# Enhancing Academic Title Generation Using SciBERT and Linguistic Rules

**Elena Callegari**
University of Iceland
Reykjavík, Iceland
ecallegari@hi.is

**Peter Vajdecka**
Prague Univ. of Economics and Business
Prague, Czechia
vajp02@vse.cz

**Desara Xhura**
SageWrite ehf.
Reykjavík, Iceland
desara@sagewrite.com

**Anton Karl Ingason**
University of Iceland
Reykjavík, Iceland
antoni@hi.is

## Abstract

This study tackles the challenge of generating appropriate academic titles based on the paper's abstract. We approach this task as a high-level text summarization problem and introduce an innovative post-processing method that combines a predictive model with a set of linguistic rules to enhance the quality of the title generation. We start by evaluating three Natural Language Generation models (BART, T5, Flan T5), by identifying the top-performing model and by configuring it to generate diverse titles. We then conduct experiments employing various post-processing strategies -using SciBERT and linguistic rules- to select the best title out of all machine-generated options. Finally, we assess our title selection methods in relation to human evaluations.

## 1 Introduction

Titles of academic articles are more than simple labels; they serve as a concise representation of the contents of the paper, providing a glimpse into its purpose. Since titles serve as an initial touchpoint, they play an indispensable role in piquing readers' interest, emphasizing the relevance of the research, and enhancing its visibility within the vast scholarly landscape. Crafting the right title can be difficult, as one must distill potentially very complex research into a single, concise statement. This can be particularly challenging as the title must reflect both the depth and breadth of the paper, while also appealing to a diverse academic audience. Selecting an appropriate title also holds significance in the context of citations: according to both Paiva et al. (2012) and Deng (2015), papers with titles that have specific characteristics, such as a certain maximum length, get cited more often than papers that do not meet such criteria.

Traditionally, researchers have relied on their own judgment and expertise to craft compelling titles that summarize the findings of their research articles. In this paper, we delve into the task of automatically generating stylistically and discipline-appropriate titles for academic articles. To do that, we thought of generating titles using an article's abstract as input, as abstracts capture key passages and findings of a paper. An alternative would be to use the full paper as input, but using only the abstract allows us to reduce run times and hence costs.

Generating a title on the basis of a paper's abstract can be thought of as a special kind of summarization process: the abstract must be condensed into a short "sentence" that is maximally descriptive of its contents. Accordingly, we approach the task of automatically generating titles for academic abstracts as a summarization task. This is in line with existing research on title generation or comparable tasks. Unlike existing methods, however, our key contribution lies in experimenting with different post-processing strategies to further refine the quality of automatically generated titles. A particularly novel approach is that of using linguistic-stylistic rules, which we use to automatically filter out generated titles that do not adhere to accepted conventions on what constitutes an optimal academic title.

### 1.1 Related Work

We will review the literature on both title generation itself as well as headline generation, which pertains to the automatic creation of news-article headlines and is thus a task similar to title generation.

In contemporary research, automatic title/headline generation is often approached as a text summarization problem. The field of text summarization is generally split into two primary categories: extractive and abstractive summariza-



*Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, pages 74–83

| Model | Rouge-1 F-score | Rouge-2 F-score | Rouge-L F-score | Rouge-1 P | Rouge-2 P | Rouge-L P | Rouge-1 R | Rouge-2 R | Rouge-L R |
|---|---|---|---|---|---|---|---|---|---|
| **BART Large** | 0.249 | 0.077 | 0.214 | 0.256 | 0.081 | 0.218 | **0.267** | 0.083 | 0.231 |
| **T5 Large** | **0.255** | **0.094** | **0.231** | **0.270** | **0.100** | **0.244** | 0.262 | **0.097** | **0.237** |
| **Flan T5 Large** | 0.242 | 0.073 | 0.213 | 0.259 | 0.078 | 0.227 | 0.245 | 0.074 | 0.215 |

