Автоматическое реферирование текстов: обзор алгоритмов и подходов к оценке качества

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

Национальный исследовательский университет «МЭИ», Москва

Аннотация: В данной работе представлен обзор задачи автоматического реферирования текстов. Проведена классификация алгоритмов автоматического реферирования текстов по типу получаемого реферата и по подходу к решению задачи. Описаны некоторые существующие проблемы в области автоматического реферирования текстов и недостатки отдельных классов алгоритмов. Определены понятия качества и информационной полноты реферата. Рассмотрены наиболее популярные подходы к оценке информационной полноты реферата и их классификация в соответствии с используемой методологией. Рассмотрены метрики семейства ROUGE применительно к задаче автоматического реферирования текстов. Отдельное внимание уделено оценке информационной полноты реферата с использованием таких метрик информационной близости, как расстояние Кульбака-Лейблера, расстояние Дженсена-Шеннона и косинусное расстояние (сходство).
Ключевые слова: автоматическое реферирование, реферат, информационная полнота, ROUGE, векторизация, TF-IDF, статическая модель векторизации, расстояние Кульбака-Лейблера, расстояние Дженсена-Шеннона, косинусное расстояние.

Введение

В последние годы человечество все стремительнее генерирует и накапливает данные. Так, например, общемировой объем накопленных данных в 2018 году оценивался на уровне 33 зеттабайтов. При этом прогнозируется, что данный показатель к 2025 году возрастет до 175 зеттабайтов [1].

По этой причине крайне важным и полезным является умение извлекать из больших объемов данных необходимую информацию. При работе с текстами такой результат может быть достигнут с использованием **рефератов** исходных текстов, то есть текстов меньшего объема, чем исходный текст, но при этом содержащих в себе все его значимые факты.

Автоматическое реферирование текстов (англ. automatic text summarization) – это процесс получения реферата исходного текста при помощи программных средств без непосредственного участия человека в

формировании текста самого реферата. Иными словами, алгоритмы автоматического реферирования способны осуществлять подготовку текстов рефератов без привлечения экспертов.

В данной работе проведен обзор алгоритмов автоматического реферирования текстов: проведена классификация алгоритмов по типу получаемого реферата и по подходу к решению задачи автоматического реферирования, объяснено понятие качества реферата и рассмотрены методы его оценки, в особенности его информационной полноты.

Классификация алгоритмов

По типу генерируемого реферата существующие алгоритмы автоматического реферирования текстов можно разделить на три класса: экстрагирующие, абстрагирующие и гибридные, а по подходу к решению задачи – на пять классов: статистические, основанные на машинном обучении, когерентные, графовые и алгебраические (рисунок 1) [2].

Рис. 1. – Классификация алгоритмов автоматического реферирования текстов

Экстрагирующие алгоритмы автоматического реферирования текстов – это алгоритмы, выполняющие извлечение из текста исходного документа некоторых наиболее значимых предложений. Реферирование с использованием экстрагирующих алгоритмов называют квазиреферированием. Значимость предложения оценивается с использованием определенных метрик, характерных для конкретного алгоритма [3].

Алгоритмы квазиреферирования, если говорить общо, состоят из двух этапов:

1. Оценка значимости каждого предложения, в результате чего получается таблица весов предложений.

2. Отбор наиболее информативных предложений, то есть сравнение весов предложений с некоторым граничным значением.

При использовании экстрагирующего подхода результатом является так называемый квазиреферат, состоящий из предложений исходного текста. Говоря о достоинствах данного подхода, стоит отметить относительную простоту его реализации. Однако, при этом необходимо понимать, что в реферате некоторые важные предложения могут быть пропущены, в результате чего часть предложений в сгенерированном реферате окажется не несущей смысла без предшествующего текста, который при автоматическом квазиреферировании был пропущен [4].

Абстрагирующий подход в автоматическом реферировании текстов заключается в генерации нового текста, отражающего содержание исходного. Указанные выше особенности подхода предполагают, что при автоматическом реферировании с использованием абстракции необходим синтаксический и семантический разбор предложений [2]. Абстрагирующий подход представляется более совершенным, нежели экстрагирующий, так как нацелен на генерацию более совершенного реферата, а не просто извлечение

отдельных предложений. Однако, стоит отметить, что абстрагирующие алгоритмы сложнее в реализации. Зачастую, крайне трудно достичь качества автоматически сгенерированного реферата, сравнимого с качеством реферата, составленного человеком, владеющим предметной областью. По этим причинам возникли гибридные алгоритмы, сочетающий в себе некоторые особенности как экстракции, так и абстракции [2].

В **гибридных** алгоритмах изначально с использованием экстрагирующих подходов извлекаются наиболее значимые предложения. После чего получившийся квазиреферат преобразуется: отдельные фрагменты опускаются, некоторые фрагменты сливаются, изменяется порядок следования фрагментов.

