Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем

Кафедра информационно-аналитических систем

Новосёлова Анастасия Максимовна

Исследование методов автоматического реферирования текстов

Выпуская квалификационная работа

Научный руководитель:
к. ф.-м. н., доцент Михайлова Е. Г.

Рецензент:
Руководитель научной лаборатории
Digital design, Ашихмин И.А.

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems

Sub-Department of Analytical Information Systems

Novosyolova Anastasiia Maksimovna

Investigation of automatic text summarization methods

Graduation Project

Scientific supervisor:
Associate Professor, Ph.D, Mikhailova E. G.

Reviewer:
Head of R&D Lab. Digital design, Ashikhmin I. A.

Saint-Petersburg
2017

Содержание

Введение **4**

1 Постановка задачи **5**

2 Обзор литературы **6**

2.1 Классификация методов автоматического реферирования . . 6

2.2 Описание используемых алгоритмов 8

2.3 Описание используемых метрик 11

3 Эксперименты **13**

3.1 Описание тестового набора данных 13

3.2 Оценка применимости алгоритмов к текстам на русском языке 15

3.3 Оценка алгоритмов и их сравнение 16

3.4 Модификации алгоритма TextRank и их сравнение 21

Заключение **23**

Список литературы **24**

Введение

Во многих документах зачастую содержится большое количество текста, который не несет существенную информацию. Хорошим примером таких документов являются различные новостные статьи. Зачастую людям, у которых нет времени на прочтение полного текста, нужно ознакомиться с кратким содержанием новости. Также очень удобно прочитать лишь краткую аннотацию новости для того, чтобы понять стоит ли читать новость полностью. Таким образом, появляется необходимость сокращать объём документа, выделяя наиболее значимую часть текста, называемую рефератом. Ручное реферирование — сложная, рутинная работа, требующая дополнительных сотрудников, поэтому целесообразно использовать системы автоматического реферирования текстов.

Задача автоматического реферирования текстов очень популярна среди исследователей. Существует большое количество публикаций, в которых описываются различные алгоритмы автоматического реферирования. Однако, различные авторы используют различные метрики для оценки предложенных ими алгоритмов. Кроме того, оценка алгоритмов производится, в основном, на англоязычных наборах документов. В связи с этим применение алгоритмов автоматического реферирования текстов к русскоязычному набору документов и их сравнение является актуальной задачей.

Данная работа выполнена при поддержке компании Digital Design, в научной лаборатории которой проводится исследование методов автоматического реферирования текстов для использования их в одном из проектов.

1 Постановка задачи

В рамках настоящей работы были поставлены следующие задачи:

- Изучить существующие подходы к автоматическому реферированию текстов и способы оценки их качества
- Реализовать работу нескольких алгоритмов автоматического реферирования и оценку их качества
- Оценить качество работы алгоритмов с помощью нескольких метрик в применении к набору русскоязычных новостных текстов
- Сравнить работу алгоритмов и исследовать возможность модификации алгоритма с лучшими показателями качества

2 Обзор литературы

2.1 Классификация методов автоматического реферирования

Первые публикации по теме методов автоматического реферирования текстов появились ещё в 1958 г. [11] С тех пор было разработано большое количество методов, среди которых можно выделить два направления: [1–3]

- реферирование одиночных документов (реферат формируется на основе одного документа);
- многодокументное реферирование (реферат формируется на основе нескольких документов, например, нескольких статей одной тематики)

Также существующие подходы можно разделить на две категории по типу реферата: [1–3]

- автоматическое реферирование, основанное на выделении из первичных документов ключевых фраз (фрагментов, предложений), которые добавляются в реферат без изменений в порядке их появления в тексте (extractive summarization);
- автоматическое реферирование, основанное на выделении наиболее существенной информации и генерации новых текстов, содержательно обобщающих оригинальные тексты (abstractive summarization).

В данной работе рассмотрены методы экстракционного реферирования одиночных документов.

