DOI: 10.18137/RNU.V9187.24.03.P.144 УДК 004.912

Э.А. Челышев, М.В. Раскатова, П. Щёголев

ВНУТРЕННИЕ МЕТОДЫ ОЦЕНКИ ИНФОРМАЦИОННОЙ ПОЛНОТЫ РЕФЕРАТОВ В ЗАДАЧЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Аннотация. В статье представлена классификация существующих методов оценки качества рефератов в задаче автоматического реферирования текстов. Представлена формальная постановка задачи автоматического реферирования текстов. Рассмотрено понятие качества реферата. Подробно рассмотрены внутренние методы оценки информационной полноты реферата, такие как метрики группы ROUGE, косинусное сходство, расстояние Кульбака – Лейблера, расстояние Дженсена – Шеннона. Представлены достоинства и недостатки рассмотренных внутренних методов и методов, использующих экспертную оценку.

Ключевые слова: автоматическое реферирование, ROUGE, косинусное сходство, расстояние Кульбака – Лейблера, расстояние Дженсена – Шеннона, векторизация.

E.A. Chelyshev, M.V. Raskatova, P. Shchegolev

INTRINSIC METHODS FOR EVALUATING INFORMATION COMPLETENESS OF SUMMARIES IN TASK OF AUTOMATIC TEXT SUMMARIZATION

Abstract. The article presents a classification of existing methods for assessing the quality of summaries in the task of automatic text summarization. A formal formulation of the problem of automatic text summarization is presented. The concept of summary quality is considered. The intrinsic methods of evaluating the information completeness of the summary, such as the metrics of the ROUGE group, cosine similarity, Kullback-Leibler distance, Jensen-Shannon distance, are considered in detail. The advantages and disadvantages of the considered intrinsic methods and methods based on expert assessment are presented.

Keywords: automatic summarization, ROUGE, cosine similarity, Kullback-Leibler distance, Jensen-Shannon distance, vectorization.

Введение

Автоматическое реферирование текстов (англ. automatic text summarization) – процесс получения из исходного текста его реферата (текста меньшего объема, содержащего основные сведения исходного) с использованием программных средств без участия человека в формировании текста реферата.

Формально задачу автоматического реферирования некоторого текста T можно представить как задачу нахождения отображения Ψ, обладающего следующими свойствами:

$$\Psi : T \to \tilde{T}, |\tilde{T}| \ll |T|, |\tilde{T}| \leq N_{max}, \quad (1)$$

где Ψ – некоторое отображение; фактически оно определяется методом автоматического реферирования;

T – исходный текст;

$\tilde{T}$ – текст реферата;

Челышев Эдуард Артурович
аспирант кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: машинное обучение по прецедентам, искусственные нейронные сети, машинная обработка текстов на естественном языке, язык программирования C++. Автор более 20 опубликованных научных работ. ORCID: 0000-0001-8417-8823, AuthorID: 1088395, SPIN-код: 5357-7604.
Электронный адрес: chel.ed@yandex.ru

Раскатова Марина Викторовна
кандидат технических наук, доцент кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: разработка программного обеспечения, информационные системы. Автор более 40 опубликованных научных работ. ORCID: 0000-0001-7671-3312, AuthorID: 609945, SPIN-код: 8053-5041.
Электронный адрес: marina@raskatova.ru

Щёголев Павел
старший преподаватель кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: языки и методы программирования, Web-разработка. Автор семи опубликованных научных работ. ORCID: 0000-0001-9954-8858, AuthorID: 1246900, SPIN-код: 6914-1637.
Электронный адрес: Shchegolevsp@mpei.ru

N_{max} – максимально возможная длина реферата.

