ПРОГРАММНАЯ СИСТЕМА АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТА

Салып Б.Ю.

Салып Богдан Юрьевич– ассистент преподавателя,
кафедра "Компьютерные системы и сети",
Московский Государственный Технический Университет им. Н.Э. Баумана, г. Москва

Аннотация: *в статье описывается разработка и тестирование программной системы автоматического реферирования текста, проводится анализ датасета для обучения, препроцессинг данных и сравнивается корреляция автоматических метрик с человеческими оценками качества реферирования текста.*
Ключевые слова: *NLP (Natural Language Processing), LLM (Large Language Models), обработка текста, суммаризация, нейронные сети, большие данные, обогащение данных.*

AUTOMATIC SOFTWARE SYSTEM FOR TEXT SUMMARIZATION

Salyp B.Yu.

Salyp Bogdan Yurievich– teaching assistant,
COMPUTER SYSTEMS AND NETWORKS DEPARTMENT,
BAUMAN MOSCOW STATE TECHNICAL UNIVERSITY, MOSCOW

Abstract: *the article describes the development and testing process of an automated text referring software, analysis of a training dataset, data preprocessing, automatic metrics versus human metrics correlations are compared.*
Keywords: *NLP (Natural Language Processing), LLM (Large Language Models), text processing, summarization, neural network, big data, data enrichment.*

УДК 004.852

Введение

В данной работе рассматриваются все классические этапы обучения модели обработки естественного языка - выбор датасета и метрик, подготовка данных, обучение модели и тестирование полученной модели на выбранных метриках.

Анализ датасета и выбор метрик

В качестве датасета для обучения модели реферирования текста был использован классический датасет новостей CNN-DailyMail, из которого случайным образом были выбраны 100 предложений на оценку. Датасет содержит созданные человеком абстракты (резюме) новостей на различную тематику. Общий объём датасета - 286817 размеченных новостей.

Далее, на тех же 100 предложениях были рассчитаны 14 автоматических метрик. Каждый из результатов метрик был проанализирован в соответствии с человеческими метриками, и по результатам их значений была построена таблица корреляции между автоматическими и человеческими метриками.

Таблица 1. Результаты корреляции метрик с человеческими оценками.

Метрика	**Согласованность**	**Последовательность**	**Осмысленность**	**Актуальность**
ROUGE-1	0.2500	0.5294	**0.5240**	0.4118
ROUGE-2	0.1618	0.5882	0.4797	0.2941
ROUGE-3	0.2206	**0.7059**	**0.5092**	0.3529
ROUGE-4	**0.3088**	0.5882	**0.5535**	0.4118
ROUGE-L	0.0735	0.1471	0.2583	0.2353
ROUGE-su*	0.1912	0.2941	0.4354	0.3235
ROUGE-w	0.0000	0.3971	0.3764	0.1618
ROUGE-we-1	**0.2647**	0.4559	**0.5092**	**0.4265**

ROUGE-we-2	−0.0147	0.5000	0.3026	0.1176
ROUGE-we-3	0.0294	0.3676	0.3026	0.1912
BertScore-p	0.0588	−0.1912	0.0074	0.1618
BertScore-r	0.1471	**0.6618**	0.4945	0.3088
BertScore-f	0.2059	0.0441	0.2435	**0.4265**
BLEU	0.1176	0.0735	0.3321	0.2206
CHRF	**0.3971**	0.5294	0.4649	**0.5882**
CIDEr	0.1176	−0.1912	−0.0221	0.1912
METEOR	0.2353	**0.6324**	**0.6126**	**0.4265**

В 5 лучших метрик попали как метрики с положительной, так и с отрицательной корреляции, которые одинаково показывают соответствие значения метрики и человеческого восприятия одного из четырёх аспектов качества текста.

Высокая корреляция означает повышенное соответствие между автоматической метрикой и человеческим восприятием, следовательно, такая автоматическая метрика может использоваться, чтобы выражать, например, согласованность текста.

