PiGGyBacking off of PEGASUS: Pre-training with Gap-sentences for Government Bills

Stanford CS224N Custom Project

Alice Chen
Department of Computer Science
Stanford University
alicezyc@stanford.edu

Evelyn Choi
Department of Computer Science
Stanford University
echoi1@stanford.edu

Karsen Wahal
Department of Computer Science
Stanford University
kwahal@stanford.edu

Abstract

Congressional bills are notoriously long and difficult to understand, limiting citizen engagement. As a result, summarizing legislation in a low-resource setting is a foundational step in making our democracy more inclusive. In this project, we propose a novel approach to bill summarization, PIGASUS, that is based off of Google's state-of-the-art summarizer, PEGASUS. PIGASUS makes government bills more accessible by creating quality summaries in low resource settings. Our model differs from PEGASUS by using a state-of-the-art extractive summarization technique, TextRank, to select masked sentences for pretraining. To experiment with this, we implement a second "pretraining" stage that builds off the pretrained PEGASUS model, mimicking the effect of pretraining a large summarization model from scratch. We find that in low-resource settings, PIGASUS is not only more efficient in learning, but also generates higher quality summaries that capture more important information than PEGASUS does in the same settings. This suggests that using TextRank to mask sentences during pretraining may improve the quality of summarization models.

1 Key Information to include

- Mentor: Eric Frankel
- External Collaborators / Sharing Project: N/A

2 Introduction

Congressional bills are notoriously complex and written in "legalese," making them incredibly difficult for citizens to understand (Institute, 2020). Bills can also be hundreds of pages long, making them even less accessible to the layperson: the Affordable Care Act, one of the most consequential pieces of healthcare legislation, is 906 pages long (Frumin, 2016). Summarizing legislation to be understandable for the ordinary citizen at a low cost is critical to make our democracy more inclusive.

However, summarizing texts is a labor-intensive and exacting task to complete manually – especially for long, jargon-laden documents. Thus, in the era of big data, text summarization has been a desirable task to automate with natural language processing. Previous work in high-quality automated text summarization has required significant computational power and resources. However, considering the rareness of large document-summary datasets, it is important to build successful summarization

models within the more realistic context of low-resources, with respect to both data and compute power.

Additionally, previous research has focused on the general task of text summarization without a particular focus to legislation. In 2019, Zhang et al. (2019) developed a novel pretraining objective for the summarization model PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization). PEGASUS was pretrained on the Colossal and Cleaned version of Common Crawl (C4), which contains 350M web-pages with a diverse-array of text styles and content. This state-of-the-art pretrained model is performs well on a variety of text styles when fine-tuned on only 100 and 1000 examples. However, PEGASUS was only fine-tuned on federal bills – not state bills.

Both state and federal bills are consequential to the average citizen. Although the general structure of state and federal bills can be similar, there are often differences in the language used, the formality of text, and the bills' specific structures. Thus, we aim to expand upon PEGASUS by 1) performing a 'second round of pretraining' using our novel method of implementing TextRank to mask sentences, and by using datasets with related, but not identical, text style as the down-stream summarization task, and by 2) fine-tuning this doubly-pretrained model in a low-resource setting on state (California) bills from the BillSum dataset (Kornilova and Eidelman, 2019).

We conduct an in-depth analysis of the quality of different summarization techniques, including: extractive summarization, a second round of our new approach to pretraining, and fine-tuning on different numbers of examples. Our results are specifically tailored to low-resource settings; we conduct additional training on only 400 unique data points and operate with limited computational resources. Using a novel approach to PEGASUS's pretraining objective, we generate a model that can summarize state-level (California) bills with a quality not seen before in previous summarization models, such as PEGASUS. Specifically, we find that using extractive summarization techniques as a tool for Gap Sentence Generation (GSG) on a related dataset to the downstream task improves the performance of the model. Given that our model achieved such success in a low-resource setting, our results indicate that given higher resources, a model that builds off PEGASUS and adopts our approach to summarization could summarize specific legislative texts with even higher quality.

3 Related Work

There are two categories of text summarization: extractive summarization involves copying the most 'important' sentences directly from the document and concatenating them to create a summary, while abstractive summarization involves generating "novel" text to summarize the input document without losing any relevant information, in a similar manner as a human would (Allahyari et al., 2017). One of the most common extractive summarization techniques uses TextRank, as detailed in section 4.2 (Mihalcea and Tarau, 2004). We focus on abstractive summarization, given that it is considerably more applicable and complex, as it requires both semantic and lexical analysis (Widyassari et al., 2022).

Specifically, we focus on automated summarization of state bills in the United States. Previously, there have been substantive efforts in the automatic summarization of legal documents – a related, but not identical, genre of text. For example, Anand and Wagh (2022) created an unsupervised, extractive summarization method tailored to legal documents that used neural networks to classify important and unimportant sentences, and ranked the important sentences with a similarity score. Huang et al. (2021b) built a summarization model that takes advantage of the global information captured by a graph representation of the input document, before this graph representation is fed into a seq2seq model for the abstractive summarization of legal public opinion news. However, building these models from scratch require immense computational power and resources that are often inaccessible. Also, these models are already trained on summarizing a specific style of text and may be difficult to generalize to legislation. Thus, fine-tuning a general, pretrained model for the abstractive summarization of US bills is the most feasible task in low-resource settings.

