Автоматическая суммаризация новостей на русском языке

Алексей Заплатин

Январь 2022

Abstract

В данной работе рассматривается применение sequence-to-sequence модели mBart для генерации заголовков новостей на русском языке. Проведен обзор существующих моделей для суммаризации текста. Полученные результаты были протестированы с помощью метрики ROUGE. Результаты работы доступны по ссылке: `https://github.com/alekz99/summarizationRussianNews/tree/main`.

1 Introduction

Суммаризация - это задача создания сжатой версии исходного текста, которая содержит его основную мысль. Данная задача является актуальной и ее значимость только растет. Книги, фильмы, новости имеют сжатую версию, чтобы человек быстро мог оценить необходимость материала для себя. Создание сжатой версии текста(заголовка, резюме) является трудоемким и долгим процессом, поэтому данная область активно развивается.

Методы, используемые для решения задачи суммаризации, можно разделить на три категории:

1. Экстрактные, смысл которых заключается в поиске наиболее "важных" информационных блоков, где блок — абзац, предложение или ключевые слова. Для определения «важности» блока используют специальные оценочные функции.

2. Абстрактные, которые создают новый текст похожий на человеческий пересказ.

3. Комбинированные, которые состоят из экстрактных и абстрактных методов.

В данной работе будет рассмотрен метод абстрактной суммаризации заголовка новостной статьи, так как такой подход позволяет создавать аннотации, которые близки к пересказу текста, составленного вручную

человеком. Из недостатков такого подхода можно выделить: сложность реализации алгоритмов, требовательность к вычислительным ресурсам при их реализации.

1.1 Team

Данная работа была выполнена студентом университета ИМТО, **Заплатиным Алексеем**.

2 Related Work

В данном разделе будут рассмотрены современные модели, которые способны генерировать новые предложения, которых не было в аннотируемом тексте.

Pointer-Generator Network — это модификация sequence-to-sequence RNN с механизмом внимания. Архитектура сети позволяет ей копировать слова из исходной последовательности. Также в модели предусмотрен алгоритм для борьбы с повторами [See et al., 2017].

mBart — sequence-to-sequence трансформер. Модель сразу предобучается для генерации текста, и поэтому лучше подходит для автоматического реферирования. Модель обучалась на многоязычном корпусе CC25, подмножестве Common Crawl из 25 языков. При предобучении использовались одноязычные части корпуса, то есть никаких переводов модель не видела. По объему данных русский язык находится в тройке лидеров, поэтому модель можно использовать для задачи суммаризации на русском языке [Liu et al., 2020].

T5 — sequence-to-sequence трансформер, глобально аналогичный BART, разработка Google Research. T5 предобучается на задаче восстановления промежутков, то есть наборов подряд идущих токенов. При обучении промежутки исходного текста скрываются, и задача модели их сгенерировать. В отличие от BART, где текст генерируется целиком, T5 нужно сгенерировать только сами скрытые промежутки [Raffel et al., 2020].

PEGASUS — sequence-to-sequence трансформер со специальной процедурой обучения, смысл которой заключается в генерации пропущенных предложений. В документе важные предложения заменяются с помощью маски, а модель обучается на неполных аннотациях текста. Авторы предлагают 3 основных стратегии выбирать важные предложения: случайно; брать первые несколько предложений; брать несколько предложения по некой мере важности. В качестве меры важности можно брать похожесть предложений на оставшийся текст по ROUGE [Zhang et al., 2020].

PEGASUS показывает лучшие результаты в сравнении с mBart и T5, но она обучена на английском корпусе, поэтому для работы была выбрана модель mBart.

3 Model Description

mBart - состоит из кодировщика BERT и декодировщика GPT. Архитектура mBart представлена на рис. 1.