На основе классификаций в [1, 2] можно представить следующую классификацию подходов к экстракционному реферированию:

- Статистические подходы

  Именно статистические подходы применялись первыми исследователями в области автоматического реферирования текстов. Они основаны на статистических характеристиках текста, которые помогают выделить наиболее важные предложения из текста. Такими характеристиками являются, например, частота встречаемости термов, расположение предложения в тексте, длина предложения, наличие в предложении

имен собственных, наличие чисел в предложении, позитивные и негативные ключевые слова, совместное появление слов, вес TF-IDF. После выделения характеристик производится подсчёт суммы их значений для каждого предложения текста. Таким образом, каждому предложению присваивается определенный вес. Реферат формируется из предложений с наибольшим весом. Длина реферата определяется желаемым количеством предложений в нём или желаемой степенью сжатия текста.

Данный подход прост в реализации и требует сравнительно мало ресурсов компьютера, однако дает не очень высокие оценки качества реферирования.

- Подходы, основанные на графах

  Эти подходы основаны на идее представления текста в виде графа. Предложения представляются вершинами графа, а связи между ними — взвешенными ребрами. С помощью различных алгоритмов каждой вершине графа, обозначающей определенное предложение текста, можно присвоить вес на основе весов ребер. Реферат формируется аналогично формированию реферата при статистическом подходе.

  Данные подходы широко используются на практике. Один из таких методов реализован в библиотеке gensim для Python.

- Подходы, использующие машинное обучение

  Для решения задачи автоматического реферирования текстов применяют и методы обучения с учителем, и методы обучения без учителя.

  В случае применения обучения с учителем классификатор определяет каждое предложение тестового документа в один из двух классов: “включить в реферат”/“не включать в реферат”. Недостаток такого подхода состоит в том, что для обучения классификатора необходим тренировочный набор документов и соответствующих им экстракционных рефератов. Для получения такого набора нужно создавать рефераты вручную, а это очень трудоемкий процесс.

  Также для решения задачи автоматического реферирования текстов применяют различные алгоритмы кластеризации.

- Подходы, использующие семантические связи

  Данные подходы используют семантические связи между словами. Чаще используются в применении к абстракционному реферированию, однако существует алгоритмы, применимые и к экстракционному реферированию текстов.

2.2 Описание используемых алгоритмов

Обобщенно схему экстракционного реферирования можно представить следующим образом:

1. Предварительная обработка изначального документа: удаление стоп-слов, стемминг, разбиение текста на предложения.

2. Формирование реферата из предложений исходного документа с помощью какого-либо алгоритма.

Описанные далее алгоритмы выполняют шаг (2) схемы.

2.2.1 Алгоритм TextRank

Данный алгоритм является представителем подходов, основанных на графах. Его применение к задаче автоматического реферирования текстов представлено в работе [4]. Он заключается в следующем:

1. По тексту строится взвешенный неориентированный граф, вершины в котором обозначают предложения текста. Весом ребра между двумя вершинами является степень схожести двух предложений, соответствующих вершинам. Она вычисляется, как количество совпадающих слов в предложениях, нормированное суммарной длиной этих предложений.

2. Исходя из весов ребер, с помощью итерационного процесса каждой вершине присваивается вес по следующей формуле:

$$W(V_i) = (1-d) + d \cdot \sum_{V_j \in Inc(V_i)} \frac{w_{ji}}{\sum_{V_k \in Inc(V_j)} w_{jk}} W(V_j),$$

   где V_i, V_j — вершины графа;
   $Inc(V_i)$ — множество вершин, смежных с вершиной V_i;

w_{ij} — вес ребра между вершинами V_i, V_j;
d — коэфициент затухания, который в данном алгоритме равен 0.85;
Итерационный процесс завершается, как только веса вершин перестают меняться более, чем на 0.0001.

3. После вычисления весов вершин они упорядочиваются по убыванию значения веса и в реферат включаются предложения, соответствующие первым n вершинам, где n — желаемое количество предложений в реферате.

2.2.2 Алгоритм k-means

Применение алгоритма кластеризации k-means к задаче автоматического реферирования было предложено в работе [5].

1. Совокупность предложений текста рассматривается как коллекция документов. Каждое предложение исходного текста представляется в виде вектора длины n, где n — количество уникальных слов в тексте. i-ый элемент вектора равен показателю TF-IDF для i-го слова текста, если оно встречается в предложении, и 0 в противном случае.

