УДК 004.8+808.2
doi:10.18720/SPBPU/2/id24-203

Медведев Данил Владимирович [1],
студент, бакалавр;
Семенов Константин Константинович [2],
доцент, канд. техн. наук;

СРЕДСТВА ОЦЕНКИ СХОЖЕСТИ РУКОПИСИ НАУЧНОЙ СТАТЬИ И КОРПУСА ПУБЛИКАЦИЙ В КОНКРЕТНОМ НАУЧНОМ ЖУРНАЛЕ

[1, 2] Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого;
[1] medvedev2.dv@spbstu.ru, [2] semenov_kk@spbstu.ru

Аннотация. В данной работе поставлена задача создания инструмента оценки схожести рукописи научной статьи корпусу работ, опубликованных в заданном научном издании. Рассмотрены существующие средства для оценки схожести научных работ. Показана реализуемость подобного интеллектуального помощника автора научных публикаций. Определены области возможного использования.

Ключевые слова: мера сходства текста, обработка естественного языка, искусственный интеллект.

Danil V. Medvedev [1],
Bachelor, Student;
Konstantin K. Semenov [2],
Candidate of Technical Sciences (PhD), Associate Professor

MEANS FOR ASSESSING THE SIMILARITY BETWEEN THE DRAFT OF A SCIENTIFIC ARTICLE AND THE SET OF PUBLICATIONS FROM A PARTICULAR SCIENTIFIC JOURNAL

[1, 2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia;
[1] medvedev2.dv@spbstu.ru, [2] semenov_kk@spbstu.ru

Abstract. This paper sets the problem of creating a tool for assessing the similarity of a scientific article manuscript to a set of articles published in a given scientific journal. We considered the existing tools for the similarity assessing of scientific works. The paper shows the feasibility of such an intellectual assistant for the scientific publications' author. We identified the areas of possible use.

Keywords: text similarity measure, natural language processing, artificial intelligence.

Введение

Рецензирование отдельной научной статьи является трудоемким процессом, выполняющимся экспертами из соответствующей научной области. Несмотря на сформированный общественный консенсус в отношении требований научной этики, многие исследователи процесса публикации научных статей отмечают, что процессу рецензирования зачастую присущи такие черты как предвзятость, конфликт интересов, противоречивость оценки рецензентами, недобросовестное поведение и т. п. [1], что в свою очередь увеличивает вероятность отклонения статьи для публикации в конкретном журнале. Несомненным является существенное проявление в публикационном процессе эффекта Матфея, указывающего на значимое смещение в результатах процесса рецензирования в зависимости от личностей авторов поданной в научной журнал статьи [2, 3]. Необходимость преодоления указанных обстоятельств для большинства авторов приводит к повышению требований к качеству самих научных статей перед подачей их в авторитетные научные издания. Вместе с тем в мировой науке наблюдается устойчивый значительный рост общего числа статей, направляемых в научные издания, при меньших темпах роста числа самих ученых: с 2018 по 2022 гг. количество публикаций увеличилось почти на 23 % и составило 5,14 млн. [4], притом что количество ученых увеличилось по сравнению с 2007 г. приблизительно на те же 20 % — до 7,8 миллиона человек [5]. Совокупность данных факторов приводит к увеличению времени прохождения публикации через весь цикл рецензирования (в том числе многократного – при повторных подачах статьи) — порой до 1–2 лет, — а также к дисбалансу между потоком поступающих статей и возможностями изданий для их качественной оценки. В таких условиях цена ошибки для авторов научных статей, связанная с неточным выбором издания для подачи работы, может оказаться непомерно дорогой. В этой связи особую значимость приобретают интеллектуальные помощники для ученых, которые позволяют автоматизировать часть публикационного процесса и повысить шансы успешного прохождения этапа рецензирования. Одним из них может стать инструмент оценки схожести подготовленной рукописи научной статьи и корпуса публикаций в конкретном научном журнале, который позволяет автору косвенно оценить шансы рассмотрения его работы (и соответственно вероятность того, что редакционная коллегия научного издания не отклонит его публикацию по формальным признакам). Текущее уровень развития технологий искусственного интеллекта делают создание такого помощника вполне возможным. В настоящей работе обсуждаются вопросы, связанные со структурой программного приложения, реализующего функции описанного инструмента, и набо-

ром существующих нейронных сетей открытого доступа, которые могут быть положены в его основу.

