Huawei NLP Course: PEGASUS in Russian

Konstantin Kotik

May 2020

Abstract

Summarization is the task of producing a shorter version document that preserves most of the input's meaning. In this project is described finetuning of PEGASUS model for Abstractive Summarization on Russian sport text broadcasts. Link to project code: `https://github.com/kotikkonstantin/pegasus-in-russian`.

1 Introduction

Summarization is the task of producing a shorter version document that preserves most of the input's meaning. Summaries as short as 17% of the full text length speed up decision making twice, with no significant degradation in accuracy. Query-focused summaries enable users to find more relevant documents more accurately, with less need to consult the full text of the document [Mani et al., 2002]. If you do not have much time to read a collection of long documents the automatic summarization of texts will allow you to highlight important documents and get a summary of texts in much less time. The better the summarization model is, the better its generalization ability is.

There is interest to use more complex approach for summarization problem. It's abstractive approach. At the time of publication PEGASUS abstractive summarization model from Google Research shows state-of-the-art results on different English datasets (X-Sum, Gigaword etc.) [Zhang et al., 2019].

But how does it work out-of-box on Russian dataset?

1.1 Team

Konstantin Kotik has done completely the project myself.

2 Related Work

Let http://nlpprogress.com/ [Ruder,] is measure of Natural Language Processing (NLP) progress. We can see here that achievements in summarization task on Russian text corpora are absent. So it arises additional motivation to

build good summarization model for language with rich morphology like Russian has. On English text corpora let's consider two previous models – BART [Lewis et al., 2019] and T5 [Raffel et al., 2019].

The main idea of T5 is illustrated on Fig. 1.

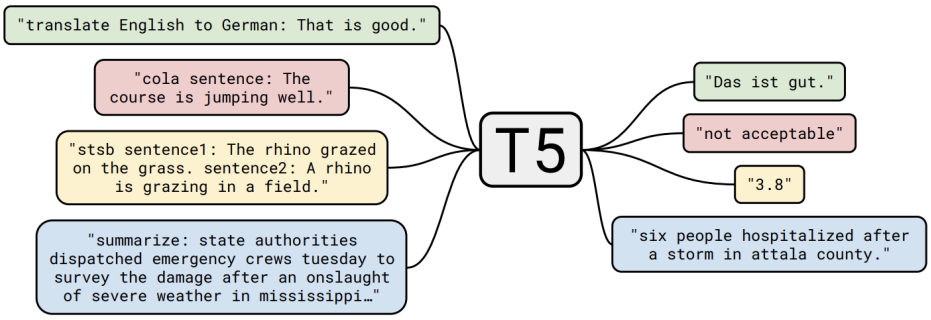

Figure 1: According to [Raffel et al., 2019]: a diagram of the text-to-text framework. Every task they consider – including translation, question answering, and classification – is cast as feeding our model text as input and training it to generate some target text. This allows them to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. "T5" refers to the model, which they dub the "Text-to-Text Transfer Transformer".

Thus the key points of T5 model:

- State each NLP problem as a text-to-text problem
- Universal approach for different NLP Deep Learning problems – translation, question answering, and classification
- Using of a simple denoising training objective function for pretraining
- Full encoder-decoder transformer architecture
- It is trained using teacher forcing

According to [Lewis et al., 2019] BART is a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT [Jacob et al., 2018] (due to the bidirectional encoder), GPT [Radford et al., 2019] (with the left-to-right decoder), and many other more recent pretraining schemes. For intuitive understanding BART represents "ensemble" between GPT and BERT models. The schematic concept of BART is illustrated on Fig. 2.

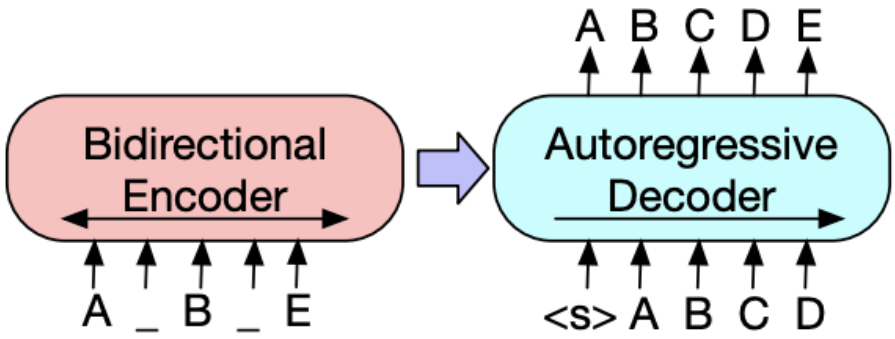

Figure 2: According to [Lewis et al., 2019]: inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and they use representations from the final hidden state of the decoder.