2. Производится кластеризация предложений с помощью алгоритма k-means, в котором количество кластеров равно желаемому количеству предложений в реферате.

3. Реферат формируется из предложений, наиболее близких к центроидам полученных кластеров.

2.2.3 Алгоритм LSA

Данный алгоритм основан на скрытом семантическом анализе (Latent Semantic Analysis), позволяющем получить неявное представление текстовой семантики на основе совместной встречаемости слов. Впервые использование этого метода для реферирования одиночных документов было предложено в [7]. В данной работе рассмотрен несколько модифицированный метод, который был описан в [8]. Он заключается в следующем:

1. Пусть A — матрица терм-предложение, полученная по исходному документу. Её размер равен $n \times m$, где n — количество термов в документе, m — количество предложений. Элемент a_{ij} этой матрицы равен частоте встречаемости терма i в тексте, если этот терм встречается в предложении j, и 0 в противном случае.

2. К полученной матрице применяется сингулярное разложение:

$$A = U\Sigma V^T,$$

где $U = [u_{ij}]$ — ортонормированная матрица размера $n \times m$, $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_m)$ — диагональная матрица, $V = [v_{ij}]$ — ортонормированная матрица размера $m \times m$.

Если $rank(A) = r$, то выполняется:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \sigma_m = \ldots = 0$$

С точки зрения семантики сингулярное разложение матрицы A интерпретируется как разбиение исходного документа на r концепций (тем). Каждый элемент v_{ij} матрицы V отражает степень информативности предложения j по теме i. При этом значение σ_i матрицы Σ отражает степень важности темы i в исходном документе.

3. Каждому предложению s_k исходного документа присваивается вес по формуле:

$$s_k = \sqrt{\sum_{i=1}^{m} {v_{ik}}^2 \cdot {\sigma_i}^2}$$

Т.о. больший вес получают предложения, наиболее информативные по одной из тем документа, при этом учитывается и степень важности концепции в документе.

4. Значения весов предложений упорядочиваются по убыванию, и в реферат включаются предложения, соответствующие первым l значениям, где l — желаемое количество предложений в реферате.

2.3 Описание используемых метрик

Для оценки качества работы алгоритма автоматического реферирования необходим набор документов с приложенными к ним рефератами, называемыми образцовыми. Рефераты, полученные с помощью алгоритмов далее для удобства будем называть автоматическими.

В работе [9] был предложен пакет для оценки качества алгоритмов автоматического реферирования текстов ROUGE, метрики которого имеют высокую корреляцию с человеческими оценками. Рассмотрим подробнее основные метрики данного пакета.

ROUGE – N

Метрика ROUGE – N основана на вычислении количества n-грамм, встречающихся и в образцовом, и в автоматическом рефератах, а именно:

$$ROUGE - N = \frac{Count_{match}(gram_n)}{Count_{ref}(gram_n)},$$

где $Count_{match}(gram_n)$ — количество n-грамм, появляющихся и в автоматическом реферате, и в образцовом; $Count_{ref}(gram_n)$ — количество n-грамм в образцовом реферате.

В данной работе использованы метрики ROUGE – 1 и ROUGE – 2, основанные на вычислении количества униграмм и биграмм соответственно.

ROUGE – L

Пусть X — последовательность слов образцового реферата длины n, Y — последовательность слов автоматического реферата длины m, $LCS(X,Y)$

— длина наибольшей общей подпоследовательности между X и Y. Тогда:

$$P_{lcs} = \frac{LCS(X,Y)}{m}$$

$$R_{lcs} = \frac{LCS(X,Y)}{n}$$

$$ROUGE - L = \frac{2P_{lcs}R_{lcs}}{P_{lcs} + R_{lcs}}$$

ROUGE – S

Пусть $SKIP2(X,Y)$ — количество биграмм с пропусками, встречающихся и в образцовом реферате X длины n (слов), и в автоматическом реферате Y длины m. Тогда:

$$P_{skip2} = \frac{SKIP2(X,Y)}{C_m^2}$$

$$R_{skip2} = \frac{SKIP2(X,Y)}{C_n^2}$$

$$ROUGE - S = \frac{2P_{skip2}R_{skip2}}{P_{skip2} + R_{skip2}}$$

3 Эксперименты

3.1 Описание тестового набора данных

Тестовый набор данных был предоставлен компанией Digital Design.

