When science journalism meets artificial intelligence : An interactive demonstration

by

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

in

2018 Conference on Empirical Methods in Natural Language Processing
(*EMNLP-2018*)

Brussels, Belgium

Report No: IIIT/TR/2018/-1

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

When science journalism meets artificial intelligence : An interactive demonstration

Raghuram Vadapalli[1]*, **Bakhtiyar Syed**[1]*, **Nishant Prabhu**[1]*, **Balaji Vasan Srinivasan**[2], **Vasudeva Varma**[3]
[1, 3] IIIT Hyderabad, [2] Adobe Research
{raghuram.vadapalli, syed.b, nishant.prabhu}@research.iiit.ac.in
balsrini@adoWhen be.com, vv@iiit.ac.in

Abstract

We present an online interactive tool[1] that generates titles of blog titles and thus take the first step toward automating science journalism. Science journalism aims to transform jargon-laden scientific articles into a form that the common reader can comprehend while ensuring that the underlying meaning of the article is retained. In this work, we present a tool, which, given the title and abstract of a research paper will generate a blog title by mimicking a human science journalist. The tool makes use of a model trained on a corpus of $87,328$ pairs of research papers and their corresponding blogs, built from two science news aggregators. The architecture of the model is a two-stage mechanism which generates blog titles. Evaluation using standard metrics indicate the viability of the proposed system.

1 Introduction

With approximately 2.5 million new scientific papers being published every year (Jinha, 2010), there is an ever growing need to make this vast trove of scientific knowledge accessible to the common man. This accessibility of scientific knowledge plays an important role in key political, economic, cultural and social policy discussions and also in public dialogue. Websites like *sciencedaily.com*, *phys.org*, *eurekalert.org* aim to address this problem by aggregating and showcasing the top science news stories from the worlds leading universities and research organizations.

News-writing bots have captured the headlines in the recent past, leading to the growing popularity of "Robo Reporting"[2,3]. However, extending this framework to be used for science journalism is a non-trivial task as that would entail understanding scientific content and translating it to simpler language without distorting its underlying semantics. To our knowledge, there have been no prior attempts within the scientific community to extend "Robo Reporting" to science journalism, and this dearth of research in this area can be partially attributed to the lack of suitable data for AI algorithms to be trained. To address this lack of an appropriate training corpus, we have created a parallel corpus of scientific paper titles and abstracts, and their corresponding blog titles with the aim of initiating this foray into automated science journalism and engendering further research.

This initiative is an initial step towards the larger goal of understanding the entire research paper and generating a complete blog. The system makes use of a pipeline-based architecture that uses a combination of the title of the research paper and its abstract to generate the title of the blog. Sample of an abstract, paper title and its corresponding blog title is given in Table 1.

Table 1: Sample - Blog Title, Paper Title and Abstract from our corpus

Blog title: Applying machine learning to the universe's mysteries
Paper title: An equation-of-state-meter of quantum chromodynamics transition from deep learning
Abstract: A primordial state of matter consisting of free quarks and gluons... Here we use supervised learning with a deep convolutional neural network to identify the EoS employed...

Our system models the blog title generation task via a two-stage process: first, it uses a heuristic function mechanism to extract relevant information from the title and abstract of the research

* The authors contributed equally.

[1] https://irel.iiit.ac.in/science-ai/

[2] Washington Post's robot reporter has published 850 articles.

[3] New York Times is using bots to create more one-to-one experiences.

