TEXT SUMMARIZATION WITH RECURRENT NEURAL NETWORK FOR HEADLINE GENERATION

Shevchuk A.A. (anthonyshevchuk@gmail.com), National Research Tomsk State University, Tomsk, Russia
Zdorovets A.I. (loferist@gmail.com), National Research Tomsk State University, Tomsk, Russia

Due to increased need of automatic text summarization, neural networks are being used more often. One of the more widely used architectures for headline generation is Seq2Seq, which handles both summarization (an integral part of headline generation), and text generation itself. However, even though this model has been used for a number of languages, its applicability and effectivity for Russian has not been researched profoundly. This paper describes implementation of a simple version of such architecture as well as analysis of (rather) preliminary results. We also try to deduce some of the factors affecting the quality of headline generation.

Keywords: natural language processing, neural network, Seq2Seq, text generation, text summarization, headline generation

СУММАРИЗАЦИЯ ТЕКСТА ПРИ ПОМОЩИ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ

Шевчук А.А. (anthonyshevchuk@gmail.com), Национальный исследовательский Томский государственный университет, Томск, Россия
Здоровец А.И. (loferist@gmail.com), Национальный исследовательский Томский государственный университет, Томск, Россия

В последнее время нейронные сети все чаще применяются для задач автоматической суммаризации текста. Одним из частных случаев суммаризации текста является генерация новостных заголовков. Среди многих моделей нейронных сетей, применяемых для генерации текста, самой распространенной является модель Seq2Seq. Несмотря на то, что данная модель активно применяется для различных языков, применимость Seq2Seq для анализа материала на русском языке на сегодняшний день остается малоисследованной. В данной работе представлен процесс реализации данной архитектуры, а также анализ полученных результатов. Кроме того, в исследовании приведен ряд факторов, которые влияют на качество генерации новостных заголовков.

Ключевые слова: обработка естественного языка, нейронная сеть, Seq2Seq, генерация текста, суммаризация текста, генерация заголовков

1. Background

Today we witness rapid increase of textual data and it's becoming more critical to build systems that can effectively perform text summarization tasks. Recent advances in language modelling [5] and problems addressed at NLP conferences confirm that text generation and summarization tasks are currently in the spotlight of computational linguistics.

Over the past years machine and deep learning methods proved to be very effective for various natural language processing tasks. These tasks include text classification, sentiment analysis, language modeling, automatic text summarization, as well as neural headline generation for news articles [4].

The purpose of any headline is to convey a meaning of an article in a contracted form. Therefore, the peculiarity of headline generation is that it in fact comprises two tasks: capturing the meaning of an article (i.e. summarization) and generating a text from the meaning (i.e. text generation.).

Naturally, a text is a sequence of language units, be it characters, words, sentences etc. Regarding the task of headline generation it means our data has two properties: it is sequential (i.e. the order matters) and its length is not predefined. Hence the best approach would be to process the input one step (i.e. token) at a time, finishing computations after reaching the end of sequence. That means that the best choice would be to use recurrent neural networks, specifically LSTM or GRU to overcome the exploding/vanishing gradient problem and to effectively allowing RNN to remember longer sequences.

As for extracting the gist of a text and wording it, we are using Seq2Seq architecture . Today, it is widely used for generating texts with the length which is not determined a-priori. These tasks include machine translation, speech generation and recognition as well as text summarization. Basically, Seq2Seq is 2 connected RNNs. The first one, the encoder, takes a sequence (e.g. text) as an input and outputs a vector sometimes called 'Thought Vector' [6]. Another one, the decoder, takes this vector as as input and outputs a target sequence (e.g. headline); in its core it is a probabilistic language model, conditioned by the vector representation of the source sequence.

Seq2Seq is not the only architecture suitable for this task and definitely not the newest. However, we chose this one for several reasons. First, it is still widely used, for example, by Google Translate [8]. Secodenly, it might be not the newest one, but still relatively young and thus not reached its full potential. Thirdly, there is little evidence of applicability of this model to the Russian language.

In this paper we tested a Seq2Seq model tailored to the Russian language on the task of neural headline generation. For decoding an output sequence we used naive geedy search as well as beam search algorithm than was implemented specifically for this task. We also analyse predicted headlines and provide evaluation of our model's predictions via BLEU [3] and ROUGE [1] score metrics, which were initially introduced for machine translation and automatic summarization evaluation. We also discuss various factors which can affect model's performance.