Рассмотрим подробнее классификацию по подходу к решению задачи автоматического реферирования. При использовании **статистического** подхода значимость предложений исходного текста рассчитывается на основе статистических характеристик входящих в предложение термов. К данному классу алгоритмов относится, например, алгоритм Луна [2].

Когерентный подход основан на учете отношений согласованности между словами, поиск которых является нетривиальной задачей и исследуется в различных направлениях обработки естественного языка [5].

Алгебраический подход использует алгебраический аппарат для определения значимости предложений исходного текста. Среди алгоритмов автоматического реферирования, относящихся к алгебраическому подходу, можно выделить латентный семантический анализ (LSA) и факторизацию симметричных неотрицательных матриц (SNMF) [6].

При использовании **графового** подхода некоторые смысловые единицы текста (термы, предложения, абзацы и т.п.) представляются как вершины некоторого графа, а отношения между смысловыми единицами обозначаются в виде ребер между соответствующими вершинами [7]. К

данному классу алгоритмов можно отнести такие алгоритмы автоматического реферирования текстов, как TextRank и LexRank.

Подход, основанный на **машинном обучении**, заключается в применении различных алгоритмов машинного обучения: наивный байесовский классификатор, логистическая регрессия, деревья решений, искусственные нейронные сети, машина опорных векторов и др. [8]. Отметим, что данные алгоритмы находят свое применение не только в задаче автоматического реферирования текстов, но и во многих других задачах обработки текстов на естественном языке, например, анализе тональности, классификации, распознавании нежелательного контента и т.п. [9, 10].

Оценка качества реферата

На практике обычно важным является не разнообразие используемых методов, а качество, с которым решается задача. Поэтому крайне важным является вопрос оценки качества реферата, получаемого при помощи алгоритмов автоматического реферирования текстов. Посредством такой оценки становится возможным понять, насколько качественно оцениваемый алгоритм решает поставленную перед ним задачу.

Качество реферата – комплексная характеристика, состоящая из большого числа отдельных критериев, среди которых можно выделить следующие: степень сжатия, логичность изложения, отсутствие искажений информации и двусмысленности, полнота раскрытия содержания исходного текста и т.д. Практически все перечисленные выше критерии трудно формализуемы. По этой причине наиболее полно и достоверно провести оценку реферата возможно только с привлечением экспертов. Однако стоит отметить, что существуют методы и алгоритмы, способные автоматически осуществлять оценку качества реферата по одному или нескольким из перечисленных критериев.

Информационной полнотой реферата назовем критерий, показывающий, насколько полно реферат раскрывает содержание исходного текста.

Методы оценки информационной полноты реферата по методологии решения данной задачи можно разделить на два класса: внутренние (англ. intrinsic) и внешние (англ. extrinsic) методы оценки [6]. Внешние методы оценки качества автоматического реферирования рассматривают, насколько качественно сгенерированный реферат позволяет решать какие-либо практические задачи. Например, ряду экспертов могут предложить дать ответы на некоторые вопросы, используя реферат, а затем проверить их корректность.

Внутренние методы оценки информационной полноты реферата ориентируются лишь на сравнение сгенерированного реферата и исходного текста либо сгенерированного реферата и другого реферата, называемого опорным. Опорный реферат воспринимается, как образцовый и полностью корректный и, как правило, составлен экспертом.

Важно отметить, что внешние методы оценки информационной полноты реферата фактически заключаются в сравнении информационной близости двух текстов: полученного реферата с опорным или исходным текстом. При этом иногда такую оценку проводят автоматически с использованием ряда метрик.

Семейство метрик **ROUGE** (англ. Recall-Oriented Understudy for Gisting Evaluation) активно используется при оценке информационной полноты автоматического реферирования текстов. Данное семейство метрик показывает достаточно высокую корреляцию с ручным оцениванием качества автоматически генерируемых рефератов с привлечением экспертов. Метрики семейства ROUGE нацелены на сравнение автоматического и опорного рефератов. Использовать данное семейство метрик для сравнения

исходного текста и автоматического реферата не представляется разумным, так как в силу своего построения метрики данного семейства не будут нести в таком случае никакой полезной информации [11].

Назовем *n*-граммой упорядоченную последовательность термов длиной *n*. Частным случаем *n*-граммы являются униграммы, учитывающие один терм, и биграммы, учитывающие пары термов. При этом, как правило, под *n*-граммой понимают именно непрерывную последовательность термов. В то же время *n*-граммой с пропусками называют не обязательно непрерывную упорядоченную последовательность термов. То есть, например, биграммой с пропусками являются два упорядоченных терма, между которыми могут стоять прочие термы.

Метрики данного семейства основаны на подсчете количества совпадений *n*-грамм в двух текстах (применительно к задаче автоматического реферирования текстов – для полученного и опорного рефератов). Данное семейство метрик включает в себя метрики, представленные в таблице 1.