Результаты корреляции показывают несколько тенденций. Было обнаружено, что большинство метрик имеют самую низкую корреляцию в измерении согласованности, где сила корреляции может быть классифицирована как слабая или умеренная. Этот вывод следует из того, что большинство метрик опираются на жесткое или мягкое выравнивание подпоследовательностей, которое плохо измеряет взаимозависимость между последовательными предложениями на более высоком уровне. Низкие и умеренные показатели корреляции были также обнаружены для измерения Актуальность. Как обсуждалось в описании метрик, такие тенденции могут быть обусловлены субъективностью измерения и сложностью сбора результатов с человеческой стороны. Корреляции моделей значительно увеличиваются в измерениях согласованности и беглости. Хотя это и неожиданно, сильная корреляция с последовательностью может быть объяснена низкой абстрактностью большинства нейронных моделей, что может повысить эффективность метрик, использующих перекрытие n-грамм более высокого порядка, таких как ROUGE-3 [1][2] или BLEU [3][4]. Возвращаясь к предыдущему подразделу, оба упомянутых измерения достигли высокого согласия между оценивающими людьми, что также могло положительно сказаться на показателях корреляции. Кроме того, результаты показывают значительно более высокую корреляцию между всеми оцененными измерениями и оценками ROUGE, рассчитанными для н-грамм более высокого порядка, по сравнению с ROUGE-L.

Подготовка данных к обучению

Данные для обучения были разбиты на три группы:

- тренировочный датасет, 80 000 экземпляров;
- валидационный датасет, 10 000 экземпляров;
- тестовый датасет, 10 000 экземпляров.

Тренировочный датасет был использован для дообучения модели. Оценка происходила только на функции ошибки, loss, кросс-энтропии.

Валидационный датасет использовался в процессе обучения для сравнения результатов модели на различных итерациях. По результатам оценок на данной части датасета принимались решения об изменении гиперпараметров в ходе обучения, например, learning_rate.

Тестовый датасет был отделён от остальных двух частей до обучения и в нём не участвовал. Следовательно, данные были для обучаемой модели новыми, и могли использоваться как некое обобщение реальных данных, которые модель будет обрабатывать после обучения и внедрения в программную систему.

