Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.822

ВАШКЕВИЧ
Евгений Кириллович

АЛГОРИТМ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ ПРЕДЛОЖЕНИЙ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Автореферат
на соискание степени магистра
по специальности 1–45 80 01 Системы и сети инфокоммуникаций
(информационные и коммуникационные технологии)

Научный руководитель
к.т.н, доцент
БОРИСКЕВИЧ Илья
Анатольевич

Минск 2021

Нормоконтролер

ВВЕДЕНИЕ

Задача суммаризации текстов (автореферирование) – одна из ключевых, широко обсуждаемых задач NLP. Она состоит в сжатии больших объемов текста до связного краткого содержания, отражающего только основные идеи.

Экономия времени на чтении актуальна и ежегодно публикуется множество статей, описывающих новые методы и улучшения существующих решений. Наибольший успех имеют нейронные сети, но есть и более простые и быстрые подходы, использующиеся в большинстве статей в качестве исходной точки для сравнения качества. Оптимального и универсального решения задачи автоматической суммаризации еще не найдено.

Объектом исследования является – текстовая информация больших объемов, требующая анализа с помощью машинного обучения

Предметом исследования является – алгоритмы обобщения и предобработки данной текстовой информации.

Целью диссертационной работы является выделение ключевых предложений на основе алгоритмов лексической центральности предложений на основе графов и латентно-семантического анализа. Для достижения поставленной цели были сформулированы и решены следующие задачи:

– анализ существующих алгоритмов обобщения текста;

– анализ существующих методов и алгоритмов предобработки текста;

– разработка алгоритмов предобработки и обобщения текста;

– программная реализация разработанных алгоритмов предобработки и обобщения текста;

– оценка эффективности разработанных алгоритмов предобработки и обобщения текста.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами

Тема диссертационной работы соответствует пункту 5 приоритетных направлений научных исследований Республики Беларусь на 2016–2020 гг., утвержденных Постановлением Совета Министров Республики Беларусь № 190 от 12 марта 2015 г. «Информатика и космические исследования».

Цель и задачи исследования

Целью диссертационной работы выделение ключевых слов на основе алгоритмов латентно-семантического анализа и центральности предложений на основе графов. Для достижения поставленной цели были сформулированы и решены следующие задачи:

– анализ существующих алгоритмов обобщения текста;

– анализ существующих методов и алгоритмов предобработки текста;

– разработка алгоритмов предобработки и обобщения текста;

– программная реализация разработанных алгоритмов предобработки и обобщения текста;

– оценка эффективности разработанных алгоритмов предобработки и обобщения текста.

Личный вклад соискателя ученой степени

Содержание диссертации отображает личный вклад автора. Он заключается в научном обосновании алгоритмов выделения ключевых предложений в тексте, оценке эффективности разработанных алгоритмов, обработке и анализе полученных результатов, формулировке выводов.

Определение целей и задач исследований, интерпретация и обобщение полученных результатов проводились совместно с научным руководителем к.т.н., доцент И.А. Борискевич.

Апробация диссертации и информация об использовании ее результатов

Основные положения и результаты диссертационной работы докладывались и обсуждались на 56-й научной конференции аспирантов, магистрантов и студентов БГУИР «Инфокоммуникации» (Минск, 2020) и на международной научно-практической конференции «Кодирование и цифровая обработка сигналов в инфокоммуникациях» (Минск, 2021).

Опубликование результатов диссертации

По результатам исследований, представленных в диссертации, опубликован 1 тезис и 1 статья в сборниках и материалах конференций.

Структура и объем диссертации

Диссертационная работа состоит из введения, общей характеристики работы, трех глав с выводами по каждой главе, заключения, библиографического списка.

Общий объем диссертационной работы составляет 72 страницы, из них 52 страницы текста, 21 рисунок на 19 страницах, 21 таблица на 14 страницах, список использованных библиографических источников (33 наименования на 3 страницах), список публикаций автора по теме диссертации (2 наименование на 1 странице), 1 приложение на 8 страницах и графический материал на 6 страницах.

