Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке

Д. В. Мельничук, А. В. Носкина

Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского

melnichukdv@sgu.ru, noskinaav@sgu.ru

Аннотация

В данном исследовании сравниваются основные NLP-модели, такие, как mBART, T5 и GPT-3, которые в своей основе имеют архитектуру трансформеров, т. е. механизм «внимания», кодирующий, декодирующий и нормализующий слои. Данные предобученные модели на задаче суммаризации русского текста были использованы для суммаризации научных статей на русском языке. Для выявления лучшей модели на данном классе задач в исследовании был использован набор данных, включающий в себя тексты научных статей и соответствующие им авторские аннотации на русском языке. Далее, стандартными для задачи суммаризации статистическими метриками, такими, как семейство метрик ROUGE (ROUGE-1, ROUGE-2 и ROUGE-L), а также BLEU и Perplexity, находилась наиболее эффективная модель в рамках поставленной задачи, т. е. сравнивались по отдельности сгенерированные варианты аннотаций с авторской. Полученные результаты имеют практическую ценность, так как суммаризация текста является важной задачей в области обработки естественного языка.

Ключевые слова: NLP, суммаризация, mBART, T5, GPT-3

Библиографическая ссылка: Мельничук Д. В., Носкина А. В. Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 54–59. DOI: 10.17586/2541-9781-2024-7-54–59

1. Введение

Основной целью данной работы является ответ на вопрос: «Какая из NLP-моделей суммаризации (Natural Language Processing, NLP — Обработка текстов на естественном языке) наиболее оптимально работает в контексте академической литературы на русском языке»? Под суммаризацией текста понимается процесс автоматического сокращения объёма исходного текста путём извлечения наиболее важных и существенных идей, фактов и информации, а также представления в форме краткого и сводного текста, который сохраняет основные аспекты исходного материала.

Для сравнения эффективности разных типов предобученных (pre-trained) NLP-моделей использовался набор статей из открытой научной электронной библиотеки CyberLeninka, из части массива доступных данных, были использованы тексты (text) научных статей и соответствующие аннотации (annotation) их авторов на русском языке. Всего 825 статей, среди которых область наук и тип журнала брались в случайном порядке.

2. Модели и данные

Для исследования были выделены наиболее популярные, по версии ресурса HuggingFace Hub, открытые NLP-модели суммаризации текста, обученные на одном и том же корпусе новостных текстов на русском языке (Gazeta) [5].

Языковая модель GPT-3 (Generative Pre-trained Transformer) использует механизмы трансформеров для анализа контекста и генерации последовательностей слов, учитывая вероятность каждого следующего слова на основе предыдущих слов в тексте. Модель также способна выполнить различные задачи, такие, как ответы на вопросы, перевод текстов на другие языки и создание текстовых статей. В нашем исследовании была использована GPT-3 модель, обученная под задачу суммаризации, под кодовым названием модели на ресурсе HuggingFace Hub: RuGPT3MediumSumGazeta [6].

Модель T5 (Text-to-Text Transfer Transformer) также использует архитектуру трансформеров и обучается на задачах преобразования текста в текст. На вход подаётся задание и исходный текст, а затем генерируется выходной текст, решающий поставленную задачу. Модель обучается на широком спектре задач, включая машинный перевод, генерацию текста, ответы на вопросы, классификацию текста и многое другое. Для нашего исследования была использована базовая модель RuT5Base, обученная на новостных текстах на русском языке (Gazeta) под задачу суммаризации: RuT5SumGazeta [7].

Модель mBART (multilingual Bidirectional and Auto-Regressive Transformer) использует технологию мультиязычного перевода, обученную на большом количестве текстов на разных языках. Каждый язык представлен в виде уникального кода, и модель может работать с несколькими языками одновременно. При обучении модели mBART используется подход обучения с подкреплением, который позволяет модели улучшать свой перевод по мере того, как она получает обратную связь. Архитектура трансформеров позволяет данной модели учитывать контекст и зависимости между словами в предложении. Аналогичным образом использована базовая модель mBART, обученная на новостных текстах на русском языке (Gazeta) под задачу суммаризации: MBARTRuSumGazeta [8].