Table 1: Initial title generation results

tion. Presently, both these categories are addressed using methodologies anchored in the Transformer architecture (Song et al., 2020; Bukhtiyarov and Gusev, 2020; Liu and Lapata, 2019). A prevalent strategy for both forms of summarization is the encoder-decoder language model, exemplified by models like BertSumExt (Liu and Lapata, 2019) and PEGASUS (Zhang et al., 2020). Viewing summarization as a seq2seq challenge aligns well with the encoder-decoder framework, given the presence of a source and target text, akin to NMT scenarios. In this configuration, the generative decoder section conducts abstractive summarization. For strictly extractive endeavors, decoders are typically substituted by a specific classifier determining which input tokens will appear in the final summary. Another strategy is to fine-tune a GPT-2 (Radford et al., 2019) style auto-regressive model for the summarization task; this approach was adopted by both Koppatz et al. (2022) for headline generation and Mishra et al. (2021) for title generation.

Many contemporary title and headline generation methods have adopted metrics like BLEU or ROUGE to assess model performance (Matsumaru et al., 2020; Bukhtiyarov and Gusev, 2020; Tilk and Alumäe, 2017; Mishra et al., 2021); these are also standard for summarization evaluation. An exception is Koppatz et al. (2022), who also rely on manual structured review by domain experts to assess the quality of their automatically generated headlines. While human evaluations (especially if by domain experts) represent a gold standard, they are both expensive and time-consuming to obtain. This is especially true for academic titles, as evaluating how well a title captures the essence of an academic paper means being able to make sense of potentially extremely technical, specialized information.

## 2 Title Generation

### 2.1 Dataset

We created an initial dataset containing 136,640 academic articles. We obtained this dataset by downloading the Huggingface ArXiv dataset (https://huggingface.co/datasets/scientific_papers) and the Kaggle ArXiv dataset (https://www.kaggle.com/datasets/Cornell-University/arxiv), by selecting those articles that appeared in both datasets (by cross-referencing article ids), and by extracting the following information for each article: title, abstract, category, and full article text. Merging the two datasets was necessary as the Huggingface ArXiv dataset does not contain the full text of a paper, nor its category. While we are not using the full text of articles for this specific study, we plan on doing so in the future for a follow-up study, hence it was important for us to have a dataset containing all parts of the articles we use.

### 2.2 Testing out Different Models

As we decided to treat title generation as a summarization task, we looked into models that could best handle summarization. We considered three different state-of-the-art language models: T5 Large (Raffel et al., 2020), Flan T5 Large (Chung et al., 2022) and BART Large (Lewis et al., 2019).

T5 treats every NLP task as a text-to-text problem, which suits title generation perfectly –the model reads in the abstract as text input and outputs the generated title as text. Flan T5 Large stands as an improved version of the T5 model, having undergone fine-tuning across a blend of tasks. Demonstrating superior performance, Flan T5 outperforms its predecessor by handling more ubiquitous tasks. However, we wanted to see how these models compare on a less common task such as summarizing academic abstracts to generate titles. On the other hand, BART, with its unique architecture that is both auto-regressive and auto-encoding, can also be used to input an abstract and output a short summary in the form of a title. BART's ability to consider the context from both directions enables the model to generate fluent and coherent titles that accurately represent the content of the abstracts.

As a first step, we tried generating titles using all three language models. To do that, we split our dataset into a training subset, a validation subset and a test subset (70:15:15 split), and trained

BART Large, T5 Large and Flan T5 Large. We employed PyTorch as the framework for training our generating models and utilized the same set of hyperparameters to train each generating model. We trained all models for 3 epochs with a learning rate of 1e-5, a batch size of 6, and using the Adam optimizer (Kingma and Ba, 2014). We set the maximum input sequence length to 512 tokens and the maximum output sequence length to 128 tokens. To promote diversity and exploration during training, we employed a sampling parameter set to true. To ensure reproducibility and control the randomization during training, we set the random seed to 42.

### 2.3 Final Model Selection

We evaluated the performance of our three models by comparing the title generated by each model to the original title of the paper, to determine how (dis)similar artificial titles were with respect to the original. While similarity to the original title is not in itself a measure of the quality of a machine-generated title (a maximally dissimilar title might still be an excellent title), we reasoned that computing similarity scores could be an at least partial indication of a machine-generated title being "human-like" (i.e. similar to what a human writer would come up with) and hence a good title. Considering that most of the evaluation mechanisms based on similarity scores are highly correlated (Fabbri et al., 2021), we decided to resort to ROUGE (Lin, 2004). The results are given in Table 1. T5 Large performed best on almost all ROUGE metrics except ROUGE 1 Recall, where BART Large performed better.

