Система автоматического аннотирования текстов с помощью стохастической модели

Т. В. Вознесенская, Д. А. Леднов
tvoznesenskaya@hse.ru; lednov59@gmail.com
Факультет компьютерных наук НИУ ВШЭ, Москва, ул. Мясницкая, 20;
ООО «DS-systems», Москва, ул. Пречистенка 40/2 стр.2

Работа посвящена системе автоматического аннотирования текста, реализованной в рамках совместного проекта компании "DC – Systems" и факультета компьютерных наук НИУ ВШЭ. Построение аннотации осуществляется с помощью синтаксически согласованных словосочетаний, наиболее близких к семантике всего текста. При этом пренебрегается возможными дополнительными смыслами отдельных фрагментов текста. Качество аннотации определяется семантической близостью к исходному тексту. Задача построения аннотации разбивается на две части: оценка семантики текста в целом, то есть без разделения на более мелкие составляющие, и преобразование текста, приводящее к построению аннотации. В работе описана структурная схема реализованной системы автоматического аннотирования и алгоритм ее работы. Система протестирована на коллекции из 50 текстов различной тематики, приведен пример построенной аннотации и дана оценка его качества с помощью набора мер качества *ROUGE* [9]. Ограничением применения текущей версии системы является наличие в тексте формул и специальных символов.

Ключевые слова: *аннотирование, автоматическая обработка текста, корпусная лингвистика.*

DOI: 10.21469/22233792.4.4.04

1 Введение

Аннотация — это сжатая, краткая характеристика текста с точки зрения его содержания и значения, т.е. с точки зрения его семантики [1–3]. Аннотация дает ответ на вопрос, о чем говорится в исходном тексте, каковы его главные вопросы и проблемы. Цель аннотирования- максимальное сокращение объема текста при существенном сохранении основного содержания.

Обзор публикаций [4–6] на эту тему показывает, что основной стратегией аннотирования является выделение ключевых фрагментов текста с различными способами представления этого списка ключевых фрагментов пользователю. Размеры текстовых фрагментов заключаются в пределах от слов до предложений, в качестве аннотации пользователю представляется список ключевых слов или фраз, либо список целых предложений в необработанном виде.

Значимость фрагмента текста для аннотирования определяется весовыми коэффициентами. В [7] эти весовые коэффициенты $W(s)$, имеют вид $W(s) = Lok(s) + Ph(s) + Pr(s) + Add(s)$, где $Lok(s)$ — коэффициент расположения фрагмента в тексте $Ph(s)$ — коэффициент значимости фразы, характеризующий аннотирующие коллокации, такие как «в заключение», «в данной статье», «согласно результатам анализа» и так далее, $Pr(s)$ — коэффициент статистической значимости фрагмента текста, $Add(s)$ — коэффициент дополнительного наличия терминов.

Если в качестве фрагментов текста использовать слова, то с помощью их взвешивания по приведенной выше формуле, можно получить список ключевых слов, на основе которого пользователь способен определить семантический класс текста, но при этом теряется

информация о том, в каком контексте ключевые слова употреблялись. Использование же в качестве фрагментов целых предложений может быть избыточным и перегружать пользователя излишней информацией. Поэтому был выбран способ построения аннотации текста с помощью синтаксически согласованных словосочетаний [8, 9] (фрагментов предложений), которые семантически наиболее близки к семантике всего текста.

Качество аннотации определяется семантической близостью к исходному тексту при минимальном объеме текста.

В [10] описаны два подхода к оцениванию качества аннотирования текста: внешний (extrinsic), – когда качество аннотирования оценивается косвенно путем оценки задачи, предположительно зависящей от качества аннотации, и внутренний (intrinsic), — при котором оценивается непосредственно качество полученной аннотации.

Задачи внешнего подхода — классификация текстов, ответы на вопросы по содержанию текста. Сложностью такого оценивания является подбор реальной задачи и способа измерения эффекта использования автоматической аннотации вместо оригинального текста. Внутренняя оценка требует некоторого стандарта или модели для оценки качества аннотирования. На практике она осуществляется путем нахождения существующих наборов текстов с аннотациями или экспертов, создающих аннотации. При этом сравнение автоматически полученной аннотации с эталоном выполняют тоже эксперты по выбранным критериям (согласованность, краткость, грамматическая правильность, понятность, смысл и т.п.).

Chin-YewLin в работе «A Package for Automatic Evaluation of Summaries» [11] предложил набор мер качества $ROUGE$ (Recall-Oriented Understudy for Gisting Evaluation), который стал стандартом «де факто» в данной области. Он включает в себя несколько автоматических методов оценки, позволяющих измерять сходство между двумя аннотациями (автоматической и выполненной экспертом). Основные меры качества: $ROUGE-N$ (N-gram Co-Occurrence Statistics) – подсчет количества пересекающихся N–грамм слов; $ROUGE-L$ (Longest Common Subsequence) – отношение длины максимальной общей подпоследовательности к общей длине предложения; $ROUGE-W$ (Weighted Longest Common Subsequence) — взвешенный $ROUGE-L$ — к каждой подпоследовательности добавляется вес, основанный на плотности последовательности (среднее расстояние появления в исходном предложении); $ROUGE-S$ (Skip-bigrams) — анализ пересечения биграмм, находящихся на некотором расстоянии друг от друга (между первым и вторым словами биграммы могут находиться другие слова). Величина окна skip-биграммы (количество слов, которое может «вклиниваться» внутрь биграммы) является параметром.

