldsymbol{bt}, \boldsymbol{bt'}) = -\sum_{t=0}^{t=T} P(bt'_t)$$

5 EVALUATION AND ANALYSIS

We evaluate the generated titles using various metrics surveyed by Sharma et al. [13] for task-oriented language generation. **BLEU** [11] uses a modified precision to compare a candidate translation against multiple reference translations. **ROUGE_L** [8] is an F-measure based on the Longest Common Subsequence (LCS) between the candidate and reference utterances, generally used for evaluating summarization. **CIDEr** [15] is based on n-gram overlap and used for evaluating image descriptions. **Skip Thought Cosine Similarity** [7] is based on a continuous representation of sentences known as skip-thought vectors and used to measure sentence-level similarity. All these metrics either measure n-gram overlap or content overlap between the generated title and the blog title. Additionally, the generated content should be in a form that non-scientists can appreciate and understand. Thus, we also report the **Flesch Reading Ease** [3], which measures the readability of the sentence based on the no. of syllables and words. Table 2 shows the performance of our proposed input functions in stage 1 of the architecture contrasted with the proposed pointer-generator model (*abbr.* $\mathcal{PG}$) for stage 2 and the vanilla attention-based model (*abbr. Attn*).

In our proposed model, $\boldsymbol{pt \mid abs}$ outperforms other input functions by a significant margin in most of the metrics. This is expected as the pointer-generator is designed for summarization and is able to learn to extract significant information from a long sequence such as $\boldsymbol{pt \mid abs}$. This also benefits from using a single embedding matrix for both the source and target vocabulary. It is important to note that it gives the best results on $\boldsymbol{pt \mid abs}$ compared to the vanilla attention mechanism on any other heuristic. Also, the copying of words from the source was with 0.20 probability in $\boldsymbol{pt \mid abs}$ which is closest to the original overlap of 0.23, which implies that the model was able to learn to compute P_{gen} efficiently. On the other hand, we see - with a vanilla attention-based mechanism as the framework in stage 2, $\boldsymbol{pt \mid \mathcal{RP}(abs)}$ gave the best BLEU score, although $\boldsymbol{pt \mid \mathcal{RD}(abs)}$ and $\boldsymbol{pt \mid \mathcal{RPD}(abs)}$ BLEU scores are very close. Across the metrics, one of these three models performed best. Interestingly, the other heuristic functions like *abs* and *pt* | *abs*, which have a lot more information performed very poorly on the vanilla attention model. This may be explained as follows. In the vanilla attention model, there are two embedding matrices for the source and target vocabulary and the model tries to learn to output appropriate phrases corresponding to the source phrases. This benefits from the source and target sequence lengths being comparable. The three heuristics which performed well here have the property that the source and target sequences are comparable in length, but other heuristics which have very long source sequences compared to the target sequences performed poorly.

Table 2: Performance of the various heuristic functions contrasted with the proposed sequence-to-sequence generation framework

	BLEU		ROUGE_L		CIDEr		SkipThought Sim.		Flesch Reading Ease	
$\mathcal{H}(pt,\ abs)$	$\mathcal{PG}$	*Attn*	$\mathcal{PG}$	*Attn*	$\mathcal{PG}$	*Attn*	$\mathcal{PG}$	*Attn*	$\mathcal{PG}$	*Attn*
pt	0.157	0.149	0.157	0.138	0.453	0.433	0.430	0.432	46.481	43.069
abs	0.135	0.095	0.136	0.089	0.359	0.123	0.444	0.119	45.484	39.021
$\mathcal{RD}(abs)$	0.091	0.142	0.09	0.137	0.161	0.450	0.403	0.431	**57.468**	**47.521**
$\mathcal{RP}(abs)$	0.101	0.143	0.105	0.134	0.183	0.571	0.414	0.429	46.338	41.179
$\mathcal{RPD}(abs)$	0.096	0.134	0.097	0.126	0.179	0.543	0.415	0.428	43.89	40.761
$pt \mid \mathcal{RD}(abs)$	0.139	0.152	0.129	**0.145**	0.351	0.754	0.435	0.444	49.504	42.964
$pt \mid \mathcal{RP}(abs)$	0.142	**0.153**	0.137	0.144	0.372	0.745	0.431	**0.448**	40.856	35.311
$pt \mid \mathcal{RPD}(abs)$	0.151	0.152	0.158	0.140	0.399	**0.759**	0.435	0.433	40.193	44.339
$pt \mid abs$	**0.171**	0.096	**0.172**	0.091	**0.523**	0.123	**0.446**	0.219	41.307	45.641