3. Метрики

Для оценки и сравнения языковых моделей используются два подхода.

Первый подход — это внешняя оценка (External evaluation), при которой оценивание модели происходит за счёт решения с её помощью задачи, на которую она рассчитана, и дальнейший анализ итоговых показателей потерь/точности, а также является лучшим подходом к оцениванию моделей, так как это единственный способ реально оценить, как разные модели справляются с интересующей нас задачей. Однако реализация данного подхода может потребовать больших вычислительных мощностей, его применение может оказаться медленным, так как для этого нужно обучение всей анализируемой системы (BLEU, ROUGE — это внешняя оценка).

Второй же подход - это внутренняя оценка (Internal evaluation), которая производит оценку самих языковых моделей, без учёта конкретных задач, для решения которых их планируется использовать; она является не столь информативной для понимания качества работы модели на конкретной задаче, как внешняя, но, если необходимо провести итоговую оценку модели, то данный подход может быть весьма эффективным для быстрого сравнения моделей (Perplexity — это внутренняя оценка).

В данной работе были использованы метрики: BLEU, семейство метрик ROUGE и Perplexity.

Метрика BLEU (Bilingual Evaluation Understudy) — это алгоритм оценки качества машинной генерации текста (в том числе перевода), основанный на сравнении выходных текстов, т. е. сгенерированных (predictions) с известными, эталонными (references) текстами. Сам подход заключается в сравнении двух вариантов текста, по совпадению слов и их

расположению, также это называют схожести n-грамм (последовательности n слов). В итоге получается количественная оценка соответствия между результатом работы NLP-модели и результатом работы человека: чем ближе машинная генерация к исходному тексту человека, тем он лучше - такова основная идея BLEU. Метрика BLEU включает корректировки весов, такие, как фактор бонуса на основе bi-граммов и сглаживание на основе ковариационной матрицы предложений, чтобы справиться с некоторыми из проблем данного подхода [3].

Пусть C — множество слов сгенерированного текста, R — множество слов эталонного текста, соответственно c_i и r_i — это i-е слова этих множеств (списков). Пусть n — максимальная длина n-грамм, которые мы рассматриваем. Тогда BLEU оценивает качество сгенерированного текста с путём вычисления взвешенного гармонического среднего точности n-грамм:

$$\text{BLEU} = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \cdot BP$$

$$p_n = \frac{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)_{r_i}}{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)}$$

$$\text{BP} = \begin{cases} 1, & if\ c > r \\ e^{(1-r/c)}, & if\ c \leq r, \end{cases}$$

- p_n— точность n-грамм (последовательности n слов) в сгенерированном тексте;
- N — количество предложений в корпусе, c_i — i-я сгенерированная фраза, r_i — эталонный перевод для i-й кандидатской фразы;
- $Count(n-gram)_{r_i}$ — количество вхождений n-грамм в r_i;
- $Count(n-gram)$ — количество вхождений n-грамм в c_i;
- BP— штрафной фактор.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) — это набор показателей (семейство метрик) для оценки автоматического суммирования текстов (в том числе машинного перевода), основанный на сравнении n-грамм сгенерированного (predictions) текста с n-граммами эталонных (references) текстов. Основная идея метрики ROUGE заключается в сравнении двух текстов и подсчёте базовых единиц (n-грамм, т. е. последовательностей слов и количества пар слов). В результате получаем количественную оценку работы NLP-модели, которая показывает, насколько сгенерированный текст совпадает с текстом, составленным человеком (экспертом). В отличие от BLEU, ROUGE использует как полноту (recall), так и точность (precision) для сравнения сгенерированных текстов с эталонными текстами, составленными человеком [2].

В ROUGE-1 сравниваются единицы (слова) между сгенерированным и эталонным текстами. В ROUGE-2 сравниваются последовательности из двух слов, взятых из сгенерированного и эталонного текста. В ряде источников ROUGE-1 и ROUGE-2 могут обозначаться общей записью ROUGE-N. ROUGE-L, в свою очередь, не сравнивает n-граммы, а обрабатывает тексты и ищет самую длинную последовательность (LCS), которая является общей для двух текстов, а затем измеряет её длину.