3 Model Description

There are two separate approaches for summarization problem – extractive and abstractive. In extractive approach the result of summarization model represents set of subsequences from input text. In abstractive approach the result of summarization model represents some new summary text.

According to [Zhang et al., 2019] authors propose 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. They evaluated 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. The model also shows surprising performance on low-resource summarization, surpassing previous state-of-the-art results on 6 datasets with only **1000 examples**.

The authors of PEGASUS suggested to use Gap Sentences Generation (GSG) objective for Transformer pre-training. It masks full sentences. So we have abstractive self-supervised objective function to recover masked sentences. The base architecture of PEGASUS is illustrated on Fig. 3

To mask some sentences decided in model there are 3 heuristic approaches to select "important" sentences:

- masking random sentences (**Random**)
- the first few pieces (**Lead**)
- selected by ROUGE1-F1 (**Principal**). It's calculated the similarity of each sentence to the entire text and is selected a certain percentage (30% stopped) of the most representative sentences so as to mask them and try to restore them.

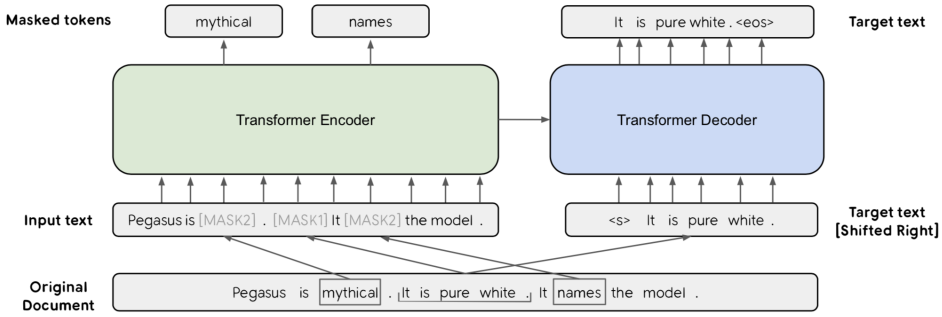

Figure 3: According to [Zhang et al., 2019]: The base architecture of PEGASUS is a standard Transformer encoder-decoder. Both GSG and Masked Language Modeling (MLM) are applied simultaneously to this example as pre-training objectives. Originally there are three sentences. One sentence is masked with [MASK1] and used as target generation text (GSG). The other two sentences remain in the input, but some tokens are randomly masked by [MASK2] (MLM).

The model was pretrained on **C4**, or the Colossal and Cleaned version of Common Crawl, introduced in [Raffel et al., 2019] and **HugeNews**, a dataset of 1.5B articles (3.8TB) collected from news and news-like websites from 2013-2019.

4 Dataset

The PEGASUS was finetuned on **6143** Russian sport text broadcasts – train data. The text broadcasts are about different football and hockey matches in Russian Language. The origin of used data is sports.ru. For validation quality procedure it was used another 763 broadcasts – validation data. Also it was used yet another 760 broadcasts for pure testing – test data.

This data is available for the research purposes.

For creating Russian Vocabulary was used by Byte Pair Encoding (BPE) in implementation of sentencepiece. Used vocabulary size was 96 000 tokens based on train data.

5 Experiments

5.1 Metrics

Metrics were used to evaluate model performance:

- **ROUGE** [Lin, 2004], **ROUGE** is the main metric for summarization quality.
- **BLEU** [Papineni et al., 2002], **BLEU** is an alternative quality metric for language generation.

5.2 Experiment Setup

Design of the PEGASUS launching on the Russian dataset:

- maximum input length: 512,
- maximum output length: 256,
- train steps: 250,
- learning rate: 0.0001,
- batch size: 16,

6 Results

The results based on test data could be found in Tab. 1. The example of model inference could be found in Tab. 2.

