Многозадачное дообучение для генерации ключевых слов к научным текстам

А. В. Глазкова
Тюменский государственный университет
Тюмень, Россия
a.v.glazkova@utmn.ru

Д. А. Морозов
Новосибирский государственный университет
Новосибирск, Россия
morozowdm@gmail.com

***Аннотация*— В работе исследуется эффективность использования заголовков научных текстов в качестве дополнительной информации при обучении модели генерации списка ключевых слов. Описывается подход к многозадачному дообучению (multi-task fine-tuning) модели BART с помощью управляющих кодов. Показано, что данный подход позволяет улучшить качество BART, обученной для решения только одной задачи. Кроме того, в ряде случаев представленная многозадачная модель превосходит другие современные модели извлечения ключевых слов.**

Ключевые слова*— *обработка естественного языка, автоматическое реферирование, генерация текста, научный текст, BART, многозадачное обучение.

1. Введение

Ключевые слова являются важным элементом научной статьи, описывающим ее тематику. Использование ключевых слов позволяет упростить поиск и систематизацию научных текстов, что является актуальной задачей в условиях постоянного увеличения объема информационных ресурсов. Кроме того, грамотно подобранные ключевые слова позитивно сказываются на видимости публикации научному сообществу и, как следствие, количестве ее цитирований [1-2].

В течение последних десятилетий исследователи предложили ряд подходов к автоматическому извлечению ключевых слов. По большей части данные подходы подразумевают выполнение следующих шагов: извлечение слов или фраз из исходного текста, их ранжирование в соответствии с некоторым алгоритмом и выбор N слов или фраз, имеющих наивысший ранг. Хотя существующие подходы к извлечению ключевых слов весьма эффективны, большинство из них обладают следующими ограничениями: 1) значение N задается вручную и определяется пользователем эмпирически; 2) модель извлекает ключевые слова непосредственно из исходного текста и не обладает возможностью генерировать ключевые слова, отсутствующие в тексте в явном виде (гиперонимы, синонимы и так далее). Указанные ограничения могут быть преодолены с помощью современных нейросетевых моделей, в том числе основанных на архитектуре Transformer [3-4]. В частности, в предыдущей работе [5] авторы данного исследования оценили эффективность моделей автоматического реферирования для генерации списка ключевых слов в виде одной строки. Было показано, что модель BART [6], обученная таким образом, демонстрирует высокое качество в сравнении с несколькими традиционными методами извлечения ключевых слов.

Данная работа посвящена оценке эффективности использования заголовков научных текстов в качестве дополнительной информации, подаваемой в модель генерации списка ключевых слов. Поскольку заголовок обычно содержит важную информацию о содержании соответствующего ему текста и часто включает в себя ключевые слова, одновременное обучение модели генерации ключевых слов и заголовков может повысить качество выполнения обеих задач. Кроме того, заголовок является неизменным атрибутом научного текста, а значит, реализация многозадачной модели не требует больших временных затрат на сбор дополнительных данных.

2. Подход к многозадачному дообучению

По аналогии с [7], в данной работе используются управляющие коды (control codes) для дообучения (fine-tuning) модели BART. Подготовка данных и дообучение модели выполняется следующим образом: 1) формируется датасет, включающий в себя пары «текст – заголовок» и «текст – список ключевых слов»; 2) к текстам из пар «текст – заголовок» добавляется управляющий код <|TITLE|>, в то время как к текстам из пар «текст – список ключевых слов» добавляется управляющий код <|KEYPHRASES|>; 3) датасет случайным образом перемешивается; 4) выполняется дообучение предобученной модели BART на датасете с управляющими кодами. Процесс дообучения проиллюстрирован на рисунке 1. В ходе тестирования модели к текстам из тестовой выборки также добавляются соответствующие управляющие коды. Так, для генерации списка ключевых слов к тексту из тестовой выборки добавляется строка “<|KEYPHRASES|>”.

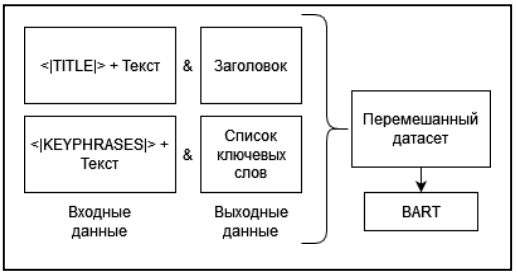

Рис. 1. Процесс дообучения модели

В данной работе дообучение BART выполнялось в течение трех эпох с помощью библиотеки Simple Transformers[1]. Была использована модель BART-base[2] [6] (по шесть слоев в энкодере и декодере, 110 млн

[1] https://simpletransformers.ai/
[2] https://huggingface.co/facebook/bart-base

параметров). Максимальная длина входной последовательности была ограничена 256 токенами.