Исследования в области автоматического реферирования текстов имеют давнюю историю и продолжаются по сей день. Так, например, первой работой, посвященной автоматическому реферированию текстов, является статья Х.П. Луна [1], опубликованная в 1958 году. В рассматриваемой работе был предложен эвристический метод, имеющий простую программную реализацию. В настоящее время для выполнения автоматического реферирования используются языковые модели, представляющие из себя искусственные нейронные сети сложной архитектуры [2].

Таким образом, можно констатировать, что методы и алгоритмы автоматического реферирования текстов имеют богатую историю. Однако на протяжении всей истории их существования стоял вопрос оценки качества получаемых таким способом рефератов.

Качество реферата является комплексным понятием и содержит целый набор критериев, среди которых можно отметить следующие:

- степень сжатия реферата относительно исходного текста;
- информационная полнота реферата (то есть то, насколько полно реферат передает содержание исходного текста);
- связность и структурированность текста реферата;
- его логическая и стилистическая целостность [3].

По степени привлеченности эксперта методы оценки качества реферата можно разделить на автоматические, полуавтоматические и методы, использующие исключительно экспертную оценку [4]. Практика показывает, что именно последний класс методов дает наилучшие результаты. Однако методы, использующие экспертную оценку, являются трудозатратными и занимают существенное время. Кроме того, они не подходят для малых

исследовательских групп.

По методологии решения задачи методы оценки качества рефератов могут быть разделены на два класса: внутренние (англ. intrinsic) и внешние (англ. extrinsic). *Внешние методы* оценки качества реферата рассматривают, насколько реферат способен помочь решить некоторую внешнюю по отношению к реферированию задачу. Примером такой внешней задачи является поиск ответов на вопросы касательно исходного текста.

Внешние методы оценки требуют привлечения экспертов. Так, например, группе экспертов можно предложить оценить тематическую принадлежность реферата, его логическую и стилистическую целостность, а также дать ответ на некоторые вопросы, опираясь на приведенные в реферате сведения. Таким образом, хоть внешние методы являются более совершенными, чем внутренние, все же они имеют существенный недостаток, а именно необходимость привлечения экспертов.

Внутренние методы оценки нацелены исключительно на сравнение реферата либо с исходным текстом, либо с другим рефератом, называемым опорным, принимаемым за образец [5]. Фактически при помощи внутренних методов возможно оценить только один критерий качества реферата, а именно его информационную полноту. Несмотря на свое несовершенство, внутренние методы оценки имеют существенное достоинство: они не требуют привлечения экспертов и могут быть осуществлены в автоматическом режиме. Это делает их более доступными и простыми в реализации [6].

В данной работе рассмотрены три внутренних метода оценки информационной полноты реферата: группа метрик ROUGE, косинусное сходство и расстояние Дженсена – Шеннона.

Группа метрик ROUGE (англ. Recall-Oriented Understudy for Gisting Evaluation) была впервые предложена для оценки качества машинного перевода, однако оказалась полезной и в автоматическом реферировании текстов. Данная группа метрик показывает высокую корреляцию с экспертной оценкой [7]. Метрики данной группы нацелены на сравнение оцениваемого реферата с опорным; использование их для сравнения реферата с исходным текстом не несет никакой полезной информации в силу особенности построения данных метрик. Вообще говоря, группа метрик ROUGE содержит несколько подгрупп метрик, в данной работе мы рассмотрим две подгруппы: ROUGE-N и ROUGE-L. Каждая из данных групп содержит в себе две метрики: точность (англ. precision) и полнота (англ. recall).

Назовем *n*-граммой упорядоченную непрерывную последовательность термов длиной *n*. Метрики ROUGE-N дают численную оценку отношения доли *n*-грамм, присутствующих как в оцениваемом, так и опорном рефератах, к количеству *n*-грамм в отдельном тексте [8]. Если обозначить опорный реферат R, а оцениваемый E, а количество *n*-грамм в них $K_n(R)$ и $K_m(E)$ соответственно, то точность и полноту ROUGE-N можно представить в виде формул (2) и (3) соответственно.

$$R^N_{precision} = \frac{\left|\{gram_n : gram_n \in E \text{ и } gram_n \in R\}\right|}{K_n(E)} \quad (2)$$

$$R^N_{recall} = \frac{\left|\{gram_n : gram_n \in E \text{ и } gram_n \in R\}\right|}{K_n(R)} \quad (3)$$

где $\left|\{gram_n : gram_n \in E \text{ и } gram_n \in R\}\right|$ – количество общих *n*-грамм в оцениваемом и опорном рефератах.

