Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

PHRASE-BASED ATTENTIONAL TRANSFORMER FOR HEADLINE GENERATION

Sokolov A. M. (sokolov.andrej.m@gmail.com)
SPBU, Saint-Petersburg, Russia

Nowadays the task of selecting key information from large amount of text data is becoming more and more relevant. This article proposes a model of deep neural network with phrase-based attentional mechanism used for automatic generation of news headlines. The proposed architecture achieves a new state-of-the-art on the RIA news dataset.

Key words: text summarization, headline generation, Russian language, neural networks, self-attention

PHRASE-BASED ATTENTIONAL TRANSFORMER ДЛЯ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ

Соколов А. М. (sokolov.andrej.m@gmail.com)
СПбГУ, Санкт-Петербург, Россия

В настоящее время задача выделения ключевой информации из больших объемов текстовых данных становится все более и более востребованной. В данной статье предлагается модель глубокой нейронной сети с фразовым механизмом внимания, применяемой для автоматической генерации новостных заголовков. Предложенная архитектура достигает наилучших на данный момент результатов на наборе новостей РИА.

Ключевые слова: автоматическое реферирование текстов, генерация заголовков, нейронные сети, механизм внимания

1. Introduction

Name of the task of generating news headlines is pretty self-explanatory: having a news text you need to generate a short title for it that reflects an essence of the news. This problem is a special case of the *abstractive summarization*. In contrast to the *extractive summarization*, where it is sufficient to select the most important words or sentences from the text, in the abstractive summarization we can use paraphrasing or words not contained in the original text.

The rapid development of recurrent networks and language models shakes-up research of abstractive summarization methods. Transformer architecture [17] became an excellent replacement of RNN and allowed us to train deep networks faster without loss of quality. A key part of the Transformer is an attention mechanism. In classic version of the Transformer, attention allows to model and recognize connections between individual tokens, ignoring connections between phrases directly. An architecture used in this article is based on *Phrase Based Attentions* [10], which allows us to fix the described drawback.

RIA Dataset proposed in [4] has been used for training. The dataset consists of approximately one million headline-news pairs of the Russian news agency "Rossiya Segodnya".

Section 2 describes the architecture of the model used. **Section 3** discusses the dataset, data preprocessing pipeline and training in more detail. **Section 4** and **5** describe the experiments and results respectively. **Section 6** is devoted to the analysis of offered approach shortcomings and reflections on ways of its solution. **Section 7** provides a brief overview of abstract summarization methods proposed by various researchers. In the last section you can find conclusions and arguments on the work done.

2. System description

Denote $\mathbf{x}$, $\mathbf{y}$ as sequences of the news text tokens vector representations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and news header $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. We will use a statistical language model $p(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i=1}^{m} p(y_i|y_{i-1}, \ldots, y_1, \mathbf{x}, \theta)$ to generate header, where θ— model parameters. It is proposed to use the classical Transformer architecture [17] with *heterogeneous* attention [10] as language model parameterization. Unlike recurrent neural networks, which have been a state-of-the-art approach to NLP problems for a long time, the use of the Transformer allows us to learn more effectively. The self-attention mechanism affords to calculate hidden representations of sequences in parallel, while RNN hidden state h_t can be obtained only after calculating the previous state h_{t-1}. One modification applied to the original transformers in our article is *heterogeneous* attention. This type of attention extends receptive fields of the model adding an ability to directly model relationships between tokens and phrases. This effect is achieved through the use combination of 1, 2 kernel convolutions applied to the input sequences of attention blocks. Then this convolved sequences are concatenated and used as input of multi-head attention. In the interest of space, we omit the details and send an interested reader to the original articles [10], [17].

3. Data and training

3.1. Dataset

We use RIA dataset[1] for train. It contains 1,003,869 news with written headings. On average, header consists of 10 words, and text of news consists of 316 words. A subset of 20,000 examples of this dataset was reserved for testing proposes. The remaining part is used for training.

3.2. Preprocessing

First of all, the entire text of the dataset was reduced to lowercase, all html tags and their contents were removed. In order to simplify model training we use only the first 3 news sentences. It is acceptable due to the fact that the main essence of news contains in the first few sentences. We also limit the maximum number of tokens processed to 150.

The next step in data preparation is to split the text into tokens. It is proposed to use *Byte Pair Encoding* [15] as a tokenization method for both news text and header. This approach is currently a state-of-the-art tokenization method for NLP problems. BPE solves a problem of out-of-vocabulary words and works well with morphologically rich languages. We also use *word2vec* [8] as a vector representation of tokens.