Метод пирамидной оценки [12] основан на ручном выделении экспертами «информационных единиц» из эталонных аннотаций – Summary Content Units (SCUs). Каждая SCU представляет собой квант информации, которая, по мнению эксперта, должна быть также отражена в автоматической аннотации. SCU получает вес, равный количеству эталонных аннотаций, где она встречается. Общая оценка автоматической аннотации определяется как отношение суммы весов SCU, которые она содержит, к общему количеству SCU для данного текста.

Таким образом, на сегодняшний день оценка качества аннотирования не обходится без работы экспертов, что безусловно дорого и требует существенных временных затрат.

2 Постановка задачи

В случае полного описания языка [13], когда синтаксические правила описывают все возможные сочетания слов, а семантический словарь описывает все возможные семантиче-

ские отношения, в котором слово используется в языке, для всех слов языка, то для любого предложения можно построить сеть, которая состоит из узлов и связей между узлами. В узлах сети находятся слова предложения со своими однозначными морфологическими характеристиками, а связи между словами определяют их синтаксическую согласованность и тип семантических отношений: валентности и семантические характеристики [14–16]. Необходимо отметить, что с одной стороны, в такой сети одному слову предложения может соответствовать более одного узла, поскольку морфологические характеристики слова не являются однозначными (как следствие омонимии), а с другой стороны, узлы в сети могут быть связаны не одной связью, это продиктовано многозначностью семантических отношений между словами. Пример морфологической многозначности (омонимии [17]) — «Эти типы стали есть в литейном цехе». Слово «стали» может быть либо глаголом, либо существительным со значением «металл». Пример семантической многозначности (полисемии [18]) — «дворник на машине». Семантически «дворник» может быть либо деталью, либо профессией. На практике описание языка не является полным, в частности:

– не все слова языка включены в семантический словарь;
– не полностью описаны семантические характеристики слова, и семантические характеристики могут быть описаны ошибочно;
– морфологический словарь является не полным: он содержит не все слова языка, и/или слово в словаре не содержит полного множества его морфологических характеристик;
– правила синтаксической согласованности не описывают всех возможных случаев сочетания слов в предложении.

Локальные семантические связи слов [16], то есть связи, которые определены в рамках отдельных предложений, заданные семантическим словарем, не определяют однозначно семантики предложения, семантика отдельного предложения не определяет семантику текста [3, 19]. Это приводит к необходимости строить семантические связи слов, выходящие за рамки предложений (глобальные). До какой степени необходимо выйти за рамки предложения для определения семантики текста? Вариантов несколько: рассматривать семантические связи между словами в рамках абзаца; рассматривать семантические связи между словами в рамках нескольких абзацев и проводить анализ с помощью скользящего окна, проводя его смещение по абзацу, или рассматривать глобальные связи слов охватывающие весь предоставленный для анализа текст, в целом.

Первая задача, которая рассматривается в данной работе, это автоматическая оценка семантики текста в целом, т.е. представленного текста, который не разделен на какие-либо более мелкие составляющие, с помощью, описанной выше сети, включающей в себя глобальные связи, при условии, что описание языка не является полным. Будем предполагать, что семантика текста является однозначной и не изменяется от одного фрагмента к другому, т.е. можно пренебречь любым отступлением от этого доминирующего значения текста.

Вторая задача состоит в преобразовании текста (удалении или добавлении слов, или целых фрагментов) таком, что при убывающем размере текста сохраняется оценка его семантики и синтаксическая согласованность. Результатом решения этих двух задач является аннотация, которая должна быть представлена в виде наиболее значимых сегментов исходного текста, связанных в согласованные предложения.

3 Описание системы

3.1 Модуль анализа текста

На рис. 1 показана структурная схема автоматической системы, реализованной в рамках данной работы. Система состоит из двух основных модулей: обучения и анализа текста.