Набор состоит из 35300 текстов новостей с приложенными к ним рефератами, которые были собраны с различных русскоязычных новостных сайтов. Пример текста и его реферата из набора представлен в таблице 3.

Статистика по набору данных представлена в таблицах 1 – 2. Для каждой характеристики были вычислены:

- avg — среднее значение
- min — минимальное значение
- max — максимальное значение
- std — среднеквадратическое отклонение

Характеристика	avg	min	max	std
Количество предложений	27	1	599	36
Количество уникальных слов	275	15	3426	245

Таблица 1: Статистика по текстам новостей

Характиристика	avg	min	max	std
Количество предложений	2	1	27	1
Количество уникальных слов	31	4	285	15
Количество уникальных слов*	22	2	195	10
* — с учетом стемминга и удаления стоп-слов				

Таблица 2: Статистика по рефератам

Текст новости
Алан Рикман скончался в Лондоне, эту новость, как передает The Guardian, уже подтвердила семья актера. В статье отмечается, что Рикман обрел целый легион поклонников, исполнив роль профессора Снейпа в фильмах о Гарри Поттере. И это несмотря на то, что вплоть до последних кадров зритель считает Снейпа отрицательным персонажем. Лишь после того, как Волан-де-Морт отдал приказ своей змее Нагайне убить Снейпа, тот, умирая, передает свои воспоминания Гарри Поттеру. Так юный волшебник узнает, что Снейп все это время хранил любовь к его матери и являлся двойным агентом, помогающим Дамблдору - директору школы магии Хогварстс - и самому Гарри Поттеру бороться с Волан-де-Мортом. Что касается Алана Рикмана, то на самом деле как актер прославился он значительно ранее, сыграв террориста в первом фильме о "Крепком Орешке вместе с Брюсом Уиллисом. Затем была роль шерифа в фильме "Робин Гуд: принц воров". Актер являлся лауреатом премий "Золотой глобус Эмми BAFTA. В личной жизни Алан Рикман хранил верность Риме Хортон, с которой познакомился, когда молодым людям было 19 и 18 лет. Но детей у пары не было.
Образцовый реферат
Один из самых любимых британских актеров Алан Рикман скончался в возрасте 69 лет. Актер страдал от онкологического заболевания. Он сыграл множество запоминающихся ролей, но все же наибольшую популярность у широкой публики снискал, снявшись в подростковом фэнтези по книгам о Гарри Поттере. Там он сыграл преподавателя зельеварения Северуса Снейпа.

Таблица 3: Пример одной записи из набора данных

Следует отметить несколько особенностей используемого набора данных:

- Образцовые рефераты имеют небольшую длину: 1-3 предложения, когда как исходные тексты имеют среднюю длину 27 предложений. Таким образом, сжатие исходного текста, в среднем, производится в 13.5 раз.
- Во многих парах реферат-текст информация, содержащаяся в реферате, упущена в самом тексте. Так, в примере, приведенном в таблице 3, в тексте новости не упоминается о причине смерти актера Алана Рикмана. Эта особенность объясняется тем, что данные были собраны автоматически с различных новостных сайтов, на которых человек

сначала получает информацию из реферата новости, а затем из полного её текста. Поэтому некоторая информация из реферата в полном тексте новости не дублируется.

- Рефераты не являются экстракционными, т.е. предложения в них переформулированы, некоторые слова заменены синонимами.

- Набор исходных текстов разнообразен: в нём присутствуют и совсем короткие тексты (длиной 15 слов), и достаточно длинные (длиной более 3400 слов).

3.2 Оценка применимости алгоритмов к текстам на русском языке

Выбранные алгоритмы и модуль оценки были реализованы на языке Python с использованием оптимизированых библиотек. Для стемминга был использован Snowball stemmer для русского языка, список русских стоп-слов был взят из библиотеки stop-words. Количество предложений для включения в автоматический реферат определялось количеством предложений в соответствующем образцовом реферате.

Для оценки применимости алгоритмов к текстам на русском языке было проведено сравнение каждого из рассмотренных алгоритмов с алгоритмом, который выбирает предложения для формирования реферата случайным образом.

