Автоматическое реферирование речевых сообщений

Кушнир Д.А., *кандидат технических наук,*

Ромашкин Ю.Н., *кандидат технических наук*

В статье представлен подход к решению задачи автоматического реферирования речевых сообщений на основе кластерного анализа и n-граммных языковых моделей, разработаны критерии и предложена методика оценки качества автоматических рефератов. Проведены экспериментальные исследования с использованием разных мер информационной значимости термов, сформулированы основные выводы о качестве автоматических рефератов в зависимости от количества ошибок распознавания, длительности реферируемых речевых сообщений и коэффициента сжатия рефератов.

• распознавание речи • автоматическое реферирование речевых сообщений • извлечение полезной информации • n-граммные модели языка • кластеризация текстовых документов.

The paper presents an approach to solving the problem of automatic summarization of voice messages based on cluster analysis and n-gram language models, and developed criteria for assessing the quality of the technique of automatic summaries. Experimental studies using different measures of information value terms, sets out the basic conclusions about the quality of automatic abstracts according to the number of recognition errors, length of peer voice and the compression ratio of abstracts.

• speech recognition • automatic speech summarization • Information retrieval • n-gram language modeling • clusterization of text documents.

Введение

Задача автоматического реферирования в целом заключается в определении тематики документа, выделении семантически наиболее значимых слов, фраз и предложений или синтезе связного текста, отражающего основное содержание документа при уменьшении его объёма. Связный текст — не единственный конечный результат реферирования.

Это может быть список ключевых слов, набор семантических отношений, заполненные поля некоторой структуры данных и др.

Автоматическое реферирование речевых сообщений сводится к реферированию текстов, полученных в результате автоматического распознавания этих сообщений. Оно относится к классу нерешённых в настоящее время задач в связи с отсутствием надёжных систем распознавания слитной речи. Дополнительно возможно привлечение просодической информации в речевом сигнале, что в некоторых случаях позволяет выделять акцентированные фрагменты речи, которые предположительно являются семантически более значимыми.

В данной статье предлагается подход к автоматическому реферированию речевых сообщений, учитывающий ошибки распознавания речи, а также отсутствие автоматической сегментации полученного текста на семантические единицы (предложения и фразы).

Чтобы уменьшить влияние ошибок распознавания, полученный в результате распознавания текст пропускается через n-граммную языковую модель, которая служит своеобразным фильтром текста от слов и словосочетаний, не связанных с общим содержанием. При этом слова и словосочетания, не распознанные в рамках языковой модели, не исключаются из последующего анализа, а отмечаются как малоинформативные. Напротив, распознанные n-программы слов служат ядром формирования семантических единиц выходного текста реферата.

Реализация такой процедуры предполагает использование большого корпуса текстовых документов, который предварительно подвергается процедуре кластеризации. Для каждого кластера документов строится своя n-граммная модель, а реферируемый текст на первом этапе обработки при помощи кластерного анализа относится к ближайшему кластеру и его n-граммной модели.

Обучение системы реферирования

Структурная схема обучения системы автоматического реферирования речевых сообщений показана на рис. 1.