One of the goals of our study was to determine how much we could improve the performance of our best-performing model through further post-processing. Based on Table 1, we thus decided to settle on T5 Large as the model to use for all additional post-processing experiments.

Our post-processing consisted of two steps: refining title generation through SciBERT, and refining title generation through linguistic-stylistic rules.

## 3 Post-Processing, Step 1: SciBERT

For the first post-processing step, we wanted to determine whether we could obtain higher ROUGE scores by generating multiple titles for each abstract using T5 Large, selecting the most representative titles out of all those generated, and creating a synthetic dataset using these most representative titles.

### 3.1 Extraction of Oracles

Using T5 Large, we generated five titles for each of the abstracts in our training and validation subsets. Below we provide an example of the types of titles that were generated using T5 Large. Using the example abstract displayed in Fig.1, originally from a paper by Mallick et al. (2017) titled *"Energy-dependent variability of the bare Seyfert 1 galaxy Ark 120"*, we generated the following five titles:

1. *A long-period XMM-Newton observation of the bare Seyfert 1 galaxy Ark 120*
2. *Ark 120: spectral-timing analysis of XMM-Newton observance over four consecutive orbits in 2014*
3. *Ark 120: spectral-timing analysis and hardness-intensity diagram*
4. *Broad-band X-ray spectroscopy of Ark 120: A spectral-timing analysis of a long 486 ks XMM-Newton observation*
5. *A spectral-timing analysis of the long 486 ks XMM-Newton observation of the bare Seyfert 1 galaxy Ark 120*

For each of the titles generated by T5 Large, we computed a ROUGE score by comparing the generated title to the paper's original title. Following this, we created a synthetic dataset with a specific labeling scheme: the title with the top ROUGE score was labeled as "1" (we refer to this as the *oracle*), while the title with the lowest score received a "0" label. Note that titles with intermediate scores were neither labeled nor included in this dataset. The purpose of this was to focus on the two most distinct title generations for a given abstract.

### 3.2 Fine-tuning SciBERT on the Synthetic Dataset

We trained SciBERT (Beltagy et al., 2019) on the obtained synthetic dataset. We decided to use SciBERT as it outperforms BERT in a variety of tasks in the scientific domain (Beltagy et al., 2019) and achieves SOTA performance in multi-class text classification on the SciCite dataset (Cohan et al., 2019).

In our study, we used a modified version of SciBERT, which was previously pre-trained to optimize

**Original Abstract (Mallick et al. 2017)**

We present results from a detailed spectral-timing analysis of a long ∼486 ks XMM-Newton observation of the bare Seyfert 1 galaxy Ark 120 which showed alternating diminution and increment in the 0.3-10 keV X-ray flux over four consecutive orbits in 2014. We study the energy-dependent variability of Ark 120 through broad-band X-ray spectroscopy, fractional root-mean-squared (rms) spectral modelling, hardness-intensity diagram and flux-flux analysis. The X-ray (0.3-10 keV) spectra are well fitted by a thermally Comptonized primary continuum with two (blurred and distant) reflection components and an optically thick, warm Comptonization component for the soft X-ray excess emission below ∼2 keV. During the first and third observations, the fractional X-ray variability amplitude decreases with energy while for second and fourth observations, X-ray variability spectra are found to be inverted-crescent and crescent shaped, respectively. The rms variability spectra are well modelled by two constant reflection components, a soft excess component with variable luminosity and a variable intrinsic continuum with the normalization and spectral slope being correlated. The spectral softening of the source with both the soft excess and UV luminosities favours Comptonization models where the soft excess and primary X-ray emission are produced through Compton up-scattering of the UV and UV/soft X-ray seed photons in the putative warm and hot coronae, respectively. Our analyses imply that the observed energy-dependent variability of Ark 120 is most likely due to variations in the spectral shape and luminosity of the hot corona and to variations in the luminosity of the warm corona, both of which are driven by variations in the seed photon flux.

Figure 1: Abstract Example

the performance of the model for scientific text analysis. This prediction model is influenced by the success of using the transformer-model architecture for the classification of sentences in extractive summarization (Liu and Lapata, 2019) or later applied in fact-checking summarization (Atanasova et al., 2020). In our experiment, we fine-tune SciBERT model to generate a probability for each generated title. This probability interprets how similar the generated title is to the original (human) title, while the original title does not enter the model in the prediction. This model learns to distinguish the titles that are most and least similar to the original, human title. Our fine-tuned SciBERT model could be applied as a classifier as well, but we only wanted to rank our generated titles by assigning a SciBERT probability value to each generated title.

To fine-tune our SciBERT model, we followed the design and optimization decisions described in Beltagy et al. (2019) and Devlin et al. (2019). Our approach involved using a linear one-layer feed-forward classifier with the ReLu activation function. The classifier took the last hidden state of the [CLS] token as input, effectively using it as the sequence's features. We conducted extensive experiments to determine the optimal hyperparameters for fine-tuning SciBERT. This included varying the number of epochs (ranging from 2 to 5), batch sizes (16, 32, or 50), learning rates (5e-5, 5e-6, 1e-5, or 2e-5), and incorporating or excluding a dropout rate of 0.1. To optimize the training process, we utilized the AdamW optimizer and cross-entropy loss. Our best results were achieved by fine-tuning the models for 3 epochs, with a batch size of 32 samples, a learning rate of 5e-5, and no dropout applied. Following this, we applied a linear warmup and linear decay technique as described in Devlin et al. (2019). We employed the softmax function to determine probabilities for predictions, which served as the initial selection or ranking score for finding the best title candidate.

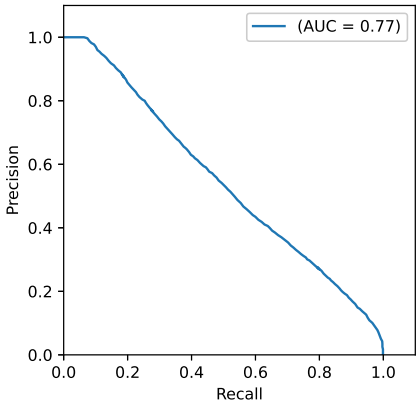


Figure 2: PR curve on testing dataset

### 3.3 Results

To evaluate the effectiveness of title generation selection, employing Precision-Recall (PR) curves and the corresponding Area Under Curve (AUC) (see Fig.2) provides comprehensive and robust testing (Boyd et al., 2013). This quality approach allows for an exhaustive evaluation of the model's performance across a broad spectrum of probability rankings.

The achieved performance of the model Area Under the Precision-Recall Curve (AUC-PR) 0.77 is particularly interesting because we always labeled the generated titles with the highest and lowest ROUGE scores in the synthetic training dataset.

In the initial row of Table 2, we consider the baseline model as a fine-tuned T5 Large, producing a single title for each abstract, identical to row 2 in Table 1. In the subsequent row of Table 2, we analyze the same fine-tuned T5 Large model, but this time generating five titles for each abstract. From this set of machine-generated titles, we select the machine-generated title with the highest SciBERT probability. Those selected titles for each

| Model | Rouge-1 F-score | Rouge-2 F-score | Rouge-L F-score | Rouge-1 Precision | Rouge-2 Precision | Rouge-L Precision | Rouge-1 Recall | Rouge-2 Recall | Rouge-L Recall |
|---|---|---|---|---|---|---|---|---|---|
| **T5 - Baseline** | 0.255 | 0.094 | 0.231 | 0.270 | 0.100 | 0.244 | 0.262 | 0.097 | 0.237 |
| **T5 + SciBERT** | **0.281** | **0.103** | **0.253** | **0.295** | **0.109** | 0.265 | **0.291** | **0.107** | **0.262** |
| **T5 + SciBERT + Linguistic rules** | **0.281** | **0.103** | **0.253** | **0.295** | **0.109** | **0.266** | 0.288 | **0.107** | 0.260 |
| **Oracles by Rouge 1 F-score** | 0.393 | 0.176 | 0.352 | 0.416 | 0.188 | 0.372 | 0.400 | 0.179 | 0.357 |