Проверка на заимствования

Проведена экспертиза диссертационной работы на корректность использования заимствованных материалов с применением сетевого ресурса «Антиплагиат» (адрес доступа: http://nlb.antiplagiat.ru) в on-line режиме 15.04.2021 г. В результате проверки установлена корректность использования заимствованных материалов (оригинальность диссертационной работы составляет 92,3 %).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** дано краткое обоснование актуальности работы, сформулированы цель работы и задачи исследования.

В **первой главе** дан обзор и классификация методов и алгоритмов предобработки и обобщения текста.

Подходы к обобщению текста различаются в зависимости от количества входных документов (один или несколько), цели (общие, предметно-ориентированные или основанные на запросах) и вывода (экстрактивные или абстрактные).

Экстрактивное обобщение означает определение важных частей текста и их дословное генерирование, создавая подмножество предложений из исходного текста

Абстрактное обобщение воспроизводит важный материал по-новому после интерпретации и изучения текста с использованием передовых методов естественного языка для создания нового более короткого текста, который передает наиболее важную информацию по сравнению с исходным.

Очевидно, что абстрактное обобщение более продвинуто и ближе к человеческой интерпретации. Хотя оно имеет больший потенциал (и в целом более интересен для исследователей и разработчиков), до сих пор более традиционные методы показали лучшие результаты.

К настоящему моменту ядро всех экстрактивных рефератов составляют три независимых задачи:

1 Построение промежуточного представления входного текста. Существует два типа подходов, основанных на представлении: представление темы и представление индикатора. Представление темы преобразует текст в промежуточное представление и интерпретирует тему (темы), обсуждаемую в тексте. Методы, используемые для этого, различаются по сложности и делятся на частотно-зависимые подходы, тематические подходы, латентно-семантический анализ и байесовские тематические модели. Индикаторное представление описывает каждое предложение как список важных формальных характеристик (индикаторов), таких как длина предложения, положение в документе, наличие определенных фраз и т.д.

2 Оценка предложений на основе представления. Когда создается промежуточное представление, каждому предложению присваивается оценка важности. В подходах к тематическому представлению оценка предложения показывает, насколько хорошо предложение объясняет некоторые из наиболее важных тем текста. В представлении индикатора оценка рассчитывается путем агрегирования доказательств из различных взвешенных индикаторов.

3 Выбор резюме, состоящего из ряда предложений. Система реферирования выбирает k самых важных предложений для составления резюме. Некоторые подходы используют жадные алгоритмы для выбора важных предложений, а некоторые подходы могут преобразовать выбор предложений в задачу оптимизации, когда выбирается набор предложений, учитывая ограничение, которое должно максимизировать общую важность и согласованность и минимизировать избыточность.

Обработка естественного языка или НЛП - это область искусственного интеллекта, которая дает машинам возможность читать, понимать и извлекать значение из человеческих языков. Это дисциплина, которая фокусируется на взаимодействии науки о данных и человеческого языка и распространяется на множество отраслей

One-hot encoding («Мешок слов») - это широко используемая модель, которая позволяет подсчитывать все слова в фрагменте текста.

Токенизация. Это процесс разбиения текста на предложения и слова. Это задача разрезать текст на части, называемые токенами, и в то же время отбросить определенные символы, такие как знаки препинания

Удаление стоп-слов включает в себя избавление от общеязыковых артиклей, местоимений и предлогов, таких как «and», «the» или «to» на английском языке. В этом процессе некоторые очень распространенные слова, которые, по-видимому, не имеют большого значения для цели НЛП или не имеют никакого значения, фильтруются и исключаются из обрабатываемого текста, тем самым удаляя распространенные и часто встречающиеся термины, которые не информативны для соответствующего текста.

Стемминг относится к процессу разрезания конца или начала слова с целью удаления аффиксов (лексических дополнений к корню слова).

Лемматизация - это приведение слова к его базовой форме и группировка различных форм одного и того же слова.

Тематическое моделирование является методом выявления скрытых структур в наборах текстов или документов. Метод группирует тексты, чтобы обнаруживать скрытые темы на основе их содержания, обрабатывая отдельные слова и присваивая им значения на основе их распределения.