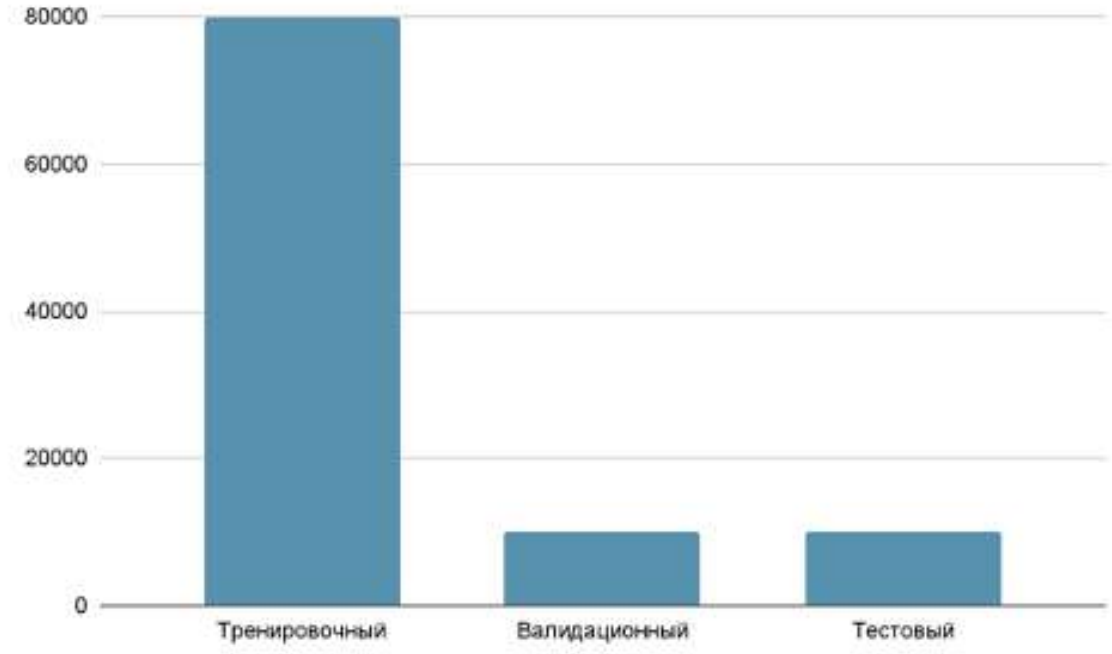

Рис. 1. Визуализация разбиения датасета.

В качестве результата модели использовались результаты метрик только на тестовом датасете. На данном датасете были рассчитаны метрики Meteor [5], BLEU и ROUGE-1/2.

Обучение модели

Обучение производилось на двух видеокартах V100 и заняло около 2 часов.

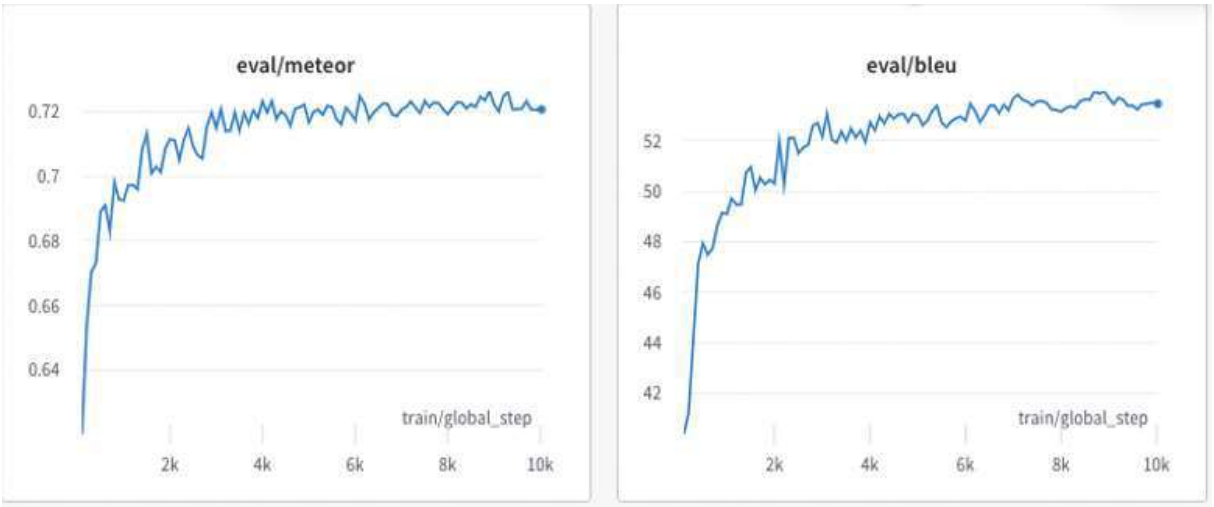

Рис. 2. Результаты обучения на метриках Meteor и BLEU.

Тестирование

Результаты работы программной системы были протестированы на осмысленность, качество сокращения и связность текста. На основе 20 тестовых текстов были сделаны выводы и возможные улучшения обученной модели для программной системы резюмирования текста.

Приведем пример исходного текста:

“Химики из Дании экспериментально подтвердили возможность образования гидротриоксидов — взрывчатых органических окислителей — из пероксидных радикалов в атмосферных условиях. Оказалось, что в атмосфере может образовываться до 10 миллионов тонн гидротриоксидов каждый год. Влияние этих необычных соединений на атмосферные процессы еще предстоит изучить, пишут ученые в Science.