Таблица № 1

Метрики семейства ROUGE

Метрика	Описание
ROUGE-1	Метрика, основанная на подсчете количества совпадений отдельных термов
ROUGE-2	Метрика, основанная на подсчете количества совпадений биграмм
ROUGE-N	Метрика, основанная на подсчете количества совпадений *n*-грамм
ROUGE-L	Метрика, основанная на определении длины наибольшей общей подпоследовательности термов
ROUGE-S	Метрика, основанная на подсчете количества совпадений *n*-грамм с пропусками

При этом стоит отметить, что каждая из вышеперечисленных метрик сама по себе является набором из трех чисел, выражающих значения точности (англ. precision), полноты (англ. recall) и F1-меры (англ. F1-score).

Предварительно термы исходного текста и реферата могут быть нормализованы с использованием таких методов, как стемминг и лемматизация [10, 12]. Нормализация может проводиться опционально в зависимости от того, насколько важным является учет форм термов, входящих в тексты.

Как уже отмечалось выше, существенным недостатком семейства метрик ROUGE является необходимость использовать опорный реферат, подготовка которого сама по себе может оказаться трудоемкой и затратной. Поэтому интерес представляют методы, позволяющие проводить сравнение реферата с исходным тестом. Для этого производится векторизация сравниваемых текстов с дальнейшей оценкой их схожести.

Векторизация текстов может быть произведена при помощи различных методов: статистическая векторизация, TF-IDF, с использованием статических моделей векторизации (Word2Vec, FastText и т.п.) [13]. Выбор метода векторизации должен производиться с учетом специфики конкретной задачи.

Схожесть полученных векторных представлений текстов может быть оценена с использованием метрик информационной близости. Так, например, в работе [14] в качестве таких метрик предлагаются расстояние Кульбака-Лейблера и расстояние Дженсена-Шеннона.

Расстояние Кульбака-Лейблера является мерой удаленности одного вероятностного распределения от другого и определяется в соответствии с формулой (1). Расстояние Кульбака-Лейблера можно понимать, как количественную меру информации, потерянной при замене вероятностного распределения P другим вероятностным распределением Q (которое может являться приближением вероятностного распределения P).

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

Расстояние Кульбака-Лейблера является несимметричным, поэтому удобным представляется использование расстояния Дженсена-Шеннона, которое является производной мерой от расстояния Кульбака-Лейблера. Расстояние Дженсена-Шеннона называют также информационным радиусом или полным отклонением от среднего. Расстояние Дженсена-Шеннона определяется в соответствии с формулой (2).

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

где M – вероятностное распределение, вычисляемое по формуле (3).

$$M(x_i) = \frac{P(x_i)+Q(x_i)}{2}, \ i = 1 \dots n \quad (3)$$

Помимо описанных выше метрик информационной близости, для задачи определения информационной полноты реферата с использованием векторных представлений может использоваться косинусное расстояние (либо косинусное сходство, которое численно равно дополнению значения косинусного расстояния до единицы) [15].

Заключение

Автоматическое реферирование текстов в современном мире является важной задачей. В данной работе в силу ограниченного объема проведен лишь частичный обзор достижений науки в этой области. Проведена классификация алгоритмов автоматического реферирования текстов по типу реферата и по подходу к решению задачи. Подробно описаны методы оценки информационной полноты реферата с использованием ряда метрик информационной близости.

Литература

1. Reinsel D., Gantz J., Rydning J. The Digitalization of the World, 2018. URL: seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (дата обращения: 11.10.2023).

2. Батура Т.В., Бакиева А.М. Методы и системы автоматического реферирования текстов. Новосибирск: Новосибирский национальный исследовательский государственный университет, 2019. 110 с.

3. Полицына Е.В., Полицын С.А., Касаткина А.О. Создание интегрального алгоритма и инструментов автоматического реферирования текстов на русском языке // Информационные технологии. 2020. №1. С. 30-38.

4. Бакиева А.М., Батура Т.В., Федотов А.М. Методы и системы автоматического реферирования текста // Вычислительные и информационные технологии в науке, технике и образовании (CITech-2015) : Совместный выпуск журналов "Вычислительные технологии" (Том 20) и "Вестник КАЗНУ им. Аль-Фараби" (Серия математика, механика и информатика №3(86)), Алматы, 24–27 сентября 2015 года. Алматы: Казахский национальный университет имени Аль-Фараби, 2015. С. 263-274.

5. Емашова О.А. Об одном подходе к автоматическому реферированию русскоязычных текстов // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. 2008. №1. С. 50-56.

6. Aries A., Zegour D.E., Hidouci W.K. Automatic text summarization: What has been done and what has to be done, 2019. URL: arxiv.org/abs/1904.00688 (дата обращения: 11.10.2023).