Во **второй главе** предложены алгоритм лексической центральности на основе графов и алгоритм на основе латентно-семантического анализа.

Алгоритм LexRank основан на лексической центральности на основе графов. В начале алгоритма составляется граф, где каждое предложение обрабатывается как узел в графе. Затем вычисляется схожесть двух предложений в тексте. Далее набор предложений характеризуется матрицей схожести, каждый элемент которой соответствует вычисленной величине схожести между парой предложений. Эту матрицу можно интерпретировать как матрицу связности взвешенного графа, содержащую веса ребер. В завершении алгоритма, веса упорядочиваются по убыванию и в реферат включаются предложения с наибольшим весом. Блок-схема алгоритма представлена на рисунке 1.

Алгоритм LSA основан на латентно-семантическом анализе. В начале выполнения алгоритма формируется матрица терм-предложение. Затем производится сингулярное разложение полученной матрице, согласно теореме о сингулярном разложении. Далее для каждого предложения изначального текста рассчитывается вес. В завершении алгоритма отбирается необходимое количество предложений с наибольшим весом. Блок-схема алгоритма представлена на рисунке 2.

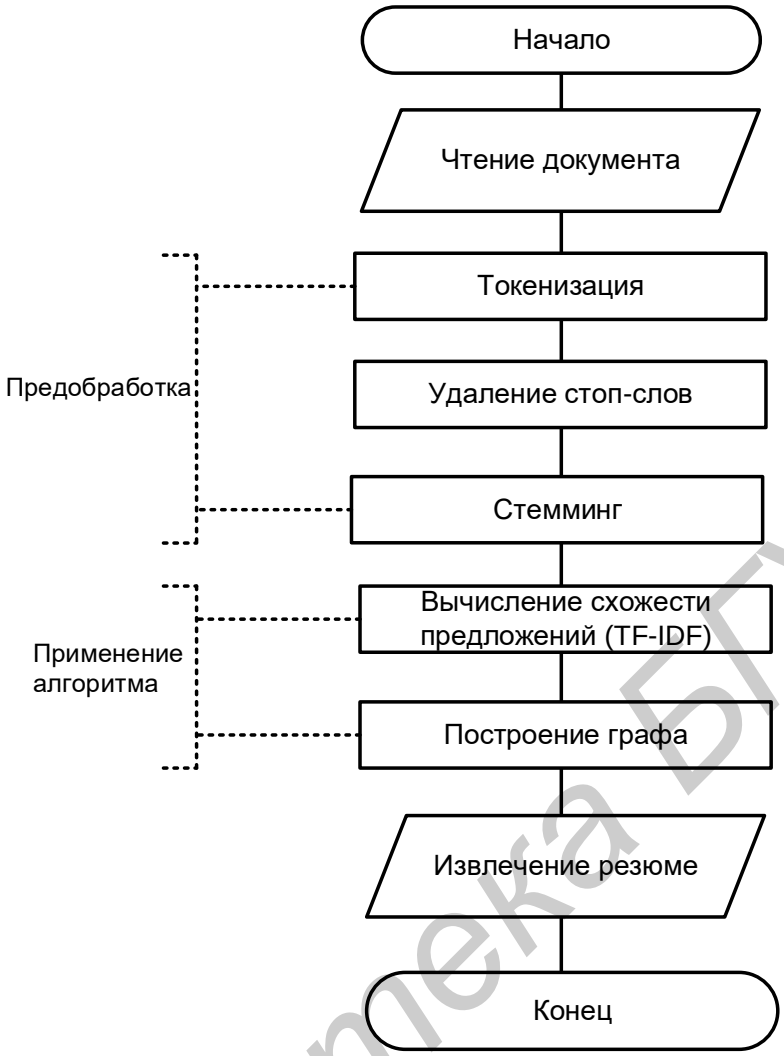

Рисунок 1 – Блок-схема алгоритма лексической центральности на основе графов LexRank