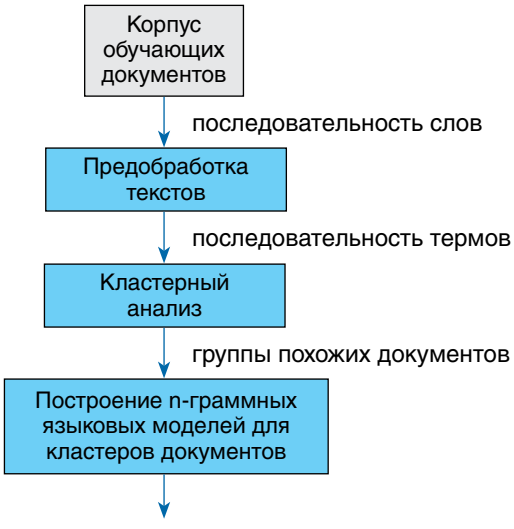

Рис. 1. Структурная схема обучения системы реферирования

Модуль предобработки текстов используется как на этапе обучения, так и на этапе реферирования речевых сообщений. Основная задача этого модуля заключается в преобразовании всех словоформ входного текста в нормальную форму и удалении малозначимых слов и словосочетаний. Лингвистическое обеспечение системы состоит из морфологического словаря, стеммера, словарей стоп-слов и синонимов. Морфологический словарь представлен различными формами слов русского языка со ссылками на их нормальную форму. Он необходим при работе с русским языком со сложной системой флексий. В данной работе словарь насчитывал более 3 млн словоформ. Если слово отсутствует в словаре, то оно поступает на вход стеммера, который осуществляет формальное выделение основы слова. Словарь стоп-слов состоял из 1200 малоинформативных слов и словосочетаний.

Сбор обучающего множества текстов проводился с помощью информационно-поисковой системы. В качестве источников были выбраны интернет-сайты (более 50, включая архивы), среди них www.rian.ru, www.rbc.ru, www.kp.ru, www.aif.ru и др. Общий объём обучающей выборки составил порядка 200 тыс. документов.

Для выполнения кластерного анализа использовалась векторно-пространственная модель представления текстов. В рамках данной модели каждому терму документа сопоставляется неотрицательный вес, и образ документа представляется в виде многомерного вектора. В качестве весов терма использовалась мера TF-IDF [1]. Элементы многомерного вектора нормировались для учёта размера документа путём деления на квадратный корень из суммы весов всех термов. Алгоритм k-средних разбивал множество текстов, представленных в векторном пространстве, на заранее известное число кластеров. В результате было получено 110 кластеров, каждый из которых включал в себя от 1000 до 3000 документов.

Языковые модели, реализованные в данной работе, представляют собой иерархические структуры из динамических ассоциативных запоминающих устройств (ИС ДАЗУ) [2]. Иерархическая структура ДАЗУ — многоуровневая однородная сеть нейроэлементов (НЭ). На первом уровне множество НЭ соответствует множеству отдельных слов в нормальной форме, на втором — биграммам слов, на третьем — триграммам слов и т.д. Таким образом, каждый НЭ соответствует n-грамме слов и однозначно идентифицируется своими координатами (уровнем и номером).

Кроме того, он содержит в себе поля, необходимые для выполнения аналитики: частоту встречаемости каждого НЭ в обучающих примерах, ссылку на НЭ-вершину, индекс текстового документа и признак того, что НЭ является вершинным, т.е. включает в себя целое предложение. Частота встречаемости показывает, насколько часто соответствующая последовательность слов встречалась в обучающих данных. Ссылка на НЭ-вершину содержит в себе механизм доступа ко всем элементам-вершинам, частью которых является данный НЭ. При такой реализации НЭ-вершина соответствует одному предложению текста. Благодаря такой ссылке в ИС ДАЗУ есть возможность быстро восстановить все предложения, в которые входит анализируемый НЭ. Индекс текстового документа используется для организации доступа к оригиналу с целью извлечения из него исходных последовательностей слов. Данная процедура необходима для того, чтобы иметь возможность получить список реализаций НЭ в исходных текстах. Признак вершины используется для выявления n-грамм целых предложений.

Процедура обучения ИС ДАЗУ сводится к запоминанию входных последовательностей термов, поступающих из обучающих текстов, после их предварительной обработки. В результате обучения n-граммная языковая модель позволяет распознавать произвольные последовательности термов в обрабатываемом тексте.

Алгоритм автоматического реферирования

Структурная схема разработанного алгоритма автоматического реферирования речевых сообщений (РС) показана на рис. 2.