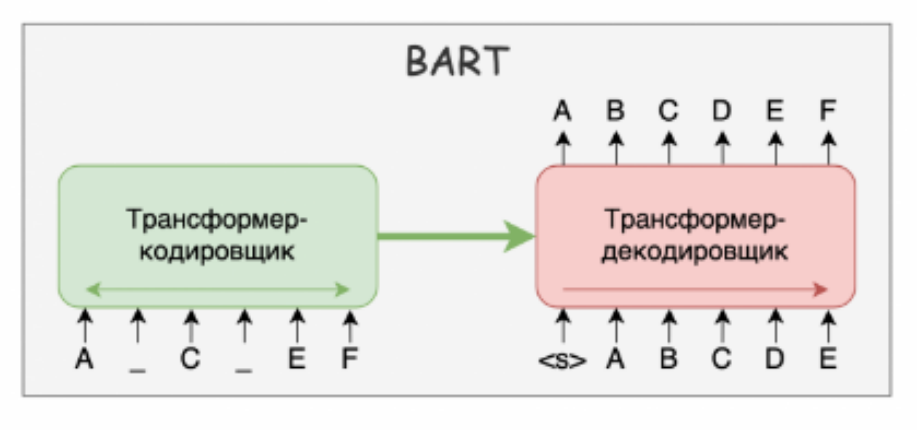

Figure 1: Архитектура mBart

Как было сказано ранее, модель обучается на зашумленных данных. Способы зашумления данных:

- перемешивание предложений,
- маскирование токенов,
- циклический сдвиг текста,
- удаление токенов.

Таким образом, в данной работе будет дообучена модель mBart на корпусе новостей на русском языке для генерации заголовков статей.

4 Dataset

Для обучения модели использовался набор данных с сайта Kaggle[1]. Он состоит из 800975 новостей на русском языке с сайта Lenta.ru. Набор данных содержит следующие атрибуты: ссылка на новость, заголовок, текст новости, тема, тег, дата.

Предобработка набора данных:

1. удаление новостей без текста;
2. преобразование заголовка и текста новости к нижнему регистру;

[1] https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta

3. удаление неразрывного пробела из данных (\xa0);

4. удаление новостей опубликованных до 2000 года.

Статистика набора данных представлена в табл. 1. Были удалены новости, текст которых не попадает в отрезок от 56 до 395 слов.

	Длина заголовка	**Длина текста**
count	797884	797884
mean	7.496639	180.315358
std	1.801768	73.651864
min	1	1
25%	6	133
50%	7	170
75%	9	216
max	18	8092

Table 1: Статистики набора данных

81.6% новостей имеют теги: все, политика, общество, Украина, происшествия, госэкономика, футбол, кино, интернет, бизнес. Количество новостей за каждый год представлено на рис. 2.

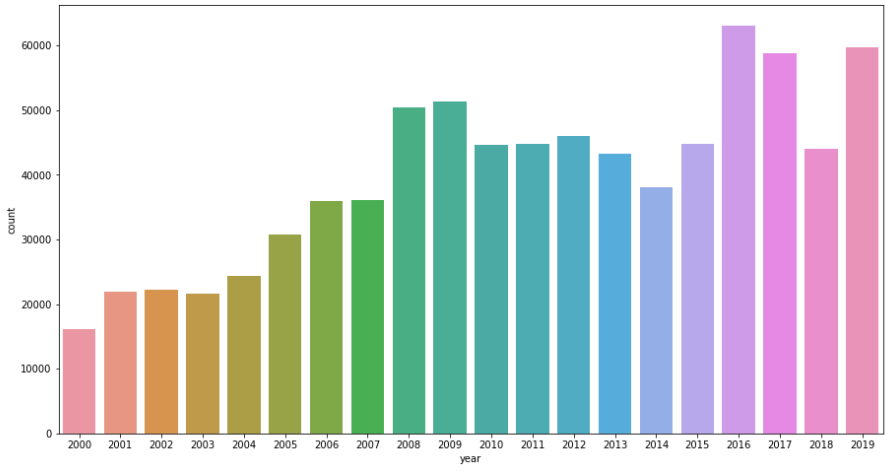

Figure 2: Количество опубликованных новостей за каждый год

Для обучения и оценки из набора данных было выбрано 40000 новостей и разделено на три части. В табл. 2 представлено разделение.

	Train	Valid	Test
Количество объектов	30000	5000	5000

Table 2: Разделение набора данных

5 Experiments

5.1 Metrics

Для оценки полученных результатов будет использоваться метрика ROUGE, которая распространена для оценки задач суммаризации и машинного перевода:

- ROUGE-1 - расчет совпавших униграм,
- ROUGE-2 - расчет совпавших биграм,
- ROUGE-L - расчет, учитывающий самую длинную общую последовательность.

5.2 Experiment Setup

Модель была обучена с помощью фреймворка Hugging Face[2].

Для обучения модели на русском языке были указаны токены языка для входной и выходной последовательностей, которые указывают, что переданные последовательности относятся к русскому языку.

Опираясь на статистику набора данных, была установлена максимальная длина входной и выходной последовательностей 600 и 60 токенов соответственно. Последовательности дополнялись до максимальной длины токенами-паддингами. Также ид токена паддинга заменялся на -100, чтобы он не учитывался при подсчете метрики.