Если обозначить длину наибольшей общей подпоследовательности термов оцениваемого и опорного рефератов как $|LCS\ (E, R)|$, то точность и полноту ROUGE-L можно записать с использованием формул (4) и (5) соответственно:

$$R^{L}_{precision} = \frac{|LCS(E, R)|}{K_1(E)} \tag{4}$$

$$R^{L}_{recall} = \frac{|LCS(E, R)|}{K_1(R)} \tag{5}$$

Несомненными достоинствами метрик группы ROUGE являются простота их использования и невысокая вычислительная сложность. Однако данная группа метрик имеет существенный недостаток: для их использования необходимо наличие опорного реферата. Применение данной группы метрик для сравнения реферата с исходным текстом является неправильным, так как в таком случае все *n*-граммы оцениваемого реферата окажутся общими для обоих текстов, а значение метрики полноты окажется равным единице. При этом метрика точности окажется численно равна отношению длины оцениваемого реферата к длине исходного текста. Кроме того, группа метрик ROUGE не учитывает синонимию, что может привести к заниженной оценке для качественного реферата.

Такие метрики, как косинусное сходство и расстояние Дженсена – Шеннона, не требуют опорного реферата и могут быть успешно применены для сравнения реферата и исходного текста напрямую. Однако важно понимать, что данные метрики оценивают схожесть числовых векторов, поэтому предварительно необходимо выполнить векторизацию текстов. Для векторизации могут применяться различные методы, среди которых можно выделить частотную векторизацию, TF-IDF и статические модели векторизации [9]. При этом наибольший интерес представляют именно два последних метода, так как первый из них позволяет назначать большие веса тем термам, которые являются ключевыми, а второй учитывает семантическую, то есть смысловую близость термов [10; 11].

Косинусное сходство (косинусное подобие, англ. cosine similarity) численно оценивает близость двух векторов в некотором векторном пространстве.

Предположим, что имеются два вектора $\vec{a}$ и $\vec{b}$, являющихся векторными представлениями текста оцениваемого реферата и исходного текста. Тогда косинусное расстояние может быть вычислено по формуле (6) как отношение скалярного произведения данных векторов к произведению их мер (в случае Евклидова пространства – их длин). Чем ближе значение векторного сходства к единице, тем ближе друг к другу расположены два вектора в векторном пространстве [12].

$$CS(\vec{a}, \vec{b}) = \cos\varphi = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\| \cdot \|\vec{b}\|} \tag{6}$$

Еще одной метрикой информационной полноты является ***расстояние Дженсена – Шеннона***. Данная метрика основана на расстоянии Кульбака – Лейблера (относительной энтропии).

Расстояние Кульбака – Лейблера $D_{KL}(P||Q)$ между двумя вероятностными распределениями P и Q численно характеризует удаленность данных вероятностных распределений. Расстояние Кульбака – Лейблера не является симметричным, так как первый его аргумент интерпретируется как постулируемое априори (то есть истинное) вероятностное распределение, а второй – как оцениваемое на близость к первому вероятностное рас-

пределение. Иными словами, можно рассматривать расстояние Кульбака – Лейблера как количественную меру информации, потерянной при замене постулируемого априори вероятностного распределения другим вероятностным распределением [13].

Если рассматривать два дискретных вероятностных распределения, то расстояние Кульбака – Лейблера между ними можно вычислить, используя формулу (4).