Table 2: Improved title generation results

abstract are compared to original human titles. Consequently, in Table 2, the ROUGE metric is consistently calculated based on the same original human titles, although the chosen artificial titles may differ across various models. When comparing the second row to the first row in the table, we observed a significant improvement in the overall title quality by utilizing SciBERT for title selection. This improvement is evident across all ROUGE metrics. However, when comparing the second row with the last row, where ROUGE utilized artificial titles for evaluation, it becomes evident that there is still considerable room for further improvement.

## 4 Post-Processing, Step 2: Linguistic Rules

To ensure the quality of artificially generated titles for academic papers, we also implemented a second post-processing step that involved evaluating each title against a set of linguistic-stylistic rules. These rules were designed to adhere to the conventions of academic titles while at the same time enhancing clarity, conciseness and outreach potential of the paper. We employed six distinct rules; titles that met all rules were assigned a score of 6, titles meeting only 5 rules were assigned a score of 5, and so on. The six rules we used are outlined below. We arrived at these rules after consulting several papers and online resources on how to write effective titles for academic papers.

I. *Title Length*: Titles should be concise, but also not so short that it is unclear what the paper is about, or how it differs from related articles discussing the same topic (Knight and Ingersoll, 1996; Paiva et al., 2012; SHU Library, 2020). Therefore, for this category, we gave a 0 score to titles longer than 16 words (Wordvice, 2023) or shorter than 5 words (USC Libraries, 2023), and a score of 1 otherwise.

II. *Geographical Locations*: Paiva et al. (2012) found a negative correlation between mentions of specific geographical locations (e.g. "Mortality Rates in Sub-Saharan Africa") in titles and number of citations per article. Accordingly, we gave a score of 0 to titles containing any reference to geographical locations, and a score of 1 otherwise.

III. *Forbidden Punctuation Marks*: Paiva et al. (2012) found a negative correlation between the number of citations and the presence of exclamation marks, question marks, and dashes in titles (see also USC Libraries (2023)). We thus gave a 0 score to titles containing these punctuation marks: '?', '-', '!'.

IV. *Suboptimal Nouns*: According to Knight and Ingersoll (1996), phrases such as "The Effects of," "A Comparison of," "The Treatment of," and "Reports of a Case of" should be avoided in titles (see also SHU Library (2020); USC Libraries (2023)). Accordingly, we gave a 0 score to titles containing the nouns "analysis," "effects," "comparison," "treatment," "report/reports".

V. *Passive Verbs*: Active voice should be preferred in academic titles (SciPress, 2017). We gave a 0 score to titles containing verbs in the passive voice, and a 1 score otherwise.

VI. *Abbreviations*: We gave a 0 score to titles that included abbreviations. This rule aimed to ensure that the titles are accessible to a wide range of readers without relying on specialized knowledge or acronyms (SHU Library, 2020; Wordvice, 2023).

To assign these linguistic scores, we wrote Python text-processing rules that would take generated titles as input and assign to each title a score from 0 to 6 based on how many of the above rules each title met. While there are many tips on how to write effective titles for scientific publications, we specifically chose the above-mentioned rules as it is easy to code text-processing scripts that check automatically whether these rules are met. The motivation for adding this additional post-processing step was thus to obtain a simple and computationally inexpensive way of further checking machine-generated titles for adherence to standard norms in academic writing. We reasoned in particular

that adding this type of post-processing could partially eliminate/reduce the scope of work of any human evaluator who was to manually check each machine-generated title for quality, which can be a lengthy and costly process.

### 4.1 Linguistic Score Results

We normalized the linguistic scores using the following formula:

$$linguistic\ score = \frac{\sum (all_{scores})}{6}$$

where $all_{scores}$ indicates the list of linguistic rules to be summed up in the equation.

