

BLOG ›

PEGASUS: A State-of-the-Art Model for Abstractive Text Summarization

TUESDAY, JUNE 09, 2020

Posted by Peter J. Liu and Yao Zhao, Software Engineers, Google Research

Students are often tasked with reading a document and producing a summary (for example, a book report) to demonstrate both reading comprehension and writing ability. This [abstractive text summarization](#) is one of the most challenging tasks in natural language processing, involving understanding of long passages, information compression, and language generation. The dominant paradigm for training machine learning models to do this is [sequence-to-sequence](#) (seq2seq) learning, where a neural network learns to map input sequences to output sequences. While these seq2seq models were initially developed using [recurrent neural networks](#), [Transformer](#) encoder-decoder models have recently become favored as they are more effective at modeling the dependencies present in the long sequences encountered in summarization.

Transformer models combined with self-supervised pre-training (e.g., [BERT](#), [GPT-2](#), [RoBERTa](#), [XLNet](#), [ALBERT](#), [T5](#), [ELECTRA](#)) have shown to be a powerful framework for producing general language learning, achieving state-of-the-art performance when fine-tuned on a wide array of language tasks. In prior work, the self-supervised objectives used in pre-training have been somewhat agnostic to the down-stream application in favor of generality; we wondered whether better performance could be achieved if the self-supervised objective more closely mirrored the final task.

In “[PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization](#)” (to appear at the [2020 International Conference on Machine Learning](#)), we designed a pre-training self-supervised objective (called *gap-sentence generation*) for Transformer encoder-decoder models to improve fine-tuning performance on abstractive summarization, achieving state-of-the-art results on 12 diverse summarization datasets. Supplementary to the paper, we are also releasing the training code and model checkpoints on [GitHub](#).

Our hypothesis is that the closer the pre-training self-supervised objective is to the final down-stream task, the better the fine-tuning performance. In PEGASUS pre-training, several whole sentences are removed from documents and the model is tasked with recovering them. An example input for pre-training is a document with missing sentences, while the output consists of the missing sentences concatenated together. This is an incredibly difficult task that may seem impossible, even for people, and we don't expect the model to solve it perfectly. However, such a challenging task encourages the model to learn about language and general facts about the world, as well as how to distill information taken from throughout a document in order to generate output that closely resembles the fine-tuning summarization task. The advantage of this self-supervision is that you can create as many examples as there are documents, without any human annotation, which is often the bottleneck in purely supervised systems.

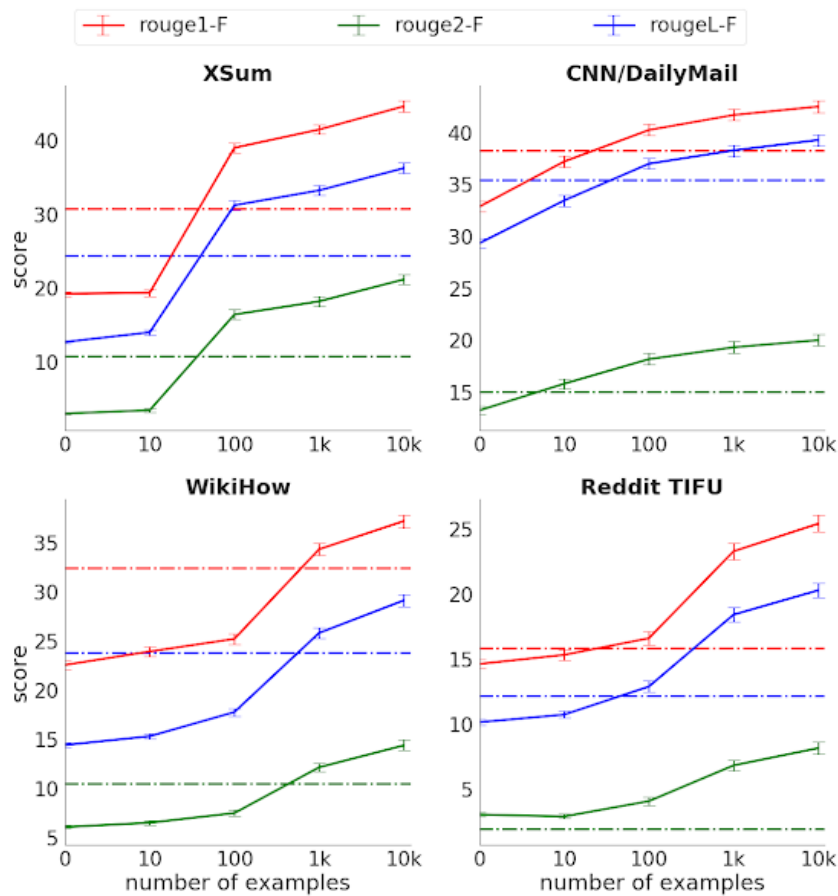
TRANSFORMER

A self-supervised example for PEGASUS during pre-training. The model is trained to output all the masked sentences.

We found that choosing "important" sentences to mask worked best, making the output of self-supervised examples even more similar to a summary. We automatically identified these sentences by finding those that were most similar to the rest of the document according to a metric called [ROUGE](#). ROUGE computes the similarity of two texts by computing [n-gram](#) overlaps using a score from 0 to 100 (ROUGE-1, ROUGE-2, and ROUGE-L are three common variants).