The current state-of-the-art pretrained summarization model is PEGASUS (Zhang et al., 2019). Notably, PEGASUS was fine-tuned on the BillSum dataset for the summarization of federal bills, but state (California) bills were purposely excluded from the fine-tuning dataset. The structure, language, and formality of a bill varies from state to state, and also from the state-level to the federal-level. Therefore, it is important to create a model that can summarize state bills.

4 Approach

4.1 Background

PEGASUS is a standard encoder-decoder transformer. Its novelty lies in its use of Gap-Sentences Generation (GSG), a self-supervised pre-training objective for abstractive summarization that involves masking out and predicting entire sentences in a given input. PEGASUS picks the masked sentences by calculating each sentence's ROUGE1-F1 score (described in section 5.2) relative to the rest of the document. This score is used to find the most "important" sentences, and is calculated based on matching n-grams between the sentence and the rest of the document. Notably, PEGASUS selects sentences independently and double-counts identical n-grams in its ROUGE1-F1 calculation.

The GSG loss used for PEGASUS's pretraining objective, is, at its core a cross-entropy loss function. That is, given an input sequence with a masked contiguous span of text with tokens Y, and context tokens X (all other tokens in the sequence besides Y),

$$L_{train} = -\log P(Y|X;\theta). \tag{1}$$

PEGASUS uses a similar cross-entropy loss for downstream tasks, to compare its tokenized output summary with a tokenization of the target summary.

4.2 Baselines

We implement two baselines. First, we implemented and evaluated TextRank, an extractive summarization algorithm, on California bills from the BillSum dataset (LIANG, 2020). TextRank is a graph-based algorithm used to rank phrases in a given body of text by importance. Our implementation uses TextRank to first rank the important phrases, before generating a phrase vector for the top 10 phrases. TextRank uses the following formula to calculate the weights, or importance of phrases, for graph-based ranking, as shown in Mihalcea and Tarau (2004):

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{2}$$

Then, the algorithm ranks sentences in the document by importance. It does so by 1) for each sentence, creating a sentence vector that keeps track of which of the top 10 phrases appear in the sentence, and 2) calculating the Euclidean distance between this sentence vector and the phrase vector. We concatenated the top 3 sentences to create the summary. Our code is modified from Nathan (2016).

Second, we evaluated PEGASUS-large (Zhang et al., 2019), the PEGASUS model pretrained on the C4 and HugeNews Datasets, without a second round of pretraining and fine-tuning, on the same bills. This model acts as a baseline for our low-resource fine-tuning. We use ROUGE scores to evaluate both baselines relative to the target summaries. We denote this model as PEGASUS-large-0.

4.3 Fine-tuning PEGASUS and Model Generalizability

To analyze PEGASUS's ability to generalize to state bills, we evaluate PEGASUS-large on California bills. To simulate low-resource settings, we fine-tune PEGASUS-large on 10 and 100 California bills (models PEGASUS-large-10 and PEGASUS-large-100, respectively), which has not been implemented previously. Additionally, we evaluate two existing fine-tuned models that were finetuned on much more, but different, data, to evaluate PEGASUS' ability to generalize past finetuning. We evaluate PEGASUS-BillSum and PEGASUS-XSum, PEGASUS models fine-tuned on around 19,000 federal legislation and around 200,000 news articles, respectively, to evaluate the quality of summarization models fine-tuned on different, but related texts (Kornilova and Eidelman, 2019; Narayan et al., 2018).

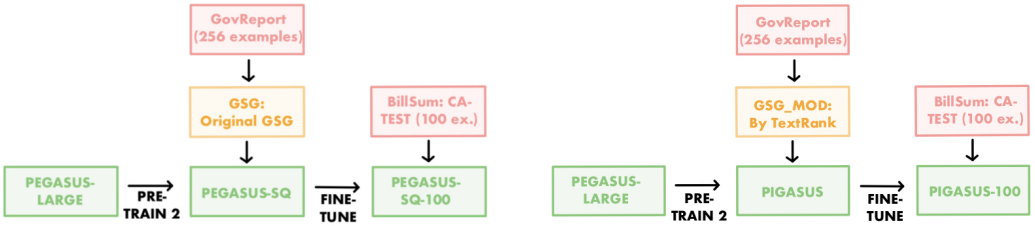
4.4 Pretraining 2.0

We experiment with a second "pretraining" round, where we pretrain PEGASUS-large on 256 government reports from the GovReport Dataset (Huang et al., 2021a). We conduct an additional

pretraining round to explore whether additional pretraining on a related, but different dataset improves the quality of the downstream task. Indeed, the GovReport dataset has similar terminology and style to the California bills, while still being more general than the data we fine-tune on.

