АННОТАЦИЯ

диссертационной работы Ахметова Искандера Рафаиловича на тему: «Разработка метода для информативного экстрактивного реферирования научных текстов на английском языке», представленной на соискание степени доктора философии (PhD) по специальности «8D06101 – Информатика, вычислительная техника и управление»

Введение

Быстрая обработка информации является жизненно важной функцией, необходимой в настоящее время каждому современному человеку. Процесс Автоматического Реферирования Текста (АРТ) сталкивается со множеством проблем несмотря на то, что технологии в этой области постоянно развиваются, и эта проблема изучается с 1958 года. Например, как оценивать качество полученных рефератов, и что должно служить эталоном для сравнения? Есть две основные задачи, которые решаются в процессе АРТ:

1. Выбор критически важной информации из заданного текста.
2. Представление этой информации в сжатом виде.

АРТ — это сложная задача в области обработки естественного языка, поскольку она включает в себя тщательный семантический и лексический анализ текста для создания обоснованного сжатого представления исходных текстовых данных. Высококачественный автореферат должен содержать основную информацию, быть точным в отношении фактов, релевантным, читаемым и не избыточным. Исследования в этой области начались в 1958 году и новые многочисленные работы и методы появляются каждый год, начиная с 2003 года, когда стали доступны большие массивы данных для этой цели и необходимое вычислительное оборудование оживившие интерес к данной теме исследований.

С самого начала исследования проблем реферирования текста было разработано множество различных методов; подробнее см. Глава 3.2. Методы различаются по количеству документов, к которым они применяются; таким образом, существует одно-документное и многодокументное автореферирование определил два класса методов обобщения текста:

1. **Экстрактивный автореферат**, который включает в себя шаги по извлечению конкретных предложений из исходного текста, без каких-либо изменений.
2. **Абстрактивный автореферат** - предполагает связное и сжатое изложение исходного текста в свободной форме.

Если сравнивать эти два метода, то второй тип больше похож на человеческое мышление, поскольку встает необходимость заменять слова синонимами и переставлять их местами. В отличие от этого, экстрактивный

метод заключается в составлении резюме из исходного текста, находя самые важные предложения. Таким образом, экстрактивные резюме легче получить и ожидается, что они дадут лучшие результаты, чем абстрактные резюме.

Из этого следует, что второй класс сложнее, так как в нем задействованы такие сложные техники, как генерация естественного языка.

В настоящее время исследования во всем мире переориентировались на абстрактивное авто-реферирование. Тем не менее, экстрактивное автоматическое реферирование все еще в тренде, как видно из научных работ последних двух лет. Помимо сложности формирования резюме, открытым вопросом в научном сообществе является его оценка. Метрика качества текстов должна учитывать неоднозначность естественного языка.

Актуальность

Работы в направлении разработки методов для автоматического информативного реферирования научных текстов сейчас являются как никогда актуальными в следствии экспоненциального роста информации в целом и научной информации в частности в нашем сегодняшнем Мире.

Самые современные модели автоматического реферирования текстов на текущий момент построены на основе сложных архитектур нейронных сетей с миллиардами параметров, и натренированные на огромных количествах данных и используют эмбеддинги из предтренированных языковых моделей таких как BERT, GPT-3 и другие. Это поднимает вопросы о возможной избыточной сложности таких моделей, их способности к обобщению, ведь миллиарды параметров нейросети позволяют моделям просто "зазубривать" правильные ответы, и в конечном счете вопросы об экономической эффективности и экологической безопасности этих моделей.

В настоящее время, каждому человеку, а научному работнику в первую очередь, остро необходимы инструменты для эффективной работы с информацией, одним из которых может быть система для автоматического реферирования текстов. Данная тема подробно исследована в классических работах.

Объект исследования

Процесс автоматического, информативного и экстрактивного реферирования научных текстов на английском языке.

Предмет исследования

Метод автоматического реферирования текстов.

Цель исследования

Цель диссертационной работы – Разработка метода для информативного экстрактивного реферирования научных текстов на английском языке, для экономии времени и сокращения объема информации для обработки.

Задачи исследования

1. Разработать метод автоматического реферирования на основе подхода жадного алгоритма.
2. Подобрать управляющие параметры разработанного метода.
3. Экспериментально оценить наивысшее значение метрики ROUGE с помощью методов экстрактивного автоматического реферирования текстов.
4. Провести сравнительный анализ результатов разработанного метода с существующими методами.