Metric name	95% lower bound	Mean	95% upper bound
rouge1-R	0.020335	0.024443	0.028950
rouge1-P	0.094834	0.111006	0.127912
rouge1-F	0.030708	0.035601	0.041000
rouge2-R	0.002816	0.004676	0.007199
rouge2-P	0.015092	0.023864	0.033897
rouge2-F	0.003743	0.005679	0.007902
rougeL-R	0.019067	0.022773	0.026978
rougeL-P	0.090116	0.106110	0.124162
rougeL-F	0.028705	0.033412	0.038519
rougeLsum-R	0.019225	0.022945	0.027161
rougeLsum-P	0.089901	0.105752	0.121960
rougeLsum-F	0.028633	0.033411	0.038445
bleu	0.725627	0.763095	0.804040

Table 1: Model performance results.

7 Conclusion

In the result of this work it was done:

- custom TensorFlow Datasets preparing
- custom vocabulary model building
- PEGASUS adapting for custom dataset
- finetuning PEGASUS model on Russian sport text broadcasts

References

[Jacob et al., 2018] Jacob, M.-W., Chang, K., Lee, and Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Lewis et al., 2019] Lewis, Y., Liu, N., Goyal, M., Ghazvininejad, A., Mohamed, O., Levy, V., Stoyanov, and Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out:74–81.

[Mani et al., 2002] Mani, G., Klein, D., House, L., Hirschman, T., Firmin, and Sundheim (2002). Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

[Papineni et al., 2002] Papineni, S., Roukos, T., Ward, and Zhu (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

[Radford et al., 2019] Radford, J., Wu, D., Amodei, D., Amodei, J., Clark, M., Brundage, and Sutskever (2019). Language models are unsupervised multitask learners.

[Raffel et al., 2019] Raffel, N., Shazeer, A., Roberts, K., Lee, S., Narang, M., Matena, Y., Zhou, W., Li, and Liu (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

[Ruder,] Ruder, S. http://nlpprogress.com/.

[Zhang et al., 2019] Zhang, Y., Zhao, M., Saleh, and Liu (2019). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Document	всем привет! 12-й в истории клубный чемпионат мира (в этом году он официально называется fifa club world cup japan 2015 presented by alibaba e-auto) вышел на финишную прямую. в решающем матче встретятся аргентинский «ривер плейт» и каталонская «барселона». наслаждаемся вместе! последние успехи «ривера» связаны с именем знаменитого в прошлом плеймейкера «монако» и сборной аргентины марсело гальярдо, под руководством которого «речники» выиграли копа судамерикана, а затем и кубок либертадорес. в 1986 году «ривер плейт» брал межконтинентальный кубок, но в клубном чемпионате мира еще никогда не побеждал. победитель лиги чемпионов одной левой нокаутировал китайский «гуанчжоу», а выигравший кубок либертадорес «ривер», который еще совсем недавно барахтался во втором дивизионе своей страны из-за финансового коллапса, разобрался с хозяевами турнира – «санфречче». финал станет первой официальной встречей команд в истории. за «барсу» сегодня сыграют все сильнейшие – урологи помогли месси раздробить и вывести камень из почки, а неймара больше не беспокоит пах. в старте также появляется экс-игрок «ривера» маскерано, ... "барса" выигрывает третий клубный чемпионат мира в истории! борьбы не получилось - слишком уж велика оказалась разница в классе команд. каталонцы могли и крупнее выиграть, но сжалились. всем спасибо, пока!
Gold (target)	«Барселона» выиграла клубный чемпионат мира, разгромив в финале аргентинский «Ривер Плейт» (3:0). Таким образом, каталонцы выиграли этот турнир в тритий раз и стали единоличными лидерами по этому показателю. Напомним, ранее «Барселона» побеждала на клубном чемпионате мира в 2009 и 2011 годах.
Model	выигравший кубокдорес «ривер», который еще совсем недавно барахтался межконтинента втором дивизионе своей страны из-за финансового коллапса, разобрался с) хозяевами турнира – «санфречче». финал станет первой официальной встречей команд в истории. за «барсу» сегодня сыграют все " – урологи помогли месси раздробить и вывести камень из почки, а

Table 2: Example of output sample.