We also experiment with two different "pretraining steps" that differ in how they choose sentences for masking in Gap Sentence Generation (GSG). The first chooses sentences for GSG using ROUGE1-F1, as in the original paper. The second represents an original approach to selecting sentences, whereby the model chooses sentences for GSG by using the TextRank algorithm, described in section 4.2, to determine the most important sentences in the text. Specifically, while the TextRank algorithm calculates the most important sentences in a text in order to concatenate them into a summary, our model uses this method to find the most important sentences in a text and masks out those sentences for GSG. In both cases, we mask out 30% of sentences. Given our low-resource setting, we conduct a second round of pretraining, rather than pretrain an entirely new model, in order to test whether our approach to masking sentences improves upon the original PEGASUS model.

We fine-tune both doubly-pretrained models on 100 California bills. Throughout this paper, we denote the model using the original pretraining method as "PEGASUS-squared-100", and the model using our novel pretraining method as "PIGASUS-100." Lastly, we evaluate the models on our test set of 186 California bills. The full approach can be seen in Figure 1.



(a) PEGASUS-squared-100 Pipeline

(b) PIGASUS-100 Pipeline

Figure 1: Pretraining 2.0 and Fine-tuning Pipeline

5 Experiments

5.1 Data

Our primary dataset is the BillSum dataset (Kornilova and Eidelman, 2019), which contains 22,218 federal and California bills, as well as their prepared summaries. The federal bills are from the 103rd-115th (1993-2018) sessions of Congress, while the California bills are from the 2015-2016 session. The BillSum corpus focuses on mid-length legislation from 5,000 to 20,000 character in length. We use the California split of the data, which contains 1,237 bill-summary pairs, to fine-tune and evaluate our various models. Details on fine-tuning can be found in sections 5.2 and 5.3.

In order to conduct the second round of pretraining, we use the GovReport dataset, consisting of about 19,500 U.S. government reports published by the Government Accountability Office and Congressional Research Service, with expert-written abstractive summaries (Huang et al., 2021a). However, we do not use these summaries, and simply use the raw text of the reports to conduct pre-training. Details can be found in section 5.3.

5.2 Evaluation method

We evaluate all models on our test set of 186 bill-summary pairs from the 'CA-TEST' split of BillSum. We use two evaluation metrics. First, we use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores, a reference used to calculate the amount of the reference summary that the system summary captures. Specifically, we use ROUGE1-F1, ROUGE2-F1, and ROUGEL-F1 scores. Scores are calculated with the assistance of the Python package rouge-score-0.1.2 (pyp). ROUGE scores are calculated as follows:

$$\text{ROUGE-F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

Where,

$$\text{recall} = \frac{\text{\# n-grams in both model and reference}}{\text{\# n-grams in reference}} \tag{4}$$

$$\text{precision} = \frac{\text{\# n-grams in both model and reference}}{\text{\# n-grams in model}} \tag{5}$$

In the case of ROUGEL, the numerator of the $recall$ and $precision$ formulas is replaced by the length of the longest common subsequence, and the denominator is the number of unigrams.

Second, we use human evaluation. We average the ROUGEL-F1 scores in the PEGASUS-Squared-100 and PIGASUS models, and sample the worst summary, the median summary, and the best summary. Given that summaries included substantial jargon, we use PEGASUS-paraphrase to paraphrase each sentence of the summary into more jargon-free and understandable language (tuner007). We use PEGASUS-paraphrase so that evaluators judge primarily based on content, rather than style, which we consider a more minor problem, given that paraphrasing could be incorporated as an additional step without excessive difficulty. Then, we survey 37 unbiased individuals to rank each PEGASUS-squared-100 and PIGASUS-100 summary (without knowing which model was which), relative to the target summary, to determine the relative quality of each model. The bills that were used can be found in Appendix A.4.

5.3 Experimental details

Table 5 in Appendix A.1 depicts all fixed hyperparameters that were used when fine-tuning and pretraining all models. For pretraining or fine-tuning for all models, we used the adafactor optimizer (Shazeer and Stern, 2018). Table 1 depicts the results of our hyperparameter search for number of epochs or training steps when fine-tuning each model. Training and validation loss over time during finetuning are illustrated in Graphs 3 and 4 of Appendix A.2. Due to computation constraints, PEGASUS-large-10 was fine-tuned for a maximum of 80 epochs, while PEGASUS-large-100, PIGASUS-100, and PEGASUS-squared-100 were fine-tuned for a maximum of 25 epochs. All models were fine-tuned on the same 100 California BillSum examples (and thus end with -100) except for PEGASUS-large-10 and PEGASUS-large-0. The validation set for fine-tuning consisted of 50 bills. We based our finetuning script on jiahao87 (2021), and modified it to create a custom loss and trainer for our pretraining code. Due to computational constraints, we froze the encoder of the model to halve the number of trainable weights during pretraining and fine-tuning. As in the original PEGASUS paper, all models tokenized input using SentencePiece and truncation, and had maximum token size of 1024 Kudo and Richardson (2018).