This allowed us to obtain a total linguistic score ranging from 0.0 to 1 for each of the generated titles, 1 being a title that meets all six linguistic rules, 0.0 being a title that flouts all rules. For each title, we then multiplied this normalized linguistic score by the SciBERT probabilities obtained in the previous post-processing step to obtain a combined SciBERT*linguistic score. Titles with the highest SciBERT*linguistic score were chosen as the best titles out of all generated options.

We also calculated the number of times a title ranked first by the combined SciBERT*linguistic score would also be the title ranked first by SciBERT probability alone. We looked at the titles generated for 20,000 abstracts, and in this sample, the highest-ranked title was the same in 18,770 cases (= 1,230 differences). If we examine these discrepancies more closely, we find that the majority occur because some of the highest-ranked titles according to SciBERT probability exceed 16 words in length.

While the addition of a linguistic post-processing step has not yielded dramatically different results, it did have an effect. It is possible that if more stylistic rules were to be implemented, or if more restrictive rules were to be adopted (for example, maximum title length could be reduced to 13 words, as suggested by different academic style guides), this type of linguistic post-processing could be useful in automatically discarding a larger chunk of title generations that do not comply with academic guidelines.

In Table 2 (third row), we also see how ROUGE metrics on SciBERT probability ranking change if linguistic scores are considered as well. We see in particular that, if linguistic scores are also applied, ROUGE scores are almost comparable to T5 model with SciBERT probability ranking only (second row). Note however that this could also be due to the original title flouting one or more of the linguistic rules we selected for this post-processing step.

## 5 Human Evaluations

As a final step of this study, we sought to understand the nuances of human evaluations vis-à-vis machine-generated academic titles. To achieve this, we asked three human annotators to evaluate the titles that our model generated for a selection of 40 abstracts from our dataset. All three evaluators were academics themselves.

We decided to include a human evaluation step for several reasons. First of all, we wanted to determine whether title evaluation is a purely subjective matter, or whether there is some consensus among different individuals concerning what constitutes a good or a bad title. Secondly, we wanted to determine how feasible of a task it is to ask human annotators to evaluate the quality of titles of academic papers. In the specialized realm of academic articles, titles generally refer to highly technical information. This raised the question of the extent to which human evaluators could accurately judge if an academic title captures the essence of a paper's technical depth: even if one only selects evaluators who are at least familiar with the field of research of a particular set of abstracts, it is impossible to expect that each evaluator will be able to fully understand all of the abstracts they are asked to review. Finally, we were interested in determining whether the subtleties introduced by the linguistic improvements in our second post-processing step might resonate more profoundly with human evaluators.

Evaluators were presented with the original abstract, five machine-generated titles, and the original title of the paper from which the abstract was derived, resulting in a total of six titles to be evaluated. Note that we randomly selected 40 abstracts from the set of 1,230 abstracts for which SciBERT and SciBERT*linguistics outputted distinct highest-ranking titles (see again section 4.1).

The sequence in which the titles were presented was randomized. Furthermore, to prevent any attempt by the evaluators to evaluate the machine-generated titles by comparing them with the original title, evaluators were told that all titles, without exception, were machine-generated.

The evaluators were asked to read the abstract, read each of the six titles, and then pick i) what they thought was the 'best title'—that is, the title

they perceived as the most fitting given the content of the abstract and the intrinsic qualities of the title itself, ii) what they thought was the second-best title, and iii) what they thought was the worst title out of all six title options. Our decision to request evaluators to pinpoint the best, second-best, and worst titles, rather than having them rank all six titles from best to worst, was twofold. Firstly, we anticipated that the deeply technical nature of some abstracts could pose challenges in the ranking process; we figured that simply selecting best, second-best and worst title would be a more feasible task. Secondly, we recognized that when presented with a set of titles potentially bearing very close similarities, distinguishing and ranking all six titles on a gradient scale might be problematic. The inclusion of the original title amidst the machine-generated ones also served a dual purpose. First of all, we wanted to assess if evaluators would rank the original title of the paper as 'best title'. Moreover, this approach also allowed us to determine how frequently machine-generated titles are perceived as superior to the original title of a given paper.

