Суммаризация научных подборок в задаче автоматической генерации обзоров

A Preprint

Дегтев Василий Денисович
МГУ им. М.В. Ломоносова
Москва, Россия
DegtevVD@my.msu.ru

Ищенко Роман Валерьевич
МГУ им. М.В. Ломоносова
Москва, Россия
ischenkorv@my.msu.ru

Abstract

Автоматическая генерация обзора литературы является важной задачей в области автоматической обработки научных публикаций и направлена на создание релевантных и контекстуально уместных цитат для поддержки аргументации в академических текстах. Появление сверхбольших языковых моделей значительно продвинуло исследования в этой области: на сегодняшний день большинство существующих подходов так или иначе основаны на использовании больших языковых моделей (БЯМ). В данной работе исследуется возможность достижения сопоставимого качества генерации обзоров литературы с помощью предварительно обученной языковой модели, имеющей значительно меньшее число параметров, при условии тщательного подбора промптов. Также обсуждаются потенциальные применения генерации цитирующего текста в системах поддержки научного письма, автоматической аннотации научных работ и интеллектуальных помощниках для исследователей.

1 Введение

Качественный обзор литературы является важной частью любой научной работы, обеспечивая контекст и обосновывая актуальность исследования. Однако его подготовка — трудоемкий процесс. В связи с этим автоматическая генерация обзора литературы, направленная на создание релевантных и контекстуально уместных цитат для поддержки аргументации, стала одной из важных задач в области обработки научных публикаций.

На сегодняшний день большинство передовых подходов в генерации текста цитирования опираются на использование Больших Языковых Моделей (БЯМ), число параметров которых может насчитывать десятки или сотни миллиардов.

В данной работе мы исследуем гипотезу о том, что добиться сопоставимого качества генерации цитирующего текста (citation text generation) возможно с помощью языковой модели, имеющей значительно меньшее число параметров. Мы предполагаем, что ключом к достижению этой цели является не экстенсивный рост модели, а интенсивный подход - дообучение и выравнивание путем комбинирования подходов дообучения с учителем (supervised fine-tunning), обучения с подкреплением и тщательного подбора промптов (prompt engineering).

Большинство работ направлено на генерацию текста цитирования отдельных статей [AbuRa'ed et al., 2020, Şahinuç et al., 2024, Gu and Hahnloser, 2023a, Arita et al., 2022], и лишь некоторые статьи предлагают генерацию текста для ссылки сразу на несколько источников [Wu et al., 2021, Anand et al., 2023]. Все работы так или иначе рассматривают генерацию лишь отдельных предложений . Мы же предлагаем инструмент для генерации целых абзацев связного текста цитирований.

2 Обзор литературы

2.1 Подходы к решению

Ранние работы, такие как [Hu and Wan, 2014], использовали тематическое моделирование и классические методы машинного обучения для экстрактивного выделения текста цитирования. Современные же решения основанны на БЯМ с трансформерной архитектурой, которые впервые были применены в [Xing et al., 2020]. Значительный прогресс в качестве генерации цитирующего текста связан с внедрением намерений цитирования (citation intents), предложенные [Wu et al., 2021] с целью повышения качества генерируемых цитат. Этот подход развит в [Jung and Lin, 2022] и [Gu and Hahnloser, 2023a] с использованием обучения с подкреплением (PPO).

Контекстное обогащение стало важным направлением: в [Arita et al., 2022] на вход модели подаются релевантные фрагменты из цитируемых статей, а в [Li et al., 2024] вводится концепция cited text spans — автоматически извлечённых информативных отрывков. Генерация обзоров для нескольких работ одновременно решена в [Anand et al., 2023] с помощью моделей с увеличенным контекстом, способных формировать связные абзацы.

С появлением БЯМ на миллиарды параметров активно применяются техники промптинга: Chain-of-Thought в [Ji et al., 2024], seq2seq-модели с графами знаний в [Anand et al., 2024]. Графовые энкодеры и контрастивное обучение использованы в [Chen et al., 2022] и [Chen et al., 2021] для моделирования связей между документами. Интегрированные пайплайны, такие как SciLit [Gu and Hahnloser, 2023b], сочетают поиск, суммаризацию и генерацию цитирований. Систематическое исследование промптинга и контроля намерений проведено в [Şahinuç et al., 2024].

2.2 Данные

Качество генерации во многом зависит от специализированных датасетов. SciReviewGen [Kasanishi et al., 2023] предоставляет масштабный корпус для генерации обзоров литературы. CORWA [Li et al., 2022] содержит аннотации связанных работ с указанием семантических ролей цитирований. В [Anand et al., 2024] представлен новый датасет с мультидокументными примерами и улучшенными аннотациями. Многие работы используют абстракты или полные тексты из открытых научных репозиториев (arXiv, ACL Anthology), а также предварительно обработанные фрагменты [Li et al., 2024]. В работе [Martin Funkquist, 2023] представлен бенчмарк, основанный на корпусе CORWA.

2.3 Критерия оценки качества

Оценка качества генерации остаётся сложной задачей. Стандартной метрикой является ROUGE [Lin, 2004], измеряющая пересечение n-грамм, однако она недостаточна для оценки семантической точности. Метрика BERTScore [Tianyi Zhang, 2019], основанная на энкодере BERT, позволяет более точно оценить семантическую близость между эталонным текстом и выходами модели. [Şahinuç et al., 2024] использовали вместо обычного BERT модель SciBERT [Iz Beltagy, 2019] для подсчета BERTScore. Работы [Anand et al., 2024],[Gu and Hahnloser, 2023a] предложили собственные метрики, учитывающие точность цитирований и релевантность контента.