	PEGASUS -large-10	**PEGASUS -large-100**	**PIGASUS-100**	**PEGASUS -squared-100**
# Epochs	80	25	25	12.4
# Steps	400	1250	1250	624
Train Loss	3.291	3.763	3.709	3.210
Val. Loss	6.357	4.995	4.910	3.654

Table 1: Final Loss and Chosen Hyperparameters Model for fine-tuning

Table 2 shows the results of our hyperparameter search when pretraining PIGASUS and PEGASUS-squared. For both models, we pretrained on 256 reports, and used a validation set of 50 reports. Our final test set for all models consisted of 186 bills. Prior to our project, PEGASUS-BillSum was finetuned on 18,949 Congressional bills, while PEGASUS-XSum was finetuned on 204,045 news articles.

	PIGASUS-100	**PEGASUS-squared-100**
# Epochs	3.6	10.8
# Steps	960	2880
Train Loss	2.565	2.388
Val. Loss	2.809	4.803

Table 2: Second "Pretraining" Hyperparameters and Loss

We observe in Table 1 that the best validation losses for both PIGASUS-100 and PEGASUS-squared-100 are lower than that of PEGASUS-large-100, even though both are fine-tuned on 100 examples. This may indicate that additional pretraining improves the model's accuracy on government documents as we start fine-tuning on weights better suited for government jargon than the PEGASUS-large model without additional pretraining. We discuss this possibility further in section 5.4.1 .

We observe in Table 2 that in the second pretraining step, PIGASUS-100 reaches a lower evaluation loss in fewer training steps than PEGASUS-squared-100. This may be because TextRank masks higher quality sentences than the original PEGASUS approach, increasing the efficiency of learning. However, it is apparent in Table 1 that PEGASUS-squared-100 reaches a lower validation loss in finetuning than PIGASUS much earlier. We discuss this in section 5.4.1.

5.4 Results

5.4.1 ROUGE Scores

Each model's performance on the California test set of 186 bills, based on ROUGE scores, can be seen in Table 3 and Figure 2.

	Mean R1/R2/RL	Min R1/R2/RL	Max R1/R2/RL
Extractive*	0.377 / 0.165 / 0.218	0.104 / 0.0000001 / 0.0860	0.695 / 0.588 / 0.507
PEGASUS-large-0*	0.353 / 0.160 / 0.225	0.0626 / 0.0000001 / 0.0620	0.681 / 0.495 / 0.527
PEGASUS-large-10	0.401 / 0.161 / 0.233	0.0615 / 0.0124 / 0.0609	0.659 / 0.453 / 0.502
PEGASUS-large-100	0.412 / 0.190 / 0.256	0.0365 / 0.0147 / 0.0292	0.759 / 0.591 / 0.647
PEGASUS-squared-100	0.340 / 0.142 / 0.231	0.051 / 0.00868 / 0.0435	0.668 / 0.484 / 0.531
PIGASUS-100	0.423 / 0.204 / 0.262	0.0306 / 0.0161 / 0.0267	0.721 / 0.603 / 0.674
PEGASUS-BillSum	0.419 / 0.197 / 0.249	0.115 / 0.0185 / 0.0782	0.735 / 0.553 / 0.493
PEGASUS-XSum	0.148 / 0.0989 / 0.290	0.0 / 0.0 / 0.0	0.788 / 0.762 / 0.759

Table 3: ROUGE-F1 scores of baselines (denoted by *) and models on test set of 186 CA bills.

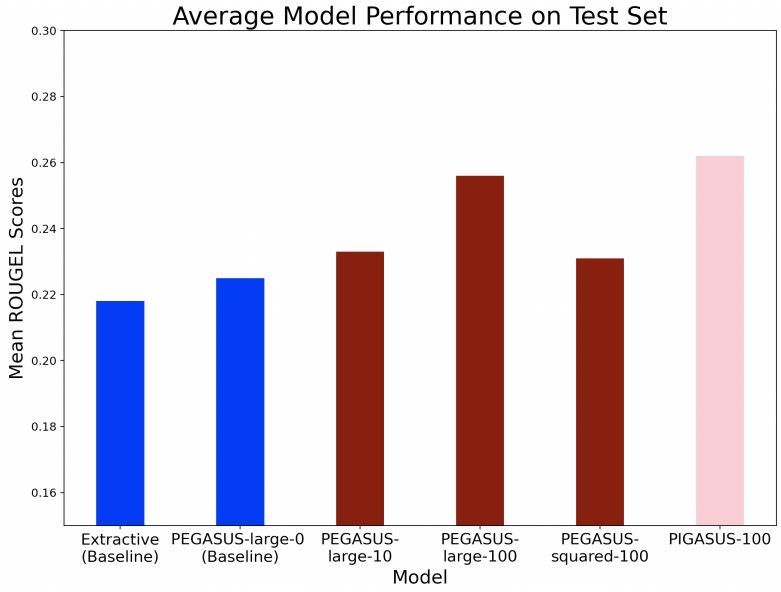


Figure 2: Mean ROUGEL-F1 scores each model achieved on the test set (n=186)

The performance of PEGASUS-XSum in Table 3 demonstrates PEGASUS's difficulty with generalizing on text of different formats than the texts it was fine-tuned on, justifying our focus on summarizing California bills. The high performance of Google's PEGASUS-BillSum shows that in high-resource settings, PEGASUS can reasonably generalize to closely related texts. However, PEGASUS-XSum's poor performance shows that this is only true for very closely related texts, and the replication of such models is infeasible in lower-resource settings. Indeed, PEGASUS-BillSum

was fine-tuned on 190 times as many bills than our models, and in our models, we froze the encoders of the initial models due to computational constraints. Yet PEGASUS-BillSum still performed worse than PEGASUS-large-100.