Гидротриоксиды — органические соединения, в которых есть три последовательно связанных друг с другом атома кислорода. Их редко применяют в лабораторной практике: они легко разлагаются со взрывом

и проявляют сильные окислительные свойства. Но химикам известно, что гидротриоксиды могут получаться в результате радикальных реакций между пероксидными (ROO·) и гидроксильными (·OH) радикалами, которые постоянно образуются в атмосфере из кислорода, воды и газообразных углеводородов. При этом за счет своих окислительных свойств гидротриоксиды могут влиять на разные химические процессы в атмосфере, но химики до сих пор не подтвердили экспериментально возможность их образования в атмосферных условиях.

Химики под руководством Хенрика Кьергора (Henrik Kjaergaard) из Копенгагенского университета решили смоделировать образование гидротриоксидов в атмосферных условиях. Они смешали изопропилнитрит — из него под действием света получаются гидроксильные радикалы — с триметиламином при комнатной температуре и атмосферном давлении воздуха в проточном трубчатом реакторе. Ход реакции химики мониторили с помощью масс-спектрометрии."

Результат резюмирования:

"Ежегодно в атмосфере может образовываться до 10 миллионов тонн гидротриоксидов. Влияние этих необычных соединений на атмосферные процессы еще предстоит изучить. Химики смешали изопропилнитрит с триметиламином при комнатной температуре и атмосферном давлении воздуха в трубчатом проточном реакторе."

Результаты резюмирования программной системы были оценены по вышеперечисленным четырём критериям (согласованность, последовательность, беглость, актуальность).

Также на основе метрик программной системы и результатов вышеупомянутого исследования были выявлены особенности резюмирующих систем. "Слабым местом" подобных моделей является актуальность, то есть правильный выбор фактов из исходного текста, которые требуется добавить в резюме, то есть самых важных фактов. При этом большинство моделей, судя по человеческим метрикам, удачно (оценка > 4.5) справляется с соответствием фактов и грамматикой генерируемого текста.

Дополнительно можно указать тот факт, что датасет CNN-DailyMail, на котором обучаются практически все модели резюмирования текста, не является идеальным примером резюмирования текста (оценки не превышают 4.5).

Выводы

В результате человеческой оценки текста было выявлено несколько закономерностей и потенциальных улучшений модели.

Во-первых, модель показала хорошее качество естественности сгенерированных текстов, которые построены грамматически и синтаксически правильно. Сгенерированный текст зачастую сложно отличить от написанного человеком.

Во-вторых, обученная модель склонна к приоритезации предложений из первого или первых абзацев. По сравнению с другими существующими моделями на русском языке (mbart_ru_sum_gazeta, rubert_ext_sum_gazeta, rugpt3medium_sum_gazeta, rut5_base_sum_gazeta) есть прогресс, так как остальные модели чаще выбирают первые предложений как резюме всего текста, однако, обученная модель проигрывает человеческому резюме в плане равномерного выбора предложений из всего текста.

Список литературы/ References

1. *Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan* RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question // arxiv.org [Электронный ресурс]. 2019. URL: https://arxiv.org/pdf/2211.01482.pdf/ (Дата доступа: 26.09.2023).
2. *Kavita Ganesan* ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks // arxiv.org [Электронный ресурс]. 2019. URL: https://arxiv.org/pdf/1803.01937.pdf/ (Дата доступа: 26.09.2023).
3. *Matt Post* A Call for Clarity in Reporting BLEU Scores // arxiv.org [Электронный ресурс]. 2019. URL: https://arxiv.org/pdf/1804.08771.pdf/ (Дата доступа: 26.09.2023).
4. *Hadeel Saadany, Constantin Orasan* BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text // arxiv.org [Электронный ресурс]. 2019. URL: https://arxiv.org/pdf/2109.14250.pdf/ (Дата доступа: 26.09.2023).
5. *Krzystof Wolk, Danijel Korzinek* Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking // arxiv.org [Электронный ресурс]. 2019. URL: https://arxiv.org/pdf/1601.02789.pdf/ (Дата доступа: 26.09.2023).