7. Кравцов А.А., Липницкий С.Ф., Степура Л.В. Система автоматического индексирования и реферирования текстовых документов // Таврический вестник информатики и математики. 2008. №1(12). С. 260-266.

8. Kumar Y.J., Goh O.S., Basiron H. A review on automatic text summarization approaches // Journal of Computer Science. 2016. № 4. pp. 178-190.

9. Максютин П.А., Шульженко С.Н. Обзор методов классификации текстов с помощью машинного обучения // Инженерный вестник Дона, 2022, № 12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8043.

10. Челышев Э.А., Оцоков Ш.А., Раскатова М.В., Щёголев П. Сравнение методов классификации русскоязычных новостных текстов с использованием алгоритмов машинного обучения // Вестник кибернетики. 2022. №1(45). С. 63-71.

11. Lin C.Y. ROUGE: A package for automatic evaluation of summaries // Proceedings of ACL Text Summarization Branches Out Workshop, Forum Convention Centre Barcelona, Spain, 2004. pp. 74–81.

12. Денисов М.Е., Катышев А.М., Сычев О.А., Аникин А.В. Извлечение ключевых понятий и связей между ними из тематических текстов на русском языке // Инженерный вестник Дона, 2022, № 12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8106.

13. 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 года. Под общей редакцией Т.Б. Казиахмедова. Нижневартовск: Нижневартовский государственный университет, 2022. С. 284-288.

14. Lin C.Y., Cao G., Gao J., Nie J.Y. An Information-Theoretic Approach to Automatic Evaluation of Summaries // Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, New York City, USA, 2006. pp. 463–470.

15. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, 2008. 569 p.

References

1. Reinsel D., Gantz J., Rydning J. The Digitalization of the World, 2018. URL: seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (date accessed 11.10.2023).

2. Batura T.V., Bakieva A.M. Metody i sistemy avtomaticheskogo referirovaniya tekstov [Methods and systems of automatic text summarization]. Novosibirsk: Novosibirskiy natsional'nyy issledovatel'skiy gosudarstvennyy universitet, 2019. 110 p.

3. Politsyna E.V., Politsyn S.A., Kasatkina A.O. Informatsionnye tekhnologii. 2020. №1. P. 30-38.

4. Bakieva A.M., Batura T.V., Fedotov A.M. Vychislitel'nye i informatsionnye tekhnologii v nauke, tekhnike i obrazovanii (CITech-2015): Sovmestnyy vypusk zhurnalov "Vychislitel'nye tekhnologii" (Tom 20) i "Vestnik KAZNU im. Al'-Farabi" (Seriya matematika, mekhanika i informatika №3 (86)), Almaty, 24–27 sentyabrya 2015 goda. Almaty: Kazakhskiy natsional'nyy universitet imeni Al'-Farabi, 2015. pp. 263-274.

5. Emashova O.A. Vestnik Moskovskogo universiteta. Seriya 15: Vychislitel'naya matematika i kibernetika. 2008. №1. pp. 50-56.

6. Aries A., Zegour D.E., Hidouci W.K. Automatic text summarization: What has been done and what has to be done, 2019. URL: arxiv.org/abs/1904.00688 (date accessed 11.10.2023).

7. Kravtsov A.A., Lipnitskiy S.F., Stepura L.V. Tavricheskiy vestnik informatiki i matematiki. 2008. №1 (12). pp. 260-266.

8. Kumar Y.J., Goh O.S., Basiron H. Journal of Computer Science. 2016. № 4. pp. 178-190.

9. Maksyutin P.A., Shul'zhenko S.N. Inzhenernyj vestnik Dona, 2022, № 12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8043.

10. Chelyshev E.A., Otsokov Sh.A., Raskatova M.V., Shchegolev P. Vestnik kibernetiki. 2022. №1 (45). pp. 63-71.

11. Lin C.Y. Proceedings of ACL Text Summarization Branches Out Workshop, Forum Convention Centre Barcelona, Spain, 2004. pp. 74–81.

12. Denisov M.E., Katyshev A.M., Sychev O.A., Anikin A.V. Inzhenernyj vestnik Dona, 2022, №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8106.

13. Raskatova M.V., Chelyshev E.A. Sovremennoe programmirovanie: Materialy IV Mezhdunarodnoy nauchno-prakticheskoy konferentsii, Nizhnevartovsk, 08 dekabrya 2021 goda. Pod obshchey redaktsiey T.B. Kaziakhmedova. Nizhnevartovsk: Nizhnevartovskiy gosudarstvennyy universitet, 2022. pp. 284-288.

14. Lin C.Y., Cao G., Gao J., Nie J.Y. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, New York City, USA, 2006. pp. 463–470.

15. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, 2008. 569 p.

Дата поступления: 21.10.2023

Дата публикации: 8.12.2023