PIGASUS-100 outperforms all other models, as seen in Figure 2, including our baselines of extractive summarization with TextRank and PEGASUS-large-0, demonstrating the effectiveness of our second round of pretraining and using TextRank to mask sentences. However, PEGASUS-squared-100, despite being fine-tuned on 100 examples, performs worse on the test set than even PEGASUS-large 0. This may be caused by overfitting during pretraining, since PEGASUS-squared-100 went through 3 times as many pretraining steps than PIGASUS-100 (as seen in Table 2) although validation loss during pretraining never indicated such a thing. The substantial difference between the validation loss during fine-tuning and the test set performance of PEGASUS-squared-100 and PIGASUS-100 could be an indication that our validation set is somehow too small to provide reliable information about model performance during training. Otherwise, this may indicate that additional pretraining on a related, but different dataset does not improve, and actually may degrade the quality of the downstream task, which at the same time demonstrates the effectiveness of our TextRank approach. This showcases the potential volatility of the second pretraining step.

Both PEGASUS-squared-100 and PIGASUS-100 vary more in their performance than PEGASUS-BillSum. For instance, PIGASUS-100 and PEGASUS-squared-100 have minimum ROUGE1 scores of 0.0306 and 0.0510 respectively, while PEGASUS-BillSum has a minimum ROUGE1 score of 0.115. This variance may be because PEGASUS-squared-100 and PIGASUS-100 were trained on only 100 examples, while PEGASUS-BillSum is trained on almost 19,000 documents.

5.4.2 Human Evaluation

The percent of respondents who prefer the PIGASUS-100 model to PEGASUS-squared-100 can be seen in Table 4. Additional explanation and analysis can be found in sections 5.2 and 6.

	Best	**Median**	**Worst**	**Total**
% who prefer PIGASUS	91.891	86.486	37.838	72.072

Table 4: Results of human evaluation (n=37). Percent of respondents preferring PIGASUS-100 summary on best, median, and worst summaries by averaged ROUGEL Score.

6 Analysis

We analyze the 10 bills that each model (PIGASUS-100 and PEGASUS-squared-100) performs best on, the 10 bills that each model performs worst on, and the median bills, based on ROUGEL scores. Examples can be found in Appendix A.3.

6.1 Common Errors

Notably, PIGASUS-100 and PEGASUS-squared-100 share a number of common "best" and "worst" bills. Our results indicate a number of shared issues.

First, the models tend to struggle, unsurprisingly, with more complex bills. The lowest-performance bills have a greater tendency to be longer, include more headings and subheadings, and include a variety of moving parts (e.g., bills that have multiple focuses, rather than an emphasis on one specific issue). The highest-performance bills tend to have the opposite characteristics. This may be partially attributable to the fact that, just as it is more difficult for humans to choose the most important issues in a more complex bill, it is more difficult for the model to decide which issues to focus on. Additionally, if a bill is focused on one topic, it is more likely that the same key n-grams appear in the target summary, original bill, and produced summaries. As a result, the ROUGE scores, which are based on identical n-grams, are necessarily higher, even if the summary's true quality is not higher.

Second, the models tend to struggle with more obscure topics. For instance, legislation relating to construction defect litigation and limits were very low-performance. In particular, these bills were summarized with much more repetition than the standard bill (e.g., each sentence in the summary covers the same content). We hypothesize that this is because these words and topics are less present

in the pretraining and fine-tuning datasets. As a result, is it much more difficult for the model to "guess" the important sentences, and instead latches onto one or two sentences that seem important.

Third, in general, the models tend to repeat phrases and sentences in the summary. According to Nair and Singh (2021), this issue is common within abstractive summarization models, and the true cause remains unknown. Possible causes include the model architecture or the nature of sampling performed by the model which is very different from human natural language.

Fourth, the summaries often directly copy sentences from the bill in a manner that is more extractive than abstractive. This may be due to the limited fine-tuning that was conducted (100 examples) due to our low-resource setting. Specifically, original bill and target summaries frequently use the phrases "Existing law requires" and "This bill would." The model summaries naturally also use these phrases, in an effort to minimize the cross-entropy loss. However, in many cases, the model may not have seen enough examples to generate fully original sentences, and instead copies sentences beginning with "Existing law requires" or "This bill would" from the original bill.

6.2 Differences between models

Although both models share some errors, there remain differences between each model's summaries. First, PIGASUS-100 summaries tend to copy more formatting from the original bill (e.g., include "(1)" at the beginning of summaries, just as bills have "(1)" before a subsection). This may be because PIGASUS-100 was fine-tuned for double the epochs as PEGASUS-squared-100. As a result, it may have overfit to the train-set texts, indicating that perhaps a larger validation set should have been used.