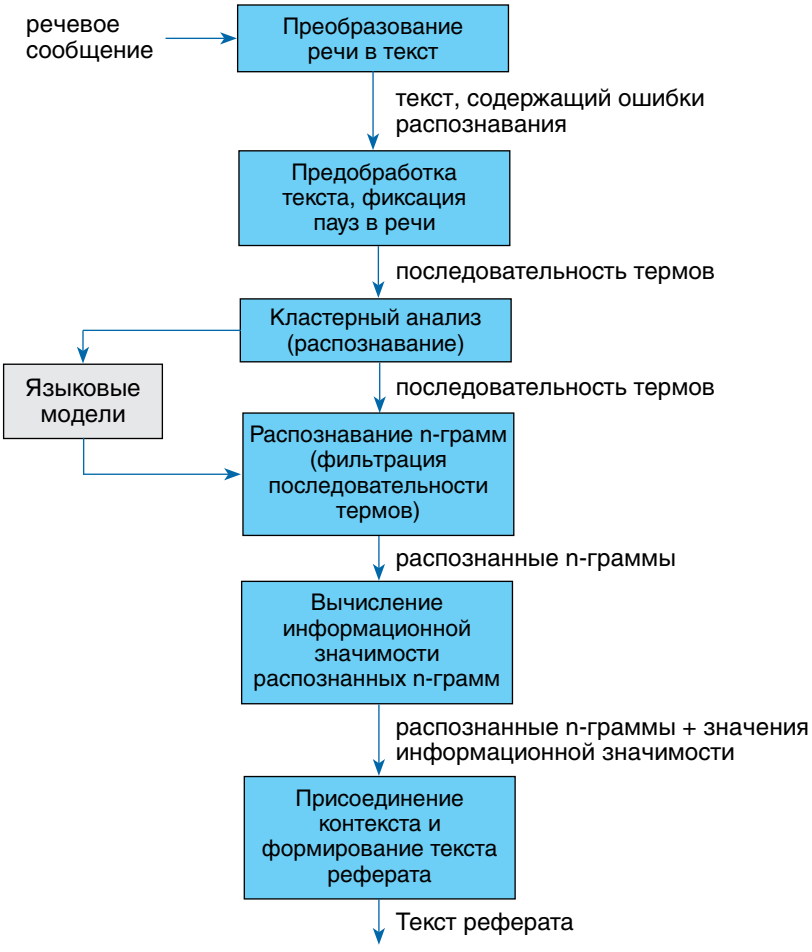

Рис. 2. Схема алгоритма автоматического реферирования РС

Результат автоматического преобразования речи в текст поступает на вход модуля, определяющего идентификатор кластера, которому больше всего соответствует входной текст. По распознанному кластеру загружается соответствующая n-граммная модель. Необходимость данного этапа реферирования обусловлена тем, что результат автоматического распознавания слитной речи содержит много ошибок. Их можно частично устранить при помощи фильтров, настроенных на заданную предметную область реферируемого текста. Роль такого фильтра выполняет ИС ДАЗУ, построенная на множестве текстов, принадлежащих распознанному кластеру. Фильтрация реализуется за счёт распознавания n-грамм в ИС ДАЗУ. Предполагается, что если языковая модель действительно соответствует входному речевому сообщению, то будут распознаны только те слова и словосочетания, которые уместны в рамках данной модели, остальные будут не распознаны и отмечены как неинформативные на этом этапе.

Одна из особенностей задачи реферирования речевых сообщений заключается в том, что в распознанном тексте недостаточно информации для его точной сегментации на такие семантические единицы, как фразы и предложения. Обучение ИС ДАЗУ происходит на предложениях. По этой причине распознанные n-граммы, как правило, являются частью одного предложения.

Для каждой распознанной n-граммы вычисляется значение информационной важности, которое складывается из значений информационной значимости входящих в неё отдельных термов. В исследованиях проверялись три способа вычисления информационной значимости термов.

Первый способ соответствовал выражению из [3, 4]:

$$I_i^{(1)} = f_i \cdot \log \frac{F_\Sigma}{F_i},$$

где f_i и F_i — частоты встречаемости определённого слова во входном тексте и в корпусе обучающих текстов соответственно, F_Σ — общее количество слов в корпусе текстов.