1. Формализация процесса рассмотрения поданной статьи научным изданием

Похожими в контексте рассматриваемой задачи следует считать две статьи, которые схожи по одному из ключевых показателей или характеристик: например, охватывают одну и ту же тему, используют общий набор ключевых терминов и оборотов или написаны одним и тем же шрифтом [6]. Если исходить из постулата о том, что вероятность принятия статьи, которая схожа с одной из ранее опубликованных в рассматриваемой издании, больше, чем у менее схожей, то оценка схожести научных публикаций становится ключевой задачей для интеллектуального помощника автора. Высокая степень схожести будет указывать на повышенные шансы прохождения этапа редакционного отбора в рамках процедуры предварительного рассмотрения, который обычно состоит из: проверки соответствия требуемому формату, обнаружения плагиата, обнаружения машинной генерации текста или данных, определения тематической области и определения типа публикации. Данные процедуры в известной степени формализуемы и, следовательно, могут и должны приниматься автором во внимание при подготовке рукописи с целью повышения шансов на ее рецензирование.

2. Средства для оценки схожести научных статей

В научной литературе, посвященной сравнению больших текстов, отмечены следующие основные типы сходства научных публикаций, которые могут быть использованы в контексте поставленной задачи:

1) *Структурное сходство*. Данный тип оценки свидетельствует о сходстве статей визуально и по составу текстовых сегментов и позволяет оценить возможность отклонения по причине несоответствия формату.

2) *Соответствие тематической области*. Одной из самых частых причин отказа автору издательством в рассмотрении поданной рукописи является ее несоответствие научным областям, охватываемым журналом, или невозможность подбора рецензентов по той же причине. На основе анализа корпуса публикаций, изданных конкретным изданием, можно очертить круг основных научных вопросов публикуемых исследований и сравнить с ними тематику подготовленной рукописи. Основным методом для сравнения тематических областей является выделение корпусов ключевых слов с последующим их сравнением с использованием инструментов Rake [7], Yake [8], TF-IDF, TextRank [9] и ряда других.

3) *Семантическое сходство*. Два документа считаются семантически похожими, если они охватывают связанные темы или имеют одинаковое семантическое значение. Наличие семантического сходства может

в некоторой степени косвенно свидетельствовать и о повышенной вероятности положительной экспертной оценки при рецензировании. Основными инструментами получения данной оценки сходства выступают сиамские нейронные сети [10], BERT [11] модели, методы Word2Vec [12] и другие.

Каждый из перечисленных основных типов сходства научных публикаций должен быть использован при оценке того, насколько сильно поданная рукопись статьи соответствует уже опубликованным в данном издании. Итоговая вероятность p принятия статьи к рассмотрению и допуску к рецензированию складывается из сочетания вышеописанных оценок как средневзвешенное

$$p = \omega_1 \cdot s_1 + \omega_2 \cdot s_2 + \omega_3 \cdot s_3,$$

где s_1, s_2 и s_3 — оценки структурного, тематического и семантического сходства, чьи значения отложены на шкале от 0 до 1, а ω_1, ω_2 и ω_3 – соответствующие неотрицательные веса, определяемые для каждого журнала в отдельности по распределению оценок схожести статей в корпусе опубликованных им научных статей, такие, что $\omega_1+\omega_2+\omega_3 = 1$.

Для получения оценок s_1, s_2 и s_3 в качестве предварительной процедуры должна осуществляться декомпозиция файла рукописи научной статьи и статей, опубликованных рассматриваемым изданием, на элементы. Последнее необходимо для обучения нейросетевых моделей, осуществляющих оценку того или иного типа схожести (для каждого издания или даже рубрики журнала в отдельности). Для рассматриваемой задачи затруднительно использование существующих наборов данных с информацией о научных публикациях, находящихся в открытом доступе – в частности, известные наборы данных PeerRead [13] и AAPR [14] не содержат информацию о таких структурных элементах научных статей как содержащиеся в них изображения, таблицы и формулы.

Заключение

В данной статье представлена постановка задачи создания инструмента оценки схожести рукописи научной статьи корпусу работ, опубликованных в заданной научном издании. Рассмотрены основные средства, необходимые для построения интеллектуального помощника автора, реализующего функционал инструмента указанной оценки. Возможными областями применения подобного программного средства могли бы быть:

– самостоятельная оценка схожести рукописи автором для указанного им издания,

– автоматический подбор рекомендуемого издания для подачи созданной рукописи путем максимизации метрик ее схожести со статьями, вышед-

шими в научных изданиях, для которых произведено обучение нейросетевых моделей,
– автоматическая оценка издательством соответствия поданной на рассмотрение рукописи предъявляемым им формальным требованиям.

Создание рассмотренного инструмента позволит авторам научных статей повысить качество создаваемых научных рукописей за счет возникновения обратной связи: предоставление автору в режиме реального времени значений интегральной метрики соответствия текущей версии рукописи указанному журналу позволит понять, как именно сказываются те или иные вносимые им изменения или дополнения на вероятности принятия рукописи издательством к рассмотрению и подобрать наиболее удачный их вариант для максимизации метрики схожести. Эта возможность обеспечивает как обучение автора, так и повышает качество рукописи.

Список литературы

1. Тамбовцев В.Л. Рецензирование в современных научных коммуникациях // Управление наукой: теория и практика. – 2021. – Т. 3. № 1. – С. 35–54. – DOI: 10.19181/smtp.2021.3.1.2.

2. Serenko A., Cox R., Bontis N., Booker L. The superstar phenomenon in the Knowledge management and intellectual capital academic discipline // Journal of Informetrics. – 2011. – Vol. 5. – Pp. 333–345. – DOI: 10.1016/j.joi.2011.01.005.

3. Москалева О.В. Научные публикации как средство коммуникации, анализа и оценки научной деятельности // Руководство по наукометрии: индикаторы развития науки и технологии. – Издательство Уральского университета, 2014. – С. 110–163.

4. Number of Academic Papers Published Per Year [Electronic Source] // WordsRated. – URL: https://wordsrated.com/number-of-academic-papers-published-per-year/ (date of access: 29.01.2024).

5. Доклад ЮНЕСКО: Наука в авангарде всемирного движения к устойчивому росту [Электронная страница]. – URL: https://www.unesco.org/ru/articles/doklad-yunesko-nauka-v-avangarde-vsemirnogo-dvizheniya-k-ustoychivomu-rostu (дата обращения: 15.01.2024).

6. Lin D. An information-theoretic definition of similarity // ICML'98: Proceedings of the Fifteenth International Conference on Machine Learning. – 1998. – Pp. 296–304. – DOI: 10.5555/645527.657297.

7. Bottomley G.E., Ottosson T., Wang Y.P.E. A generalized RAKE receiver for interference suppression // IEEE Journal on selected areas in Communications. – 2000. – Vol. 18. No. 8. – Pp. 1536–1545. – DOI: 10.1109/49.864017.

8. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. – 2020. – Vol. 509. – Pp. 257–289. – DOI: 10.1016/j.ins.2019.09.013.

9. Mihalcea R., Tarau P. Textrank: Bringing order into text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. July 2004, Barcelona, Spain. – 2004. – Pp. 404–411. – URL: https://aclanthology.org/W04-3252 (date of access: 29.01.2024).

10. Bromley J., Guyon I., LeCun Y., Säckinger E., Shah R. Signature verification using a “siamese” time delay neural network // NIPS’93: Proceedings of the 6th International Conference on Neural Information Processing Systems. – San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. – Pp. 737–744. – DOI: 10.5555/2987189.2987282.

11. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // NAACL-HLT. – 2018. – Pp. 4171–4186. – DOI: 10.18653/V1/N19-1423.

12. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. – 2013. – ArXiv preprint. ArXiv:1301.3781.

13. Kang D., Ammar W., Dalvi B., Van Zuylen M., Kohlmeier S., Hovy E., Schwartz R. A dataset of peer reviews (peerread): Collection, insights and nlp applications // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2018. – Vol. 1. – Pp. 1647–1661. – DOI: 10.18653/v1/N18-1149.

14. Yang P., Sun X., Li W., Ma S. Automatic academic paper rating based on modularized hierarchical convolutional neural network // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – 2018. – Vol. 2. – Pp. 496–502. – DOI: 10.18653/v1/P18-2079.