3 Постановка задачи

3.1 Дано

Набор данных $\{(A_j, a_{1j}, \ldots, a_{kj}, c_j, G_j), j_{=1}^{N}\}$ состоит из A_j - аннотации цитирующей статьи, $a_{1j}, \ldots, a_{kj}$ - аннотаций цитируемых статей, c_j (citation intent) - цели цитирования, G_j - эталонных текстов цитирований.

3.2 Найти

Обозначим модель за $M = M(\Theta)$, $\Theta = (\theta_1, \ldots, \theta_s)$ - вектор параметров модели, размер словаря $|V|$ и длину выходной последовательности T. Отображение модели M из пространства входов X в пространство предсказанных распределений вероятностей Y:

$$M: \quad X \longrightarrow Y$$

где X – пространство входных данных, а Y – последовательность векторов вероятностей, предсказанных моделью:

$$Y = (y_1, y_2, \ldots, y_T)$$

Каждый элемент y_t в этой последовательности является вектором (распределением) вероятностей по всему словарю V:

$$y_t \in \mathbb{R}^{|V|}$$

3.3 Критерии

В качестве функции потерь будем использовать кросс-энтропию для casual language modeling. Пусть $Y = (y_1, \ldots, y_t), y_t \in \mathbb{R}^{|V|}$ - вероятности для t-ого сгенерированного токена, а $G = (g_1, \ldots, g_T), g_t \in \{0, 1\}^{|V|}$ - эталонный текст цитирования. При этом $g_{ti} = 1$, если токен имеет i-й номер в словаре, и $g_{ti} = 0$ иначе. Определим наш функционал потерь как:

$$\mathcal{L}_{CE} = -\frac{1}{T}\sum_{i=1}^{T} log(y_i) \cdot g_t \tag{1}$$

3.4 Внешние критерии оценки качества

Для оценки качества сгенерированных ответов будем использовать BERTScore [Tianyi Zhang, 2019], аналогично [Şahinuç et al., 2024] в качестве эмебеддера будем использовать SciBERT[Iz Beltagy, 2019]. Пусть:

- $G = (g_1, \ldots, g_t)$ — эталонная последовательность.
- $Y = (y_1, \ldots, y_t)$ — сгенерированная последовательность.
- $\hat{y}_j, \hat{g}_j \in \mathbb{R}^{|V|}$ — векторы эмбеддингов для токенов y_i и g_j.
- $w(t)$ — IDF-вес для токена t.
- $s(\hat{a}, \hat{b}) = \frac{\hat{a} \cdot \hat{b}}{\|\hat{a}\| \|\hat{b}\|}$.

$$R_{BERT} = \frac{\sum_{i=1}^{m} w(y_i) \max_{j=1}^{n} s(\hat{y}_i, \hat{g}_j)}{\sum_{i=1}^{m} w(y_i)} \tag{2}$$

$$P_{BERT} = \frac{\sum_{j=1}^{n} w(g_j) \max_{i=1}^{m} s(\hat{y}_i, \hat{g}_j)}{\sum_{j=1}^{n} w(g_j)} \tag{3}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{4}$$

Список литературы

Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Àlex Bravo. Automatic related work section generation: experiments in scientific document abstracting. 2020.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. Systematic task exploration with llms: A study in citation text generation. 2024.

Nianlong Gu and Richard H. R. Hahnloser. Controllable citation sentence generation with language models. 2023a.

Akito Arita, Hiroaki Sugiyama, Kohji Dohsaka, Rikuto Tanaka, and Hirotoshi Taira. Citation sentence generation leveraging the content of cited papers. 2022.

Jia-Yan Wu, Alexander Te-Wei Shieh, Shih-Ju Hsu, and Yun-Nung Chen. Towards generating citation sentences for multiple references with intent control. 2021.

Avinash Anand, Kritarth Prasad, Ujjwal Goel, Mohit Gupta, Naman Lal, Astha Verma, and Rajiv Ratn Shah. Context-enhanced language models for generating multi-paper citations. 2023.

Yue Hu and Xiaojun Wan. Automatic generation of related work sections in scientific papers: An optimization approach. 2014.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic generation of citation texts in scholarly papers: A pilot study. 2020.

Shing-Yun Jung and Ting-Han Lin. Intent-controllable citation text generation. 2022.

Xiangci Li, Yi-Hui Lee, and Jessica Ouyang. Cited text spans for scientific citation text generation. 2024.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. Chain-of-thought improves text generation with citations in large language models. 2024.

Avinash Anand, Ashwin R. Nair, Kritarth Prasad, Vrinda Narayan, Naman Lal, Debanjan Mahata, Yaman K. Singla, and Rajiv Ratn Shah. Advances in citation text generation: Leveraging multi-source seq2seq models and large language models. 2024.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. Target-aware abstractive related work generation with contrastive learning. 2022.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. Capturing relations between scientific papers: An abstractive model for related work section generation. 2021.

Nianlong Gu and Richard H. R. Hahnloser. Scilit: A platform for joint scientific literature discovery, summarization and citation generation. 2023b.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Scireviewgen: A large-scale dataset for automatic literature review generation. 2023.

Xiangci Li, Biswadip Mandal, and Jessica Ouyang. Corwa: A citation-oriented related work annotation dataset. 2022.

Yufang Hou Iryna Gurevych Martin Funkquist, Ilia Kuznetsov. Citebench: A benchmark for scientific citation text generation. 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. 2004.

Felix Wu Kilian Q. Weinberger Yoav Artzi Tianyi Zhang, Varsha Kishore. Bertscore: Evaluating text generation with bert. 2019.

Arman Cohan Iz Beltagy, Kyle Lo. Scibert: A pretrained language model for scientific text. 2019.