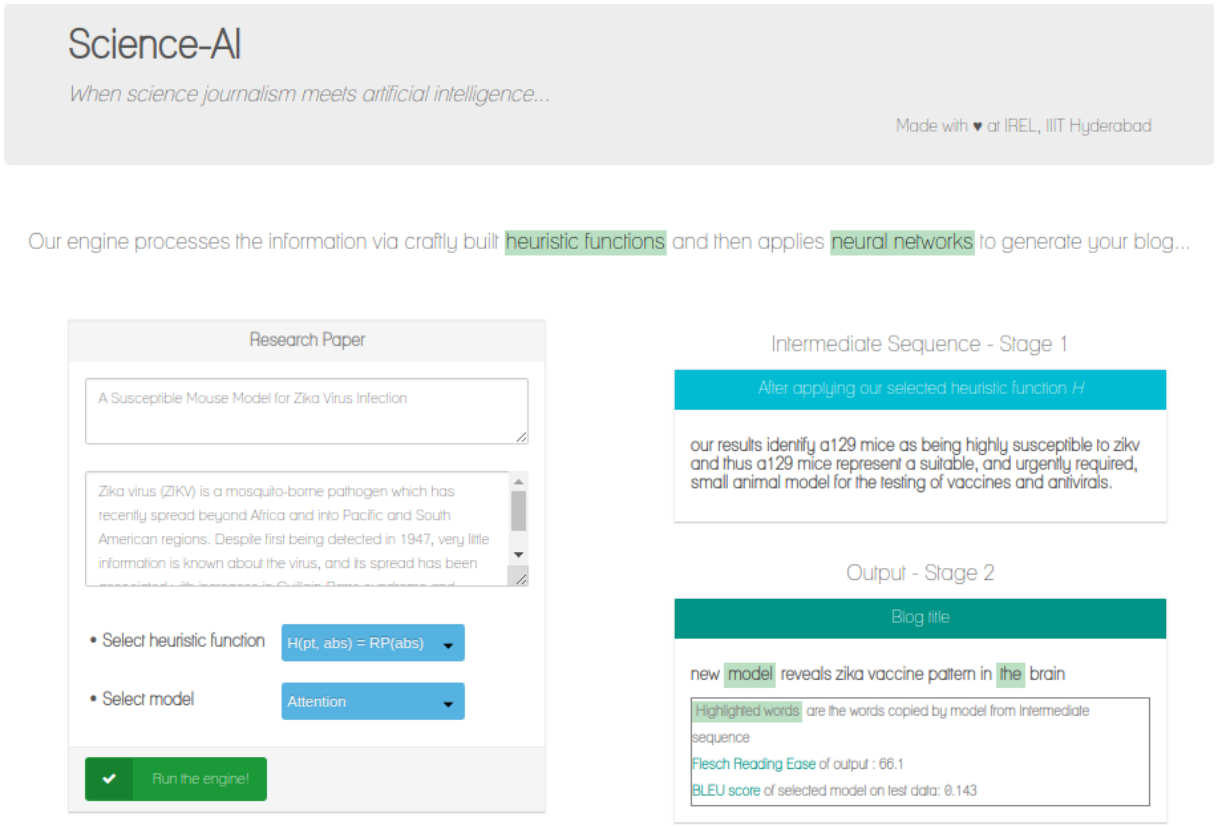

Figure 1: Layout of the web application for our prototype, demonstrating blog title generation

paper and then it it uses the extracted information to generate the blog title. The state-of-the-art sequence-to-sequence neural networks for natural language generation like the Pointer Generator Network (See et al., 2017) are used in the second stage of the pipeline. The generated blog titles are evaluated using all standard metrics for natural language generation tasks and the results indicate the viability of the proposed model to produce semantically sound blog titles. Our contributions can be summed up as follows:

1. A new parallel corpus of $87,328$ pairs of research paper titles and abstracts and their corresponding blog titles.
2. Demonstrating the web application, which uses a pipeline-based architecture that can generate blog titles in a step-by-step fashion, while enabling the user to choose between various heuristic functions as well as the neural model to be used for generating the blog title.
3. Analyzing the outcomes of the experiments conducted to find the best heuristic function as well as network architecture.

We have thus taken the first steps towards building an automated science journalism system by generating blog titles with a long-term vision of generating an entire blog from a given research paper - thereby paving the way for future research in the area.