В таблицах 4 — 6 представлена разность между оценками выбранного и случайного алгоритмов. Оценка была проведена по метрикам, описанным в разделе 2.3. Каждая метрика была вычислена тремя способами:

- basic — вычисление на автоматическом и образцовом рефератах в их изначальном виде

- stem — с проведением стемминга в рефератах

- stem, stop-words — с проведением стемминга и удалением стоп-слов

	TextRank		
	basic	stem	stem, stop-words
ROUGE – 1	0.08	0.11	0.11
ROUGE – 2	0.04	0.05	0.04
ROUGE – L	0.03	0.04	0.04
ROUGE – S	0.01	0.02	0.02

Таблица 4: Оценка алгоритма TextRank

	k-means		
	basic	stem	stem, stop-words
ROUGE – 1	0.04	0.06	0.07
ROUGE – 2	0.03	0.03	0.02
ROUGE – L	0.02	0.03	0.04
ROUGE – S	0.01	0.01	0.02

Таблица 5: Оценка алгоритма k-means

	LSA		
	basic	stem	stem, stop-words
ROUGE – 1	0.08	0.10	0.10
ROUGE – 2	0.04	0.04	0.04
ROUGE – L	0.02	0.03	0.03
ROUGE – S	0.01	0.02	0.01

Таблица 6: Оценка алгоритма LSA

Из полученных значений видно, что выбранные алгоритмы превосходят алгоритм, случайно выбирающий предложения. Кроме того, наибольшая разница в оценках наблюдается при оценке с удалением стоп-слов и проведением стемминга, а именно эта оценка наиболее объективна с точки зрения человека.

3.3 Оценка алгоритмов и их сравнение

Сравнение алгоритмов было проведено по метрикам, вычисленным на рефератах с проведением стемминга и удалением стоп-слов.

Результаты оценки алгоритмов представлены в таблице 7.

	TextRank	LSA	k-means
ROUGE – 1	0.25	0.24	0.21
ROUGE – 2	0.08	0.08	0.06
ROUGE – L	0.12	0.11	0.12
ROUGE – S	0.04	0.03	0.04

Таблица 7: Сравнение алгоритмов

В таблице 8 представлено значение метрики ROUGE – 1, вычисленной с проведением стемминга и удалением стоп-слов, из оригинальной статьи об алгоритме TextRank. В ней оценка проводилась на англоязычном наборе данных, состоящем из 567 новостных статей с приложенными рефератами. Следует отметить некоторые особенности этого набора данных [4, 5]:

- Набор был подготовлен экспертами, которым было дано задание составить экстракционные рефераты к статьям. На самом деле, они получились не строго экстракционными, однако, более экстракционными, чем в используемом в настоящей работе наборе.
- Длина рефератов в наборе в среднем в 5 раз меньше исходных текстов. Т.е. степень сжатия текста в наборе существенно меньше, чем в используемом.
- Размер этого набора данных существенно меньше размера используемого набора.

	ROUGE – 1
Полученная оценка	0.25
Оценка из статьи [4]	0.42

Таблица 8: Сравнение полученной оценки алгоритма TextRank с оценкой авторов статьи

Значение полученной оценки алгоритма ниже, чем значение оценки из статьи. И, в целом, полученные оценки (см. табл. 7) нельзя назвать высокими. Это объясняется рядом причин:

- Русский язык очень разнообразен, в нём присутствует большое количество синонимичных слов и выражений, а метрики, по которым оценивались алгоритмы не учитывают возможную синонимию слов.

- Высокая степень сжатия текста в используемом для оценки наборе данных. Она более чем в 2 раза превышает степень сжатия исходных статей из набора, использованного авторами статьи [4]. Очевидно, что при формировании рефератов, которые значительно короче исходного текста, велик риск потери важной информации из текста.

- В используемом наборе данных некоторая информация из образцовых рефератов может быть не продублирована в тексте. Таким образом, алгоритм не имеет возможности сгенерировать реферат, содержащий часть информации из образцового реферата, что снижает значения оценок.