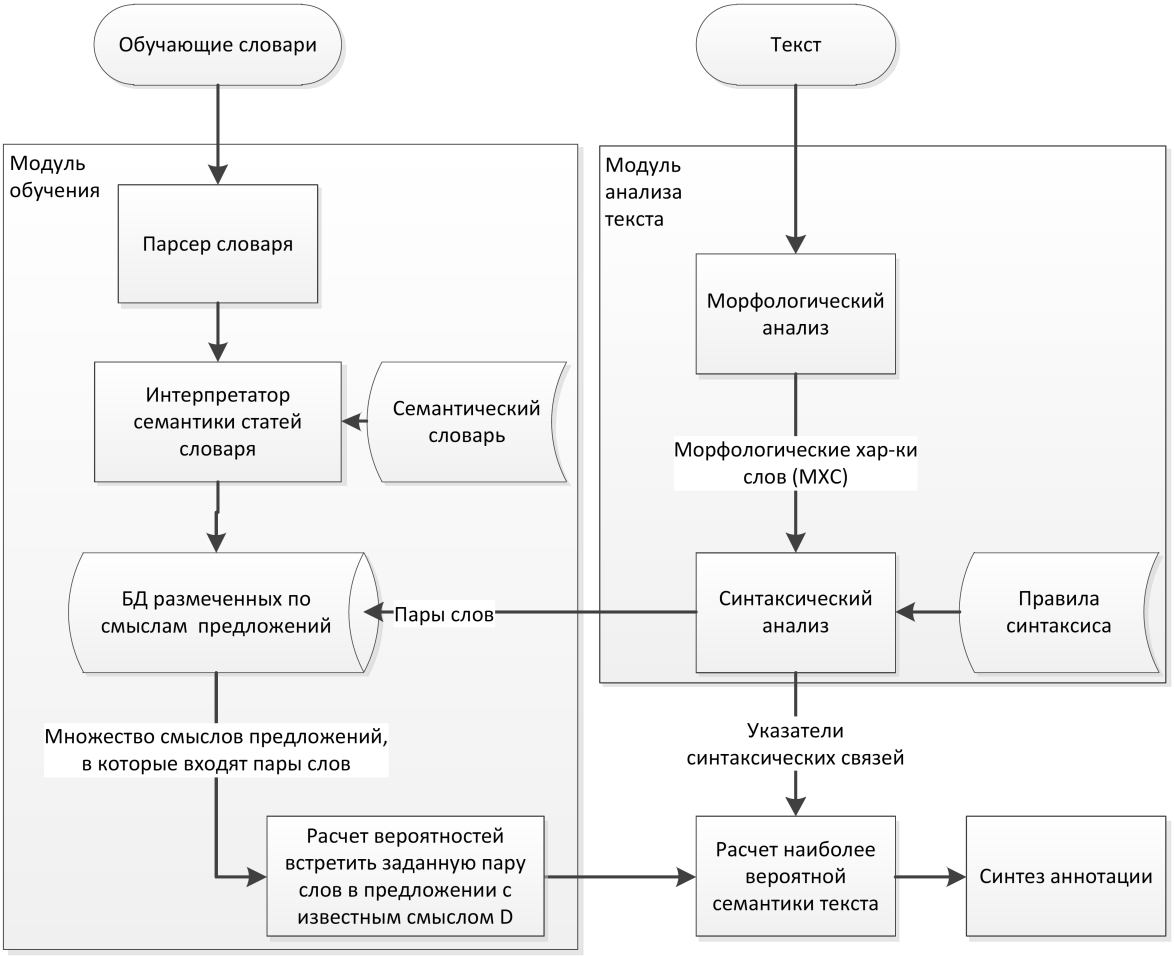

Рис. 1 Структурная схема системы, для решения задач поиска семантики текста и формирования аннотации

Модуль анализа текста содержит морфологический анализатор, который получает на своем входе слово текста, а на выходе представляет множество вариантов морфологических характеристик этого слова, включая его базовую форму.

Морфологический анализатор построен на основе свободно распространяемого морфологического словаря М.Хагена «Полная парадигма. Морфология» [20]. В системе используется морфологическое описание слов, приведенное в Таблице 1, не зависимо от частей речи, которые они выражают.

Результаты работы морфологического анализатора поступают на вход синтаксического анализатора, который на основе морфологических характеристик слов и правил синтаксического согласования, пример реализации которых приведен в Таблице 2, строит дерево

Таблица 1 Морфологическое описание слов

Падеж	**Число**	**Род**	**Время**	**Часть речи**	**Одуш.**	**Лицо**
именительный	нет	средний	прош.	глагол	одуш.	1-е
родительный	ед.	женский	наст.	сущ.	неодуш.	2-ое
творительный	множ.	мужской	будущее	прилаг.	неизв.	3-е
дательный				местоимение		
винительный						
предложный						

Таблица 2 Примеры правил синтаксического согласования

С1+ПРЕДЛ1+С2=(С1.падеж:им+С2.падеж:рд)	//дом возле озера
С1+ПРЕДЛ1+С2=(С1.падеж:им+С2.падеж:дт)	//направление к врачу
С1+ПРЕДЛ1+С2=(С1.падеж:им+С2.падеж:вн)	//право на молчание
С1+ПРЕДЛ1+С2=(С1.падеж:им+С2.падеж:тв)	//комната под крышей
С1+ПРЕДЛ1+С2=(С1.падеж:им+С2.падеж:пр)	//сообщение об ошибке

Таблица 3 Пример устройства семантических классов в словаре Н. Ю. Шведовой

Слово	**Индексы классов**
Кляча	1.1.2.1.1.1.2.5.5., 1.1.2.1.1.1.3.1.4., 1.1.2.2.1.2.4.4.
Индекс	**Название семантического класса**
1.1.2.1.1.1.2.5.5.	характеристика по физическому, физиологическому состоянию
1.1.2.1.1.1.3.1.4.	брань, хула
1.1.2.2.1.2.4.4.	обиходные, сниженные названия, названия домашних животных по выполняемой ими функции

синтаксических связей предложения, а также разбивает предложение на именные, глагольные и дополнительные именные группы.

Для синтаксической связи между словами w_n, w_m, определенной k-м правилом синтаксического согласования, введем обозначение $\mu(rule(k)|w_n, w_m)$. Эта величина принимает значение 1 или 0 в зависимости от того, найдено правило согласования или нет.

Значение этой величины и пары слов, для которых она получена, поступают далее либо в модуль обучения, либо в блок расчета наиболее вероятной тематики, в зависимости от фазы работы системы.

3.2 Модуль обучения

Цель модуля обучения, состоит в том, чтобы на основе анализа толковых словарей [21–23], которые на Рисунке1 названы обучающими, вычислить значение вероятности $p(D|w_n, w_m)$ того, что пара синтаксически согласованных слов встретилась в предложении, которое отнесено к определенному семантическому классу D и совместной вероятности синтаксической согласованности данной пары слов $p(w_n, w_m)$.

Для вычисления этой вероятности используем семантический словарь под общей редакцией Н. Ю. Шведовой [24], который состоит из слов и соответствующих ссылок на их семантические классы. Пример устройства словаря для слова «кляча» приведен в Таблице 3.