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(x_i) \log \frac{P(x_i)}{Q(x_i)} \tag{7}$$

Хотя при помощи расстояния Кульбака – Лейблера возможно оценить сходство двух вероятностных распределений, оно не является метрикой в пространстве вероятностных распределений в силу своей антисимметричности. Расстояние Дженсена – Шеннона, определяемое при помощи формулы (8), лишено данного недостатка.

$$D_{JS}(P \| Q) = \frac{1}{2} D_{KL}(P \| M) + \frac{1}{2} D_{KL}(Q \| M) \tag{8}$$

где M – вероятностное распределение, вычисляемое в соответствии с формулой (9).

$$M(x_i) = \frac{P(x_i) + Q(x_i)}{2}, i = 1 \ldots n \tag{9}$$

При этом чем более схожи два вероятностных распределения, тем ближе значение расстояния Дженсена – Шеннона к нулю. Учитывая, что прочие метрики, рассмотренные в данной работе, трактуются иначе (чем ближе их значение к единице, тем более схожими являются сравниваемые объекты), представляется полезным ввести метрику, производную от расстояния Дженсена – Шеннона, определяемую по формуле (10) [14].

$$\overline{D}_{JS}(P \| Q) = 1 - D_{JS}(P \| Q) \tag{10}$$

Если определить все возможные термы словаря, составленного из всех термов двух сравниваемых текстов, как все возможные исходы, а в качестве вероятностей данных исходов использовать частоту термов в каждом из текстов, то в таком случае расстояния Кульбака – Лейблера и Дженсена – Шеннона окажутся пригодными для сравнения текстов, а следовательно, и для оценки реферата. Существует ряд работ, использующих расстояние Дженсена – Шеннона для оценки информационной близости текстов. В частности, в работе [15] представлено применение расстояния Дженсена – Шеннона для задачи оценки информационной полноты реферата. В работе также отмечается высокая корреляция между результатами оценки информационной полноты реферата с использованием расстояния Дженсена – Шеннона и оценкой, полученной при помощи группы метрик ROUGE.

В заключение хочется отметить, что внутренние методы оценки информационной полноты рефератов в задаче автоматического реферирования текстов, такие как группа метрик ROUGE, косинусное сходство и расстояние Дженсена-Шеннона, являются крайне полезными, так как позволяют проводить такую оценку в автоматическом режиме. При этом все же стоит помнить, что рассматриваемые методы не дают столь качественного результата, как экспертная оценка. Однако существенная ресурсозатратность привлечения экспертов делает внутренние методы единственным доступным способом оценки информационной полноты рефератов в условиях ограниченных ресурсов.

Литература

1. *Luhn H.* The automatic creation of literature abstracts // IBM Journal of Research and Development. New York, 1958. Vol. 2 (2). P. 159–165. DOI: 10.1147/rd.22.0159

2. *Бабуркин Э.В., Нестругина Е.С.* Разработка языковых моделей для различных жанров текста и языка текста на основе глубокого обучения в системе автоматического реферирования текста // Донецкие чтения 2022: образование, наука, инновации, культура и вызовы современности : Материалы VII Международной научной конференции, посвящённой 85-летию Донецкого национального университета, Донецк, 27–28 октября 2022 г. / Под общ. ред. С.В. Беспаловой. Т. 2. Донецк : Донецкий национальный университет, 2022. С. 224–226. EDN MFEGTH.

3. *Челышев Э.А., Раскатова М.В., Мишин А.А., Щёголев П.В.* Автоматическое реферирование текстов: обзор алгоритмов и подходов к оценке качества // Инженерный вестник Дона. 2023. № 12 (108). С. 11–23. EDN NKOZOS.

4. *Батура Т.В., Бакиева А.М.* Методы и системы автоматического реферирования текстов. Новосибирск : Новосибирский национальный исследовательский государственный университет, 2019. 110 с. ISBN 978-5-4437-0974-1.

