**PEGASUS Abstract**

PEGASUS stands for Pre-training with Extracted Gap-sentences for Abstractive Summarization.

**Abstract from the research paper:**

Recent work pre-training Transformers with self-supervised objectives on large text corpora has shown great success when fine-tuned on downstream NLP tasks including text summarization.

However, pre-training objectives tailored for *abstractive text summarization* have not been explored. Furthermore, there is a lack of systematic evaluation across diverse domains.

In this work, the researchers proposed pre-training large Transformer-based encoder-decoder models on massive text corpora with a new self- supervised objective.

In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an *extractive summary*.

Researchers has evaluated their best PEGASUS model on 12 downstream summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills.

Experiments demonstrate it achieves state-of-the-art performance on all 12 downstream datasets measured by ROUGE scores.

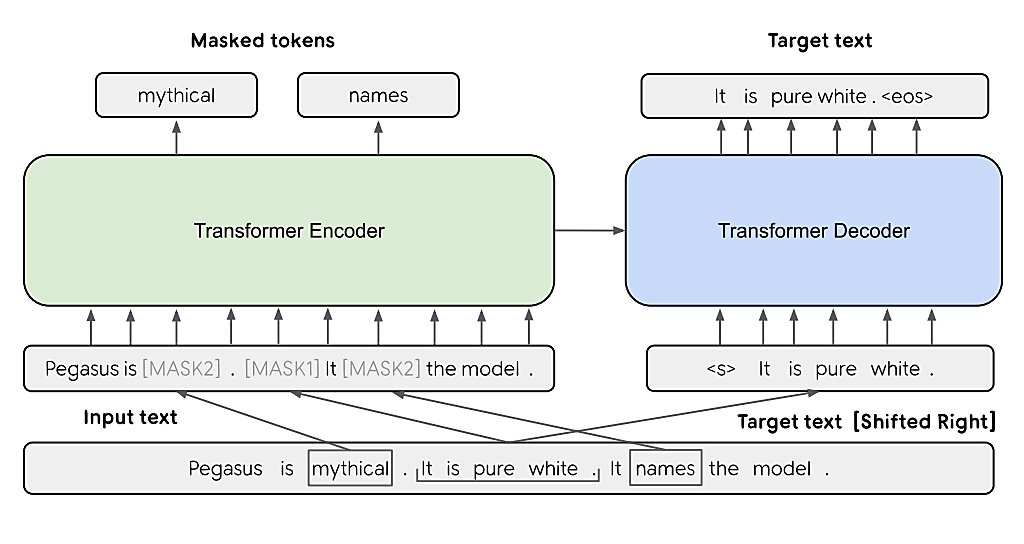
Their model also shows surprising performance on low-resource summarization, surpassing previous state-of-the-art results on 6 datasets with only 1000 examples.

Finally, they validated the results using human evaluation and show that our model summaries achieve human performance on multiple datasets.

**A small note on PEGASUS**

<https://paperswithcode.com/paper/pegasus-pre-training-with-extracted-gap#code>

PEGASUS proposes a transformer-based model for abstractive summarization. It uses a special self-supervised pre-training objective called gap-sentences generation (GSG) that's designed to perform well on summarization-related downstream tasks. As reported in the paper, "both GSG and MLM (*masked language modelling*) are applied simultaneously to this example as pre-training objectives. Originally there are three sentences. One sentence is masked with [MASK1] and used as target generation text (GSG). The other two sentences remain in the input, but some tokens are randomly masked by [MASK2]."



**Abstractive vs Extractive Summarization**

Text summarization condenses a lengthy piece of text into a shorter version that captures the main points. There are two main approaches to this: extractive summarization and abstractive summarization.

PEGASUS is a model for *abstractive summarization*, a type of summarization that creates summaries that are not just *extractive* (i.e. copying sentences from the original text) but that rephrase and condense the information using their own words. PEGASUS uses a special technique called *gap-sentence generation* (GSG) to pre-train a Transformer-based model.

*Extractive summarization* acts like a highlighter. It scans the original text and identifies the most important sentences. These sentences are then copied and assembled to form the summary. Techniques for identifying important sentences can involve factors like sentence position, keyword frequency, and statistical analysis.

*Abstractive summarization* is more sophisticated. It reads and comprehends the original text, then rephrases it in a new, concise form. This involves techniques from natural language processing (NLP) like machine translation and text generation. Abstractive summaries can include new sentences and phrases that are not present in the original text, but that capture the overall meaning.