2. Model

The Seq2Seq model used for our task had one LSTM layer for encoder and one for decoder with 256 hidden units each. We used word embeddings for our input layers for both encoder and decoder with the dimension size of 256. The training process employed categorical cross-entropy loss and rmsprop optimizer. Model's learning rate was set to 0.001.

Seq2seq inference is usually implemented in two ways. The naive approach, often referred to as "greedy" considers only the best word prediction from the output dense layer of the model, which gives the probability distribution amongst the whole vocabulary. Another approach is called beam search, which expands all possible next steps and keeps the *n* most likely predicted words. In general, beam search is considered to be a more optimal solution for Seq2Seq output decoding.

For our task we implemented both, greedy and beam search inference, and tested performed evaluation using BLEU and ROUGE score. For our beam search inference we set the beam size equal to 3.

3. Data

In order to train our model we used the Russian news articles corpus collected from Lenta.ru [9]. Initially, the corpus consisted of 699777 news articles along with their urls, titles, topics and tags. However, for our task we considered only articles and their headlines. We divided our data 80/20% for training/test sets, then we used 20% of the training set to form a validation set.

Since Seq2Seq models rely on vector representations of words, we used a fixed vocabulary for both input and output sequences. All text data was converted into lower case. For our input vocabulary 30000 of the most frequent news article words were selected. The same process was applied for compiling a target vocabulary. Each unknown word was replaced by a special "UNK" token. We also added "START" token at the beginning and "END" token at the end of each headline. Maximum input length was set to 550 words and maximum target sequence to 13 words. Each sequence that exceeded its maximum length was clipped.

4. Experiment

We trained our model for 12 epochs with the batch size of 256 which took us 12 hours on Tesla K80 GPU with 11 GB of memory.

Our model achieved 1.8678 loss on training set and 2.3024 loss on validation set. Alteration of training and validation loss during training is depicted on the Fig.1.

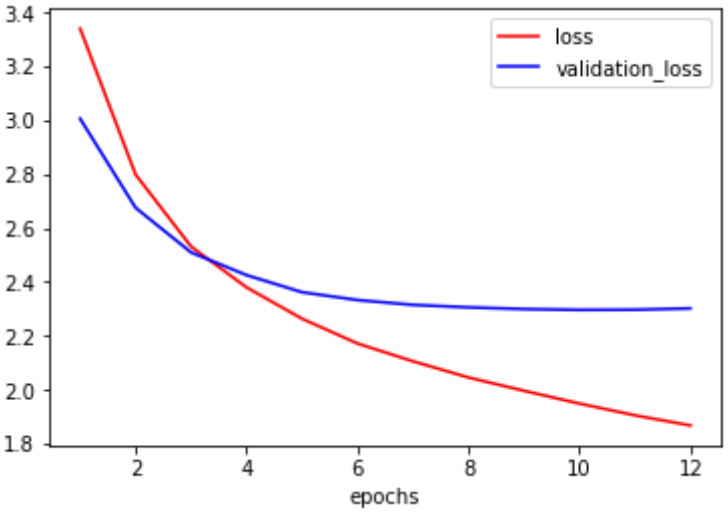

Fig. 1. Training and validation loss

5. Evaluation

For the purpose of evaluation we used BLEU score metric, which stands for. Due to complexity of BLEU score and beam search computing, evaluation process is time consuming. Initially we planned to test our model's performance on the entire test set, but so far we have finished evaluation on 5000 examples. BLEU score for greedy inference achieved the value of 0.10875064633778518, whereas the score for beam search inference reached 0.05100243652838388.

Moreover, one of the shared tasks, organized by Dialogue 2019 conference, was devoted to the task of automatic news headline generation. Evaluation method proposed by organizers (based on another metric for evaluating automatic summarization and machine translation) is an arithmetic mean of Rouge-1-F, Rouge-2-F, Rouge-L-F. Based on this

method, our model achieved 0.05800769685750703 on greedy inference, and 0.04194463545277877 using beam search, as compared to 0.23141523643530523 which is a highest score achieved by the winner of this competition.