2 Related Work

Recently, a lot of activity in the space of advances in natural language generation has resulted from pioneering works in building sequence-to-sequence neural networks. Among these advances, two particular areas relevant to the problem we have formulated are *neural headline generation* and *style transfer*.

In the space of **Neural Headline Generation**, *Long Short Term Memory* (LSTM) based sequence-to-sequence architectures for headline generation using the attention mechanism have been explored (Ayana et al., 2017). However, the authors generate headlines for the same domain which effectively means we cannot apply the architectures directly to our problem where the domains and vocabulary are very different. While directly using seq2seq architectures was somewhat helpful in our case - as we will show later; cross domain headline generation requires the consideration of aspects such as style, readability, etc in the two different domains of study.

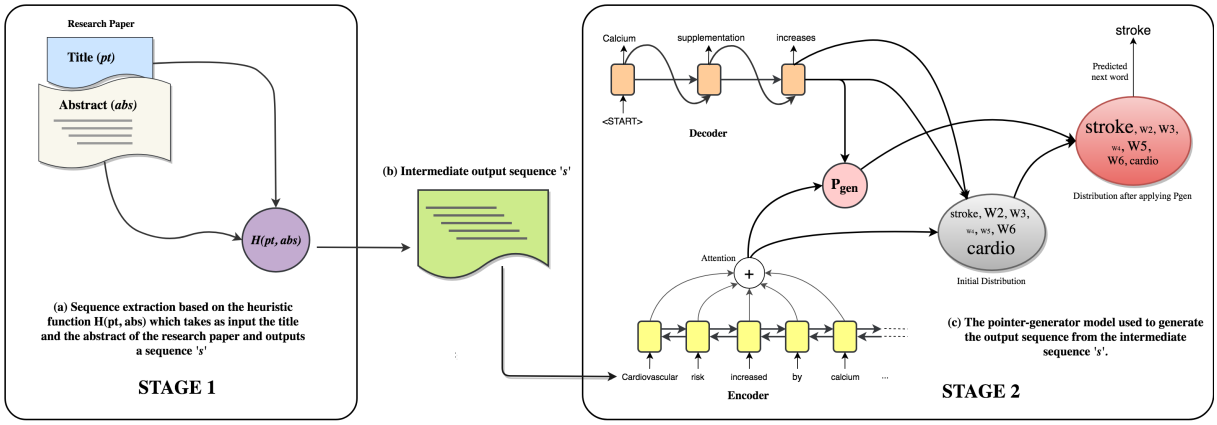

Figure 2: Model of pipeline architecture, describing the two stages in which we model blog title generation

Existing literature in **non-parallel style transfer** assumes the unavailability of sufficient parallel data (Shen et al., 2017; Fu et al., 2018; Kabbara and Cheung, 2016). In first trying to address the problem of style transfer on non-parallel data, Shen et al. (2017) tried to separate the content from the style of the article. It was assumed that a shared latent content distribution exists across different text corpora, and proposed a method that leveraged refined alignment of latent representations to perform style transfer. While Shen et al. (2017) demonstrated their results on sentiment transfer, this cannot be accepted as style transfer from a linguistic point of view.

In other recent works, Fu et al. (2018) address the style-transfer problem by learning separate content and style representations using adversarial networks. Their reported results are on their custom Paper-News Title dataset and the samples reported by the authors either copy the entire source text or replace a 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. A lack of parallel data again presents a drawback. While Kabbara and Cheung (2016) presented a variant of an auto-encoder where the latent representation had two separate components: one for style and one for content, the authors do not report results on any dataset and hence is not useful in our context.

One key assumption across all the non-parallel style transfer works is a significant overlap between the vocabulary of the source and target style. On the other hand, in the context of science journalism - the overlap in vocabulary between the source and the target is not significant which is one of the prime reasons why the non-parallel style transfer methods cannot be directly extended to our problem. This puts our problem in the bracket of content re-purposing, for which we give a demonstrable prototype.