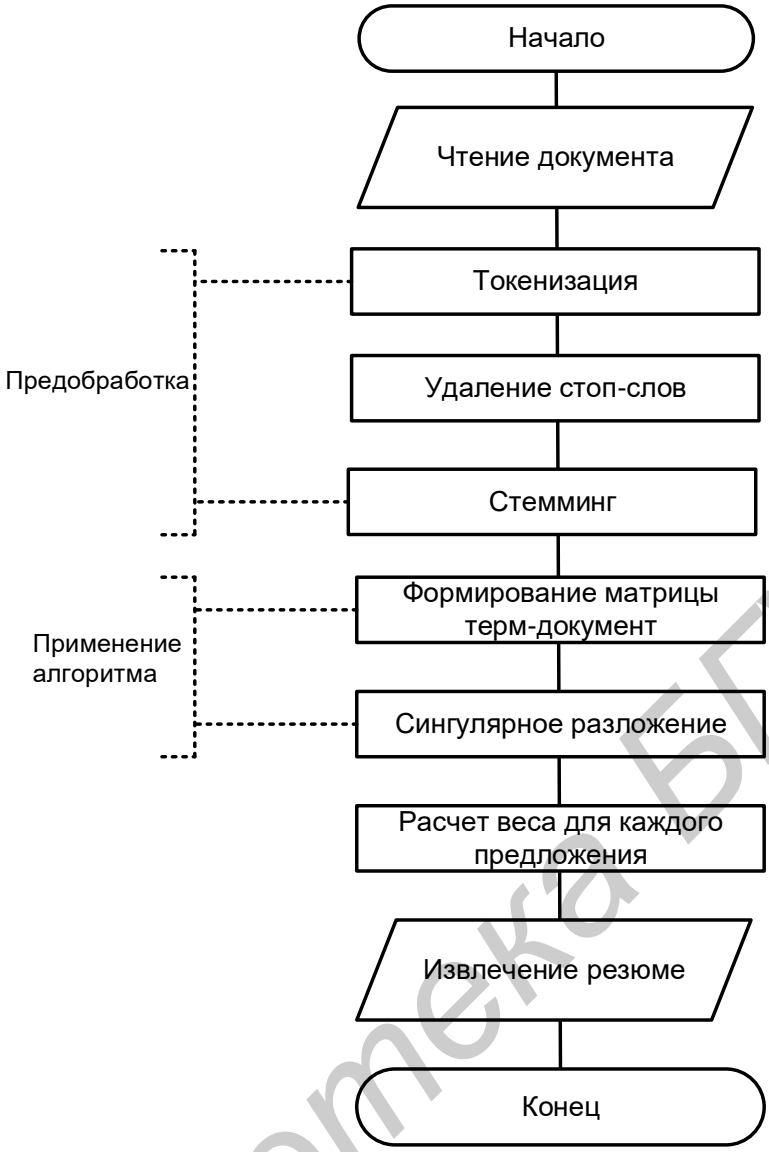

Рисунок 2 – Блок-схема алгоритм на основе латентно-семантического анализа LSA

В **третьей главе** произведена оценка эффективности предложенных алгоритмов.

Для оценки эффективности использовались следующие метрики оценки: на основе совместного отбора, на основе содержания, на основе документов.

1 Оценка, основанная на совместном отборе, основывается на одновременном появлении терминов в системной сводке и требует справочной сводки документа для сравнения. Оценка выполняется путем выбора общих терминов из краткого описания системы и справочного обзора. Соответствующие показатели для оценки на основе совместного отбора – это отзыв, точность и *F*-балл:

Отзыв: это отношение общего количества найденных правильных предложений к общему количеству найденных правильных предложений и не

найденных правильных предложений в документе. Его можно оценить следующим образом:

$$\mathrm{Re}call = \frac{\sum_{S \in ModelSummary} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ModelSummary} \sum_{gram_n \in S} Count(gram_n)}$$

где, S обозначает предложения, n – длина из n-грамм, $gram_n$ и $Count_{match}(gram_n)$ – максимальное количество n-грамм, одновременно встречающихся в резюме кандидата и резюме модели. $Count(gram_n)$ – это количество n-грамм в сводке модели.

Точность: это отношение общего количества найденных правильных предложений к общему количеству найденных правильных предложений и найденных неправильных предложений в документе. Ее можно оценить следующим образом:

$$Percision = \frac{\sum_{S \in SystemSummary} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in SystemSummary} \sum_{gram_n \in S} Count(gram_n)}$$

где, S относится к предложениям, n – к длине n-грамм, $gram_n$ и $Count_{match}(gram_n)$ – максимальное количество n-грамм, одновременно встречающихся в резюме кандидата и в наборе или единственной справочной сводке. $Count(gram_n)$ – это число n-грамм в сводке системы.

F-оценка: измеряет эффективность поиска по отношению к пользователю, который придает в β раз большее значение отзыву, чем точности. Оценка F для неотрицательного действительного β $(0 \le \beta < \infty)$ рассчитывается следующим образом:

$$F_\beta \frac{(1+\beta^2)(\mathrm{Pr}ecision * \mathrm{Re}call)}{(\beta^2 * \mathrm{Pr}ecision + \mathrm{Re}call)}$$

2 Метод совместного отбора оценивает систему реферирования на общих основаниях. Он не может получить связность идей, поток предложений, связь предложений с их предыдущими предложениями, новизну содержания в резюме. Метод, основанный на содержании, может решить все эти проблемы. Здесь описываются некоторые методы оценки, основанные на содержании, которые учитывают различные свойства текста. Для оценки на

основе содержания требуется только системное резюме. Соответствующие показатели для оценки на основе содержания следующие:

Связность: это важный параметр, отражающий отношения между концепциями в тексте. Определено пять общих категорий отношений связности: союз, ссылка, многоточие, подстановка и лексика. Соединение «и» является самым основным и наименее связным отношением между предложениями, и ссылочное отношение может быть либо анафорическим, либо катафорическим. Анафорическое отношение возникает, когда предложение относится к чему-то, что было объяснено ранее, в то время как катафорическое отношение прямо противоположно анафорическому отношению. Отношение многоточия возникает, когда после более конкретного объяснения слова опускаются в повторяющейся фразе. В отношении замены слова не опускаются, как в многоточии, а заменяются более общими словами вместо повторения слов. Лексическое отношение относится к выбору связанных слов для соединения содержимого текста. Связность по отношению к лексическому отношению с использованием вложений слов может быть вычислена следующим образом:

$$Cohension(Sum_s) = \frac{\log(Avg_{S_x \in \{Sum_s\}}(Sim(S_x)) \times 9 + 1)}{\log(\max_{S_x \in Sum_s}(Sim(S_x)) \times 9 + 1)}$$

где, $Avg_{S_x \in \{Sum_s\}}$ – это среднее значение подобия всех предложений, принадлежащих системной сводке, $\max_{S_x \in Sum_s}(Sim(S_x))$ – максимальное сходство в системной сводке, $Sim(S_x)$ – среднее значение подобия S_x со всеми предложениями в сумме. Сходство между двумя предложениями рассчитывается по четырем.

Отсутствие избыточности: отсутствие избыточности относится к новизне в резюме. Предполагается, что резюме не должно быть избыточным, чтобы увеличить охват информации, содержащейся в документе. Новизну сформированного резюме можно рассчитать следующим образом:

$$Novelty(Sum_s) = 1 - \max_{y \in Sum_s}(sim(S_x, S_y))$$

где, Sum_s обозначает сводку системы, $sim(S_x, S_y)$ – сходство между S_x и S_y.

Удобочитаемость: удобочитаемость также является важным параметром для измерения производительности сумматора. Критерий показывает, насколько легко текстовый контент можно прочитать и понять. Читаемость

текста можно измерить по двум аспектам: содержание и связь предложения с предыдущим предложением. Читаемость по отношению к содержанию зависит от сложности словаря и синтаксиса, тогда как связь с предыдущим предложением показывает беглость чтения. Читаемость на основе контента рассчитывается с использованием пяти популярных формул удобочитаемости, а именно:

Flesch Kincaid Grade level (FKGL):

$$FKGL = 0,39(\frac{W}{Sen}) + 11,8(\frac{Syl}{W}) - 15,59$$

Gunning Fog Score (GFS):

$$GFS = 0,4[(\frac{W}{Sen}) + 100(\frac{CW}{W})]$$

Coleman Liau Index(CLI):

$$CLI = 0,0588L - 0,296S - 15,8$$

Automated Readability Index (ARI):

$$ARI = 4,71(\frac{C}{W}) + 0,5(\frac{W}{Sen}) - 21,43$$

SMOG Index (SMOI):

$$SMOGI = 1,0430 \times \sqrt{(Syl \times \frac{30}{Sen})} + 3,1291$$

где W обозначает общее количество слов, Sen, Syl и CW – общее количество предложений, слогов и сложных слов соответственно; L – среднее количество символов C на 100 слов, а S – среднее количество предложений на 100 слов.

Связь с удобочитаемостью на основе предыдущего предложения рассчитывается следующим образом:

$$RPS = \frac{\sum_{0 \le x < D} Sim(S_x, S_{x+1})}{\max Sim(S_x)}$$

где *D* относится к документу, используемому для расчета удобочитаемости.

3 Метрики оценки на основе документов. Метрики оценки, основанные на совместном выборе, не позволяют правильно оценить системное резюме, когда два предложения имеют одинаковую важность в документе, а справочная сводка не содержит ни одного из них. Оценка на основе документов может помочь избавиться от этой ситуации. Здесь ранжируются предложения документа по их значимости. По ранжированию предложений оценивается сводка системы следующим образом:

предложения обозначаются (*sen_significance*) вручную между 0 и 1. Самому значимому предложению присваивается наивысшее значащее значение.

Summary_significance сводной информации о системе рассчитывается следующим образом:

$$Summary\_significance = \frac{\sum_{S \in sum_s} sen\_significance}{\max sen\_significance}$$

где, *Summary_significance* – ручная значимость системной сводки, *S* относится к предложениям системной сводки.

ЗАКЛЮЧЕНИЕ

В ходе работы был осуществлен анализ существующих алгоритмов обобщения: на основе графов, нейронных сетей и метаэвристики. Также проанализированы существующие методы и алгоритмы предобработки текста: токенизация, лемматизация, стемминг, удаление стоп-слов.

Разработаны алгоритмы предобработки и обобщения текста, основанные на лексической центральности на основе графов и латентно-семантическом анализе. Осуществлена программная реализация разработанных алгоритмов в среде Anaconda на языке Python.

Оценка эффективности проводилась на тринадцати алгоритмах автоматического реферирования с аналогичными настройками для новостных наборов данных на английском языке.

Произведена оценка производительности с использованием показателей точности, отзыва и показателей F_1 на пяти различных уровнях отсечения суммарной длины для разных n-грамм. Обнаружено, что показатель точности уменьшается с увеличением длины сводки, а также с увеличением значений слова n-грамм. Оценки запоминания увеличиваются с увеличением длины сводки, но уменьшаются по сравнению с n-значениями в n-граммах слов, максимальные значения наблюдались при длине сводки 40 %. Ограничение данных алгоритмов обусловлено тем, что их эффективность зависит от суммарной длины. Непосредственная близость значений F1 для алгоритмов SummaRuNNer, NN-SE и NN-ED обусловлена тем, что они являются нейронными сетями, основанными на инструменте word2vec. При этом алгоритмы на основе нейронных сетей не показали лучшей производительности по сравнению с разработанными алгоритмами. Также было показано, что почти все алгоритмы генерируют неизбыточные, удобочитаемые, значимые резюме.

Список публикаций автора

1-А. Вашкевич, Е. К. Токенезация в NLP / Вашкевич Е. К. // Инфокоммуникации : сборник тезисов докладов 56-ой научной конференции аспирантов, магистрантов и студентов БГУИР, Минск, – 2020. – С. 67–68.

2-А. Вашкевич, Е. К. Сравнительный анализ алгоритмов автоматического обобщения текста / Е.К. Вашкевич, И.А. Борискевич // Международная научно-практическая конференция «Кодирование и цифровая обработка сигналов в инфокоммуникациях». – 2021. – С. 13–18.