Пусть S — сгенерированный текст, G — эталонный текст, соответственно s_i и g_i — это i-е слова в S и G. ROUGE-N оценивает качество генерации из S путём вычисления точности совпадения слов в S с G, подсчитывает количество совпадающих (co-occurrences) n-грамм (для ROUGE-1 это одно слово, для ROUGE-2 это последовательность из двух слов), найденных как в выходных данных модели, так и в эталоне, а затем делит это число на общее количество n-грамм в S:

$$\text{ROUGE} - 1 = \frac{\sum_i Count_{natch}(s_i)}{\sum_i Count(s_i - 1, s_i)}$$

$$\text{ROUGE} - 2 = \frac{\sum_i Count_{match}(s_i - 1, s_i)}{\sum_i Count(s_i - 1, s_i)}$$

$$\text{ROUGE} - \text{L} = \frac{LCS(S,G)}{max(|S|,|G|)},$$

- *S* — сгенерированный текст, *G* — эталонный текст, s_i — i-е слово в *S*;
- $Count_{match}(s_i)$ — число вхождений слова s_i в обоих текстах *S* и *G;*
- $Count(s_i)$ —общее число n-грамм s_i в *S;*
- *LCS(S, G)* — самая длинная общая последовательность слов в *S* и *G;*
- *max(|S|, |G|)* — максимальное значение между количеством слов в *S* и *G*.

Метрика Perplexity в языковых моделях используется для оценки того, насколько хорошо модель может предсказать следующее слово в тексте. Для хорошей NLP-модели метрика Perplexity будет давать высокие вероятности синтаксически корректным предложениям, а предложениям некорректным (или очень редко встречающимся) — низкие вероятности. При условии, что набор данных состоит из корректных предложений, лучшей моделью будет та, которая назначит наивысшую вероятность этому тестовому набору, что означает то, что модель обладает хорошим пониманием того, как устроен язык [4].

$$\text{P(W)} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_{i-1,\dots,}w_{i-n+1})}$$

$$\text{Perplexity(W)} = \sqrt[N]{\frac{1}{P(w_i, w_2, \dots, w_N)}},$$

- W — набор слов в предложении;
- $p(w_i|w_{i-1})$ — вероятность того, что слово i будет следовать за словом i — 1.

4. Методология

Методология исследования выглядит следующим образом: изначально выделяются данные, включающие в себя авторскую аннотацию и три варианта автоматически сгенерированных для каждого отдельно взятого текста статьи NLP-моделями суммаризаций, а затем проводится сравнение на близость сгенерированных текстов с исходным текстом авторской аннотации.

В результате для каждой исходной статьи формируется оценка по пяти метрикам для трех NLP-моделей. Далее, имея данные результаты, находится среднее по каждому показателю, что и является итоговой оценкой эффективности работы данных NLP-моделей на задаче суммаризации.

5. Результаты и выводы

В результате, на задаче суммаризации академических текстов на русском языке, наилучшим образом проявила себя модель T5, которая показала наибольшую эффективность на основе статистических метрик. Данный результат может быть обусловлен тем, что модель T5 обеспечивает лучшую производительность и точность на задаче суммаризации текста, благодаря своей более общей и гибкой архитектуре, а также улучшенным параметрам и настройкам, в отличие от mBART и GPT-3.

Таблица. Результаты исследования на всем объёме данных

Модель	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity
mBART	9.1	**28.3**	13.3	27.0	39.84
T5	**10.1**	24.2	**13.4**	**27.6**	**30.7**
GPT-3	4.3	21.1	6.9	19.7	42.1

В дальнейших исследованиях планируется расширение проверочного набора статей, сравнение большего числа моделей, разделение проверочного датасета на области наук и сравнение результатов дискретно по научным областям.

Исследование выполнено за счёт гранта Российского научного фонда № 22-18-00153 «Образ СССР в исторической памяти: исследование медиастратегий воспроизводства представлений о прошлом в России и зарубежных странах» (https://rscf.ru/project/22-18-00153/).