При втором способе вычислялось произведение

$$I_i^{(2)} = TF_i \cdot IDF_i = \frac{n_i}{\sum_k n_k} \cdot \log \frac{|D|}{|d_i \supset t_i|},$$

где TF — отношение числа вхождений i-го терма к общему количеству слов в тексте, IDF — обратная частота документа, определяемая как логарифм отношения общего количества документов в кластере к количеству документов кластера, в которых встречается терм t_i.

Третий способ (TF-IDF-LEN) основан на мере TF-IDF с добавлением длины терма в символах (LEN) [5]:

$$I_i^{(3)} = TF_i \cdot IDF_i \cdot LEN.$$

Идея его базируется на предположении, что часто встречающиеся слова стремятся быть краткими, т.е. являются стоп-словами.

Многие термы реферируемого текста не несут существенной информационной нагрузки, но они необходимы в качестве контекста к информационно значимым n-граммам для повышения связности изложения. Однако определить автоматически, где начинается и заканчивается нужный контекст информационно значимой n-граммы, задача нетривиальная.

Для её решения был разработан критерий *равномерного распределения информационной важности* по реферату. Идея предлагаемого подхода состоит в следующем: значение информационной значимости n-граммы пропорционально «силе притяжения» контекста, т.е. чем важнее распознанная n-грамма, тем более широким контекстом целесообразно её снабдить при составлении текста реферата. При этом в процессе расширения контекста происходит постоянный пересчёт общей информационной важности формируемой семантической единицы. Как только она достигает некоторого среднего значения, наращивание контекста текущей n-граммы прекращается и происходит переход к следующей. Все n-граммы обрабатываются в порядке уменьшения их информационной важности.

Процедуру выравнивания информационной важности (путём присоединения контекста к информационно значимым n-граммам) можно представить в такой последовательности:

1) все распознанные n-граммы ранжируются в порядке уменьшения их информационной значимости (ИЗ);

2) по заданному объёму реферата определяется количество n-грамм, которые должны попасть в реферат из ранжированного списка, начиная с самой важной;

3) для полученного в п.2 подсписка n-грамм вычисляется среднее значение ($\overline{I}$) ИЗ;

4) задаётся желаемое значение ИЗ фрагментов исходного текста, из которых будет складываться реферат: $I_0 = \overline{I} / k$, где коэффициент *k* определяет значимость контекста и варьируется от 1 до 2;

5) к каждой n-грамме присоединяются их левые и правые контексты в порядке уменьшения ИЗ до тех пор, пока для полученного фрагмента текста ИЗ не достигнет желаемого значения I_0.

Методики оценки качества

В настоящее время применительно к обработке текстовых документов разработан ряд методик оценки качества алгоритмов их автоматического реферирования [6–10]. Однако к задаче реферирования речевых сообщений, когда отдельные слова распознаются неверно и информация о границах предложений отсутствует, они не подходят.

Качество реферирования речевых сообщений зависит от точности работы модуля автоматического распознавания речи и эффективности модуля автоматического реферирования результата распознавания. Поскольку в данной работе функционирование этих модулей осуществлялось независимо друг от друга, желательно провести оценку качества как для системы в целом, так и для модуля реферирования отдельно. При этом целесообразно использовать два способа оценки качества алгоритмов реферирования:

1) объективная оценка на основе вычисления некоторой меры подобия рефератов, подготовленных экспертами и полученных автоматически;

2) субъективная оценка с помощью метода экспертных оценок качества автоматических рефератов.

Учитывая указанную выше специфику реферируемого текста, подобие рефератов характеризовалось размером области пересечения множеств лексических единиц (n-грамм слов) рефератов, составленного экспертом и полученного автоматически.

Процедура получения объективной оценки качества реферирования состоит из пяти шагов:

1. Отбирается заданное количество тестовых речевых сообщений с представительным разбросом по точности автоматического преобразования в текст (значения ошибки правильного распознавания слов от 0.1 до 0.5).