Second, PEGASUS-squared-100 contains far more repetition in summaries than PIGASUS-100. For PEGASUS-squared-100, even the "best" summaries by ROUGE score contain repetition. This may be because ROUGE scores are based upon n-gram similarity; if a very common n-gram is repeated throughout the summary and used in the label summary, then the corresponding ROUGE score is high. As a result, PEGASUS-squared-100 would learn to repeat common sentences, so its summaries are optimized for ROUGE, rather than for grammatical structure, quality, or uniqueness. In comparison, PIGASUS-100's graph algorithm in TextRank acts as an extra buffer to prevent such simplistic word similarity degeneration that occured with PEGASUS-squared-100.

In general, we believe that PIGASUS-100's outperformance of PEGASUS-squared-100 may be because our pretraining method, TextRank, is a preferable algorithm for determining masked sentences than ROUGE scores. ROUGE1 scores, which are used in PEGASUS-squared-100 and the original PEGASUS model, are based on 1-gram similarities. Such similarities are not necessarily representative of the importance of a sentence, and hold a bias towards shorter sentences which include very common words. In contrast, TextRank takes co-occurrence of words into account, as well as Parts-of-Speech tags. TextRank's inclusion of additional features ensures that the masked sentences are more accurate proxy for important sentences, and reduces the bias toward shorter, "common word" sentences.

7 Conclusion

In this project, we aimed to summarize California bills, in the hopes of making legislation more accessible to the public, in a low resource setting. In doing so, we propose a new pretraining objective approach for choosing masked sentences using TextRank for the PEGASUS model. Our results in implementing this in a second pretraining step have highlighted the potential and effectiveness of our new approach when compared to pretraining using the original PEGASUS objective. Specifically, we propose using a related but not identical dataset for a second round of pre-training in situations where there is limited data available for the specific text-style of the downstream summarization task. We have also demonstrated PEGASUS's lack of generalizability, given its poor performance on bills with a different format than the ones it is finetuned on.

Future work could test the PIGASUS approach using more compute power and training data points. For instance, training PIGASUS without freezing the encoder in pretraining and finetuning, and using all California bills in the BillSum dataset could better highlight PIGASUS's strengths. Instead of implementing a second pretraining round, pretraining from scratch on C4 and HugeNews datasets with the PIGASUS masking approach could give more insight into the full potential of this method and its effects.

References

Python package index - rouge-score-0.1.2.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey.

Deepa Anand and Rupali Wagh. 2022. Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University - Computer and Information Sciences*, 34(5):2141–2150.

Alan Frumin. 2016. Obamacare was not passed using budget reconciliation.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021a. Efficient attentions for long document summarization.

Yuxin Huang, Zhengtao Yu, Junjun Guo, Yan Xiang, and Yantuan Xian. 2021b. Element graph-augmented abstractive summarization for legal public opinion news with graph transformer. *Neurocomputing*, 460:166–180.

Legal Information Institute. 2020. Legalese.

jiahao87. 2021. pegasus$_finetune.py$..

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of US legislation. *CoRR*, abs/1910.00523.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Xu LIANG. 2020. Textrank for keyword extraction by python.

Rada Mihalcea and Paul Tarau. 2004. Pdf.

Pranav Nair and Anil Kumar Singh. 2021. On reducing repetition in abstractive summarization. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 126–134, Online. INCOMA Ltd.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Paco Nathan. 2016. PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

tuner007. pegasus-paraphrase. `https://huggingface.co/tuner007/pegasus_paraphrase`.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.