3.3. Training

Using a given model parameterization $p(\mathbf{y}|\mathbf{x}, \theta)$, we will minimize the *negative log likelihood* function $NLL(\theta) = -\log \mathcal{L}(\theta) = -\sum_i \log p(\mathbf{y}_i|\mathbf{x}_i, \theta)$ during training, where $(\mathbf{x}_i, \mathbf{y}_i)$ are training pairs of the dataset.

Models were trained using *Adam* [5] optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ with learning rate $lr = 10^{-4}$. Each iteration of the optimizer uses a batch of data of size equal to 16. Training continues for 10 epochs.

3.4. Inference and evaluation

During inference we will use a *beam-search* algorithm. Unlike a greedy approach, where a token with the maximum probability is selected for each decoding step, *beam-search* uses m most likely independent header generations, trying to maximize resulting likelihood.

The most popular summarization quality metric is ROUGE [6]. We use ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-l (longest common subsequence) version, scored by F-measure.

[1] The dataset is available at https://vk.cc/8W0l5P

4. Experiments

In this article[2], *BPE* and *word2vec* were trained on the training part of the dataset used. The generated vocabulary contains 30,000 tokens, the dimension of word vector representations is 300. Decoding beam-search size equal to 5.

First Sentence: The simplest and most naive approach to generating headlines. The first sentence of the text is used as a summarization. This approach is used with the assumption that the main point is contained in the beginning of the news.

RNN: We will use a classic seq2seq [16] architecture with attention as a baseline. Encoder and decoder are five-layer bidirectional GRU [2] with hidden size equal to 500. A dropout with probability equal to 0.1 after each layer is used for regularization. As attention we use attention via dot-product [7].

Universal Transformer: For comparison, we use the results of Gavrilov et al. [4], based on Universal Transformer. They used 4 layers in encoder and decoder with 8 heads of attention.

Vanilla Transformer: In this case, we will use the classic Transformer architecture with default settings for both encoder and decoder: 6 layers, 8 attention heads, model hidden size is 512, position-wise block hidden size is 2048, dropout equal to 0.1.

Phrase Based Attentional Transformer: PBA transformer settings are very similar to the classic Transformer, except for the attention block. It uses the *heterogeneous* approach: one and two kernel convolution applied to key, value and query. Next, we concatenate this convolved sequence representations and calculate scaled dot-product attention.

5. Results

Table 1: Evaluation results on RIA dataset

Model	ROUGE-1-f	ROUGE-2-f	ROUGE-l-f
First Sentence	24.08	10.57	16.70
RNN	37.98	20.51	35.36
Universal Transformer	39.75	22.15	36.81
Vanilla Transformer	42.42	25.06	39.50
PBA Transformer	**42.96**	**25.43**	**40.02**

[2] The source code for all experiments is available at https://github.com/gooppe/deep-summarization-toolkit

Table 2: PBA Trabsformer generation sample

Text: к 2016 году 20 % школ должны быть доступными для обучения инвалидов, в настоящее время этот показатель составляет чуть больше 2 %, сообщил министр труда и социальной защиты максим топилин на заседании правительства рф. «к 2016 году 20 % школ, не коррекционных, а обычных, должны быть приведены в доступный вид для обучения инвалидов, сегодня этот показатель на начало реализации программы (по доступной среде для инвалидов) составляет 2,5 %», — сказал топилин. по его словам, увеличение в 10 раз — это неплохой показатель, хотя в дальнейшем доступными для обучения инвалидов должны быть все школы. **Original summary**: топилин: к 2016 году 20 % школ должны быть доступными для инвалидов **Generated summary**: к 2016 году 20 % школ должны быть доступны для инвалидов — минтруд
Text: бригада сахалинского бассейнового аварийно-спасательного управления из-за непогоды приостановила работы на аварийном судне мр-150–289 у южного побережья сахалина, сообщили риа новости в главном управлении мчс рф по региону. судно мр-150–289 шло в портпункт озерск корсаковского района. но из-за поломки в системе теплоподачи судно зашло в бывший портпункт новиково. **Original summary**: спасатели из-за непогоды приостановили работы на подтопленном судне **Generated summary**: спасатели приостановили работу на аварийном судне у сахалина

As you can see from Table 1, PBA Transformer shows the best result. The classical recurrent sequence-to-sequence approach is slow. A recurrent network needs much more time to achieve the quality of Transformers. Vanilla Transformer has better results than Universal Transformer presumably due to greater depths. PBA transformer shows interesting results. Qualitatively, it has a small increase, but its ability to generate abbreviations seems to be quite interesting. Table 2 shows example of PBA Transformer generation sample.

Unfortunately, this model is not perfect. Sometimes it makes mistakes, such as usage of incorrect forms of words, confuses with key figures or repeating them. However, the model almost always highlights relevant information, which is inspiring.