Таблица 4 Пример иерархического устройства семантического словаря [24]

1. имена существительные с конкретным значением
1.1. все живое: человек, животные, растения
1.1.1 общие для всех живых организмов, нескольких их классов
1.1.1.1. собственно организмы
1.1.1.1.1. общие обозначения
1.1.1.1.2. по характерному признаку
1.1.1.1.3. совокупности живых организмов
1.1.1.2. части организмов
1.1.1.2.1. клетки, ткани, биологические элементы организма
1.1.1.2.2. части организмов, зачатки

Индексы семантических классов в словаре [24] имеют иерархическую структуру. В Таблице 4 приведен пример этой структуры из части словаря – существительные.

Максимально словарь [24] имеет 16 уровней вложенности и всего 2546 семантических классов существительных и 4 уровня вложенности и 206 классов для глаголов, т.е. общее количество семантических классов $Q = 2752$. С помощью этого словаря каждой статье толкового словаря, описывающей семантический класс какого-либо слова, припишем индекс семантического класса этого слова. Будем считать, что каждое предложение статьи толкового словаря сохраняет значение семантического класса.

Введем следующие обозначения: $\omega(w_n, w_m)$ – это количество синтаксически согласованных слов w_n и w_m во всех предложениях толковых словарей, используемых для обучения; $\omega(D|w_n, w_m)$ - это количество синтаксически согласованных слов w_n и w_m во всех предложениях статей толкового словаря, которые отнесены к семантическому классу D, тогда вычислить значение вероятности $p(D|w_n, w_m)$ можно с помощью отношения

$$p(D|w_n, w_m) = \omega(D|w_n, w_m)/\omega(w_n, w_m).$$

Очевидно, что матрица $p(D|w_n, w_m)$ имеет размер равный $N \times N \times Q$, где N – размер словаря языка. Построение всех элементов матрицы требует значительных вычислительных затрат и большого количества обучающего материала. При недостатке обучающего материала матрица будет сильно разряженной (содержать большое количество нулевых элементов) и требовать использования методов сглаживания, работа которых при построении моделей языка не является корректной [25, 26].

С целью понижения размерности $p(D|w_n, w_m)$ перейдем к матрице вида $p(D|d(w_n), d(w_m))$, где $d(w_n)$– семантический класс слова w_n, определенный с помощью словаря [24]. Тогда размер новой матрицы $p(D|d(w_n), d(w_m))$ будет Q^3, что значительно меньше размера исходной матрицы. Если учитывать все семантические классы, представленные в словаре, то размер этой матрицы составляет около 10Гб. Для большего понижения размера новой матрицы можно использовать иерархическое устройство словаря (см. Таблицу 4) и ограничивать точность классификации существительных уровнем вложенности иерархии. Так, при ограничении вложенности 7-м уровнем объем матрицы сокращается в 15 раз, безусловно, при этом падает точность семантической классификации. Вопрос снижения размерности матрицы представляет вопрос отдельных исследований. Здесь авторы ограничиваются именно 7-м уровнем вложенности словаря.

3.3 Модель расчета семантического класса текста

Обозначим $p(D|\{w_{nj}, w_{nj-1}, \ldots, w_1\}_j)$– вероятность того, что j-ое предложение, состоящее из последовательности семантически согласованных слов $w_{nj}, w_{nj-1}, \ldots, w_1$, относится к семантическому классу D, где n_j – количество согласованных слов в j-ом предложении.

Логарифм вероятности того, что текст $Text$ принадлежит семантическому классу D можно представить в виде:

$$\log P(D|Text) = \sum_{j=1}^{J} \log p(D|\{w_{nj}, w_{nj-1}, \ldots, w_1\}_j)$$

$$\sum_{j=1}^{J} \sum_{k=2}^{n_j} \log\, p(D|w_k, w_{k-1})$$

$$\sum_{j=1}^{J} \sum_{k=2}^{n_j} \log\, p(D|d(w_k), d(w_{k-1})) \tag{1}$$

Решение задачи поиска семантического класса текста состоит в вычислении наиболее вероятного семантического класса, т.е. $\Theta = \arg\max_D p(D|Text)$

3.4 Алгоритм синтеза аннотации

Проведенный синтаксический анализ позволяет разбить все предложения текста на именные (ИГ), глагольные (ГГ) и дополнительные именные группы (ДИГ). Без потери общности, для простоты представления допустим, что текст состоит из простых предложений в виде:

ИГ(1)+ГГ(1)+ДИГ(1)... ИГ(i-1)+ГГ(i-1)+ДИГ(i-1)... ИГ(i)+ГГ(i)+ДИГ(i)...

ИГ(i+1)+ГГ(i+1)+ДИГ(i+1)...ИГ(M)+ГГ(M)+ДИГ(M), где i – номер предложения, M – количество предложений.

Для построения модели используем следующие понятия: Показатель эквивалентности именных групп различных предложений, который будем считать равным:

$$q = \beta + \frac{2n(i,j)}{n_c(i) + n_c(j)}, \tag{2}$$

где $n(i,j)$ – число совпавших в i-ой и j-ой именных группах основных форм существительных, $n_c(i)$ – количество существительных в i-ой именной группе, $\beta \in (0,1)$ и характеризует совпадение основной формы главных слов именных групп или тот факт, что главные слова являются синонимами. Принадлежность слов одному синонимическому ряду определяется с помощью словаря [27], в случае совпадения $\beta = 1$.

Логарифм перплексии [28, 29] каждой синтаксической группы, которая характеризует то, что группа относится к семантическому классу Θ:

$$Z_s = \frac{1}{n_s} \sum_{k=2}^{n_s} \log\, p(\Theta|d(w_k), d(w_{k-1})), \tag{3}$$

где n_s – количество слов в синтаксической группе.