Similar to other recent methods, such as [T5](#), we pre-trained our model on a very large corpus of web-crawled documents, then we fine-tuned the model on 12 public down-stream abstractive summarization datasets, resulting in new state-of-the-art results as measured by automatic metrics, while using only 5% of the number of parameters of T5. The datasets were chosen to be diverse, including news articles, scientific papers, patents, short stories, e-mails, legal documents, and how-to directions, showing that the model framework is adaptive to a wide-variety of topics.

were surprised to learn that the model didn't require a large number of examples for fine-tuning to get near state-of-the-art performance:



ROUGE scores (three variants, higher is better) vs. the number of supervised examples across four selected summarization datasets. The dotted-line shows the Transformer encoder-decoder performance with full-supervision, but without pre-training.

With only 1000 fine-tuning examples, we were able to perform better in most tasks than a strong baseline (Transformer encoder-decoder) that used the full supervised data, which in some cases had many orders of magnitude more examples. This “sample efficiency” greatly increases the usefulness of text summarization models as it significantly lowers the scale and cost of supervised data collection, which in the case of summarization is very expensive.

Human-Quality summaries


While we find automatic metrics such as ROUGE are useful proxies for measuring progress during model development, they only provide limited information and don't tell us the whole story, such as fluency or a comparison to human performance. To this end, we conducted a human evaluation, where raters were asked to compare summaries from our model with human ones (without knowing which is which). This has some similarities to the [Turing test](#).

Document:

Evans recorded a time of 55 minutes 40 seconds to take the yellow jersey from Andy Schleck before Sunday's largely ceremonial final stage in Paris. Germany's Tony Martin won the time-trial from Evans, who moved 1:34 ahead of Schleck in the overall standings. Schleck began the day with a 53-second lead in the general classification. [truncated for brevity]


Summary:

Cadel Evans is all but certain to become Australia's first Tour de France winner after a stunning time trial-victory in the suburbs of Grenoble.




Summary:

Australia's Cadel Evans is on the brink of winning his first Tour de France title after finishing second in the penultimate stage time-trial.




Summary:

Britain's Steve Evans clinched his first Tour de France title with a stunning victory in the time-trial on stage 11.



Summary:

Steve Evans won the Tour de France as fellow Briton Chris Froome retained the leader's yellow jersey.



Human raters were asked to rate model and human-written summaries without knowing which was which. The document is truncated here for illustration, but raters see the full text.

We performed the experiment with 3 different datasets and found that human raters do not consistently prefer the human summaries to those from our model. Furthermore, our models trained with only 1000 examples performed nearly as well. In particular, with the much studied [XSum](#) and [CNN/Dailymail](#) datasets, the model achieves human-like performance using only 1000 examples. This suggests large datasets of supervised examples are no longer necessary for summarization, opening up many low-cost use-cases.

A Test of Comprehension: Counting Ships

Following this post is an [example article](#) from the XSum dataset along with the model-generated abstractive summary. The model correctly abstracts and paraphrases four named frigates (HMS Cumberland, HMS Campbeltown, HMS Chatham and HMS Cornwall) as “four Royal Navy frigates”, something an extractive approach could not do since “four” is not mentioned anywhere. Was this a fluke or did the model actually count? One way to find out is to add and remove ships to see if the count changes.

As can be seen below, the model successfully “counts” ships from 2 to 5. However, when we add a sixth ship, the “HMS Alphabet”, it miscounts it as “seven”. So it appears the model has learned to count small numbers of items in a list, but does not yet generalize as elegantly as we would hope. Still, we think this rudimentary counting ability is impressive as it was not explicitly programmed into the model, and it demonstrates a limited amount of “symbolic reasoning” by the model.

To support ongoing research in this field and ensure reproducibility, we are releasing the PEGASUS code and model checkpoints on [GitHub](#). This includes fine-tuning code which can be used to adapt PEGASUS to other summarization datasets.

Acknowledgements

This work has been a collaborative effort involving Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. We thank the T5 and Google News teams for providing datasets for pre-training PEGASUS.

**Four
ships**

Five
ships

Two
ships

Three
ships

Six
ships

The decommissioned Type 22 frigates
HMS Cumberland, HMS Campbeltown, HMS Chatham and HMS Cornwall
are currently moored in Portsmouth Harbour.
Bidders had until 23 January to register an interest
in the former Devonport-based ships. The BBC
understands no proposals to preserve the ships have
been submitted. Those who have registered an
interest are finalising their bids with viewings set
to take place in late February and March. A final
decision is not expected until the spring. The
government's Disposal Services Authority, which is
handling the sale, wants to award at least one of
the frigates to a UK ship recycler to determine the
capacity of the UK's industry in the field. Penny
Mordaunt, Conservative MP for Portsmouth North, said
it was important UK recyclers had the chance to
prove themselves in the field but she was also keen
to see at least one of them saved from the
scrapyard. She added: "For anyone that has served on
a ship it's your home, you've literally been through
the wars with it... and you want them to have a
noble second life. "My preference is to go for the
reef and diving attraction. "We've got to get best





JUN 9, 2020

Recent Advances
in Google Translate



MAY 30, 2020

DADS:
Unsupervised
Reinforcement



MAY 28, 2020

Federated
Analytics:
Collaborative Data