Для обучения использовался бесплатный сервис Colab. Были выбраны следующие параметры для обучения модели:

- num_train_epochs = 3,
- per_device_train_batch_size = 4,
- per_device_eval_batch_size = 4,
- gradient_accumulation_steps = 8,
- gradient_checkpointing = True,
- fp16 = True.

[2]https://huggingface.co/facebook/mbart-large-cc25

gradient_accumulation_steps и gradient_checkpointing необходимы для того, чтобы хватило оперативной памяти для обучения модели. **Gradient Accumulation** позволяет вместо вычисления градиентов для всего батча сразу, делать это более мелкими шагами. **Gradient Checkpointing** оптимизирует количество необходимой памяти при прямом и обратном проходе. **fp16** - уменьшается размерность данных, что увеличивает скорость обучения модели [3].

Обучение заняло 4,5 часа. Результаты обучения представлены в табл. 3.

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel
0	1.938100	1.285532	34.808800	17.808800	32.907900
1	1.038900	1.222873	35.582200	18.410300	33.631100
2	0.744700	1.251436	35.794900	18.600500	33.828500

Table 3: Процесс обучения модели

5.3 Baselines

Для сравнения полученных результатов будет использоваться алгоритм для экстрактной суммаризации LexRank, [Erkan and Radev, 2004]. LexRank — алгоритм обучения без учителя на основе графов, который использует модифицированный косинус обратной частоты встречи слова, как мера похожести двух предложений. Похожесть используется как вес грани графа между двумя предложениями. LexRank также внедряет шаг умной постобработки, которая убеждается, что главные предложения не слишком похожи друг на друга. Была использована готовая реализация[4].

6 Results

После обучения модели были сгенерированы 5000 заголовков на тестовом наборе данных, которые доступны в репозитории. Значения метрики ROUGE представлены в табл. 4. Видно, что mBart намного лучше справляется с генерацией заголовков. Также можно косвенно оценить полученные результаты по статье [Gusev, 2020]. В данной статье использовался набор данных с сайта gazeta.ru. В статье генерировали резюме новости, состоящее из нескольких предложений, с помощью mBART*. Это более сложная задача, поэтому значения метрик меньше. Таким образом, можно сделать вывод, что удалось получить высокое качество сгенерированных заголовков.

В табл. 5 представлены оригинальные и сгенерированные заголовки. В первом сгенерированном заголовке модель фактически ошибается, но

[3] https://huggingface.co/docs/transformers/v4.23.1/en/perf_train_gpu_one

[4] https://pypi.org/project/lexrank/

	Rouge-1	**Rouge-2**	**Rouge-L**
LexRank	12.9408	3.9882	11.9508
mBart*	32.1	14.2	27.9
mBart	35.9931	18.6076	34.1661

Table 4: Значения метрики ROUGE

Оригинальный заголовок	Сгенерированный заголовок
бабочку хофштадтера впервые «поймали» вне магнитного поля	бабочку хофштадтера впервые продемонстрировали в магнитном поле
youtube занялся гражданской журналистикой	youtube запустил сервис видеохостинга для сми
минздрав пообещал заменить импортные лекарства безболезненно для пациентов	минздрав рассказал о внедрении новой методики перевода пациентов с зарубежных лекарств на российские аналоги
конгресс дал обаме денег на войны и борьбу с гриппом	американские законодатели одобрили экстренные бюджетные расходы на борьбу с эпидемией гриппа

Table 5: Примеры оригинальных и сгенерированных заголовков

в целом заголовок соответствует новости. В остальных трех заголовках модель верно передала суть статьи. Из интересного можно заметить, что модель часто опирается на первые предложения статьи, но не берет целиком их, а легко и точно использует аббревиатуры, синонимы или перефразирует предложения.

7 Conclusion

Таким образом, в данной работе была проведена предобработка набора данных, получены базовые версии заголовков с помощью алгоритма LexRank, дообучена модель mBart для генерации новостных заголовков на русском языке. Было показано, что при небольшой обработке входных данных, модель способна выдавать высокое качество на русском языке в задаче суммаризации.

References

[Erkan and Radev, 2004] Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

[Gusev, 2020] Gusev, I. (2020). Dataset for automatic summarization of russian news. In *Conference on Artificial Intelligence and Natural Language*, pages 122–134. Springer.

[Liu et al., 2020] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

[See et al., 2017] See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

[Zhang et al., 2020] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.