Несократимый текст — это такой текст, удаление любой группы (именной, дополнительной именной, глагольной) из которого приводит к потере смысла текста. Приведем несколько примеров несократимых тестов.

Пример 1: Во дворе стоял дуб. Рядом с дубом находился колодец. У колодца сидела лягушка.

Пример 2: Соседская дача была огромной площади. Она была огорожена забором и пустовала. Ей суждено стать предметом зависти.

Если к обоим примерам применить понятие эквивалентности именных групп 2 и объединить эти эквивалентные группы, то для первого примера будет характерен Рисунок 2, а для второго примера Рисунок 3.

Объединение именных групп на основе понятия эквивалентности назовем ядром текста (на рисунках ядро текста выделено овалом с пунктирной линией). Очевидно, что текст может обладать значительным количеством ядер и разветвленной системой связей между такими ядрами.

Здесь авторы предполагают, что задача аннотирования состоит в представлении текста в виде множества несократимых текстов.

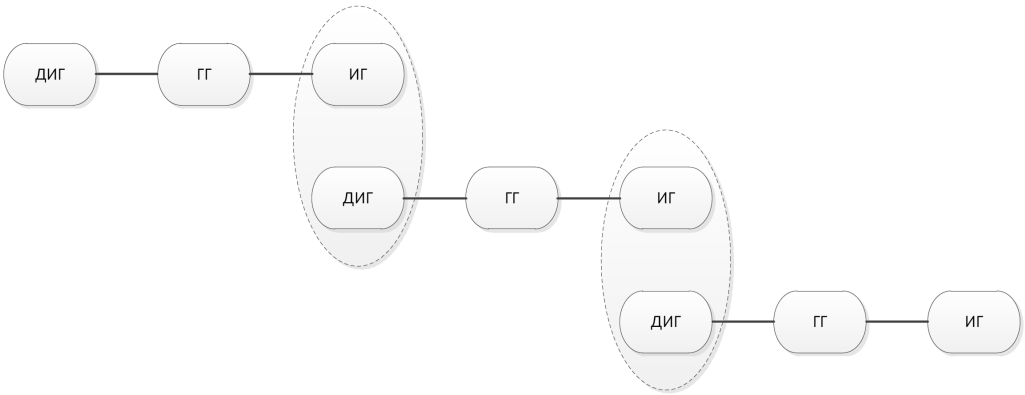

Рис. 2 Графическое представление примера 1

В рамках данной работы для построения аннотации ограничимся следующим алгоритмом:

1. С помощью показателя эквивалентности и перплексии построим ядра текста, которые принадлежат одной семантической категории и выражены множеством одинаковых слов. Минимальное количество одинаковых слов в именных группах в рамках одного ядра будет определяться порогом, который устанавливается для значения показателя эквивалентности.
2. Ограничим аннотацию предложениями, именные группы которых входят в наиболее мощное ядро текста. Мощность ядра определяется количеством входящих в него именных групп. Удалим из этих предложений именные и связанные с ними глагольные группы, которые не входят в это ядро.

Описанный алгоритм строит несократимый текст, соответствующий графическому представлению Рисунка 3. Способов приведения текста к несократимому виду можно предложить довольно много, в рамках данной статьи ограничимся приведенным выше.

4 Пример построения аннотации

Тексты, которые были использованы для проверки качества аннотирования были случайным образом взяты из Интернета. Полученная коллекция состояла 7-и тематических

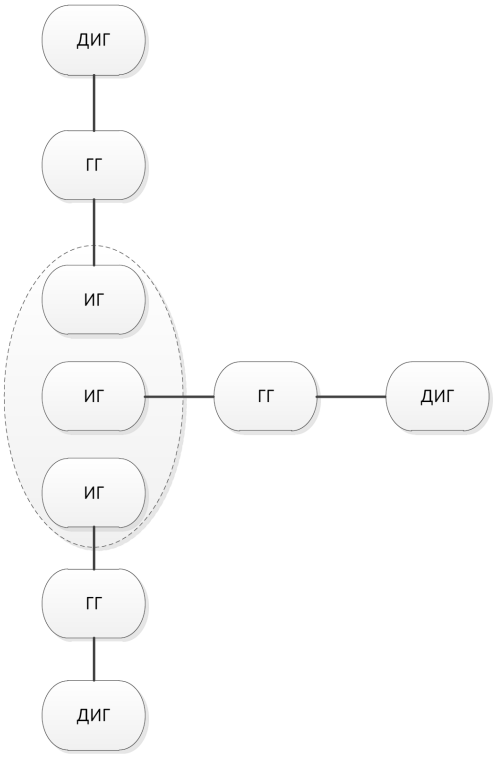

Рис. 3 Графическое представление примера 2

категорий: физика, экономика, лингвистика, IT, автомобили, медицина, химия. К текстам предъявлялось единственное требование – они не должны содержать математических и химических формул и других специальных символов.