Examples of predicted headlines are shown in the Table 1.

Table 1: Predictions from sources different from training

Original Headline: Российская теннисистка вышла в полуфинал турнира в Сеуле **Generated greedy headline:** российская теннисистка вышла в полуфинал турнира **Generated beam-search headline:** российская теннисистка вышла замуж
Original Headline: Новую часть WWE выпустят в октябре **Generated greedy headline:** названа дата выхода новой части про **Generated beam-search headline:** названа дата выхода новой игры
Original Headline: В США арестовали одного из самых плодовитых спаммеров **Generated greedy headline:** в сша арестован самый большой в мире **Generated Beam Headline:** в сша sony
Original Headline: Немцы заработали на продаже акций "Газпрома" 3,4 миллиарда евро **Generated greedy headline:** "газпром" продал долю в европе **Generated beam-search headline:** "газпром" продал долю "газпрома"
Original Headline: Курс юаня вырос на 7588 пунктов **Generated greedy headline:** курс рубля вырос впервые с начала года **Generated beam-search headline:** курс рубля вырос
Original Headline: ЕС не стал открывать границу для украинцев **Generated greedy headline:** евросоюз отказался ввести визовый режим для украины **Generated beam-search headline:** евросоюз и сша
Original Headline: LiveJournal решил сократить начальников **Generated greedy headline:** основатель livejournal объявил об отставке **Generated beam-search headline:** основатель livejournal

Evaluation process revealed that currently naive greedy inference outperformes beam search. This may be due to lack of training data, insufficient model complexity or need for more training epochs. However, we can see that generated headlines (generated via greedy inference in particular) resembled original ones in terms of vocabulary and succeeded in learning word order.

In addition, competition results on the same task suggest that more advanced models (such as seq2seq with attention) can yield much better results. This points out that pure seq2seq isn't an optimal approach for the task.

6. Conclusion

In this paper, we explored the application of Seq2Seq model for the task of neural headline generation as well as tested two inference algorithms (greedy and beam search) via BLEU score metric. Experiments showed that the model is capable of generating headlines

given a news article. Predicted examples (generated by greedy inference) had coherent word order and resembled original news headlines. However, these model predictions are still far from handcrafted ones, which indicates the importance of further investigation of this problem. Moreover, recent progress of similar shared tasks points out the number of more advanced neural architectures and encoding techniques which can be applied for the same problem and lead to a better results.

7. Future plans

We plan to continue working on the problem of neural headline generation and currently we are training a multilayer Seq2Seq model and expect new results for our research. In our next experiments we aim at employing attention mechanism [7] as well as different data encoding algorithms such as byte pair encoding.

We are also going to extend our training data by merging our dataset with the recently introduced corpus Rossiya Segodnya [2]. One of our objectives is to investigate how model performance depends on its complexity and amount of training data.

Acknowledgements

This work has been done in the framework of TSU Competitiveness Enhancement Programme 2013-2020.

References

1. Chin-Yew Lin (2004), ROUGE: A Package for Automatic Evaluation of Summaries, available at: https://www.aclweb.org/anthology/W04-1013
2. Gavrilov D., Kalaidin P., Malykh V. (2019), Self-Attentive Model for Headline Generation, available at: https://arxiv.org/abs/1901.07786
3. Kishore Papineni, Roukos S., Ward T., Zhu. W. Bleu (2002), A method for automatic evaluation of machine translation [In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02], Stroudsburg, PA, USA, pp. 311–318.
4. Lopyrev K. (2015), Generating News Headlines with Recurrent Neural Networks, available at: https://arxiv.org/abs/1512.01712
5. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. (2019) Language Models are Unsupervised Multitask Learners, available at: https://openai.com/blog/better-language-models/
6. Sutskever I., Vinyals O., Le Q.V. (2014) Sequence to sequence learning with neural networks, available at: https://arxiv.org/abs/1409.3215
7. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017), Attention is all you need [Advances in Neural Information Processing Systems], pp. 5998–6008.
8. Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M. (2016), Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, available at: https://arxiv.org/pdf/1609.08144.pdf
9. Yutkin D. (2018), Corpus of Russian news articles collected from Lenta.Ru, available at: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta