Модификация графового метода для задач автоматического реферирования с учетом синонимии

И.Н. Полякова, И.О. Зайцев

Аннотация **- В статье рассмотрены существующие подходы к автоматическому реферированию текста. Создание реферата, очевидно, требует понимания текста на уровне прагматики. Однако, на данный момент, надежный семантический анализ все еще не доступен для ЭВМ, тем более не доступен анализ прагматики. Широко распространенные методы, основанные на нейронных сетях, могут учитывать семантику благодаря специальным векторным представлениям слов. Но остальные методы автоматического реферирования опираются на морфологию и синтаксис. Однако часть семантики все же доступна и им - существуют тезаурусы и сети, а также алгоритмы, позволяющие установить семантическую связь между отдельными словами, такую как их семантическое сходство, в частности - синонимию.**

Предлагается метод, учитывающий семантическое сходство слов. Разработана модификация графового метода, учитывающая синонимию и позволяющая получить более качественный реферат. Программно реализованы базовая и модифицированная версии графового метода для задач автоматического реферирования с учетом синонимии. Проведено сравнение их качества на русских и английских текстах при помощи метрик автоматической оценки рефератов. По результатам оценки видно улучшение у модифицированного графового метода по сравнению с обычным, особенно на русскоязычных текстах.

Ключевые слова **- автоматическое реферирование текста, графовый метод, метрики автоматической оценки рефератов, модификация с учетом синонимии, семантическое сходство слов, синонимия**

Статья получена 23 марта 2022.
Полякова И.Н., Московский государственный университет имени М.В. Ломоносова (email: polyakova@cs.msu.ru)
Зайцев И.О., магистрант, Московский государственный университет имени М.В. Ломоносова (email: igorza97@mail.ru)

I. ВВЕДЕНИЕ

В наше перегруженное информацией время актуальность рефератов и аннотаций как никогда высока. От них требуется краткая, но полная передача сути рассматриваемого документа. Рефераты и аннотации широко используются в новостных порталах и научных статьях – для быстрого погружения читателя в суть дела, в поисковых системах - для вывода краткой информации о найденных документах. И, как и для любой другой часто встречающейся задачи, возникла необходимость ее автоматизации.

Задача автоматического реферирования заключается в создании нового текста на основе исходного документа, в идеале максимально удовлетворяющего следующим требованиям [1]:

- имеет меньший объем, чем исходный документ;
- полностью передает основные идеи оригинала;
- является связным, удобно читаемым текстом.

В зависимости от решаемой задачи предпочтение обычно отдается какой-то части требований.

Выделяют следующие виды задач автоматического реферирования:

- ✓ реферирование одного документа - самая типичная задача, создание реферата для одного документа;
- ✓ реферирование по многим документам - создание обзорного реферата для группы документов на одну тему;
- ✓ реферирование на основе запроса - создание реферата одного или нескольких документов, содержащего только релевантную информацию для данного запроса;
- ✓ выделение ключевых слов - поиск в тексте слов или фраз, передающих основную тему документа.

В предлагаемой работе, если не сказано иное, будет рассматриваться задача реферирования одного документа.

II. МЕТОДЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Все методы автоматического реферирования (автореферирования) можно разделить на две категории - извлекающие и обобщающие (иначе - генерирующие) [2]. Обобщающие методы порождают совершенно новый текст, не опираясь на элементы старого документа, а только лишь на его смысл. Такие рефераты называют абстрактами. С развитием методов машинного обучения появляются обобщающие методы, основанные на искусственных нейронных сетях [3]. Такие методы, благодаря использованию технологий, основанных на векторном представлении слов [4], способны отчасти учитывать семантику. Промежуточное положение между извлекающими и обобщающими занимают методы, сокращающие текст посредством удаления малозначимых синтаксических элементов, например, деепричастных оборотов или обстоятельств. Такие методы могут быть эффективно использованы совместно с любыми другими. Извлекающие методы выделяют из текста самые ценные фрагменты (предложения) и на их основе создают реферат. Рефераты, полученные таким образом, называются экстрактами. Среди извлекающих методов выделяют следующие [2]:

- Основанные на частоте слов, базирующиеся на предположении, что часто встречающиеся слова в тексте, имеют большую значимость для данного текста.
- Основанные на графах, рассматривающие текст как граф, узлами которого являются его части(чаще всего - предложения), а ребрами -- связи между частями. Оценивают эти части исходя из структуры графа.
- Основанные на машинном обучении - идея этих методов в том, чтобы задачу автореферирования рассматривать как задачу бинарной классификации предложений: на те, которые будут, и те, которые не будут принадлежать реферату.

Чаще всего перечисленные методы опираются исключительно на морфологию слов текста, тогда как для обобщающих методов нужен синтаксичесий и семанический анализ, который является сложной, а в случае семантического анализа еще и не полностью решенной задачей. Однако возможно модифицировать существующие извлекающие методы так, чтобы они могли учитывать часть семантики, а именно семантическое сходство слов, при этом не сильно усложняя эти методы.

III. АНАЛИЗ СЕМАНТИКИ СЛОВ

Анализ семантики - сложная задача для компьютера, так как для этого необходимо сопоставлять слова с образами, которые, в свою очередь, надо как-то представить в машине. Но есть методы и инструменты, позволяющие в некоторой степени работать с семантикой слов.

Семантическая сеть - представление некоторой предметной области в виде графа, вершинами которого являются объекты этой области, а ребрами - различные связи между объектами. Семантическую сеть можно построить и для естественного языка, тогда вершинами будут слова (или объединения слов, словосочетания), а ребрами - семантические отношения. Такой семантической сетью/тезаурусом естественного языка является WordNet [5], словарь английского языка. Вершинами в WordNet выступают синсеты (от англ. synset, Synonym set) - объединения слов и словосочетаний с одинаковым значением, синонимов. Одно и то же слово может входить в разные синсеты, так как слова могут быть многозначными. Пример синсета: "peak, crown, crest, top, tip, summit" - множество слов, обозначающих вершину, наивысшую точку чего-либо. WordNet поддерживает достаточно много семантических отношений между синсетами, из которых для наших целей интерес представляют:

- ✓ Гипоним - понятие, обозначающее частную сущность по отношению к другому, более общему понятию. Например, термин "такса" гипоним для термина "собака";
- ✓ Гипероним - наоборот, общее понятие для данного слова (обратное гипонимии). Например, "транспортное средство" гипероним для термина "мотоцикл";

Они позволяют организовать слова в иерархию понятий, от более общих к частным. На основе этой иерархии можно оценить сходство слов, чем ближе слова в иерархии, тем более они похожи. Существует достаточно много способов оценки сходства на иерархии гиперонимов [6].

Помимо методов, использующих тезаурусы и сети, существуют методы [4], основанные на распределении слов (дистрибутивной семантике). Основа таких методов - предположение, что слова,

встречающиеся в схожем контексте (окружении слов), семантически связаны. Ресурс Word2vec [4] - один из самых известных инструментов, основанных на дистрибутивной семантике. В целом, Word2vec инструмент мощный, но требовательный и громоздкий - для его работы нужно время, вычислительные мощности и очень много данных. Поэтому в качестве инструмента для анализа семантики в дальнейшем будем использовать WordNet.

IV. ОЦЕНКА ПОЛУЧАЕМЫХ РЕФЕРАТОВ

Чтобы говорить о качестве какого либо алгоритма, нужно ввести способ оценки этого качества. Человеческие представления о "хорошем" реферате сильно разнятся, что усложняет автоматизацию этой задачи. При ручной оценке эксперты читают и реферат, и оригинал и оценивают содержание и читаемость на некой численной шкале. Оценки экспертов затем усредняются. Ручная оценка результатов позволяет достоверно оценить и содержание, и читаемость текста, однако очень трудоемка и времезатратна.

Существуют автоматические метрики оценки. Среди автоматических метрик самыми распространенными являются метрики семейства ROUGE [7]. Чаще всего используются ROUGE-1 и ROUGE-2. Все они, так или иначе, сравнивают данный реферат с эталонами - рефератами, созданными людьми. Метрики вида ROUGE-N (где N - натуральное число) оценивают совпадение N-грамм в данном реферате и эталоне. В данном случае, N-грамма - N последовательных лемм (слов в нормальной форме) из текста. Например, текст: "Мама мыла раму. Коля шел домой" содержит четыре биграммы: (мама мыть), (мыть рама), (Коля идти), (идти дом). *Полнота* при данной метрике отражает то, насколько полно полученный реферат передает эталонный. *Точность* при данной метрике отражает то, какая часть сравниваемого реферата действительно актуальна, является частью эталона. На основе полноты R и точности P можно вычислять F-меру - удобную универсальную оценку, учитывающую и полноту и точность. А часто вовсе ограничиваются вычислением одной полноты - для задач автореферирования она важнее точности.

Метрика ROUGE-L оценивает на основе наибольшей общей подпоследовательности (LCS - longest common subseqence) лемм реферата и эталона. Данный способ лучше подстраивается под структуру предложений, однако зависим от перестановки слов в предложении. Логическим развитием ROUGE-L является ROUGE-W. Например, если мы имеем эталонное предложение R и сравниваемые предложения S1 и S2:

R = {w1 w2 w3 w4}

S1 = {w1 w2 w3 w4 w5 w6 w7 w8}

S2 = {w1 w7 w2 w8 w5 w3 w6 w4},

то у S1 и S2 метрика ROUGE-L будет одинакова [7], несмотря на то, что S1 имеет последовательные совпадения. ROUGE-W учитывает это, отдавая больший вес менее разреженным подпоследовательностям. В результате, ROUGE-W лучше коррелирует с человеческими оценками на задачах реферирования одного документа.

V. ПРЕДОБРАБОТКА ТЕКСТА

Текст перед применением алгоритмов автореферирования необходимо преобразовать к удобному для работы формату. Этапы предобработки текста:

- токенизация
- определение части речи
- лемматизация
- фильтрация стоп-слов

Сначала текст разбивается на предложения и отдельные слова (токенизация). После этого для слов необходимо определить части речи. И в русском и в английском языке есть слова, которые в зависимости от контекста, могут быть разными частями речи, и соответственно иметь разный смысл. В английском это более частое явление. Например слово "low" может быть глаголом со значением "мычать", существительным "низина", прилагательным "низкий", и наречием "низко". И это далеко не все возможные значения данного слова. Учет части речи позволяет сузить возможный спектр значениий слов. Следующий шаг, лемматизация - приведение слова к нормальной форме. Это необходимо для языков подобных русскому, в которых слова могут иметь много морфологических форм. Например слова "хорошие" и "хорошая" разные формы одной и той же леммы "хороший". Так же, к предобработке текста можно отнести фильтрацию стоп-слов (иначе - шумовых слов) - служебных и высокочастотных слов, несущих малую смысловую нагрузку, и встречающихся почти в каждом документе. Типичными стоп-словами являются предлоги, частицы, междометия, и местоимения.

VI. ГРАФОВЫЕ МЕТОДЫ

Графовые методы [8, 9] основаны на алгоритме PageRank [10], ранжирующим веб-страницы по значимости на основе ссылок между ними. Текст можно представить в виде графа, вершинами которого будут его предложения. Предложения соединяются ребрами, с весом, соответствующим сходству предложений. Для сравнения предложений обычно используют косинусную меру, совместно с мерой веса слов tf.idf, либо меры, основанные на совпадении слов. Так, сходство предложений X и Y (предложения рассматриваются как мультимножества слов), может быть вычислено следующим способом [9, 10]:

$$\text{sim}(X,Y) = \frac{2*|X \cap Y|}{|X|+|Y|}$$

Если сходство между предложениями меньше определенного порога (например, 0.1), то такие вершины оставляют не соединенными. В результате, в графе будут появляться кластеры, подграфы схожих предложений соответствующих определенной теме, а также центральные вершины кластеров - предложения, связанные с большим числом других вершин, и, скорее всего, наиболее важные в тексте. И для оценки вершин графа, на основе их центральности, как раз может быть применен алгоритм PageRank -- он может быть применен к любому графу, в том числе и взвешенному. Это итеративный алгоритм, основанный на случайных блужданиях, принимающий на вход граф и вычисляющий веса его вершин, на основе количества входящих и выходящих ребер. Наибольший вес получают центральные вершины. На основе полученных весов вершин-предложений, отбираем наиболее ценные, и из них составляем результирующий реферат.

VII. РАЗРАБОТКА МЕТРИКИ СХОДСТВА ПРЕДЛОЖЕНИЙ ДЛЯ МОДИФИКАЦИИ ГРАФОВОГО МЕТОДА, УЧИТЫВАЮЩЕЙ СЕМАНТИЧЕСКОЕ СХОДСТВО

При вычислении сходства предложений, согласно описанному подходу, никак не учитывается семантическая информация. Сходство зависит только от того, есть ли в предложениях одинаковые слова. Однако в предложении могут быть разные слова, но с одинаковым смыслом. Например, если заменить в предложении все слова на их синонимы, то сходство исходного и предложения из синонимов будет, согласно метрике из предыдущей главы, нулевое. Тогда как семантически предложения будут идентичны. Соответственно, необходимо учитывать сходство слов при вычислении сходства предложений. Нужно ввести метрику сходства мультимножеств (мы рассматриваем предложения как множества слов с повторениями), основанную не на полном совпадении, а на сходстве элементов этих мультимножеств. И, в идеале, на граничных значениях сходства, когда все слова либо полностью не похожи, либо совпадают, данная метрика должна выдавать те же результаты, что и выбранная в базовом алгоритме, основанная на полностью совпадающих словах.

Данную проблему можно частично решить, если вычислять сходство предложений на основе семантического сходства. Семантическое сходство слов можно оценить на основе их близости в семантической сети – сети, вершинами которой являются слова или группы слов, а дугами - семантические связи. Примером такой сети является *wordNet* для английского языка.

Метрику сходства введем следующим образом: будем взаимно однозначно сопоставлять слова одного предложения словам другого. Если предложения разной длины, то часть слов будет без пары и не входить в сопоставление. Для всех уникальных сопоставлений вычислим сходства входящих в него пар слов и просуммируем эти значения - для каждого сопоставления получим вес. Выберем из этих весов максимальный и поделим его на суммарную длину предложений, для нормализации. Таким образом, сходство предложений X и Y, согласно этой метрике, вычислим следующим образом:

$$\text{sent\_sim}(X,Y) = \frac{\max_{M \in \text{Matches}(X,Y)} (\sum_{w,v \in M} \text{sim}(w,v))}{|X| + |Y|}$$

где Matches(X,Y) - множества всех возможных сопоставлений слов X и Y, sim(x,y) - сходство между словами x и y [6], принимает значения на интервале от 0 до 1, единица соответствует полной синонимии, ноль - отсутствию какой бы то ни было семантической связи.

Такая метрика достаточно понятна, однако у нее есть существенный недостаток: сложность вычисления. Алгоритмическая сложность растет как факториал от длины предложений. Поэтому будем использовать упрощенную версию:

$$\text{sent}_{\text{sim}}(X,Y) = \frac{2 \cdot \min(\sum_{w \in X} \max_{v \in Y} (\text{sim}(w,v)), \sum_{v \in Y} \max_{w \in X} (\text{sim}(w,v)))}{|X| + |Y|}$$

Она не всегда выдает те же результаты, что и исходная, но она значительно проще для вычисления - ее алгоритмическая сложность растет как произведение длин предложений. Остальная часть алгоритма остается без изменений.

VIII. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ РАЗРАБОТАННОГО МЕТОДА АВТОРЕФЕРИРОВАНИЯ

Программная реализация выполнена на языке программирования python. Он достаточно удобен для научного программирования - интерпретируемый, имеет хорошую читаемость и огромный набор пользовательских пакетов, в том числе и для обработки естественного языка.

Библиотеки NLTK предоставляли методы для вычисления сходства слов, однако в этих методах не было ограничения на максимальное расстояние поиска. Методы NLTK честно обходят всю сеть и ищут путь, даже если слова удалены на огромные расстояния. Из-за этого эти методы были исключительно медленными. К тому же в wiki_ru_wordnet эти методы отсутствовали полностью. Для решения этой проблемы был реализован восходящий поиск по иерерхии гиперонимов, с ограничением на максимальное расстояние, после которого сходство слов считается нулевым.

B. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МОДИФИЦИРОВАННОГО ГРАФОВОГО МЕТОДА

Программа реализована в виде отдельного модуля. Вначале происходит предобработка – текст разбивается на предложения и слова, выводятся части речи слов, проводится лемматизация и удаление стоп-слов. В итоге, во внутреннем представлении слова хранятся как кортеж из нормальной формы слова и его части речи. Для английского языка NLTK предоставляет все необходимые методы. Для русского используется комбинация NLTK, pymorphy2 и регулярных выражений. Из-за этого, для русского и английского языка сделаны отдельные модули.

Перед работой методов, использующих семантическое сходство, строится матрица сходства слов. Матрица сходства A имеет размер n*n, где n - количество уникальных слов в тексте. Элементом a_{ij} матрицы выступает сходство между i-ым и j-ым словом. Заметим, что эта матрица симметрична, и на ее главной диагонали исключительно единицы.

A. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ВСПОМОГАТЕЛЬНЫХ СРЕДСТВ

Были использованы следующие сторонние пакеты и библиотеки:

- ✓ NLTK (Natural Language ToolKit) - набор библиотек для обработки естественных языков [11], частично пригодный для обработки русскоязычных текстов;
- ✓ pymorphy2 - морфологический анализатор русского языка [12];
- ✓ wiki_ru_wordnet - семантическая сеть типа WordNet для русского языка, составленная из данных русского Викисловаря;
- ✓ rouge - пакет для оценки аннотаций метриками ROUGE;
- ✓ networkx - пакет для работы с графами [13].

Поэтому, достаточно хранить n(n-1)/2 значений. Предложения попарно оцениваются на сходство. После чего на основе полученных оценок строится полный граф, вершинами которого являются предложения текста, а каждое ребро между вершинами имеет вес, равный сходству предложений. Если же сходство нулевое, то ребро удаляется из графа. Вершины полученного графа ранжируют при помощи алгоритма PageRank, доступного в пакте networkx. В результате работы описанных методов предложения получают определенный вес, описывающий степень их значимости. Из этих предложений отбираются N наиболее значимых, где число N задается пользователем, отобранные предложения упорядочиваются согласно их позиции в тексте и выдаются как итоговая аннотация.

IX. ПРИМЕР РАБОТЫ

Для составления реферата разработанному методу был предложен следующий текст.

Исходный текст:

Система отслеживания эффективности, применяемая на складах Amazon, за год автоматически уволила несколько сотен работников компании в американском городе Балтимор. Об этом сообщает издание The Verge, получившее доступ к письму представителя Amazon, направленному в Национальное управление по вопросам трудовых отношений США.

Amazon, один из крупнейших онлайн-ритейлеров мира, владеет десятками складов для сортировки товаров. После получения заказа компания отправляет данные о нем на склад, после чего один из

работников начинает работу с ним. Сначала он собирает товары для заказа, затем сканирует их для учета в системе и загружает их в коробку или другую упаковку. Для оценки эффективности работников Amazon использует несколько метрик, в том числе «Time Off Task», отражающую время, в течение которого работник не выполнял свои обязанности. Оно рассчитывается исходя из перерывов в сканировании товаров.

Ранее Amazon уже подвергали критике за слишком жесткие требования к работникам складов. Журналисты The Verge выяснили, что отслеживанием эффективности и принятием решений о расторжении контрактов с неэффективными работниками в Amazon занимается специальное программное обеспечение. Издание получило доступ к письму адвоката компании в Национальное управление по вопросам трудовых отношений США, воспользовавшись законом о свободе информации, обязывающим органы исполнительной власти США предоставлять гражданам документы и другую информацию по их запросу.

В письме, касающемся обвинения компании в нарушении трудового законодательства одним из работников, представитель Amazon рассказал о том, как именно система формирует уведомления о разрыве контрактов. Система автоматически отслеживает эффективность выполнения работы всех сотрудников складов и формирует выговоры или уведомления о расторжении контракта без участия начальства сотрудников, хотя они имеют право отменить решение системы, если посчитают его ошибочным. Кроме того, уволенные сотрудники могут обжаловать свое увольнение. В период с августа 2017 года по сентябрь 2018 года из-за работы системы были уволены около 300 сотрудников склада Amazon в городе Балтимор.

В письме отмечается, что, если время, проведенное не за работой (метрика «Time Off Task») составляет от 30 минут до часа в день, работник получает обычное письменное предупреждение. Если от часа до двух — он получает последнее письменное предупреждение. А если он не выполняет работу на протяжении двух и более часов, за этим следует увольнение. За год сотрудник может получить до пяти обычных письменных выговоров и не более одного последнего, иначе система так же автоматически формирует уведомление об увольнении.

Сейчас Amazon активно автоматизирует склады и заменяет работников на роботов. На складах компании уже работают роботы-транспортировщики, выполняющие свою работу в 4-5 раз эффективнее людей. Кроме того, до 2018 года компания проводила соревнования Amazon Picking Challenge среди команд инженеров, разрабатывающих роботов для захвата и перемещения предметов. Предполагается, что наработки, полученные во время конкурса, помогут Amazon автоматизировать сортировку, а также упаковку товаров в коробки перед доставкой клиентам.

В результате работы графового алгоритма (см. Рис. 7) был получен реферат:

Система отслеживания эффективности, применяемая на складах Amazon, за год автоматически уволила несколько сотен работников компании в американском городе Балтимор. После получения заказа компания отправляет данные о нем на склад, после чего один из работников начинает работу с ним. В письме, касающемся обвинения компании в нарушении трудового законодательства одним из работников, представитель Amazon рассказал о том, как именно система формирует уведомления о разрыве контрактов. Система автоматически отслеживает эффективность выполнения работы всех сотрудников складов и формирует выговоры или уведомления о расторжении контракта без участия начальства сотрудников, хотя они имеют право отменить решение системы, если посчитают его ошибочным. Сейчас Amazon активно автоматизирует склады и заменяет работников на роботов.

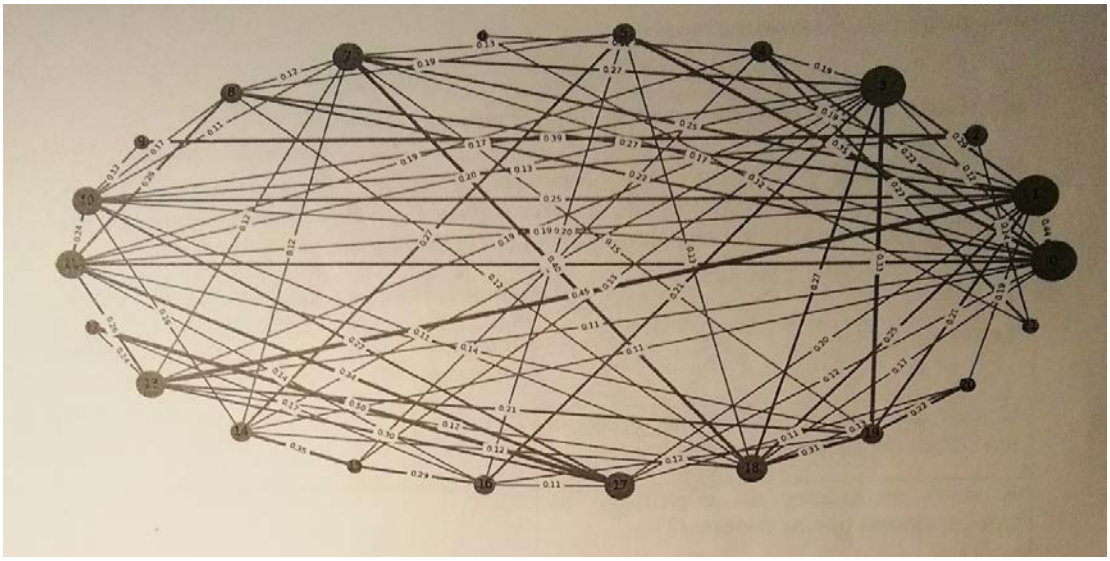

Рис. 7 - визуализация графового метода

А это - результат работы модифицированного графового алгоритма (см. Рис. 8):

Система отслеживания эффективности, применяемая на складах Amazon, за год автоматически уволила несколько сотен работников компании в американском городе Балтимор. После получения заказа компания отправляет данные о нем на склад, после чего один из работников начинает работу с ним. Система автоматически отслеживает эффективность выполнения работы всех сотрудников складов и формирует выговоры или уведомления о расторжении контракта без участия начальства сотрудников, хотя они имеют право отменить решение системы, если посчитают его ошибочным. В период с августа 2017 года по сентябрь 2018 года из-за работы системы были уволены около 300 сотрудников склада Amazon в городе Балтимор. За год сотрудник может получить до пяти обычных письменных выговоров и не более одного последнего, иначе система так же автоматически формирует уведомление об увольнении.

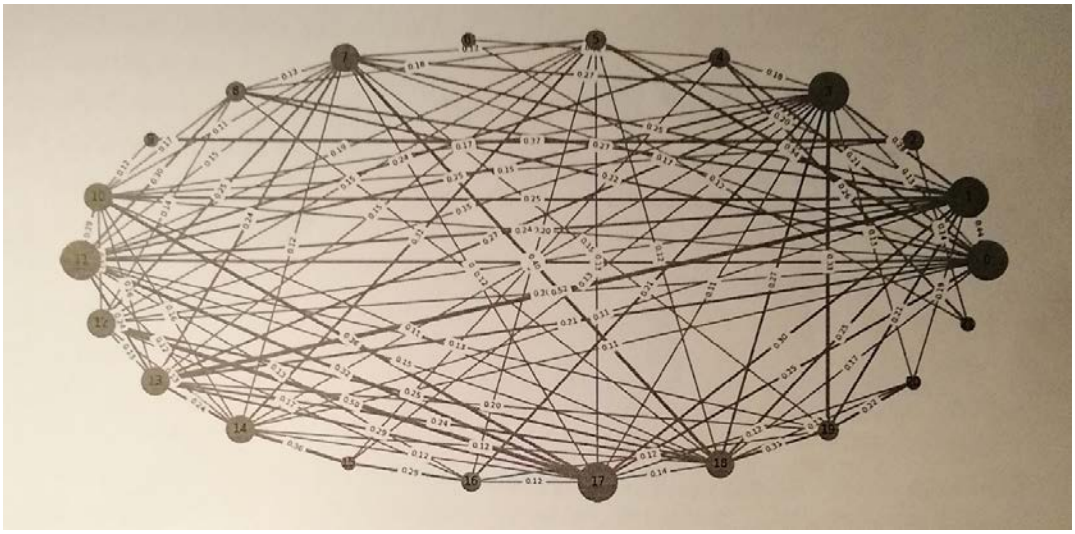

Рис. 8 - визуализация разработанного модифицированного графового метода

X. ОЦЕНКА РЕАЛИЗОВАННОГО МЕТОДА

Для оценки качества получаемых рефератов использованы следующие наборы документов с рефератами:

- англоязычные новости BBC, порядка 2500, поделенные на 5 категорий
- англоязычные новости из различных индийских источников, порядка 5000 новостей
- аннотации русских новостей, порядка 150+

Реализованные алгоритмы оценивались при помощи метрик ROUGE, основанных на сравнении полученных рефератов с эталонными. А именно, вычислялась полнота согласно метрике ROUGE-W. Для графового метода была реализована мера Сёренсена. Результаты по всем документам представлены в таблицах 1,2,3.

Табл. 1 Результаты оценки, метрика ROUGE-1

	Новости BBC	Индийские новости	Русские новости
Графовый метод (мера Сёренсена)	54.82	41.20	46.13
Модифицированный графовый метод	56.10	42.54	47.97

Табл. 2 Результаты оценки, метрика ROUGE-2

	Новости BBC	Индийские новости	Русские новости
Графовый метод (мера Сёренсена)	48.15	19.23	37.46
Модифицированный графовый метод	47.97	19.41	39.78

Табл. 3 Результаты оценки, метрика ROUGE-W

	Новости BBC	Индийские новости	Русские новости
Графовый метод (мера Сёренсена)	40.28	28.18	43.87
Модифицированный графовый метод	41.02	28.64	45.85

XI. ЗАКЛЮЧЕНИЕ

Работа выполнена в рамках кафедральной НИР “Математическое и программное обеспечение перспективных систем обработки символьной информации”. В работе рассмотрены существующие подходы к автоматическому реферированию текста. На их основе предложен метод, учитывающий семантическое сходство слов. Разработана модификация графового метода, учитывающая синонимию и позволяющая получить более качественный реферат. Программно реализованы как базовая, так и модифицированная версии, а также проведено сравнение их качества на русских и английских текстах при помощи метрик автоматической оценки рефератов. По результатам оценки (см. табл. 1, 2, 3) видно улучшение у модифицированного графового метода по сравнению с обычным, особенно на русскоязычных текстах.

БЛАГОДАРНОСТИ

Авторы выражают благодарность профессору МГУ имени М.В.Ломоносова Сергею Юрьевичу Соловьеву за полезные замечания, сделанные в ходе подготовки статьи.

БИБЛИОГРАФИЯ

[1] Автоматическое реферирование и аннотирование [Электронный ресурс]. – Электрон. дан. – URL: https://refdb.ru/look/1532518.html. (дата обращения 16.10.2021)

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Saeid Safaeii др. Text Summarization Techniques: A Brief Survey [Электронный ресурс]. – Электрон. дан. – URL: https://arxiv.org/pdf/1707.02268.pdf. (дата обращения 15.12.2021)

[3] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, Chandan K. Reddy. Neural Abstractive Text Summarizationwith Sequence-to-Sequence Models [Электронный ресурс]. – Электрон.дан. – URL: https://arxiv.org/pdf/1812.02303.pdf. (дата обращения 20.07.2021)
[4] Tomas Mikolov, Ilya Sutskever. Distributed Representations of Words and Phrasesand their Compositionality [Электронный ресурс]. – Электрон.дан. – URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf. (дата обращения 20.11.2021)
[5] George A. Miller, Richard Beckwith. Introduction to WordNet: An On-line Lexical Database [Электронный ресурс]. – Электрон.дан. – URL: http://wordnetcode.princeton.edu/5papers.pdf. (дата обращения 11.11.2021)
[6] Dan Jurafsky. Word Meaning and Similarity [Электронный ресурс]. – Электрон.дан. – URL: https://web.stanford.edu/class/cs124/lec/sem. (дата обращения 21.03.2019)
[7] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries [Электронный ресурс]. - Электрон.дан. - URL: https://www.aclweb.org/anthology/W04-1013. (дата обращения 08.11.2020)
[8] Rada Mihalcea, Paul Tarau. TextRank: Bringing Order into Texts [Электронный ресурс]. – Электрон.дан. – URL: https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf. (дата обращения 22.10.2021)
[9] Güneş Erkan, Dragomir R. Radev LexRank: Graph-based Lexical Centrality as Salience in Text Summarization [Электронный ресурс]. – Электрон.дан. – URL: https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume22/erkan04a-html/erkan04a.html. (дата обращения 02.11.2021)
[10] The PageRank Citation Ranking: Bringing Order to the Web [Электронный ресурс]. – Электрон.дан. – URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf. (дата обращения 10.12.2021)
[11] NLTK documentation [Электронный ресурс]. – Электрон.дан. - URL: https://www.nltk.org. (дата обращения 09.12.2020)
[12] Морфологический анализатор pymorphy2 [Электронный ресурс]. – Электрон.дан. - URL: https://pymorphy2.readthedocs.io/en/latest/ (дата обращения 01.12.2021)
[13] Networkx documentation [Электронный ресурс]. – Электрон.дан. - URL: https://networkx.github.io/documentation/stable. (дата обращения 09.09.2021)

Modification of the graph method for automatic abstraction tasks taking into account synonymy

Irina Polyakova, Igor Zaitsev

Abstract **- The article discusses existing approaches to automatic text abstracting. The creation of an abstract obviously requires an understanding of the text at the level of pragmatics. However, at the moment, reliable semantic analysis is still not available for computers, especially not the analysis of pragmatics. Widespread methods based on neural networks can take into account semantics thanks to special vector representations of words. But the rest of the automatic referencing methods rely on morphology and syntax. However, part of the semantics is still available to them - there are thesauruses and networks, as well as algorithms that allow you to establish a semantic connection between individual words, such as their semantic similarity, in particular, synonymy.**

A method is proposed that takes into account the semantic similarity of words. A modification of the graph method has been developed that takes into account synonymy and allows for a better abstract. The basic and modified versions of the graph method are implemented programmatically for automatic referencing tasks, taking into account synonymy. The comparison of their quality in Russian and English texts using automatic evaluation metrics of abstracts is carried out. The evaluation results show an improvement in the modified graph method compared to the usual one, especially in Russian-language texts.

Keywords **- automatic text abstracting, graph method, metrics of automatic evaluation of abstracts, modification taking into account synonymy, semantic similarity of words, synonymy**

REFERENCES

[1] Automatic abstracting and annotation [Electronic resource]. - URL: https://refdb.ru/look/1532518.html . (Accessed 10/16/2021)

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Saeid Safaei, and others Text Summarization Techniques: A Brief Survey [Electronic resource]. - URL: https://arxiv.org/pdf/1707.02268.pdf. (Accessed 15.12.2021)

[3] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, Chandan K. Reddy. Neural Abstract Text Summarizationwith Sequence-to-Sequence Models [Electronic resource]. - URL: https://arxiv.org/pdf/1812.02303.pdf. (Accessed 20.07.2021)

[4] Tomas Mikolov, Ilya Sutskever. Distributed Representations of Words and Phrasesand their Compositionality [Electronic resource]. - URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf. (Accessed 20.11.2021)

[5] George A. Miller, Richard Beckwith. Introduction to WordNet: An On-line Lexical Database [Electronic resource]. - URL: http://wordnetcode.princeton.edu/5papers.pdf. (Accessed 11.11.2021)

[6] Dan Jurafsky. Word Meaning and Similarity [Electronic resource]. - URL: https://web.stanford.edu/class/cs124/lec/sem . (Accessed 21.03.2019)

[7] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries [Electronic resource]. - URL: https://www.aclweb.org/anthology/W04-1013. (Accessed 08.11.2020)

[8] Rada Mihalcea, Paul Tarau. TextRank: Bringing Order into Texts - URL: https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf. (Accessed 10/22/2021)

[9] Guneş Erkan, Dragomir R. Radev LexRank: Graph-based Lexical Centrality as Salience in Text Summarization [Electronic resource]. - URL: https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume22/erkan04a-html/erkan04a.html. (Accessed 02.11.2021)

[10] The PageRank Citation Ranking: Bringing Order to the Web [Electronic resource]. - URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf. (Accessed 10.12.2021)

[11] NLTK documentation [Electronic resource]. - URL: https://www.nltk.org . (Accessed 09.12.2020)

[12] Morphological analyzer pymorphy2 [Electronic resource]. - URL: https://pymorphy2.readthedocs.io/en/latest / (Accessed 01.12.2021)

[13] Networkx documentation [Electronic resource]. - URL: https://networkx.github.io/documentation/stable . (Accessed 09.09.2021)