2. Рефераты составляются экспертами путём анализа результата автоматического распознавания речи и последующего выбора n-грамм слов, которые наиболее точно характеризуют семантику каждого речевого сообщения (для заданных коэффициентов сжатия текста в 2–10 раз с шагом 2).

3. Автоматическое формирование рефератов в соответствии с разработанным алгоритмом и с заданными коэффициентами сжатия.

4. Сравнение составленных рефератов по полноте и точности в соответствии со следующими выражениями:

$$\textit{Полнота} = R^{(n)} = \frac{\{N_{\dot{Y}}^{(n)}\} \cap \{N_A^{(n)}\}}{N_{\dot{Y}}^{(n)}}$$

$$\textit{Точность} = P^{(n)} = \frac{\{N_{\dot{Y}}^{(n)}\} \cap \{N_A^{(n)}\}}{N_A^{(n)}},$$

где $N_{\dot{Y}}^{(n)}$ — количество n-грамм слов в реферате, подготовленном экспертом;

$N_{\dot{A}}^{(n)}$ — количество n-грамм слов в реферате, полученном автоматически.

Т.е. полнота равна отношению числа n-грамм автоматического реферата, совпавших с n-граммами реферата эксперта, к общему количеству n-грамм в реферате эксперта. Точность равна отношению числа n-грамм автоматического реферата, совпавших с n-граммами реферата эксперта, к общему количеству n-грамм в автоматическом реферате. Для проверки совпадения выбирались униграммы и биграммы слов (т.е. n=1, 2).

Для получения одной оценки качества автоматического реферата использовалась F-мера — гармоническое среднее полноты и точности:

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}}, \text{ где } \alpha = 0{,}5$$

5. Усреднение полученных оценок качества по всем тестовым речевым сообщениям (отдельно по точности автоматического распознавания речи и заданному коэффициенту сжатия текста).

При субъективной оценке качества автоматических рефератов группа из пяти экспертов выставляла баллы в соответствии со шкалой качества (таблица 1). Затем по среднему значению проставленных баллов определялся класс качества реферата для заданного коэффициента сжатия текста.

Таблица 1

Класс качества	Характеристика класса	Балльная оценка
Высший	В реферате отражены все основные факты и упоминаются все личности из речевого сообщения (при данном объёме реферата)	≥ 4,5
I	По реферату можно определить суть речевого сообщения. Он содержит несколько (более одного) важных фактов, включая основной, или упоминание о ключевой личности	от 3,5 до 4,4
II	По реферату можно определить общий смысл речевого сообщения. Реферат содержит один важный, не обязательно основной, факт или упоминание об одной, не обязательно ключевой, личности	от 2,5 до 3,4
III	По реферату понятна только тема речевого сообщения. Суть сообщения определить невозможно. Реферат не содержит ни одного важного факта или упоминания о какой-либо личности	от 1,5 до 2,4
Срыв реферирования	По реферату невозможно определить даже тему речевого сообщения	<1,5

Результаты экспериментов

В результате работы группы из пяти экспертов была получена база рефератов со следующими характеристиками:

— общее количество обработанных речевых сообщений на русском языке из новостных передач радиовещательных станций = 176;

— количество рефератов на каждое сообщение = 5 (с коэффициентами сжатия текста 2, 4, 6, 8 и 10);

— общее количество рефератов на одного эксперта: = 880;

— ошибка алгоритма автоматического распознавания слов (WER) в слитной речи (дикторской или спонтанной) от 0,1 до 0,5;

— распределение числа реферируемых текстов в зависимости от WER приведено в таблице 2.

Таблица 2

WER	0.1	0.2	0.3	0.4	0.5
Количество реферируемых текстов	6	27	46	59	38

С целью определения лучшего способа вычисления информационной значимости термов и значения коэффициента значимости контекста были получены объективные оценки качества рефератов для различных возможных комбинаций этих параметров (для каждого варианта создавалось 880 автоматических рефератов). Анализ этих оценок показал, что наилучшим способом вычисления информационной значимости n-грамм слов, включаемых в автоматический реферат, является TF-IDF-LEN с коэффициентом значимости контекста равным 2.