3. Эксперименты и результаты

Представленный подход был протестирован на следующих текстовых корпусах: 1) Inspec [8], англоязычный корпус, состоящий из 2000 заголовков и аннотаций научных статей, а также соответствующих им ключевых слов; 2) Krapivin2009 [9], корпус, включающий в себя заголовки 2293 англоязычных статей, их полные тексты, аннотации и списки ключевых слов. Корпус Krapivin2009 в ходе экспериментов был разделен на две части: Krapivin-A (в качестве исходного текста для генерации списка ключевых слов использовались аннотации статей) и Krapivin-T (исходным текстом являлся полный текст статьи). Таким образом, исследование проводилось на материале трех коллекций текстов. Полученные результаты сравнены с результатами модели BART, дообученной только на списках ключевых слов, а также ряда моделей для извлечения ключевых слов: 1) TFIDF; 2) TopicRank [10]; 3) KeyBERT [11]; 4) KeyBART [3]. Для реализации моделей использовались библиотеки PKE [12] и Transformers [13]. Качество моделей было оценено с помощью F-меры, рассчитанной на основе количества полностью совпадающих ключевых слов в оригинальном и сгенерированном списке ключевых слов, и BERTScore [14]. Полученные результаты представлены в таблице 1. Наиболее высокие показатели метрик для обоих корпусов выделены полужирным шрифтом.

Таблица I. Результаты (качество генерации ключевых слов)

Модель	F-мера			BERTScore		
	Inspec	***Krapivin-A***	***Krapivin-T***	***Inspec***	***Krapivin-A***	***Krapivin-T***
TFIDF	14,81	10,8	**8,59**	84,28	85,91	85,32
TopicRank	16	7,73	5,72	85,45	86,53	85,95
KeyBERT	11,6	9,46	5,54	84,64	86,43	85,45
KeyBART	12,66	8,74	5,29	87,25	**88,49**	**87,23**
BART (только списки ключевых слов)	14,45	9,19	5,55	**87,72**	87,9	86,59
BART (списки ключевых слов + заголовки)	**17,05**	**11,06**	7,39	87,64	88,22	87,11

Использование многозадачного дообучения в большинстве случаев позволило улучшить качество модели BART для генерации списков ключевых слов к научным статьям. Кроме того, на датасетах Inspec и Krapivin-A представленная модель показала лучшие показатели F-меры среди всех рассмотренных моделей. Самое высокое значение F-меры для Krapivin-T получено с помощью TFIDF. На датасете Inspec BART превзошла KeyBART по значению BERTScore, в то время как KeyBART показала лучшие результаты для Krapivin2009.

Заключение

В работе предложен подход к генерации ключевых слов с помощью модели автоматического реферирования с использованием многозадачного дообучения. Предложенный подход показал высокое качество в сравнении с несколькими существующими моделями извлечения ключевых слов на трех коллекциях научных текстов.

Благодарности

Работа выполнена в рамках проекта №МК-3118.2022.4, поддержанного грантом Президента Российской Федерации для поддержки молодых ученых – кандидатов наук.

Литература

[1] Тихонова, Е.В. Эффективные ключевые слова: стратегии формулирования / Е.В. Тихонова, М.А. Косычева // Health, food & biotechnology. – 2021. – Т. 3, № 4. – С. 7–15. DOI: 10.36107/hfb.2021.i4.s122.

[2] Ghanbarpour, A. A model-based method to improve the quality of ranking in keyword search systems using pseudo-relevance feedback / A. Ghanbarpour, H. Naderi // Journal of Information Science. – 2019. – Vol. 45(4). – P. 473–487. DOI: 10.1177/0165551518799637.

[3] Ye, J. One2Set: Generating Diverse Keyphrases as a Set / J. Ye, T. Gui, Y. Luo, Y. Xu, Q. Zhang // Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 1: Long Papers). – 2021. – P. 4598-4608. DOI: 10.18653/v1/2021.acl-long.354.

[4] Kulkarni, M. Learning Rich Representation of Keyphrases from Text / M. Kulkarni, D. Mahata, R. Arora, R. Bhowmik// Findings of the Association for Computational Linguistics: NAACL. – 2022. – P. 891–906. DOI: 10.18653/v1/2022.findings-naacl.67.

[5] Glazkova, A. Applying Transformer-based Text Summarization for Keyphrase Generation / A. Glazkova, D. Morozov // arXiv preprint. – 2022. DOI: 10.48550/arXiv.2209.03791.

[6] Lewis, M. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension / M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer // Proceedings of the 58th Annual Meeting of the ACL. – 2020. – P. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703.

[7] Cachola, I. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension / I. Cachola, K. Lo, A. Cohan, D. Weld// Findings of the Association for Computational Linguistics: EMNLP. – 2020. – P. 4766–4777. DOI: 10.18653/v1/2020.findings-emnlp.428.

[8] Hulth, A. Improved automatic keyword extraction given more linguistic knowledge / A. Hulth // Proceedings of the 2003 conference on EMNLP. – 2003. – P. 216–223. DOI: 10.3115/1119355.1119383.

[9] Krapivin, M. Large dataset for keyphrases extraction / M. Krapivin, A. Autaeu, M. Marchese // ICADL 2010. – 2010.

[10] Bougouin, A. TopicRank: Graph-based topic ranking for keyphrase extraction / A. Bougouin, F. Boudin, B. Daille // IJCNLP. – 2013. – P. 543–551.

[11] Grootendorst, M. KeyBERT: Minimal keyword extraction with BERT / M. Grootendorst // Zenodo. – 2020. DOI: 10.5281/zenodo.4461265.

[12] Boudin, F. PKE: an open source python-based keyphrase extraction toolkit / F. Boudin // Proceedings of COLING 2016: system demonstrations. – 2016. – P. 69–73.

[13] Wolf, T. Transformers: State-of-the-art natural language processing / T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush // Proceedings of the 2020 conference on EMNLP: system demonstrations. – 2020. – P. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

[14] Zhang, T. BERTScore: Evaluating Text Generation with BERT / T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi //. International Conference on Learning Representations. – 2019.