Литература

[1] Gusev I. Dataset for Automatic Summarization of Russian News // Artificial Intelligence and Natural Language. AINL 2020 / Filchenkov A., Kauttonen J., Pivovarova L. (eds). Communications in Computer and Information Science. Vol. 1292. 2020. P. 122–134.

[2] Lin C. ROUGE: a package for automatic evaluation of summaries // Text Summarization Branches Out / Association for Computational Linguistics. Barcelona. 2004. P. 74–81.

[3] Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation // 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311–318.

[4] Nallapati R., Zhai F., Zhou B. SummaRuNNer: a recurrent neural network-based sequence model for extractive summarization of documents // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligenc. 2017. P. 3075–3081.

[5] Gusev I. Gazeta - Dataset for Automatic Summarization of Russian News // Hugging Face. 2021. URL: https://huggingface.co/datasets/IlyaGusev/gazeta (дата обращения: 17.05.2023).

[6] Gusev I. RuGPT3MediumSumGazeta — Model for abstractive summarization for Russian based on rugpt3medium // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta (дата обращения: 17.05.2023).

[7] Gusev I. RuT5SumGazeta — Model for abstractive summarization for Russian based on rut5-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (дата обращения: 17.05.2023).

[8] Gusev I. MBARTRuSumGazeta — Model for abstractive summarization for Russian based on rumbart-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta (дата обращения: 17.05.2023).

Comparison of NLP-models on the Task of Summarizing Academic Texts in Russian Language

Dmitriy V. Melnichuk, Anastasia V. Noskina

Saratov State University

This study compares major NLP models such as mBART, T5 and GPT-3, which have at their core a transformer architecture, i.e., an "attention" mechanism that encodes, decodes and normalizes layers. These pre-trained models on the task of summarizing Russian text, were used to summarize scientific articles in Russian. To identify the best model on this class of tasks, the study used a dataset including the text of scientific articles and their corresponding author's annotations in Russian. Then, using standard statistical metrics, such as the ROUGE family of metrics (ROUGE-1, ROUGE-2 and ROUGE-L), BLEU and Perplexity, the most effective model was found for the task, i.e., the generated annotation variants were compared separately with the author's annotation. The results obtained are of practical value, as text summarization is an important task in the field of natural language processing.

Keywords: NLP, summarization, mBART, T5, GPT-3

Reference for citation: Melnichuk D. V., Noskina A. V. Comparison of NLP-models on the Task of Summarizing Academic Texts in Russian Language // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St.
Petersburg: ITMO University, 2024. P. 54–59. DOI: 10.17586/2541-9781-2024-7-54–59

Reference

[1] Gusev I. Dataset for Automatic Summarization of Russian News // Artificial Intelligence and Natural Language. AINL 2020 / Filchenkov A., Kauttonen J., Pivovarova L. (eds).
Communications in Computer and Information Science. Vol. 1292. 2020. P. 122-134.[2]
Lin C. ROUGE: a package for automatic evaluation of summaries // Text Summarization Branches Out / Association for Computational Linguistics. Barcelona. 2004. P. 74–81.
[3] Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation // 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311-318.
[4] Nallapati R., Zhai F., Zhou B. SummaRuNNer: a recurrent neural network-based sequence model for extractive summarization of documents // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligenc. 2017. P. 3075–3081.
[5] Gusev I. Gazeta - Dataset for Automatic Summarization of Russian News // Hugging Face. 2021. URL: https://huggingface.co/datasets/IlyaGusev/gazeta (access date: 17.05.2023).
[6] Gusev I. RuGPT3MediumSumGazeta — Model for abstractive summarization for Russian based on rugpt3medium // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta (access date: 17.05.2023).
[7] Gusev I. RuT5SumGazeta — Model for abstractive summarization for Russian based on rut5-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (access date: 17.05.2023).
[8] Gusev I. MBARTRuSumGazeta — Model for abstractive summarization for Russian based on rumbart-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta (access date: 17.05.2023).