Сбор экспертных оценок качества автоматического реферирования проводился с помощью специальной программы, вид главного окна которой показан на рис. 3. Справа в окне отображается шкала допустимых балльных оценок (шаг оценивания составляет 0,1) и качественное описание каждого из 5 возможных классов. В левой верхней части окна отображается полный текст речевого сообщения (без ошибок распознавания), а ниже — автоматически сформированный реферат для этого сообщения с заданным коэффициентом сжатия.

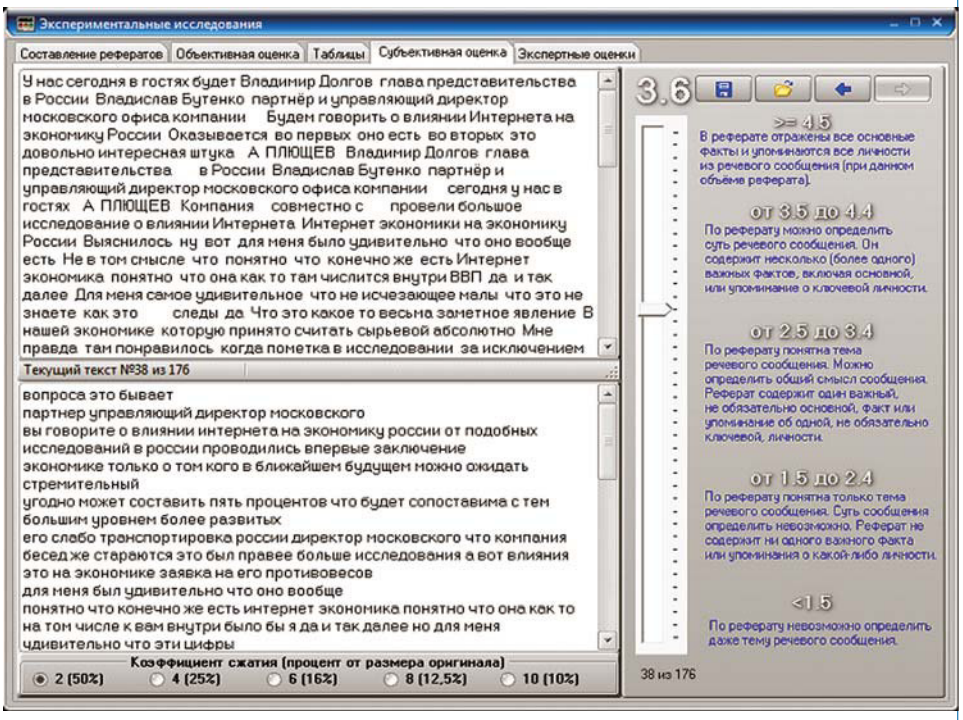

Рис. 3. Внешний вид интерфейса программы для экспертных оценок класса качества реферирования

Средние значения балльных оценок, выставленные всеми экспертами, и соответствующие им номера классов качества автоматического реферирования в зависимости от величины ошибки распознавания речи и коэффициента сжатия текста приведены в таблице 3, а в зависимости от длительности (T) реферируемых речевых сообщений — в таблице 4.

Они позволяют сделать выводы:

— качество рефератов слабо зависит от вероятности ошибок распознавания слов в слитной речи (в пределах от 0.1 до 0.5), оставаясь преимущественно в рамках одного класса;

— качество рефератов монотонно возрастает с увеличением длительности речевого сообщения. В основном, это связано с используемой мерой информационной значимости, в которой частота встречаемости терма пропорциональна его значимости;

— для коротких речевых сообщений (длительностью менее 1 мин.) при коэффициенте сжатия 10 наблюдается срыв реферирования.