Из полученных оценок (см. табл. 7) видно, что алгоритм TextRank, хоть и незначительно, но превосходит остальные алгоритмы по метрике ROUGE – 1. Например, оценка алгоритма LSA меньше всего на 0.01. Среднее количество уникальных слов в образцовых рефератах с учетом стемминга и удаления стоп-слов равно 22. Т.е. в среднем количество совпадающих слов между автоматическим рефератом, сгенерированным алгоритмом TextRank, и образцовым рефератом всего на 0.22 слова больше, чем между рефератом, полученным алгоритмом LSA, и образцовым рефератом. Другими словами, примерно для каждого пятого текста реферат, полученный алгоритмом TextRank, имеет на одно совпадение с образцовым рефератом больше, чем реферат, полученный LSA.

По метрике ROUGE – 2 алгоритмы TextRank и LSA показали одинаковые результаты, алгоритм k-means показал результат на 0.02 хуже. Исходя из средней длины образцовых рефератов, это означает, что примерно для каждого третьего текста реферат, полученный с помощью алгоритмов TextRank или LSA, имеет на одну совпадающую с образцовым рефератом биграмму больше, чем реферат, полученный с помощью алгоритма k-means.

По метрикам ROUGE – L и ROUGE – S, наоборот, алгоритмы TextRank и k-means показали одинаковые результаты, а алгоритм LSA показал чуть

худший результат.

Таким образом, алгоритм TextRank превосходит алгоритмы LSA и k-means по используемым метрикам.

Ниже представлены некоторые примеры работы алгоритмов.

Пример 1

Текст новости:

О предоставлении российским самолетам воздушного коридора сообщает Интерфакс со ссылкой на представителя пресс-службы посольства России в Афинах. Собеседник агентства отметил, что разрешение на пролет через воздушное пространство Греции «оформлено нотой МИД Греции от 31 августа». Представитель греческого МИД Константинос Котрас на вопрос журналиста о предоставлении гражданским самолетам из России воздушного коридора ответил, что «ни одна из сторон не заявила о претензиях и негативных комментариях в связи с данным вопросом, напротив, действия Греции были высоко оценены», сообщается на сайте МИД Греции. В понедельник представитель МИД Греции сообщил, что США обратились к Афинам с просьбой отказать российским самолетам с грузом для Сирии в предоставлении воздушного коридора. Решение не пускать самолеты из России в Сирию через свою воздушную территорию ранее на этой неделе также приняла Болгария. МИД страны считает, что сейчас «есть основания сомневаться в характере груза» на российских самолетах. Замглавы российского МИД Михаил Богданов заявил, что Болгария и Греция должны представить Москве объяснения по поводу отказа в воздушном коридоре. Другой замглавы МИД, Сергей Рябков, сказал, что Россия найдет для поставок в Сирию альтернативные маршруты. Как заявил Интерфаксу пресс-атташе посольства России в Иране Максим Суслов, Тегеран уже дал свое разрешение на пролет российских самолетов.

Образцовый реферат:

Греция предоставила воздушный коридор для пролета в Сирию российской гражданской авиации с гуманитарным грузом. Ранее сообщалось, что США попросили Афины не пускать российские самолеты.

Реферат TextRank:

О предоставлении российским самолетам воздушного коридора сообщает Интерфакс со ссылкой на представителя пресс-службы посольства России в Афинах. В понедельник представитель МИД Греции сообщил, что США обратились к Афинам с просьбой отказать российским самолетам с грузом для Сирии в предоставлении воздушного коридора.

Реферат LSA:

Представитель греческого МИД Константинос Котрас на вопрос журналиста о предоставлении гражданским самолетам из России воздушного коридора ответил, что «ни одна из сторон не заявила о претензиях и негативных комментариях в связи с данным вопросом, напротив, действия Греции были высоко оценены», сообщается на сайте МИД Греции. В понедельник представитель МИД Греции сообщил, что США обратились к

Афинам с просьбой отказать российским самолетам с грузом для Сирии в предоставлении воздушного коридора.

Реферат k-means:

О предоставлении российским самолетам воздушного коридора сообщает Интерфакс со ссылкой на представителя пресс-службы посольства России в Афинах. Другой замглавы МИД, Сергей Рябков, сказал, что Россия найдет для поставок в Сирию альтернативные маршруты.