Возьмем исходный текст из области физики:

Для деления с большой вероятностью тяжелое ядро должно получить энергию извне, превышающую значение барьера деления. Так, после присоединения нейтрона ядро обладает энергией возбуждения, равной сумме энергии отделения нейтрона и кинетической энергии захваченного нейтрона. Этой дополнительной энергии может быть достаточно, чтобы ядро перешло в возбужденное состояние с интенсивными колебаниями. Физически аналогичную ситуацию можно получить, если поместить каплю воды на горячую горизонтальную поверхность. Если поверхность достаточно горячая, то капля будет плавать на изолирующем слое пара, поддерживающем ее над поверхностью в свободном состоянии. При этом могут возникнуть колебания формы капли, при которых она примет последовательно шарообразную и эллипсоидальную форму. Такое колебательное движение представляет собой состояние динамического равновесия между инерционным движением вещества капли и поверхностным натяжением, которое стремится поддерживать сферически симметричную форму капли. Если силы поверхностного натяжения достаточно велики, то процесс вытягивания капли прекратится раньше, чем капля разделится. Если же кинетическая энергия инерционного движения вещества капли окажется большой, то капля может принять гантелеобразную форму и при своем дальнейшем движении разделиться на две части. В случае ядра процесс происходит аналогично, только к нему добавляется электростатическое отталкивание протонов, действующее как дополнительный фактор против ядерных сил, удерживающих нуклоны в ядре. Если ядро находится в возбужденном состоянии, то оно совершает колебательные движения, связанные с отклоне-

ниями его формы от сферической. Максимальная деформация увеличивается с ростом энергии возбуждения и при некотором ее значении может превысить критическое значение, что приведет к разрыву исходной капли и образованию двух новых. Колебательные движения возможны под действием сил поверхностного натяжения (аналог ядерных сил в капельной модели ядра) и кулоновских. На поясняющем рисунке показано изменение потенциальной энергии и отдельных ее составляющих в процессе деления заряженной капли. Энергия поверхностного натяжения резко возрастает с ростом малых деформаций и остается практически неизменной после того, как капля приобретает гантелевидную форму. Энергия кулоновского взаимодействия плавно уменьшается с ростом деформаций практически во всем диапазоне состояний. Ядра, образовавшиеся после деления исходного ядра, разлетаются в противоположные стороны под действием кулоновских сил, и потенциальная энергия превращается в кинетическую. В итоге суммарная потенциальная энергия возрастает до момента деления капли, а затем уменьшается.

Аннотация:

Для деления с большой вероятностью тяжелое ядро должно получить энергию извне. После присоединения нейтрона ядро обладает энергией возбуждения. Если поверхность достаточно горячая, то капля будет плавать на изолирующем слое пара. Могут возникнуть колебания формы капли, при которых она примет и эллипсоидальную форм. Кинетическая энергия инерционного движения вещества капли окажется высокой то капля может принять гантелеобразную форму. В случае ядра процесс происходит аналогично только добавляется электростатическое отталкивание протонов. Энергия поверхностного натяжения резко возрастает с ростом малых деформаций и остается как капля приобретает гантелевидную форму. Энергия кулоновского взаимодействия плавно уменьшается с ростом деформаций практически во всем диапазоне состояний. Ядра, образовавшиеся после деления исходного ядра, разлетаются в противоположные стороны под действием кулоновских сил. В итоге суммарная потенциальная энергия возрастает до момента деления капли.

4.1 Оценка результата

Приведем одну из экспертных аннотаций, выполненную для текста из предыдущего пункта:

Для деления с большой вероятностью тяжелое ядро должно получить энергию извне. После присоединения нейтрона ядро обладает энергией возбуждения. Данная ситуация похожа на поведение капли воды, которая, будучи помещенной на достаточно горячую горизонтальную поверхность, плавает на изолирующем слое пара. Могут возникнуть колебания формы капли, при которых она примет эллипсоидальную форму. Если кинетическая энергия инерционного движения вещества капли окажется большой, то капля может стать гантелеобразной и разделиться на две части. В случае ядра процесс происходит аналогично, только добавляется электростатическое отталкивание протонов. С ростом энергии возбуждения исходная капля может разорваться на две новых. Колебательные движения возможны под действием сил поверхностного натяжения и кулоновских. Ядра, образовавшиеся после деления исходного ядра, разлетаются в противоположные стороны. Суммарная потенциальная энергия возрастает до момента деления капли, а затем уменьшается.

Рассчитаем меру качества $ROUGE-1$: сравнение пересечения монограмм слов автоматической и экспертной аннотаций. Синим выделены монограммы (слова), встречающиеся в обеих аннотациях (автоматической и экспертной). Тогда

$$ROUGE-1 = \frac{70(\text{монограмм в пересечении})}{105(\text{монограмм в экспертной аннотации})} = 0.67 \qquad (4)$$

Таблица 5 Результаты оценки качества аннотаций

Мера	**Среднее значение**	**Среднеквадратичное отклонение**
$ROUGE-1$	0.68	0.16
$ROUGE-2$	0.71	0.12
эксперты	0.73	0.21

Меры $ROUGE-1$ и $ROUGE-2$ были рассчитаны для коллекции из 50 текстов. Также была проведена ручная оценка качества полученных аннотаций с помощью экспертов.

5 Заключение

Таким образом, разработана система автоматического аннотирования, позволяющая генерировать качественные, пригодные для использования аннотации в виде наиболее значимых сегментов исходного текста, связанных в согласованные предложения. При этом семантика текста оценивается в целом в предположении, что она является однозначной и не изменяется от одного фрагмента к другому, либо, что можно пренебречь любым отступлением от доминирующего значения всего текста.

В качестве дальнейшего развития авторы видят исключение ручных операции из процедуры оценки качества аннотации. Для этого необходимо

1. Определить автоматически множество D – смыслов полного текста,
2. Определить автоматически множество D_A — смыслов аннотации.

И, затем оценить качество аннотации по следующим параметрам: отсутствие искажения семантики — аннотация должна попасть в ту же смысловую категорию, что и основной текст; отсутствие нарушения синтаксиса — текст аннотации должен быть согласован; процент сжатия текста.

