Using BART-like model for the russian language summarization

Alexander Kazakov

31 May 2020

Abstract

BART-like model is used for the task of abstractive summarization of the russian text. To speed up training process, embedding and encoder part of the BART model are borrowed from a publicly available BERT model pretrained on russian texts. Project code is here `https://github.com/AlexanderKazakov/rubart`.

1 Introduction

The task of automatic abstractive summarization is important in the era of Internet, when the quantity of text information is growing very fast (what cannot be said about quality). The approach of this paper is to use BART model [Lewis et al., 2019] for russian texts summarization. Embedding and encoder part of the BART model are borrowed from the publicly available BERT model pretrained on russian texts by DeepPavlov [Kuratov and Arkhipov, 2019]. We show results on Lenta.Ru-News-Dataset [len,] and on a non-public dataset of sport matches broadcasts and related news.

1.1 Team

All the work: adaptation of the model to the task, programming, experiments, and report preparation was done by Alexander Kazakov.

2 Related Work

Text summarization problem already has many approaches listed on [nlp,] with ROUGE metrics reported for different tasks and datasets.

Today, the number of works devoted to summarization of the russian texts increases. Most of them make use of the popular 'transformer-like' architectures, which make gain of unsupervised pretraining on a large corpora of texts.

[V.A. and P.S,] reports results of a competition on the summarization on Lenta.Ru-News-Dataset and RIA dataset. Different approaches are described. [Gusev, 2019] describes a leading approach in this competition.

3 Model Description

In this work, BART model [Lewis et al., 2019] architecture from Huggingface transformers library [Wolf et al., 2019] is used.

The main limitation of the transformer models approach is that they are very expensive to train from scratch. However, pretrained model can be trained on a specific task in a short time. Author does not know about pretrained BART model for the russian language. So, the only way to make use of some pretrained model is to borrow embedding and encoder layers weights from the RuBERT model by DeepPavlov [Kuratov and Arkhipov, 2019].

Minor modifications of the BART model are done to be able to incorporate RuBERT's weights:

- Changing some dimensions and parameters of the model (hidden size, etc)
- Incorporating token-type embeddings (dummy all-zeros) into BART model
- Manual weights binding, as codes of these two models are slightly different

Embeddings are shared between encoder, decoder and output classifier. As embeddings are from the pretrained model, the tokenizer (WordPiece) is also borrowed from the DeepPavlov's RuBERT model. So, the only initially unoptimized parameters of the model are decoder layers.

4 Dataset

There are two public datasets for abstractive summarization in the russian language:

- Lenta.Ru-News-Dataset [len,] – 337 Mb (2 Gb uncompressed), 800K+ news articles, dates: 30/08/1999 - 14/12/2019
- RossiyaSegodnya/ria_news_dataset [Gavrilov et al., 2019] – 1M+ (4 Gb uncompressed) Russian language news documents from January, 2010 to December, 2014.

Both of them contain texts of news articles and their titles. Generation of the title from an article text can be considered as a summarization task.

There is also a non-public dataset by sports.ru: sport matches broadcasts as texts and short articles about these matches as summarizations. However, this dataset is much smaller and is not very consistent. Training on it from scratch is not possible – a pretrained model should be used.

5 Experiments

5.1 Metrics

In this work, the widely known ROUGE metric [Lin, 2004] is used.

5.2 Experiment Setup

Experiments were done on Lenta dataset and on sports.ru dataset.

On Lenta dataset, tokenized input texts were truncated by max length 256, tokenized titles – by max length 24. Firstly, just model decoder's layers were trained for 3 epochs while other weights are frozen. Then, all model's weights were trained for 5 epochs with reduced learning rate. Results were obtained using top-4 beam-search decoding.

After training the model on Lenta dataset, weights of this model were used to initialize the model for training on sports.ru dataset. For sports.ru dataset, both texts and summaries were truncated by length 256. And texts were truncated from the end, not from the start of the text (is is better for summarization to see the end of a match, not the beginning).

Due to time and hardware usage limitations, all experiments were done just once, without averaging. Lenta dataset was split into train and validation parts randomly in ratio 9/1. For sports.ru dataset, the train-validation-test split was already provided.

5.3 Baselines

Best known to author result on the Lenta dataset is from [V.A. and P.S,]:

R1f	R2f	Rlf
42.96	25.43	40.02

6 Results

6.1 Results on Lenta dataset

Results of this work on the Lenta dataset (validation part) is:

R1f	R2f	Rlf
37.7	21.1	36.6

Results do not show improvements over [V.A. and P.S,]. However, manual check of the results on the validation set shows reasonable quality of summarization:

orig/pred	Title
orig	США проследят за выборами президента Тайваня с двух авианосцев
pred	США направили к берегам Тайваня два авианосца
orig	Французскую судью с 2000 года заставляли подсуживать канадцам
pred	Судья по фигурному катанию рассказала о давлении на судей в США
orig	Два смертных приговора в США исполнили в один день впервые за 17 лет
pred	В США впервые с 2000 года казнили двух заключенных
orig	Бекмамбетов доверит спасение Москвы американской туристке
pred	У Тимура Бекмамбетова появится актриса из " самого темного часа "
orig	Преступник запрыгнул в клетку ко львам и отделался укусом пальца
pred	Беглый преступник прыгнул в вольер ко львам и отделался укушенным пальцем