5. *Aries A., Zegour D.E., Hidouci W.-K.* Automatic text summarization: What has been done and what has to be done // arXiv. 2019. April. URL: https://arxiv.org/abs/1904.00688 (дата обращения: 11.08.2024).

6. *Louis A., Nenkova A.* Automatic Summary Evaluation without Human Models // Theory and Applications of Categories. 2008. URL: https://tac.nist.gov/publications/2008/additional.papers/Penn.proceedings.pdf (дата обращения: 11.08.2024).

7. *Lin C.-Y.* ROUGE: A package for automatic evaluation of summaries // Proceedings of ACL Text Summarization Branches Out Workshop. Barcelona, Spain, 2004. P. 74–81. URL: https://typeset.io/papers/rouge-a-package-for-automatic-evaluation-of-summaries-2tymbd14i8?ysclid=m0bkmbvfwi294060368 (дата обращения: 11.08.2024).

8. *Aliguliyev R. M.* Using the F-measure as similarity measure for automatictext summarization // Computational Technologies. 2008. Vol. 13. No. 3. P. 5–14. EDN KMKOFD.

9. *Раскатова М.В., Челышев Э.А.* Векторизация текстов в задачах обработки естественного языка: история и развитие // Современное программирование : Материалы IV Международной научно-практической конференции, Нижневартовск, 08 декабря 2021 г. / Под общ. ред. Т.Б. Казиахмедова. Нижневартовск : Нижневартовский государственный университет, 2022. С. 284–288. EDN BZQQVZ. DOI: 10.36906/AP-2022/47

10. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* (2013). Distributed Representations of Words and Phrases and their compositionality // NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013. Vol. 2. Pp. 3111–3119. URL: https://dl.acm.org/doi/abs/10.5555/2999792.2999959?cookieSet=1 (дата обращения: 11.08.2024).

11. *Савченко Т.Ю.* Обработка естественного языка для использования в машинном обучении: частотная векторизация, TF-IDF, word2vec // Аллея науки. 2018. Т. 4. № 6 (22). С. 1000–1002. EDN UVDSCA.

12. *Manning C.D., Raghavan P., Schütze H.* Introduction to Information Retrieval. Cambridge, England : Cambridge University Press, 2008. 482 p. ISBN 0521865719.

13. *Брюховецкий А.А.* Модель обнаружения аномальных данных на основе информационного критерия // Дневник науки. 2021. № 4 (52). EDN UOQDOY.

14. *Челышев Э.А., Раскатова М.В., Маковец А.С.* Сравнительный анализ алгоритмов автоматического квазиреферирования текстов // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2023. № 4. С. 176–184. EDN YQIJFY. DOI: 10.18137/RNU.V9187.23.04.P.176

15. *Lin C.-Y., Cao G., Gao J., Nie J.-Y.* An Information-Theoretic Approach to Automatic Evaluation of Summaries // Moore R.C., Bilmes J., Chu-Carroll J., Sanderson M. (Eds) Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. New York City, USA, 2006. P. 463–470. URL: https://aclanthology.org/N06-1059 (дата обращения: 11.08.2024).

References

1. Luhn H. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*. New York. Vol. 2 (2). Pp. 159–165. DOI: 10.1147/rd.22.0159

2. Baburkin E.V., Nestrugina E.S. (2022) Development of language models for different text genres and text language based on in-depth learning in the automatic text referencing system. In: Bespalova S.V. (Ed) *Donetskie chteniya 2022: obrazovanie, nauka, innovatsii, kul'tura i vyzovy sovremennosti* [Donetsk readings 2022: Education, science, innovation, culture and challenges of modernity] : Proceedings of the VII International Scientific Conference on the 85th anniversary of the Donetsk National University, Donetsk, October 27–28, 2022. Vol. 2. Donetsk : Donetsk National University Publ. Pp. 224–226. (In Russian).