Литература

[1] *Л. П. Маркушевская, Ю. А. Цапаева.* Аннотирование и реферирование. СПбГУ ИТМО, 2008.

[2] *А. И. Новиков.* Семантика текста и ее формализация. М.: Наука, 1983.

[3] *Н. П. Пешкова.* Семантика и смысл текста: >, <, =, #?(экспериментальный подход к теоретическим проблемам). // Вестник Челябинского государственного университета. Филология. Искусствоведение, 15(370):69–77, 2015.

[4] *П.Г. Осминин.* Модель автоматического реферирования на основе базы знаний, ориентированная на автоматический перевод. // Вестник ЮУрГУ. Серия «Лингвистика». Том 11, 2, 2014.

[5] *П. Браславский, И. Колычев.* Автоматическое реферирование веб-документов с учетом запроса. // Автоматическая обработка веб-данных, 2005.

[6] *A. Nenkova, K. McKeown, C. C. Aggarwal, C. X. Zhai.* A survey of text summarization techniques. // Mining Text Data, 2012.

[7] *У. Хан, И. Мани.* Системы автоматического реферирования. // Открытые системы, 12, 2000.

[8] *Н. С. Валгина.* Современный русский язык: Синтаксис: Учебник. М.: Высшая школа, 2003.

[9] *П. Г. Осминин.* Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод. PhD thesis, 2016.

[10] *D. Harman, P. Over.* The effects of human variation in duc summarization evaluation. // In *Proceedings of the Text Summarization Branches Out Workshop*, Barcelona, Spain, 2004.

[11] *Chin-Yew Lin.* A package for automatic evaluation of summaries. // In Proceedings of the Text Summarization Branches Out Workshop, Barcelona, Spain, 2004.

[12] *A. Nenkova, R. Passonneau, K. Mckeown.* The pyramid method: Incorporating human content selection variation in summarization evaluation. // ACM Transactions on Speech and Language Processing, 4, 2007.

[13] *Н. С. Балашова.* Семантика. // Институт филологии и журналистики СГУ им. Чернышевского, 2011.

[14] *Н. Н. Леонтьева.* К теории автоматического понимания текста. Ч.1. Моделирование системы «мягкого понимания» текста: Информационно-лингвистическая модель. М.: Изд. МГУ, 2000.

[15] *Н. Н. Леонтьева.* К теории автоматического понимания текста. Ч. 2. Семантические словари: состав, структура, методика создания. М.: Изд. МГУ, 2001.

[16] *Н. Н. Леонтьева.* К теории автоматического понимания текста. Ч. 3. Семантический компонент. Локальный семантический анализ. М.: Изд. МГУ, 2002.

[17] *А. И. Головня.* Словарь лексико-грамматических омонимов. Мн.: БГУ, 2007.

[18] *Л. М. Лещева.* Лексическая полисемия в когнитивном аспекте. М.:Языки славянской культуры. Знак., 2014.

[19] *А. И. Новиков.* Текст и его смысловые доминанты. М.: Институт языкознания РАН, 2007.

[20] *М. Хаген.* Полная парадигма. Морфология. Частотный словарь. Совмещенный словарь.

[21] *С. И. Ожегов, Н. Ю. Шведова.* Толковый словарь русского языка. М.: ООО «А ТЕМП», 2006.

[22] *В. Даль.* Толковый словарь живого великорусского языка. Дрофа, 2011.

[23] Большая советская энциклопедия. Советская энциклопедия, 1978.

[24] *Н. Ю. Шведова.* 19. Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений. М.: Азбуковник, 1998.

[25] *А. П. Зыков.* Метод сглаживания вероятностей n-грамм на основе моделирования математического ожидания их встречаемости. // Труды СПИИРАН, 4(19), 2015.

[26] *R. Kneser, H. Ney.* Improved backing-off for m-gram language modeling. // In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 181–184, 1995.

[27] *З. Е. Александрова.* Словарь синонимов русского языка. Практический справочник. Русский язык, 2001.

[28] *C. Yanshuai, W. Luyu.* Automatic selection of t-sne perplexity. // In ICML 2017 AutoML Workshop and Conference Proceedings, 2017.

[29] *Е. А. Будников.* Оценивание вероятностей появления строк в естественном языке. // Машинное обучение и анализ данных, 1(3), 2012.

Поступила в редакцию 23.12.2018

Automatic text summarization system using a stochastic model

T. V. Voznesenskaya, D. A. Lednov

tvoznesenskaya@hse.ru; lednov59@gmail.com

Computer Science Faculty NRU HSE, Moscow, 20 Myasnitskaya street;

«DS-systems», Moscow, 40/2 b.2 Prechistenka street

This paper is toward the system of automatic text summarization developed by «DC – Systems» company in cooperation with the faculty of computer science at HSE. The summary is a concise description of the text in terms of its content and meaning, i.e. from the point of view of its semantics. The purpose of the summarization is to reduce the text as much as possible while maintaining the main content. A summary in this article is built using syntactically correlated word combinations. In this case, the possible additional meanings of separate fragments of the text are neglected. The quality of the summary is evaluated by a matching to the source text in terms of semantics. The main problem is split into two parts: an evaluation of the whole text semantics, without subdivision into parts, and the text transformation to derive an annotation. The architecture of the developed system and the main algorithm are described. An example of summary derived by the system and its quality evaluation has been provided. The current version of the system has following restrictions: it does not permit any formulas and special signs.