Таблица 3

WER	Балльные оценки / классы качества автоматического реферирования				
	Коэффициент сжатия				
	2	4	6	8	10
0,1	3,3/ II	2,6/ II	2,0/ III	1,8/ III	1,7/ III
0,2	3,6/ I	3,0/ II	2,3/ III	2,1/ III	1,8/ III
0,3	3,9/ I	3,3/ II	2,7/ II	2,5/ III	2,2/ III
0,4	3,5/ II	3,0/ II	2,5/ II	2,2/ III	2,0/ III
0,5	3,2/ II	2,7/ II	2,3/ III	2,1/ III	1,8/ III

Таблица 4

T, c	Балльные оценки / классы качества автоматического реферирования				
	Коэффициент сжатия				
	2	4	6	8	10
< 45	3,1/ II	2,4/ III	1,8/ III	1,6/ III	1,1/ C
45–60	3,2/ II	2,5/ II	1,8/ III	1,6/ III	1,3/ C
60–120	3,3/ II	2,7/ II	2,0/ III	1,8/ III	1,6/ III
121–180	3,4/ II	2,8/ II	2,3/ III	2,0/ III	1,7/ III
181–240	3,5/ I	2,9/ II	2,4/ III	2,2/ III	2,0/ III
> 240	3,6/ I	3,1/ II	2,7/ II	2,4/ III	2,1/ III

Заключение

Разработан метод автоматического реферирования речевых сообщений на русском языке, обеспечивающий максимальное согласование экспертных и автоматических рефератов по выбранным мерам (полнота и точность для униграмм и биграмм). Предложены методики оценки качества автоматического реферирования, использующие расчётные (объективные) и экспертные (субъективные) количественные показатели.

С использованием тестовых записей речевых сообщений новостных программ радиовещания на русском языке получены оценки качества их автоматического реферирования для различных длительностей (T) сообщений и коэффициентов (K) сжатия текста. Экспериментально установлено, что при T=1–3 мин. разработанный алгоритм способен обеспечивать I–III классы качества автоматического реферирования. При T<1 мин. и K=10 происходит срыв автоматического реферирования.

С увеличением ошибки алгоритма распознавания слов в речи от 0.1 до 0.5 класс качества рефератов преимущественно не изменяется.

Список литературы

1. *Debole F. and Sebastiani F.* Supervised term weighting for automated text categorization. In the Proceedings of SAC-03, 18th ACM Symposium on Applied Computing, Melbourne, US: ACM Press, New York, US, 2003. Pp. 784–788.

2. *Харламов А.А.* Нейроподобные элементы с временной суммацией входного сигнала и блоки ассоциативной памяти на основе этих элементов // Вопросы кибернетики. Устройства и системы. М.: МИРЭА, 1983. С. 57–68.

3. *Kikuchi T., Furui S., Hori C.* Automatic Speech Summarization Based on Sentence Extraction and Compaction // ICASSP 2003. Pp. 384–387.

4. *Furui S.* Recent Advances in Automatic Speech Summarization // Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007.

5. *Luhn H.* The automatic creation of literature abstracts. In IBM Journal of Research and Development, Vol. 2(2), 1958. Pp. 159–165.

6. *Salton G. et al.* Automatic Text Structuring and Summarization // Information Processing & Management, Vol. 33, No. 2, 1997. Pp. 193–207.

7. *Radev D. and Tam D.* Single-document and multidocument summary evaluation via relative utility // Poster session, *CIKM'03*, 2003.

8. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute. University of Southern California 2004.

9. *Lin C.-Y.* ROUGE: A package for automatic evaluation of summaries// Proc. of the Workshop on Text Summarization Branches Out (ACL'2004). Barcelona, Spain, 2004. Pp. 74–81.

10. *Тарасов С.Д.* Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS // Труды 11-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2009. Петрозаводск, Россия, 2009. С. 86–93.

Сведения об авторах

Кушнир Д.А. —
кандидат технических наук, ведущий научный сотрудник ЗАО «НТЦ «Поиск-ИТ».
kushnir@speechtechnology.ru

Ромашкин Ю.Н. —
кандидат технических наук.