3. Chelyshev E.A., Raskatova M.V., Mishin A.A., Shchegolev P.V. (2023) Automatic text summarization: Overview of algorithms and approaches to quality assessment. *Engineering journal of Don.* No. 12 (108). Pp. 11–23. (In Russian).

4. Batura T.V., Bakieva A.M. (2019) *Metody i sistemy avtomaticheskogo referirovaniya tekstov* [Automatic text referencing methods and systems]. Novosibirsk : Novosibirsk National Research State University Publ. 110 p. ISBN 978-5-4437-0974-1. (In Russian).

5. Aries A., Zegour D.E., Hidouci W.-K. (2019) Automatic text summarization: What has been done and what has to be done. *arXiv*. April. URL: https://arxiv.org/abs/1904.00688 (accessed 11.08.2024).

6. Louis A., Nenkova A. (2008) Automatic Summary Evaluation without Human Models. *Theory and Applications of Categories*. URL: https://tac.nist.gov/publications/2008/additional.papers/Penn.proceedings.pdf (accessed 11.08.2024).

7. Lin C.-Y. (2004) ROUGE: A package for automatic evaluation of summaries. In: *Proceedings of ACL Text Summarization Branches Out Workshop*. Barcelona, Spain. Pp. 74–81. URL: https://typeset.io/papers/rouge-a-package-for-automatic-evaluation-of-summaries-2tymbd14i8?ysclid=m0bkmbvfwi294060368 (accessed 11.08.2024).

8. Aliguliyev R.M. (2008) Using the F-measure as similarity measure for automatic text summarization. *Computational Technologies*. Vol. 13. No. 3. Pp. 5–14. URL: https://elibrary.ru/KMKOFD (accessed 11.08.2024).

9. Raskatova M.V., Chelyshev E.A. (2022) Vectorization of texts in natural language processing tasks: history and development. In: Kaziakhmedova T.B. (Ed) *Sovremennoe programmirovanie* [Modern programming] : Proceedings of the IV International Scientific and Practical Conference, Nizhnevartovsk, December 08, 2021. Nizhnevartovsk : Nizhnevartovsk State University Publ. Pp. 284–288. DOI: 10.36906/AP-2022/47 (In Russian).

10. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013) Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13: Proceedings of the 26th International Conference on*

Neural Information Processing Systems. Vol. 2. Pp. 3111–3119. URL: https://dl.acm.org/doi/abs/10.5555/2999792.2999959?cookieSet=1 (accessed 11.08.2024).

11. Savchenko T.Yu. (2018) Processing of natural language for use in machine learning: frequency vectorization, TF-IDF, word2vec. *Alleya nauki*. Vol. 4. No. 6 (22). Pp. 1000–1002. (In Russian).

12. Manning C.D., Raghavan P., Schütze H. (2008) *Introduction to Information Retrieval*. Cambridge, England : Cambridge University Press. 482 p. ISBN 0521865719.

13. Bryukhovetskii A.A. (2021) Anomalous data detection model based on information criterion. *Dnevnik nauki*. No. 4 (52). URL: https://elibrary.ru/UOQDOY (accessed 11.08.2024). (In Russian).

14. Chelyshev E.A., Raskatova M.V., Makovets A.S. (2023) Comparative analysis of automatic text quasi-summarization algorithms. *Vestnik of the Russian New University. Series: Complex Systems: Models, analysis and management*. No. 4. Pp. 176–184. DOI: 10.18137/RNU.V9187.23.04.P.176 (In Russian).

15. Lin C.-Y., Cao G., Gao J., Nie J.-Y. (2006) An Information-Theoretic Approach to Automatic Evaluation of Summaries. In: Moore R.C., Bilmes J., Chu-Carroll J., Sanderson M. (Eds) *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA. Pp. 463–470. URL: https://aclanthology.org/N06-1059 (accessed 11.08.2024).