6.2 Results on sports.ru dataset

The average rouge results on sports.ru dataset (test part) is:

R1f	R2f	Rlf
24.3	11.1	21.6

Manual check:

orig/pred	Title
orig	В рамках регулярного чемпионата КХЛ СКА обыграл«Торпедо» (5:2). Sports.
pred	Сегодня в рамках регулярного чемпионата КХЛ « Торпедо » уступило СКА (5 : 2). Sports. Календарь Кубка ГагаринаСтатистика Кубка ГагаринаКалендарь Кубка Гагарина завершилась со счетом 5 : 0. Счет в серии : 2 - 2 - 1. ПРИМЕЧАНИЕ : время начала матчей – московское.
orig	Сегодня в регулярном чемпионате КХЛ ЦСКА в Нижнем Новгороде проиграл «Торпедо». Sports. ru провел текстовую трансляцию этого матча.
pred	Сегодня в рамках регулярного чемпионата КХЛ московское « Динамо » обыграло ЦСКА (3 : 0). Sports. Счет в серии : 0 - 1. ПРИМЕЧАНИЕ : время начала матчей – московское. Календарь Кубка ГагаринаСтатистика Кубка ГагаринаКалендарь Кубка ГагаринаПервый матч – московское « Торпедо ».
orig	В третьем периоде четвертого матча серии первого раунда плей-офф Кубка Гагарина между «Ак Барсом» и «Салаватом Юлаевым» (1:4) произошла драка между защитниками Райаном Уилсоном и Максимом Гончаровым. В итоге Уилсон получил 5 минут штрафа, а Гончаров – 2+5+10. Дисциплинарный штраф ему был выписан за то, что он бросил в Уилсона шлем.
pred	Главный тренер « Ак Барса » Сергей Мозякин прокомментировал победу над « Ак Барсом » (4 : 1) в четвертом матче серии плей - офф Кубка Гагарина. « В первом периоде мы играли очень хорошо, но в третьем периоде мы не смогли забить гол, но не забили. В третьем периоде у нас было много моментов, но у нас не было шансов забить, но мы не сумели забить. Мы не смогли сравнять счет в серии 2 : 1, а в серии – 1 - 3 - 3. В серии – 2 - 2 - 3 в серии : 1. Счет в серии 4 - 1 - 1. – 0. – 1. « Салават Юлаев » – 2 : 0. Как это было
orig	Главный тренер СКА Милош Ржига остался доволен крупной победой над «Спартаком» (8:1) в матче регулярного чемпионата КХЛ. Я рад такому результату. Правда, в первом периоде мы выступили очень плохо не могли раскататься, у хозяев были какие-то моменты. А через три минуты после начала второго периода матч практически закончился. Дальше мы играли в свой хоккей, а хозяева что-то развалились. Я рад, что в конце матча ребята не отпустили игру, а наоборот добавили и порадовались хоккею, приводит слова Ржиги официальный сайт Спартака.
pred	Главный тренер Спартака Олег Знарок после победы над СКА (7 : 8) в матче регулярного чемпионата КХЛ отметил, что команда играла хорошо. В первом периоде мы играли очень хорошо, но не смогли забить, но мы не забили гол, но в третьем периоде не забили. В целом мы играли хорошо, а в большинстве, не смогли сравнять счет, цитирует Билялетдинова официальный сайт СКА.

7 Acknowledgments

Author is thankful to Valentin Malykh and other Huawei NLP course stuff for the great educational course.

8 Conclusion

In this work, RuBERT model weights were reused in a modified version of BART model. Experiments were done on Lenta.ru dataset and on non-public sports.ru dataset. Results on Lenta.ru dataset look reasonable. Results on sports.ru dataset are not very consistent. However, this dataset itself is not very big and consistent, so in future work, some additional preprocessing and the dataset extension should be done.

References

[len,] Corpus of news articles of lenta.ru.

[nlp,] Nlp-progress/summarization.

[Gavrilov et al., 2019] Gavrilov, D., Kalaidin, P., and Malykh, V. (2019). Self-attentive model for headline generation. In *Proceedings of the 41st European Conference on Information Retrieval*.

[Gusev, 2019] Gusev, I. (2019). Importance of copying mechanism for news headline generation. *ArXiv*, abs/1904.11475.

[Kuratov and Arkhipov, 2019] Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language.

[Lewis et al., 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[V.A. and P.S,] V.A., M. and P.S, K. Headline generation shared task on dialogue'2019.

[Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing.