Типовая архитектура системы реферирования звучащей речи

Ермоленко Т.В.

В статье рассмотрена типовая архитектура системы реферирования звучащей речи, описаны особенности, отличающие её от систем реферирования текста. Обязательными элементами системы реферирования речевых сообщений являются блоки автоматического распознавания и сегментации. Корректно работающая процедура сегментации позволит повысить точность распознавания и определить границы предложений. При выделении наиболее информативных предложений, входящих в текст реферата, на этапе взвешивания элементов сообщения (слов, фраз) необходимо учитывать точность их распознавания.

● системы автоматического реферирования ● распознавание речи ● сегментация речевого сигнала ● акустическая модель ● лингвистическая модель ● просодические характеристики речи ● извлечение значимых предложений

The article describes the typical architecture of the speech summarization system. The features that disting uish it from text summarization systems have been described. Mandatory elements of the summarization system voice messages are units of automatic recognition and segmentation. Working correctly segmentation procedure will improve the accuracy of recognition and define the boundaries of the sentences. When you select the most informative sentences included in the text of the abstract, in the step of elements discourse weighing (words, phrases) must take into account the accuracy of their recognition.

● automatic summarization systems ● speech recognition ● speech signal segmentation ● acoustic model ● linguistic model ● prosodic features of speech ● important sentence extraction

Введение

Автоматическое реферирование или составление аннотаций, иными словами, извлечение наиболее важных или характерных фрагментов из одного или многих источников информации, относится к задачам автоматизации процессов аналитико-синтетической обработки информации. Развитие информационных ресурсов Интернет многократно усугубило проблему информационной перегрузки, оперативно получать корректные информационные сводки становится всё сложнее, что приводит

к необходимости создания эффективных средств реферирования, которые могут стать неотъемлемой частью повседневной жизни.

Растущий объём мультимедийной информации делает её едва ли не самым важным объектом для обработки средствами реферирования, хотя исследования в этой области находятся ещё на очень ранней стадии.

Данная статья посвящена вопросам реферирования слитной речи. Чтобы понять, о чём идёт разговор, необязательно распознавать всё сказанное, достаточно выделять наиболее важные слова, словосочетания и предложения в слитной речи, исходя из структуры текста, интонации. Распознавание ключевых слов и фраз – первый шаг к автоматическому составлению текстовых резюме звукозаписей. Успешное функционирование такой системы напрямую зависит от качества работы систем автоматического распознавания естественной человеческой речи, поскольку реферат составляется на основе текста, полученного по распознанному звуковому сообщению.

Существующие методы работы со звуком позволяют вычленять из потока аудиоинформации законченные фрагменты (иными словами, распознавать периоды тишины в разговоре, смену говорящего, снятие телефонной трубки, а также осуществлять контентный анализ). Прогресс в области автоматического распознавания речи из аудиоисточников должен стимулировать развитие этих средств реферирования.

Общая схема функционирования системы автоматического реферирования речи

История применения вычислительной техники для реферирования насчитывает уже более 50 лет и связана с именами таких исследователей, как Г.П. Лун [1], И.П. Севбо [2], Э.Ф. Скороходько [3], Д.Г. Лахути [4, 5], Р.Г. Пиотровский [6]. Основной областью применения инструментов реферирования являются естественно-языковые тексты.

Хорошее представление о состоянии исследований и разработок в области автоматического реферирования дают материалы ежегодной конференции DUC (Document Understanding Conference) [Document Understanding Conferences, http://duc.nist.gov], в рамках которой проводится сравнительная оценка методов реферирования отдельного документа на стандартном наборе данных.

В общей структуре систем автоматического реферирования текстов (САРТ) можно выделить три взаимосвязанных блока: блок анализа входного текста, блок взвешивания (оценивания) элементов текста (слов, словосочетаний, предложений и др.) и блок генерации реферата. Самым трудоёмким процессом является автоматический анализ текста, который может состоять из нескольких стадий: лексического, лексико-грамматического, синтаксического и семантического анализа. Современное состояние разработок в области автоматического анализа текста таково, что полностью успешно реализуется только лексический и лексико-грамматический анализ.

Отличием системы автоматического реферирования речевых сообщений от САРТ являются обязательное включение в эту систему блока автоматического распознавания (ASR) и необходимость при взвешивании элементов сообщения (слов, фраз) учитывать точность их распознавания. Общая функциональная схема такой системы показана на рис. 1.

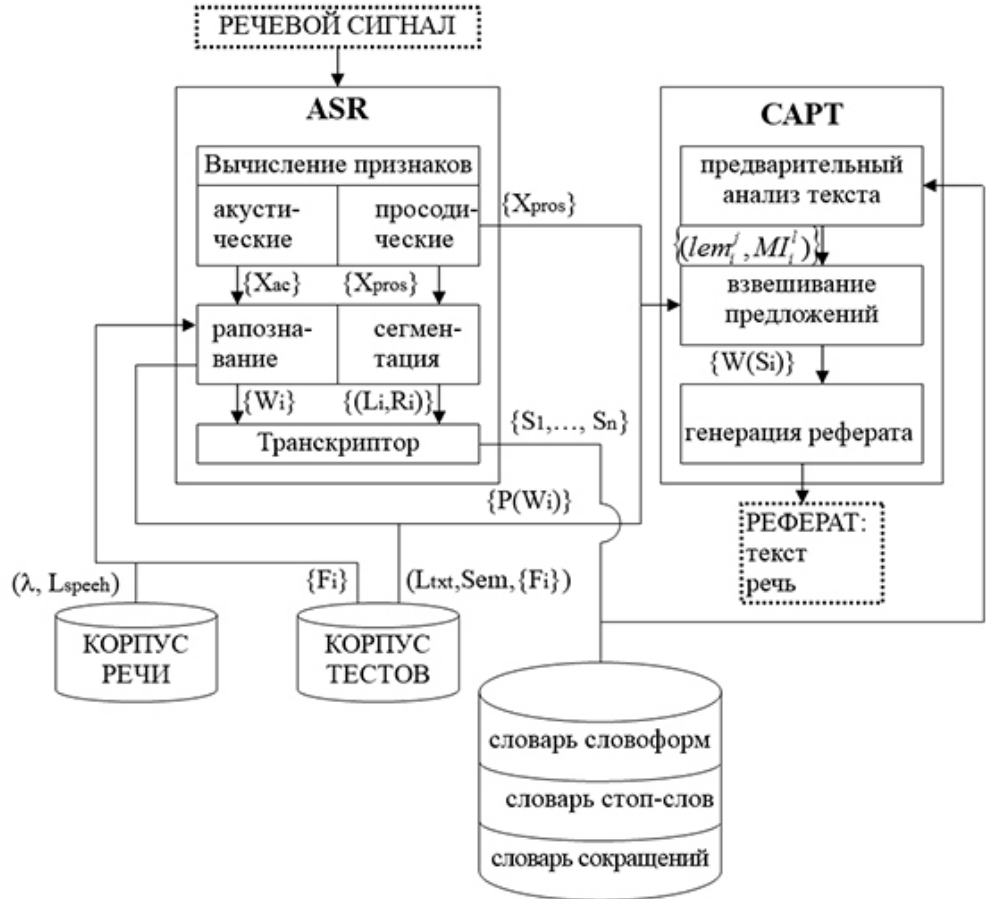

Рис 1. Функциональная схема системы реферирования речевых сообщений

Обозначения, используемые на рисунке:

(λ, Lspeech) – акустическая и лингвистическая модель, полученная по речевому корпусу;

{S1,..., Sn} – массив предложений в текстовом виде, полученный на выходе транскриптора;

Li, Ri – левая и правая границы предложения Si;

{Xac} – массив акустических признаков распознавания, полученных по фреймам входного речевого сигнала;

{Xpros} – массив просодических характеристик сигнала, полученного на участках, соответствующих произнесённому предложению;

{Wi} – массив распознанных слов;

P(Wi) – точность распознавания слова Wi;

W(Si) – вес предложения Si (в зависимости от метода взвешивания может быть векторной или скалярной величиной);

($lem^i_j MI^i_j$) – лемма и морфологическая информация слова Wj из предложения Si;

{Ltxt, Sem, TF} – полученные по корпусу текстов лингвистическая, семантическая модели и массив весов термов.

Как и в системах реферирования текстов, работу системы автоматического реферирования речевых сообщений можно разбить на три этапа: анализ распознанного текста, выделение значимых предложений из аудиоисточника и генерация на их основе реферата. При этом реферат может быть представлен как в тестовом виде, так и в виде звучащей речи. В первом случае налицо такие преимущества, как лёгкая навигация по документу и его обработка, однако недостатками данного представления являются возможность ошибок распознавания, а также потеря информации об эмоциональном состоянии диктора и его просодических особенностях произношения.

Предварительный анализ текста распознанного сообщения

Распознавание слитной речи является задачей нахождения последовательности слов W согласно критерию максимума апостериорной вероятности:

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(O \mid W)P(W), \quad (1)$$

где $P(O|W)$ – вероятность того, что входная последовательность акустических признаков O была порождена реализацией слова W; $P(W)$ – лексическая вероятность слова W [7].

Для оценки этих вероятностей ASR-система должна располагать акустическими и лингвистическими источниками знаний, то есть используемая модель распознавания должна быть предварительно обучена на размеченном речевом и текстовом корпусах.

Вид акустической модели λ существенно зависит от метода распознавания. Например, в случае использования CMM в качестве λ используются: матрица вероятностей переходов, распределение вероятности символов и начального состояния, в случае применения нейросети акустической моделью будет массив её весовых коэффициентов.

На основе лингвистической модели вычисляется вероятность появления цепочки слов в предложении. Как правило, эта вероятность считается на основе модели n-грамм, полученной по тексту транскрипций фраз из речевого кор-

$$Lspeech(W_i) = P(W_i \mid W_1 ..., W_{i-1}).$$

пуса:

На результат распознавания слова также влияет его значимость. Оценить её возможно только по данным, полученным из текстового корпуса. Так, в

$$I(W_i) = f_i \log \frac{F_A}{F_i},$$

работах [8, 9] коэффициент значимости вычисляется как:

где f_i – частота встречаемости слова W_i в распознанной фразе, F_i – частота встречаемости слова W_i в корпусе текстов, F_A – число всех слов в корпусе текстов.

В работе [10] коэффициент значимости вычисляется без учета f_i .

Распознанная цепочка слов подаётся на вход блока предварительного анализа текста. В ходе его работы удаляются неинформативные слова, выделяются термины и определяются границы предложений. Для этого требуется некоторый набор экспертных данных:

- словарь сокращений, который содержит часто употребляемые сокращения;
- словарь стоп-слов, включающий служебные части речи (предлоги, союзы, вспомогательные глаголы и пр.), так как заранее известно, что они не являются терминами, а также неинформативные слова и словосочетания;
- словарь словоформ языка, в котором каждой словоформе соответствует множество лексико-грамматических классов, которые могут иметься у данной словоформы.

В словаре словоформ для каждого лексико-грамматического класса указывается частота его встречаемости относительно других лексико-грамматических классов данной словоформы. Частота обычно подсчитывается на корпусе текстов, в котором предварительно вручную каждому слову приведен в соответствие лексико-грамматический класс.

Помимо блока распознавания неотъемлемой частью ASR-системы в составе системы автоматического реферирования речи является блок сегментации. Его задача – определить с максимальной точностью границы предложений в произнесённой фразе. При распознавании слитной речи сегментация речевого сигнала на участки, содержащие единицы распознавания (фонемы или слоги), также необходима. Корректно работающий блок сегментации позволит повысить точность распознавания. Однако в случае неподготовленной речи, когда вместо малоамплитудных сигналов, соответствующих паузам между предложениями, могут быть паузы хезитации, эффективность работы сегментатора значительно ухудшается. Кроме того, может встречаться некорректное деление на синтагматические компоненты и т.п.

Обзор методов сегментации слитной речи выходит за рамки данной статьи, необходимо лишь заметить, что в общем случае границы предложений определяют на основе многочисленных критериев (например, максимума энтропии), используя как акустические, так и просодические характеристики. В связи с чем численные значения этих характеристик поступают на вход блока сегментации (см. рис. 1).

Результатом работы блока предварительного анализа текста распознанного сообщения является массив предложений с выделенными словами, для каждого из которых известны его морфологическая информация и лемма. Эти данные используются для выделения значимых предложений на основе определения информативности слов (терминов).

Выделение значимых предложений

Обязательным параметром, который необходимо учесть при взвешивании предложений, помимо коэффициента значимости слова, является значение апостериорной вероятности $P(W|O)$, полученной по акустическим признакам [11]. Кроме того, учитываются структурные и лексические признаки предложения при его взвешивании. К структурным относятся распределение в нём существительных, которое считается, исходя из предположении о том, что распределение слов в тексте можно оценить, используя распределение Пуассона. Распределение существительных отражает структуру дискурса и оценивается как [12]:

$$PoissonNoun_j(i) = \frac{\sum_{k=1}^{N_i} pois(p,\lambda)TF(k)}{N_i},$$

где N_i – количество существительных в предложении i документа j; $TF(k)$ – частота встречаемости слова k в документе j; p – частота встречаемости слова k в остальных документах коллекции.

Таким образом, если предложение содержит новое существительное, то, вероятно, оно содержит и новую информацию; с другой стороны, часто встречаемое существительное является значимым, а следовательно, повышается значимость предложения. И в том, и в другом случае увеличивается вероятность для этого предложения войти в реферат. Также учитывается положение предложения в тексте, его близость к началу.

К лексическим признакам предложения относятся:

- количество слов в предыдущем предложении;
- количество слов в следующем предложении;
- количество в предложении одушевлённых объектов или имён и фамилий, которые упоминаются в предложении впервые;
- сумма весов слов (TF*IDF) [13], входящих в предложение.

Также важную информацию о лингвистических особенностях речи позволяют получить просодические характеристики, входящие в акустические признаки. В частности, они незаменимы при выявлении ключевых элементов в структуре мелодического и энергетического контуров, акцентных единиц высказывания. Ещё одним их преимуществом является устойчивость к шумам, что особенно актуально при записи в сложных акустических условиях. В связи с этим просодические признаки вносят немалый вклад в вес предложения. Именно по мелодическому контуру можно определить интонационную конструкцию фразы и отнести её к повествовательной, вопросительной или восклицательной.

Авторы работ [10, 11, 12] в качестве просодических характеристик участка речевого сигнала, соответствующего предложению, использовали следующие:

- максимальное, минимальное и среднее значение частоты основного тона по словам предложения;
- вариация частоты основного тона;
- промежуток убывания частоты основного тона;
- максимальное, минимальное и среднее значения энергии по словам предложениями;
- вариация энергии;
- промежуток убывания энергии;
- средняя длительность фонемы;
- средняя длительность слога.

Необходимо заметить, что информационно значимые предложения в высказывании выделяются при помощи интонации, что приводит к росту значений перечисленных характеристик и необходимости учитывать их при взвешивании предложений.

В табл. 1 обобщены результаты исследований качества реферата речевых сообщений для коэффициента сжатия исходного текста – 10%, используемый классификатор – SVM. Для взвешивания предложений использовались акустические, структурные, лексические группы признаков. В качестве данных для проведения исследования в [10–12] использовались корпуса из документов новостных ресурсов (репортажи и интервью) и лекционного материала (презентации), язык сообщений – китайский, средняя длительность сообщения – 15 минут.

Таблица 1

Результаты исследований качества реферата

Оценка качества реферата	Акустические признаки	Значение оценки	Объем речевого материала (в сообщениях)	Ссылка на лит. источник
ROUGE-1	Значения 1,2 формант и ЧОТ; пересечение 0	0,6	205(НР)	11
ROUGE-2		0,532		
ROUGE-L		0,527		
ROUGE-1	Среднее значение ЧОТ и энергии; максимум энергии	0,4198	200(НР)	10
ROUGE-2		0,3425		
ROUGE-3		0,3056		
ROUGE-4		0,271		
F-мера	Длительность предложения; средняя длительность фонемы и слога; максимальное, минимальное и среднее значение ЧОТ; вариация ЧОТ; максимальное, минимальное и среднее значение энергии; вариация энергии	0,5989	347(НР)	12
		0,6277	60(Л)	

Обозначения в таблице: ЧОТ – частота основного тона; НР – новостные ресурсы; Л – лекционный материал.

В [12] сравнение оценки качества реферата проведено по 3 группам признаков (акустические, структурные, лексические) как отдельно по каждой из групп, так и в различных их комбинациях. Наиболее существенными являются признаки структурные и акустические. На материале новостных ресурсов исследования показали, что вклад лексических признаков относительно мал.

Взвешивание распознанных предложений на основе их лексических и структурных признаков, а также генерация реферата из самых значимых предложений проводится согласно методам реферирования естественно-языкового текста [14].