6. System and error analysis

In this section, we would like to draw reader's attention to one important problem that appears during testing of trained models on another datasets. The essence of the problem is that a model trained on the dataset of one news Agency cannot be applied to generate news headlines from another source. It seems that data structure should be the same for different news Agencies, but alas, models confuse and generate bad

headlines. During testing on different datasets, we made sure that naive use of the first news sentence as a title shows results better than generations of trained models. This effect can not be considered as overfitting, because on large test subsets from the same dataset, the model achieves good metric estimates. Firstly, this problem may occure due to strong variability in a style of news and headlines. Each Agency uses its own writing style, and model is strongly attached to it. Secondly, there may be shifts in news domains, because of them model focuses on some specific topics and can not cover all areas of text news. Third, it is hard to find supervised summarization dataset, that covers all aspects of human life. There will be always some aspect missing from the training dataset that will be processed with difficulties by proposed model. More formally, it is impossible to use proposed model on different datasets due to Distribution Assumption: there is one probability distribution D that governs both training and testing examples.

The first thing that comes to mind for solving this problem is the use of pre-trained language models on large corpora [3], [11], [12]. These models parametrize language prior distribution. Language model can be fine-tuned on a specific dataset, directly to solve summarization problem. It stands to reason that having some General language representation we can more accurately distinguish information from texts that are not even present in the task-specific dataset. This approach can be used as a baseline for further research.

7. Related work

The task of abstract summarization and the task of generating news headlines in particular deserve much attention of researchers. So, Rush et al. [13] were the first, who proposed to use a deep, fully connected neural network with an attention mechanism as a language model for generating news headlines. Later, approaches based on recurrent neural networks were proposed: Chopra et al. [1] suggested to use the classic recurrent sequence-to-sequence architecture with attention. Other researchers tried to adapt such approaches to the specifics of automatic summarization. Nallapati et al. [9] offered several ideas, potentially improving previously proposed models: large vocabulary trick, hieracical attention and copy-from-text approach. [See et al. 2017] developed the idea of copying some text from the original and proposed to use the Pointer-Generator Network for modelling rare or unseen words. [Gavrilov et al. 2019] applied Universal Transformer with BPE tokenization and offered a new Russian dataset for the research of automatic referencing methods.

8. Conclusion and future work

Under this article, an attempt was made to use a modified Transformer with a phrase based attention mechanism. This modification has improved the quality of the base model and achieved a new state-of-the-art in the task of headline generation on the RIA dataset. Experiments with testing of trained models on another datasets have led us to the problem of model dependence on the news Agency. In the future the results can be improved by using language models or unsupervised approaches.

References

1. *Chopra, S. et al.:* Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 93–98 Association for Computational Linguistics, San Diego, California (2016).
2. *Chung, J. et al.:* Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR. abs/1412.3555, (2014).
3. *Devlin, J. et al.:* BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. abs/1810.04805, (2018).
4. *Gavrilov, D. et al.:* Self-attentive model for headline generation. In: Proceedings of the 41st european conference on information retrieval. (2019).
5. *Kingma, D. P., Ba, J.:* Adam: A method for stochastic optimization. CoRR. abs/1412.6980, (2015).
6. *Lin, C.-Y.:* ROUGE: A package for automatic evaluation of summaries. In: ACL 2004. (2004).
7. *Luong, T. et al.:* Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1412–1421 Association for Computational Linguistics, Lisbon, Portugal (2015).
8. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems - volume 2. pp. 3111–3119 Curran Associates Inc., USA (2013).
9. *Nallapati, R. et al.:* Sequence-to-sequence rnns for text summarization. CoRR. abs/1602.06023, (2016).
10. *Nguyen, P. X., Joty, S.:* Phrase-based attentions. CoRR. abs/1810.03444, (2018).
11. *Peters, M. et al.:* Deep contextualized word representations. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers). pp. 2227–2237 Association for Computational Linguistics, New Orleans, Louisiana (2018).
12. *Radford, A. et al.:* Improving language understanding by generative pre-training. In: Technical report, openai. (2019).
13. *Rush, A. M. et al.:* A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 379–389 Association for Computational Linguistics, Lisbon, Portugal (2015).
14. *See, A. et al.:* Get to the point: Summarization with pointer-generator networks. CoRR. abs/1704.04368, (2017).
15. *Sennrich, R. et al.:* Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 1715–1725 Association for Computational Linguistics, Berlin, Germany (2016).
16. *Sutskever, I. et al.:* Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014).
17. *Vaswani, A. et al.:* Attention is all you need. CoRR. abs/1706.03762, (2017).