3 Parallel Corpus for Science Journalism

In the process of building a solution to address the problem of automated science journalism, we built a corpus of parallel data consisting of scientific papers and their corresponding blog articles from two science news aggregation websites: *sciencedaily.com* and *phys.org*. Both these websites publish articles explaining the latest scientific advancements and are rich sources of parallel data. Though we were able to obtain over $300,000$ blog titles, only around $100,000$ of those articles had links to original research papers. These $100,000$ or so research papers were published on over 1000 different research publication websites and we used manual rules to extract abstracts and titles from the research papers that were published on the more frequent research publication websites like *nature.com*, *pnas.org*. Our final dataset comprises 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*. The statistical analysis of the dataset is as presented below:

1. Average length of blog titles: 9.55 words
2. Average length of research paper titles: 12.07 words
3. Average length of research paper abstracts:

179.54 words

4. Average word overlap between blog titles and paper titles: 1.93 words
5. Average word overlap between the paper abstracts and blog titles: 3.64 words.

The models must therefore learn which words in the target vocabulary correspond to which words in the source, so that the generated output adheres to the target style.

4 Blog Title Generation

Figure 2 illustrates the proposed architecture. Our demonstrable prototype consists of a two-stage pipeline, which is described in detail as follows:

1. A heuristic function takes the title and abstract of the research paper and extracts relevant information which is then used for further processing.
2. The output of the previous step is fed into a sequence-to-sequence neural generation model in order to generate the title of the blog post.

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

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

$\mathcal{H}(\boldsymbol{pt}, \boldsymbol{abs}) = \mathcal{RP}(\boldsymbol{abs})$: TF-IDF based measure that selects the sentence that best represents the abstract. (Allahyari et al., 2017)

$\mathcal{H}(\boldsymbol{pt}, \boldsymbol{abs}) = \mathcal{RD}(\boldsymbol{abs})$: Flesch Reading Ease based measure that selects the most readable sentence in the abstract.

$\mathcal{H}(\boldsymbol{pt}, \boldsymbol{abs}) = \mathcal{RPD}(\boldsymbol{abs})$: Selects the sentence that maximizes the product of normalized $\mathcal{RD}(\boldsymbol{abs})$ and $\mathcal{RP}(\boldsymbol{abs})$ scores, where normalization is performed across all sentences.

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

In **stage 2**, neural natural language generation models are used to generate the blog title. The system provides a baseline **attention network** which defines 'attention' over the input sequence to allow the network to focus on specific parts of the input text and the **pointer-generator** (See et al., 2017) network which extends the attention-network to compute a probability P_{gen} that decides whether the next word in sequence should be copied from the source or generated from the rest of the vocabulary. The pointer-generator aids in copying factual information from the source, and we hypothesize that this will be useful when generating blog titles. Formally, the sequence $\boldsymbol{s}$ obtained from the first stage is the input to the neural natural language generation 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 Demonstration

Figure 1 illustrates the layout of our demonstrable web application. It can be accessed publicly at the following URL: https://irel.iiit.ac.in/science-ai. The layout of the web application is broadly divided into two parts. the left half of the page has the necessary text fields and drop down menus to accept inputs from the user and the right half of the page displays the outputs- both the intermediate sequence, which is the output of the first stage of the pipeline, and the blog title, which is the output of the second stage of the pipeline. The application accepts two text inputs from the user: the **title** of the research paper and the **abstract** of the research paper. The application also allows the user to select the heuristic function to be used in the first stage of the pipeline as well as the neural generation model to be used in the second stage of the pipeline. Running the engine will parse the inputs and pass them on to the appropriate heuristic function, which will produce an intermediate sequence viewable on the right side of the page. This intermediate sequence is then passed on to the neural generation model that is selected by the user which then generates the final output, which can be viewed below the intermediate sequence.

	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

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

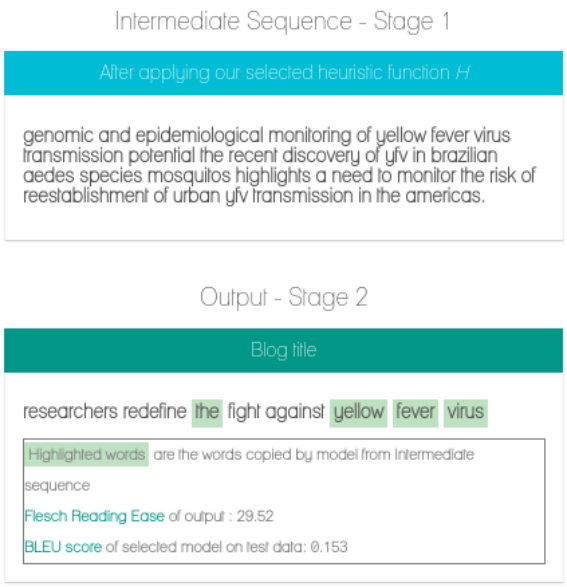

Figure 3: Results in the application

It is important to note that our research prototype is still in a nascent stage and the problem of automated science journalism is far from being solved. The same heuristic function or neural generation model might not exhibit the best results for all possible inputs. Thus, it is of exceptional importance to provide the users fine grained control over the individual components of the model. The system does this by allowing the user the freedom to select a heuristic function and neural generation model of their choice. This allows for more flexibility for the users to experiment with various heuristic functions and neural generation models and ensures better results than forcing the user to use one particular configuration for all inputs. If not selected by the user, the heuristic function $\mathcal{H}(pt,\ abs) = pt$ and the **attention network** neural generation model are used as defaults as this configuration has consistently exhibited good results.

In order to further facilitate experimentation by the user, the web application shows the performance of the user-selected configuration on our test dataset, it displays the readability scores of the generated output, and also highlights the words in the output that were copied from the source. All these features give anyone using this system a detailed view of how various configurations work and provide the flexibility to select the one that works best of their use-case. Figure 3 showcases the above mentioned features.

6 Evaluation and Analysis

We evaluate the generated titles using various metrics surveyed by Sharma et al. (2017) for task-oriented language generation.

1. **BLEU** (Papineni et al., 2002): It uses a modified precision to compare generated text against multiple reference texts
2. **ROUGE_L** (Lin, 2004): It is an F-measure that is based on the Longest Common Subsequence (LCS) between the candidate and reference utterances
3. **CIDEr** (Vedantam et al., 2015): It is based on n-gram overlap
4. **Skip Thought Cosine Similarity** (Kiros et al., 2015): It is based on a continuous representation of sentences known as skip-thought vectors

5. **Flesch Reading Ease** (Flesch, 1948): It measures the readability of the sentence based on the number of syllables and words

Table 2 shows the performance of our proposed input functions of the architecture contrasted with the proposed neural generation models pointer-generator (*abbr.* $\mathcal{PG}$) and the attention-network (*abbr.* $Attn$).

The blogs had a *Flesch Reading Ease* of around 30-35, while the research paper's reading ease was between 15-20. Our generated samples have a reading ease *(>30)* highlighting the transfer in style from research paper to the blog. The higher *FRE* indicates that the generated titles are easier to understand than the paper titles.

To further shed some light on the quality of the generated blog titles, Table 3 shows a few sampled sentences generated by the best performing models in our architecture. Based on our experiments, we conclude that our system learns to generate titles similar to a human expert for scientific blogs.

Table 3: Samples generated by our prototype

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

7 Conclusion and Future Work

This work serves as a baseline first attempt toward automating science journalism. We proposed an architecture with a two stage pipeline and have developed a demonstrable web application that accepts the title and abstract of a research paper and outputs a blog title, while also giving the user the flexibility to tinker with the individual components of the system. Future work would include using more advanced architectures to generate the body of the blog.

References

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*.

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*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *AAAI*.

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

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*.

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*.

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

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*.

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*.

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*.

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*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*.