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

Batch Size	Warmup Steps	Weight Decay	Learning Rate
2	500	0.01	1e-4

Table 5: Fixed Hyperparameters for fine-tuning and Pretraining all Models

A Appendix

A.1 Fixed Hyperparameters in Pretraining and Fine-tuning

A.2 Evaluation Losses

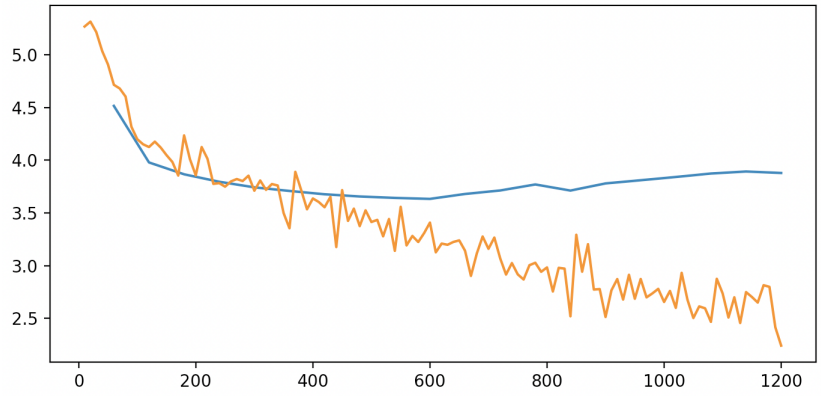


Figure 3: Evaluation loss vs Number of Epochs for fine-tuning PEGASUS-squared

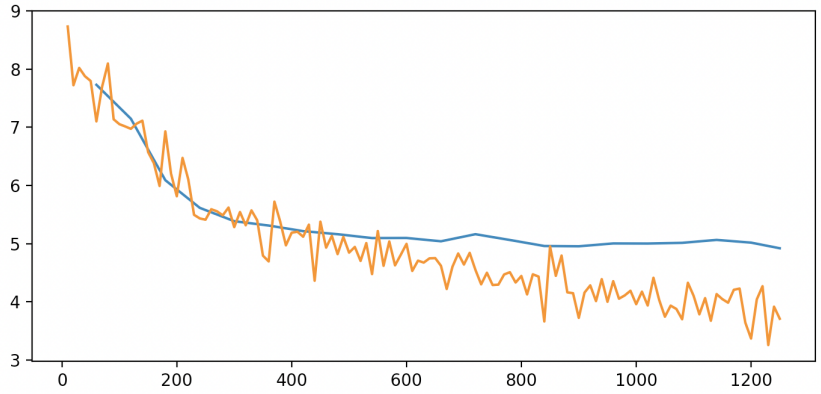


Figure 4: Evaluation loss vs Number of Epochs for fine-tuning PIGASUS

A.3 Original Summaries

	Summary
Target	Existing law authorizes the State Water Resources Control Board to adopt regulations requiring measurement and reporting of water diversion and use by persons including, among others, those authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, or a certification for livestock stockpond use. This bill would restrict the state board's authorization to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a registration or certification to uses that are year-round. Existing law requires a person who diverts 10 acre-feet of water per year or more under a permit or license to install and maintain a device or employ a method capable of measuring the rate of direct diversion, rate of collection to storage, and rate of withdrawal or release from storage, as specified, and with certain exceptions. This bill would delay those requirements for 2 years.
PEGASUS-squared-100	Existing law requires the board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. Existing law requires the board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. Existing law requires the board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. This bill would instead require the board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use.
PIGASUS-100	(1)Existing law requires the State Water Resources Control Board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, and livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. Existing law requires the initial regulations to be adopted as emergency regulations and to be considered by the Office of Administrative Law as necessary for the immediate preservation of the public peace, health, safety, and general welfare. This bill would adopt emergency regulations requiring the board to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, and livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. The initial regulations would require the board to adopt regulations requiring a person who, on or after January 1, 2018, diverts 10 acre-feet of water per year or more under a permit or license to install and maintain a device or employ a method capable of measuring the rate of direct diversion, rate of collection to storage, and rate of withdrawal or release from storage, as specified. The initial regulations would require the board to consider

Table 6: Best summary by average ROUGEL score

	Summary
Target	The California Constitution requires that the water resources of the state be put to beneficial use to the fullest extent of which they are capable and that the waste or unreasonable use or unreasonable method of use of water be prevented. Existing law, the Water Conservation in Landscaping Act, requires the Department of Water Resources to update its model water-efficient landscape ordinance by regulation and prescribes various requirements for the updated model ordinance. Existing law requires each local agency to adopt either the updated model water-efficient landscape ordinance or an ordinance that is at least as effective in conserving water as the updated model ordinance. If the local agency does not make a selection, the model ordinance shall apply within the jurisdiction of the local agency. This bill would require the department, upon identification of a funding source, to create the California Water Efficient Landscaping Program for the purpose of encouraging local agencies and water purveyors to use economic incentives that promote the efficient use of water, promote the benefits of consistent landscape ordinances, and support and enhance water inefficient grass replacement. This bill would create the Water Efficient Landscaping Fund and provide that moneys in the fund are available, upon appropriation by the Legislature, to the department for certain purposes.
PEGASUS-squared-100	(1) Existing law establishes the California Water Efficient Landscaping Program, which is administered by the State Water Resources Control Board, to promote the enhancement of the quality of life of the state by reducing the use of water by reducing the use
PIGASUS-100	(1) Existing law establishes the California Water Efficient Landscaping Program, which is administered by the Department of Water Resources, in partnership with the State Water Resources Control Board, the Central Valley Water Quality Control Board, and the Central Valley Water Quality Control Board, to promote water conservation, energy efficiency, and climate adaptation through the design and installation of landscape irrigation systems that reduce reliance on petroleum-based fertilizers and pesticides, as specified. The program requires, upon appropriation by the Legislature, 5% of the annual proceeds of the Water Efficient Landscaping Program Fund to be distributed to the Central Valley Water Quality Control Board for the purpose of implementing the program. This bill would increase the allocation of the program from 5% to 10%. The bill would create the California Water Efficient Landscaping Program Technical Advisory Council and would require the council, among other things, to prepare and provide to the council a standard operating procedure for the implementation of the program, as specified. The bill would also require the council to consult with the council's membership and recommend the council's next steps in implementing the program. (2) The California Environmental Quality Act (CEQA) requires all water agencies, except as specified, to comply with a water quality control ordinance adopted by the Central Valley Water Quality Control Board and promulgated by the Central Valley Water Quality Control