### 5.1 Inter-Annotator Agreement

In order to ascertain the inter-annotator agreement rate, we calculated Fleiss' kappa (Fleiss, 1971). The results are reported in Table 3:

| Title | Fleiss' Kappa |
|---|---|
| Best Title | 0.5805 |
| Second Best Title | 0.5195 |
| Worst Title | 0.5962 |

Table 3: Fleiss' Kappa Results

For the interpretation of Fleiss' kappa values, the following ranges are generally used:

| Range | Interpretation |
|---|---|
| $\kappa > 0.75$ | Excellent agreement |
| $0.40 < \kappa \leq 0.75$ | Fair to good agreement |
| $\kappa \leq 0.40$ | Poor agreement |

Table 4: Interpretation of Fleiss' Kappa Values

We further investigated the degree of consensus among evaluators by calculating how many times out of 40 (i.e. the total number of abstracts evaluated by our annotators) at least two reviewers both picked the same title as best, second-best, or worst title:

- Number of times at least 2 reviewers agreed on best title: 31 times
- Number of times at least 2 reviewers agreed on second best title: 17 times
- Number of times at least 2 reviewers agreed on worst title: 29 times

Based on these results, we can conclude that evaluators seemed to frequently agree on what they deemed to be the best and worst titles, even despite the very technical nature of the abstracts and titles they were asked to evaluate. This challenges the notion that title evaluation is purely subjective, suggesting that consensus among different individuals is in fact quite attainable. Furthermore, these results also indicate that the evaluators' rankings were deliberate and informed, rather than random.

### 5.2 Human Evaluation vs. Different Methods

As a final step, we wanted to determine how human evaluations relate to the different title selection methods we explored in this paper. To do so, we went through the selections made by our three evaluators, and created a set of so-called *strong candidate* machine-generated titles. A machine-generated title was deemed to be a *strong candidate* if either of the following conditions were met:

i. At least two evaluators selected that specific machine-generated title as their "best title" or "second best title" choice.

ii. The machine-generated title was selected by an evaluator who also selected the original title for that abstract as their "best title" or "second-best title" selection. E.g. if an evaluator selected the original title as their "second-best" choice, the machine-generated title that they selected as their "best" choice was considered to be a *strong candidate*.

These two conditions rested on the following assumptions:

- Some machine-generated titles might be perceived by evaluators as being of higher quality than the original paper title.
- If an evaluator chooses the original title as their "best title" selection, we assume they understand the contents of the abstract well enough, and thus that any title that they rank as "second-best title" must also be a good title for that specific abstract.

- If an evaluator chooses the original title as their "second-best title" selection for a given article, we assume they understand the contents of the abstract well enough, and thus that any title that they rank as "best title" must also be both appropriate for that specific abstract, and possibly a better title than the original title.

- If at least two evaluators select a given machine-generated title as their "best" or "second-best" selection, the title must be a good title for that abstract.

Based on these criteria, we compiled a set of *strong candidate* machine-generated titles for each of the 40 abstracts evaluated by human evaluators. The set typically comprised a maximum of two candidate titles per abstract.

After identifying the *strong candidate* titles for each of the 40 abstracts, we compared how effective each of the three selection methods used in this paper (Rouge, SciBERT alone and SciBERT* Linguistics) was in capturing human rankings. Specifically, we checked if the title ranked as highest by each of these three methods was part of the *strong candidates* list. If the title ranked as highest by a given method was part of the *strong candidates* list, it was marked as a "correct selection".

Our aim was to ascertain the number of correct selections each method achieves out of 40 trials (i.e. our forty abstracts). The results are reported below:

- Rouge (Oracles) made a correct selection 8 times,

- SciBERT made a correct selection 7 times,

- SciBERT*Linguistics topped the list with 10 correct selections.

Although the frequency of correct selections is not particularly high, likely due to the challenging nature of the task, it is interesting to see that Rouge outperformed SciBERT, especially since we trained SciBERT using similarities identified by Rouge. Furthermore, it is noteworthy that the integration of linguistic principles with SciBERT elevated the number of correct selections from 7 to 10, making this the most successful method when considering human evaluations.