Keywords: *automatic summarization, automatic text processing, corpus linguistics.*

DOI: 10.21469/22233792.4.4.04

References

[1] L. P. Markushevskaya, Yu. A. Tsapaeva. *Annotirovanie i referirovanie*. SPbGU ITMO, 2008.(In Russian)

[2] A. I. Novikov. *Semantika teksta i ee formalizaciya*. M.: Nauka, 1983.(In Russian)

[3] N. P. Peshkova. Semantika i smysl teksta: $>$, $<$, $=$, #?. // *Vestnik Chelyabinskogo gosudarstvennogo universiteta. Filologiya. Iskusstvovedenie*, 15(370):69–77, 2015.(In Russian)

[4] P. G. Osminin. Model' avtomaticheskogo referirovaniya na osnove bazy znanij, orientirovannaya na avtomaticheskij perevod. // *Vestnik YUUrGU. Seriya «Lingvistika»*. Tom 11, 2, 2014.(In Russian)

[5] P. Braslavskij, I. Kolychev. Avtomaticheskoe referirovanie veb-dokumentov s uchetom zaprosa. // *Avtomaticheskaya obrabotka veb-dannyh*, 2005.(In Russian)

[6] A. Nenkova, K. McKeown, C. C. Aggarwal, C. X. Zhai. A survey of text summarization techniques. // *Mining Text Data*, 2012.

[7] U. Han, I. Mani. Sistemy avtomaticheskogo referirovaniya. // *Otkrytye sistemy*, 12, 2000.(In Russian)

[8] N. S. Valgina. *Sovremennyj russkij yazyk: Sintaksis: Uchebnik*. M.: Vysshaya shkola, 2003.(In Russian)

[9] P. G. Osminin. *Postroenie modeli referirovaniya i annotirovaniya nauchno-tekhnicheskih tekstov, orientirovannoj na avtomaticheskij perevod*. PhD thesis, 2016.(In Russian)

[10] D. Harman, P. Over. The effects of human variation in duc summarization evaluation. // *In Proceedings of the Text Summarization Branches Out Workshop*, Barcelona, Spain, 2004.

[11] Chin-Yew Lin. A package for automatic evaluation of summaries. // *In Proceedings of the Text Summarization Branches Out Workshop*, Barcelona, Spain, 2004.

[12] A. Nenkova, R. Passonneau, K. Mckeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. // *ACM Transactions on Speech and Language Processing*, 4, 2007.

[13] N. S. Balashova. Semantika. // *Institut filologii i zhurnalistiki SGU im. Chernyshevskogo*, 2011.(In Russian)

[14] N. N. Leont'eva. *K teorii avtomaticheskogo ponimaniya teksta. CH.1. Modelirovanie sistemy «myagkogo ponimaniya» teksta: Informacionno-lingvisticheskaya model'*. M.: Izd. MGU, 2000.(In Russian)

[15] N. N. Leont'eva. *K teorii avtomaticheskogo ponimaniya teksta. CH. 2. Semanticheskie slovari: sostav, struktura, metodika sozdaniya*. M.: Izd. MGU, 2001.(In Russian)

[16] N. N. Leont'eva. *K teorii avtomaticheskogo ponimaniya teksta. CH. 3. Semanticheskij komponent. Lokal'nyj semanticheskij analiz*. M.: Izd. MGU, 2002.(In Russian)

[17] A. I. Golovnya. *Slovar' leksiko-grammaticheskih omonimov*. Mn.: BGU, 2007.(In Russian)

[18] L. M. Leshcheva. *Leksicheskaya polisemiya v kognitivnom aspekte*. M.:Yazyki slavyanskoj kul'tury. Znak., 2014.(In Russian)

[19] A. I. Novikov. *Tekst i ego smyslovye dominanty*. M.: Institut yazykoznaniya RAN, 2007.(In Russian)

[20] M. Hagen. *Polnaya paradigma. Morfologiya. CHastotnyj slovar'. Sovmeshchennyj slovar'*.(In Russian)

[21] S. I. Ozhegov, N. YU. Shvedova. *Tolkovyj slovar' russkogo yazyka*. M.: OOO «A TEMP», 2006.(In Russian)

[22] V. Dal'. *Tolkovyj slovar' zhivogo velikorusskogo yazyka*. Drofa, 2011.(In Russian)

[23] *Bol'shaya sovetskaya ehnciklopediya*. Sovetskaya ehnciklopediya, 1978.(In Russian)

[24] N. YU. Shvedova. *19. Russkij semanticheskij slovar'. Tolkovyj slovar', sistematizirovannyj po klassam slov i znachenij*. M.: Azbukovnik, 1998.(In Russian)

[25] A. P. Zykov. Metod sglazhivaniya veroyatnostej n-gramm na osnove modelirovaniya matematicheskogo ozhidaniya ih vstrechaemosti. // *Trudy SPIIRAN*, 4(19), 2015.(In Russian)

[26] R. Kneser, H. Ney. Improved backing-off for m-gram language modeling. // *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181–184, 1995.

[27] Z. E. Aleksandrova. *Slovar' sinonimov russkogo yazyka. Prakticheskij spravochnik*. Russkij yazyk, 2001.(In Russian)

[28] C. Yanshuai, W. Luyu. Automatic selection of t-sne perplexity. // *In ICML 2017 AutoML Workshop and Conference Proceedings*, 2017.

[29] E. A. Budnikov. Ocenivanie veroyatnostej poyavleniya strok v estestvennom yazyke. // *Mashinnoe obuchenie i analiz dannyh*, 1(3), 2012.(In Russian)

Received December 23, 2018