Пример 2

Текст новости:

В шести странах Восточной Европы в четверг, 3 сентября, официально открылись штабы Североатлантического альянса, в каждом из которых будут работать около 40 офицеров. Новые структуры НАТО, призванные координировать подготовку сил особо быстрого реагирования, открылись в Эстонии, Латвии, Литве, Польше, Болгарии и Румынии. Генсек НАТО Йенс Столтенберг принял участие в церемонии открытия штаба в Вильнюсе, передает портал Delfi. По его словам, открытие штабов является символом солидарности НАТО. «Флаги НАТО поднялись в шести странах союзников, весь альянс делает большой шаг к еще большей солидарности, усилению мощи и лучшей подготовленности», — цитирует издание Столтенберга. Президент Литвы Даля Грибаускайте, которая присутствовала на церемонии открытия, заявила, что Россия не может быть партнером НАТО. «Агрессия России на Украине представляет собой опасность для всего европейского континента. Вот почему агрессоры не могут быть партнерами. Именно поэтому мы решили укрепить свою безопасность», — отметила она. О создании сети новых командных центров в Восточной Европе генсек НАТО объявил в феврале 2015 года. Столтенберг тогда заявил, что действия альянса являются реакцией на «агрессивные действия» со стороны России. Как ранее писал РБК, силы быстрого реагирования, создать которые было решено на саммите НАТО в Уэльсе в сентябре 2014 года, будут насчитывать 30 тыс. человек. Предполагалось, что на первом этапе силы быстрого реагирования будут укомплектованы военнослужащими из Германии, Нидерландов и Норвегии в размере армейской бригады (от 1 тыс. до 4 тыс. человек). Однако позже контингенты вооруженных сил стран НАТО будут нести вахту в этих подразделениях посменно. Их численность будет постепенно увеличиваться, полной оперативной готовности планируется достичь к 2017 году. Прежние силы реагирования НАТО насчитывали 13 тыс. человек. По новой концепции им потребуется до семи дней, чтобы выдвинуться с баз в зону конфликта и вступить в бой. При этом в альянсе подчеркнули, что он будет придерживаться договора об ограничении вооружений в Европе 1997 года, по которому НАТО обязалось не размещать вблизи границ России «значительные боевые силы».

Образцовый реферат:

В шести странах Восточной Европы официально открылись штабы НАТО, связанные с развертыванием сил быстрого реагирования. В каждом из них будут работать около

40 офицеров, о создании этих структур НАТО объявило в феврале.

Реферат TextRank:

Новые структуры НАТО, призванные координировать подготовку сил особо быстрого реагирования, открылись в Эстонии, Латвии, Литве, Польше, Болгарии и Румынии. Как ранее писал РБК, силы быстрого реагирования, создать которые было решено на саммите НАТО в Уэльсе в сентябре 2014 года, будут насчитывать 30 тыс. человек.

Реферат LSA:

Как ранее писал РБК, силы быстрого реагирования, создать которые было решено на саммите НАТО в Уэльсе в сентябре 2014 года, будут насчитывать 30 тыс. человек. Прежние силы реагирования НАТО насчитывали 13 тыс. человек.

Реферат k-means:

Президент Литвы Даля Грибаускайте, которая присутствовала на церемонии открытия, заявила, что Россия не может быть партнером НАТО. Прежние силы реагирования НАТО насчитывали 13 тыс. человек.

Данные примеры иллюстрируют, что алгоритм TextRank генерирует рефераты, наиболее близкие по смыслу к образцовым, чем другие алгоритмы.

3.4 Модификации алгоритма TextRank и их сравнение

По результатам сравнения алгоритмов, алгоритм TextRank показал лучшие результаты, поэтому я решила рассмотреть и сравнить различные его модификации.

Данный алгоритм можно модифицировать засчет изменения способа вычисления степени схожести двух предложений при построении графа. В оригинальном алгоритме схожесть определяется как количество совпадающих слов в предложениях, нормированное суммарной длиной этих предложений. Однако, существуют и другие способы измерения схожести двух предложений. [10]

Косинусная мера схожести. С помощью модели TF-IDF предложения текста представляются в виде векторов. Затем вычисляется косинус угла между двумя векторами, соответствующими двум предложениям. Это значение принимается в качестве меры схожести двух предложений.