Выводы

Системы реферирования речевых сообщений, имеющие рассмотренную типовую архитектуру, были разработаны и протестированы на корпусах китайских, японских и английских текстов (в основном новостных). Была проведена оценка качества реферата с использованием при взвешивании предложения различных признаков: просодических, лексических и структурных. Проведённые эксперименты показали, что наиболее значимыми являются структурные признаки, вклад лексических относительно мал. Наилучшие результаты получаются при сочетании структурных и просодических признаков. Эти результаты остались стабильны даже после добавлении шума, точность распознавания при этом составляла менее 80%.

К сожалению, до сих пор экспериментов для систем, ориентированных на русский язык, не проводят. Существующие алгоритмы распознавания речи с большим словарём имеют ряд серьёзных недостатков, связанных с невысоким быстродействием, зависимостью от диктора и большим количеством ошибок распознавания, особенно при низком отношении сигнал/шум. В связи с этим автоматическое реферирование слитной речи относится к классу нерешенных задач.

В целом отрасль средств реферирования находится в самом начале своего развития. Существует единое мнение о необходимости лучших методов оценки, однако большинство задач ещё не решено, в том числе сохраняется необходимость в масштабируемых методологиях создания аннотаций. Тем не менее можно ожидать, что инструменты реферирования будут играть решающую роль в завоевании широких информационных пространств в будущем.

Литература

1. *Luhn H.* The automatic creation of literature abstracts. In IBM Journal of Research and Development. — 1958. — Vol. 2(2). — P. 159–165.
2. *Севбо И.П.* Структура связного текста и автоматизация реферирования. — М.: Наука, 1969. — 135 с.
3. *Скороходько Э.Ф.* Семантические сети и автоматическая обработка текста. — Киев: Наукова думка, 1983. — 219с.
4. *Лахути Д.Г.* Формализованное реферирование с использованием словесных клише (маркеров) / Д.Г. Лахути, Д.И. Блюменау, Н.И. Гендина // НТИ. Сер.2. 1981. — №2. — С. 16–20.
5. *ЛахутиД.Г.* Экстрагирование как один из подходов к автоматизации реферирования / Д.Г. Лахути, Д.И. Блюменау, И.С. Добронравов // Теория и практика механизации библ. и информ.-библиогр. процессов. — Л., 1982. — С 108–128.
6. *Пиотровский Р.Г.* Текст, машина, человек. — М: Наука, 1975. — 327 с.
7. *Takaaki Hori, Chiori Hori, Yasuhiro Minami.* Speech summarization using weighted finite-state transducers // In proceeding of 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003). — Geneva. — 2003. — P. 2817–2820.
8. *Tomonori Kikuchi, Sadaoki Furui, Chiori Hori.* Automatic speech summarization based on sentence extraction and compaction // in Proceedings of ICASSP. — Hongkon. — 2003. — Vol. 1. — P. 384–387.
9. *Sadaoki Furui.* Recent Advances in Automatic Speech Summarization // in Proceedings of conference Large Scale Semantic Access to Content (Text, Image, Video, and Sound). — Pittsburgh. — 2007. — P. 90–101.
10. *Yi-Ting Chen, Berlin Chen, Hsin-Min Wang.* A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization // IEEE Trans. on Audio, Speech, and Language Processing. — 2009. — Vol. 17. —№ 1. — P. 95–106.
11. *Shih-Hsiang Lin, Berlin Chen.* A Risk Minimization Framework for Extractive Speech Summarization // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Uppsala, Sweden, 11-16 July 2010). — Uppsala, 2010, — P. 79–87.

12. *Jian Zhang, Ho Yin Chan, Pascale Fung, Lu Cao.* A Comparative Study on Speech Summarization of broadcast news // In proceeding INTERSPEECH 2007. — Antwerp, 2007. — P. 2781–2784.
13. *Рязанова Н.Ю.* Анализ вопросов автоматизации поиска информации // Вестник МГТУ им. Н.Э. Баумана: электронное издание. — 2013. — С. 1–5.
14. *Тарасов С.Д.* Современные методы автоматического реферирования // Научно-технические ведомости СПБГПУ. — СПб: СПБГПУ, 2010. — № 6 (113). — С. 59–74.

Сведения об авторе:

Ермоленко Татьяна Владимировна,
кандидат технических наук, научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта МОНМС и НАН Украины. Распознаванием и обработкой речевых сигналов занимается с 2002 года. К области интересов также относится автоматическая обработка ЕЯ-текстов.