Table 7: Median summary by average ROUGEL score

	Summary
Target	Existing law provides for the administration and operation of juvenile court schools by the county board of education. This bill would express the Legislature's intent that juvenile court schools have a rigorous curriculum that includes a course of study that prepares pupils for high school graduation and career entry and fulfills the requirements for admission to the California State University and the University of California. Existing law prescribes the course of study a pupil is required to complete while in grades 9 to 12, inclusive, in order to receive a diploma of graduation, and authorizes the governing board of a school district to prescribe other coursework requirements that are in addition to the statewide requirements. Existing law exempts pupils in foster care and pupils who are homeless children or youths from local graduation requirements and also requires a school district and county office of education to accept coursework satisfactorily completed by those pupils while attending another public school, a juvenile court school, or a nonpublic, nonsectarian school. This bill would make that exemption and requirement to accept coursework satisfactorily completed applicable to former juvenile court school pupils, as defined. The bill would also require a county office of education to issue a diploma of graduation to a pupil who completes statewide coursework requirements for graduation while attending a juvenile court school but does not complete coursework and other requirements that are in addition to the statewide graduation requirements. By placing additional requirements on school districts and county offices of education, the bill would impose a state-mandated local program. The California Constitution requires the state to reimburse local agencies and school districts.
PEGASUS-squared-100	Existing law requires the county board of education to adopt and enforce a course of study that enhances instruction in mathematics and English language arts for pupils attending juvenile court schools, as determined by statewide assessments or objective local evaluations and assessments as approved by the county superintendent of schools. This bill would instead require the county board of education to adopt and enforce a course of study that enhances instruction in mathematics and English language arts for pupils attending juvenile court schools, as determined by statewide assessments or objective local evaluations and assessments as approved by the county superintendent of schools.
PIGASUS-100	The California Constitution requires the state to reimburse local agencies and school districts for certain costs mandated by the state. Statutory provisions establish procedures for making that reimbursement. This bill would provide that, if the Commission on State Mandates determines that the bill contains costs mandated by the state, reimbursement for those costs shall be made pursuant to these statutory provisions.

Table 8: Worst summary by average ROUGEL score

A.4 Paraphrased Summaries

	Summary
PEGASUS-squared-100	The board must adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. The board must adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. The board must adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stockpond use. The board would have to adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, or livestock stockpond use, if the registered use is year-round, or certification for livestock stock.
PIGASUS-100	The State Water Resources Control Board must adopt regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, and livestock stockpond use, if the registered use is year-round. The Office of Administrative Law must consider the initial regulations as emergency regulations in order to preserve the public peace, health, safety, and general welfare. The board would have to adopt emergency regulations requiring measurement and reporting of water diversion and use by persons authorized to appropriate water under a permit, license, registration for small domestic, small irrigation, and livestock stockpond use, if the registered use is year-round. The board would have to adopt regulations requiring a person who diverts 10 acre-feet of water per year or more under a permit or license to install and maintain a device or use a method capable of measuring the rate of direct diversion, by January 1, 2018, The board would have to consider the initial regulations.

Table 9: Paraphrased best summary by average ROUGEL score

	Summary
PEGASUS-squared-100	The California Water Efficient Landscaping Program is administered by the State Water Resources Control Board and is designed to promote the enhancement of the quality of life of the state by reducing the use of water by reducing the use of water by reducing the use of water.
PIGASUS-100	The California Water Efficient Landscaping Program is administered by the Department of Water Resources in partnership with the State Water Resources Control Board, the Central Valley Water Quality Control Board, and the Central Valley Water Quality Control Board to promote water efficiency, energy efficiency, and climate adaptation. 5% of the annual proceeds of the Water Efficient Landscaping Program Fund must be distributed to the Central Valley Water Quality Control Board for the purpose of implementing the program. The bill would increase the allocation of the program. The bill would create a technical advisory council for the California Water Efficient Landscaping Program and would require it to prepare and provide a standard operating procedure for the program. The bill would require the council to consult with the membership of the council and recommend the next steps in implementing the program. The Central Valley Water Quality Control Board adopted a water quality control ordinance in order to comply with the California Environmental Quality Act.

Table 10: Paraphrased median summary by average ROUGEL score

	Summary
PEGASUS-squared-100	The law requires the county board of education to adopt and enforce a course of study that enhances instruction in mathematics and English language arts for children attending juvenile court schools if they are assessed by statewide or local evaluations. The bill would require the county board of education to adopt and enforce a course of study that enhances instruction in mathematics and English language arts for pupils attending juvenile court schools if they are assessed by statewide or local evaluations.
PIGASUS-100	The state of California is required to reimburse local agencies and school districts for certain costs. Statutory provisions make it easier to make that reimbursement. The bill would provide that if the Commission on State Mandates determines that the bill contains costs mandated by the state, reimbursement for those costs shall be made.

Table 11: Paraphrased worst summary by average ROUGEL score