Наибольшая общая подпоследовательность. В качестве меры схожести двух предложений принимается длина наибольшей общей подпоследовательности между этими предложениями, которые рассматриваются

как последовательности слов.

Описанные меры схожести были использованы для алгоритма TextRank. Модифицированные версии алгоритма были оценены по метрикам, описанным в разделе 2.3, с проведением стемминга и удалением стоп-слов в рефератах. В таблице 9 представлены полученные результаты. В данной таблице приняты следующие обозначения:

- baseline — изначальный вариант алгоритма
- cosine — вариант алгоритма с косинусной мерой схожести
- lcs – вариант алгоритма с длиной наибольшей общей подпоследовательности в качестве меры схожести

	baseline	cosine	lcs
ROUGE – 1	0.25	0.23	0.26
ROUGE – 2	0.08	0.07	0.09
ROUGE – L	0.12	0.12	0.13
ROUGE – S	0.04	0.04	0.05

Таблица 9: Сравнение вариантов алгоритма TextRank

По результатам оценки алгоритм с косинусной мерой схожести показал оценки ниже, чем изначальный алгоритм по метрикам ROUGE – 1 и ROUGE – 2.

Алгоритм, использующий длину наибольшей общей подпоследовательности, напротив, показал более достойные результаты по всем используемым метрикам.

Заключение

В ходе проделанной работы были изучены основные подходы к экстракционному реферированию текстов и способы оценки их качества.

На языке Python с использованием оптимизированных библиотек были реализованы модули работы алгоритмов экстракционного реферирования TextRank, LSA, k-means, модуль предобработки текстов, модуль сбора статистики по тестовому набору данных, модули оценки алгоритмов и другие вспомогательные модули. Исходный код доступен по ссылке: `https://github.com/novonastya/summarization`

Были проведены эксперименты, в ходе которых алгоритмы были оценены по четырем метрикам на русскоязычном наборе данных, состоящем из 35300 новостных статей. В ходе оценки алгоритм TextRank показал лучшие результаты среди рассмотренных алгоритмов.

Также была проведена оценка двух возможных модификаций алгоритма TextRank, одна из которых показала более высокие оценки по сравнению с оригинальным алгоритмом.

По материалам работы был подготовлен доклад на всероссийской научной конференции по проблемам информатики СПИСОК-2017.

Список литературы

[1] Gambhir and V. Gupta. Recent automatic text summarization techniques: a survey// Artificial Intelligence Review. – 2016. – С. 1-66.

[2] A. Nenkova and K. McKeown. Automatic summarization// Foundations and Trends in Information Retrieval Vol. 5. – 2011. – С. 103–233.

[3] D. Das and A. Martins. A Survey on Automatic Text Summarization// Literature Survey for the Language and Statistics II course at Carnegie Mellon University. – 2007. – С. 1-31.

[4] R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts// Proc. of the 9th Conf. on Empirical Methods in Natural Language Processing. – 2004. – С. 404–411.

[5] R. García-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh and R. Cruz. Text Summarization by Sentence Extraction Using Unsupervised Learning// In Proceedings of the 7th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence/ Alexander Gelbukh and Eduardo F. Morales (Eds.) – Springer-Verlag, Berlin, Heidelberg, 2008. – С. 133-143

[6] Y. Kumar Meena and D. Gopalani. Analysis of Sentence Scoring Methods for Extractive Automatic Text Summarization// Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies. – ACM, New York, NY, USA, 2014.

[7] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis// Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2001. – С. 19-25.

[8] J. Steinberger and K. Jezek. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation// Proc. of ISIM. – 2004. – С. 93–100.

[9] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries// Proceedings of ACL Text Summarization Branches Out Workshop. – 2004. – C. 74–81,

[10] Federico Barrios, Federico López, Luis Argerich, Rosa Wachenchauzer. Variations of the Similarity Function of TextRank for Automated Summarization. – 44 JAIIO - ASAI 2015 - ISSN: 2451-7585, 2015. – C. 65-72

[11] H. P. Luhn. The automatic creation of literature abstracts. – IBM Journal of Research and Development, vol. 2, no. 2, 1958. – C. 159–165