Материалы исследования

Решение поставленных в работе задач осуществлялось на основе применения общенаучных методов исследования в рамках проведения экспериментов с текстовыми данными и количественной оценки полученных результатов. Программирование и исходные коды были выполнены на языке Python 3.6 (Pandas, Numpy).

Научная новизна

Новизна предложенной модели заключается в уникальном применении жадного алгоритма в методе экстрактивного, информативного реферирования текстов.

Также, модель демонстрирует производительность на уровне современных реферативных моделей, в разработке которых использовались нейронные сети и колоссальный объем данных для обучения. При этом, предлагаемая модель, относительно проста, и требует гораздо меньше времени и данных для обучения.

Вклад нашего исследования в научные знания заключается в следующем: 1) выявление верхней границы для методов экстрактивного суммирования (VNS, жадный алгоритм, генетический алгоритм) и обнаружение того, что VNS, инициализированный жадным алгоритмом, работает даже лучше, чем любой из алгоритмов самостоятельно для данной задачи, 2) предложение метода экстрактивного суммирования, основанного на алгоритме жадности, который работает на высоком уровне, несмотря на свою относительную простоту, 3) очищенный набор данных с различными типами резюме с высоким ROUGE и полезной статистикой текста.

Практическая значимость

Мы представляем подход экстрактивного суммирования[1], который использует простые и старые методы, но при этом работает на уровне современных моделей, использующих сложные архитектуры нейронных сетей

[1] Исходный код доступен на GitHub по адресу https://github.com/iskander-akhmetov/Greedy-Summarization

и огромные объемы данных для обучения; см. Рис. 1.1 для краткого описания нашего подхода. Набор данных arXiv extract из 17 тысяч статей, которые мы использовали в наших экспериментах, доступен по адресу https://data.mendeley.com/datasets/nvsxfcbzdk/1. Некоторые из других преимуществ предложенного подхода включают:

- Вычислительная простота.
- Не требуется обучение модели Machine Learning, но используется статистический вывод.
- Резюме, сгенерированные алгоритмом, богаты полезной информацией из текста.

Разработанная модель автоматического экстрактивного реферирования текста имеет самый широкий спектр практического применения в науке, образовании и бизнесе.

Наука:

- Автоматизация обзора литературы.
- Генерация аннотации статьи.
- Мультимодальное автореферирование.
- Популяризация науки.
- Обновление научной информации.
- Использование автреферирования в других задачах NLP.
- Тематическое моделирование.
- Анализ тональности.

Образование:

- Автоматическое создание заметок.
- Памятки.
- Интеллект-карты.
- Генерация слайдов презентации.
- Генерация тестовых вопросов.
- Написание эссе.

Бизнес:

- Резюме больших объемов текста (отчеты, исследования, бизнес-планы).
- Формирование протокола собрания.
- Реферирование, ориентированное на запрос.
- Оптимизация контекстной рекламы.

Основные положения выносимые на защиту

На защиту выносятся:

1. Метод эвристической оценки уровня качества реферата метрикой ROUGE-1, достижимого при помощи экстрактивных методов

автоматического реферирования текстов дает результат в 0.59, что значимо выше текущего уровня в 0.46 у самых современных методов использующих нейронные сети.

2. Разработанный метод Автоматического Экстрактивного Реферирования (АЭР) GreedSum, который показывает результат 0.42 по метрике ROUGE-1 на датасете arXiv.
3. Техника тонкой настройки управляющего параметра minimum document frequency (min_df) работы GreedSum , или минимальная частота вхождения слов в предложения реферируемого текста для их учета в создании словаря для построения TFIDF матрицы. На выборке в 376 текстов из датасета arXiv путем простого перебора оптимальное значение min_df было определено как 0.042 (т.е. слово должно появляться как минимум в 4.2% предложениях).

Апробация полученных результатов

По итогам диссертационного исследования опубликовано 13 научных статей, в том числе 10 статей опубликованы в зарубежных изданиях (4 журнальных и 6 конференций), входящих в международную базу цитирования Scopus, с процентилями от 27 до 80, получено одно авторское свидетельство.

Структура

Структура диссертационного исследования обусловливается его целью и задачами: диссертация состоит из введения, обзора литературы, основной части, заключения и списка использованной литературы.

В «Основной части» помимо данных и методов, пописываются эксперименты по оценке верхней границы качества авторефератов достижимой с применением экстрактивных методов автоматического реферирования текстов, а также модель автореферирования на основе жадного алгоритма.

Общий объем исследования 147 страниц, список литературы включает 144 наименования.