The Flesch Reading Ease (*FRE*) of the generated samples is above 35. *FRE* > 30 indicates that high-school students will be able to read and understand it successfully. The increase in readability highlights the transfer in style from research paper to the blog. The blogs had a *FRE* around 30-35, while the research paper's *FRE* was between 15-20. Our generated samples had a reading ease (>30) along the lines of the reading ease of a blog. Best scores were obtained by $\boldsymbol{\mathcal{RD}(abs)}$ which is expected as we are using the most readable sentence from the abstract as the source. But, this doesn't correspond to a good score in the other metrics which implies that although $\boldsymbol{\mathcal{RD}(abs)}$ produces the most readable results, it is not very good at mimicking the human writer.

We also report the performances had we used the intermediate output sequence of stage 1 (without using stage 2) as our generated blog title for various input functions in Table 3. As we can see, there is a clear benefit of using the two-stage pipeline against just the intermediate output sequence from stage 1. The poorer results are expected given the lack of overlap between the vocabulary of the two domains.

Table 3: Output of stage 1 directly compared with the reference blog title

$Seq.s$	BLEU	ROUGE_L	CIDEr	SkipThought Sim.	Flesch Reading Ease
pt	**0.098**	0.089	**0.201**	0.395	24.291
$\mathcal{RD}(abs)$	0.064	0.080	0.175	**0.401**	**32.560**
$\mathcal{RP}(abs)$	0.075	**0.101**	0.166	0.386	23.432
$\mathcal{RPD}(abs)$	0.073	0.095	0.143	0.390	26.675

To further shed some light on the quality of the generated blog titles, Table 4 shows a few sampled sentences generated by the best performing models in our architecture. Based on our experiments with multiple hypothesis, we conclude that our system learns to generate titles similar to a human expert for scientific blogs.

6 CONCLUSION AND FUTURE WORK

In this work, we introduced the problem of automating science journalism. We proposed an architecture in which we employ a two-stage pipeline to tackle the problem. We conducted an extensive study evaluating the performance of the proposed model indicating its viability. The desired future direction would be to generate an entire blog which may require additional external knowledge and advanced architectures. We hope that this work serves as a good start towards automating science journalism.

Table 4: Samples generated by the proposed model

Blog title: Safer alternatives to nonsteroidal antinflamatory pain killers **Model-generated title:** New hope for treating inflammatory cardiovascular disease
Blog title: The effects of soy and whey protein supplementation on acute hormonal responses to resistance exercise in men **Model-generated title:** Soy protein supplementation linked to resistance exercise in men
Blog title: Scientists reconcile three unrelated theories of schizophrenia **Model-generated title:** A new way to fight psychiatric disorders

REFERENCES

[1] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications* (2017).

[2] Ayana, Shi-Qi Shen, Yan-Kai Lin, Cun-Chao Tu, Yu Zhao, Zhi-Yuan Liu, and Mao-Song Sun. 2017. Recent Advances on Neural Headline Generation. *Journal of Computer Science and Technology* (2017).

[3] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* (1948).

[4] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style Transfer in Text: Exploration and Evaluation. *AAAI* (2018).

[5] Arif E. Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* (2010).

[6] Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic Transfer in Natural Language Generation Systems Using Recurrent Neural Networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods.*

[7] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28.*

[8] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out.*

[9] Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712* (2015).

[10] Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural Machine Translation (seq2seq) Tutorial. *https://github.com/tensorflow/nmt* (2017).

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting on Association for Computational Linguistics.*

[12] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *55th Annual Meeting of the Association for Computational Linguistics.*

[13] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation. *CoRR* (2017).

[14] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. *Advances in Neural Information Processing Systems* (2017).

[15] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition.*