## 6 Concluding remarks

We hypothesized that automatically generating an adequate research paper title can be treated as a high-level text summarization problem: a title can be seen as a very condensed summary of the paper's abstract. In this context, we have presented a novel post-processing approach that combines a SciBERT prediction model enhanced with linguistic-stylistic rules to tackle the problem of finding adequate titles for research papers.

We started by considering three powerful NLG models (BART Large, T5 Large, FLAN T5 Large) and evaluating their text-generation results against the original titles. Out of these models, we chose the best-performing one: T5 Large. T5 Large was then set up to generate multiple diverse titles for the same abstract. For each abstract, we generated five different titles and again compared them against the original title of the paper using ROUGE. Subsequently, we created a synthetic dataset by labeling the title with the top ROUGE score as "1", and the title with the lowest ROUGE score as "0"; we then trained SciBERT on this synthetic dataset. In addition, we defined a set of linguistic rules a title should adhere to. Based on these rules, we calculated a normalized score between 0 and 1 for each generated title. We then multiplied this normalized linguistic score by the SciBERT probabilities obtained in the previous post-processing step.

We also assessed our title selection methods in relation to human evaluations. The human evaluations section was instrumental in providing insights into the nuances of human perspectives concerning machine-generated academic titles. Our findings revealed that while title evaluation can be subjective to some extent, there exists a noticeable degree of consensus among evaluators about what constitutes a quality title. The performance comparison between various methods, with the linguistics-enhanced SciBERT emerging as the most successful in capturing human evaluations, further underscores the effectiveness of our proposed approach.

In the future, we would like to experiment with generating titles using a paper's conclusion section rather than its abstract. Working with conclusions is more complicated than working with abstracts, as not all papers have a self-standing conclusion section, yet an improvement of our results might be obtained as conclusions often define in more detail what the key contributions of a paper are.

## 7 Acknowledgements

This study was partly supported by a grant from Rannís, the Icelandic Institute for Research, and a grant by the European Union (Women TechEU, European Innovation Ecosystems programme, Horizon Europe). Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This work has been partially supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

## References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Kendrick Boyd, Kevin H Eng, and C David Page. 2013. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer.

Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of transformer-based models for news headline generation. In *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, pages 54–61. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Boer Deng. 2015. Papers with shorter titles get more citations. *Nature*, 2(8):150266.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenneth L Knight and Christopher D Ingersoll. 1996. Structure of a scholarly manuscript: 66 tips for what goes where. *Journal of Athletic Training*, 31(3):201.

Maximilian Koppatz, Khalid Alnajjar, Mika Hämäläinen, and Thierry Poibeau. 2022. Automatic generation of factual news headlines in finnish. *arXiv preprint arXiv:2212.02170*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Labani Mallick, Gulab C Dewangan, IM McHardy, and Mayukh Pahari. 2017. Energy-dependent variability of the bare seyfert 1 galaxy ark 120. *Monthly Notices of the Royal Astronomical Society*, 472(1):174–188.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. *arXiv preprint arXiv:2005.00882*.

Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and G Srinivasaraghavan. 2021. Automatic title generation for text with pre-trained transformer language model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 17–24. IEEE.

Carlos Eduardo Paiva, João Paulo da Silveira Nogueira Lima, and Bianca Sakamoto Ribeiro Paiva. 2012. Articles with short titles describing the results are cited more often. *Clinics*, 67:509–513.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

SciPress. 2017. 12 key tips on how to write a good research paper title. https://shorturl.at/cMQVW. Last Accessed: 14th of October 2023.

SHU Library. 2020. Organizing academic research papers: Choosing a title. https://library.sacredheart.edu/c.php?g=29803&p=185911. Last Accessed: 14th of October 2023.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8902–8909.

Ottokar Tilk and Tanel Alumäe. 2017. Low-resource neural headline generation. *arXiv preprint arXiv:1707.09769*.

USC Libraries. 2023. Research guides: Organizing your social sciences research paper: Choosing a title. https://libguides.usc.edu/writingguide/title. Last Accessed: 14th of October 2023.

Wordvice. 2023. How to write a research paper title with examples. https://blog.wordvice.com/how-to-write-the-perfect-title-for-your-\research-paper/. Last Accessed: 14th of October 2023.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.