Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2017»

Москва, 31 мая — 3 июня 2017

МЕТОДИКА СОЗДАНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ФОРМИРОВАНИЯ РЕЧЕВОГО КОРПУСА

Пискунова В. Ю. (piskunova.victoria@gmail.com),
Бирин Д. А. (d.birin@kvant-rdi.spb.ru)

ФГУП НИИ «КВАНТ», Санкт-Петербург, Россия

AUTOMATED SYSTEM OF SPEECH CORPORA CREATION

Piskunova V. Y. (piskunova.victoria@gmail.com),
Birin D. A. (d.birin@kvant-rdi.spb.ru)

FSUE RDI “KVANT”, Saint-Petersburg, Russia

The focus of this paper is the development of the system that enables automatized speech and text corpora creation. Speech and text data play the key role in speech recognition research. However, it turns out often that the amount of data or the content of open source corpora are not sufficient to serve as basis for reliable analysis.

Automated system of speech corpora creation simplifies the creation of big text and speech corpora by downloading them from Internet resources such as newspapers, radio broadcasting and television sites. The developed system deals with patterns of web pages and media links created by an operator with the help of regular expression language. The system thus downloads media (audio and video fragments of radio and TV programs) and extracts texts from HTML-code of the web pages. It is also possible to make an automatic data extraction according to timetable set by the user that reduces to a minimum manual control.

Key words: speech corpora, text corpora, big data, speech recognition, corpus forming

Введение

В статье отражаются основные этапы создания специализированного программного обеспечения для сбора текстовой и медиа информации в сети Интернет, а также приводится обоснование необходимости разработки данной системы, связанное с отсутствием программных продуктов, которые могут выполнять подобные функции. Создание системы сбора текстовой и медиа информации, как будет показано ниже, было продиктовано необходимостью выполнения сотрудниками ФГУП НИИ «Квант» работ по сбору большого объема текстов и звуковых данных для формирования речевых и текстовых корпусов. Созданное программное обеспечение применяется для создания различных видов корпусов в зависимости от практических задач. К таким задачам относится, например, сбор корпуса новостных текстов, необходимого для формирования языковой модели на основе новостной лексики, а также формирования корпуса речи большого количества дикторов с текстовыми подстрочниками для построения акустических моделей и улучшения качества распознавания. Система сбора файлов, таким образом, входит в комплекс программных продуктов, образующих систему, которая используется для проведения работ по увеличению качества распознавания. Помимо описания процесса разработки системы в статье приводится методика отбора источников для создания речевого корпуса.

1. Опыт создания речевых корпусов

Целью создания речевых и текстовых баз является проведение работ в области лингвистики и всестороннего изучения языковых явлений. Анализ литературы по вопросу создания речевых корпусов показывает, что существует достаточно большое количество методик формирования речевых корпусов. К таким методикам относятся, например, сбор и последующее ручное аннотирование данных из общедоступных открытых источников, например, кинофильмов, записей выступлений и т. п., как это было реализовано в «Национальном корпусе русского языка» (http://ruscorpora.ru/). Создание корпусов спонтанной речи или чтения возможно также и в лабораторных условиях, когда респондентов приглашают на запись их речи на заранее подготовленную тему или чтения составленного исследователями текста [3]. Сбор речевого материала с последующей расшифровкой также может производиться путем записи речи в полевых условиях,как это было реализовано в рамках проекта “Один речевой день” в Санкт-Петербургском Государственном университете [5].

Однако, как показывает практика работы по улучшению качества распознавания и созданию языковых моделей для различных типов речи, несмотря на достаточно обширный выбор корпусных материалов, представленных в настоящее время для открытого доступа в сети Интернет, ни один из них не дает достаточного объема данных для повышения точности распознавания как подготовленной, так и спонтанной речи [4]. Другой проблемой создания речевых и текстовых корпусов является отсутствие возможности их постоянного пополнения, следовательно, не представляется возможным говорить

об их использовании в качестве материала для улучшения языковых моделей, необходимых для повышения качества распознавания, поскольку данный процесс требует постоянной актуализации лексики.

Проекты, работа над которыми ведется в ФГУП НИИ «Квант», требуют постоянно обновляемых и автоматически пополняемых баз данных речи различных стилей и типов произнесения, обязательно имеющих текстовую расшифровку, объем которых в несколько раз превосходит имеющиеся в свободном доступе речевые корпуса. Создание таких объемных корпусов вручную за короткий промежуток времени не представляется возможным.

В связи с этим встал вопрос о поиске источников большого количества текстовых и звуковых данных, а также о разработке специализированного программного обеспечения для сбора этих данных и формирования на их основе текстовых или речевых корпусов с текстовыми подстрочниками. В рамках решения вопроса о повышения точности распознавания новостных текстов для большого количества дикторов было необходимо разработать программный продукт, способный формировать корпус медиа-записей с текстовым подстрочником-расшифровкой на основании постоянно обновляемых в сети Интернет выпусков новостей. Очевидно, что для формирования аннотированного речевого корпуса необходимо создание целого комплекса программных продуктов, позволяющих в автоматическом режиме по заданным критериям собирать большие объемы текстовых и аудиоданных из любый источников, разбивать тексты по группам в зависимости от источника и тематики, осуществлять чистку текстов, построение языковой модели на основании собранных текстов, создавать автоматически фонетическую транскрипцию текстов и соотносить ее с результатами распознавания аудиоданных для построения акустической модели. Для новостных текстов оказалось возможным разработать систему сбора текстовых и медиа данных, речь о которой пойдет в следующем разделе статьи. Однако, как будет показано ниже, данный программный продукт возможно использовать для формирования текстовых или речевых корпусов различной тематики, то есть осуществлять сбор не только с новостных ресурсов.

2. Методика создания системы сбора текстовой и медиа информации

Основным требованием к системе сбора текстовых и медиа данных (ССТиМ) стала реализация возможности извлечения текста из HTML-разметки веб-страниц, а также осуществление загрузки медиа файлов, находящихся на веб-страницах.

Для разработки понадобилось решить ряд теоретических и практических задач:

- Провести анализ существующих ресурсов в сети Интернет на предмет соответствия задачам создания речевых корпусов;
- Провести анализ видов организации и представления материала на странице ресурса, установить правила выделения текстового материала, определения метаданных, а также правил оформления ссылок на медиафайлы (аудиофайлы и видеофайлы на страницах);

- Сформулировать основные требования к разрабатываемой системе (для реализации алгоритмов сбора и сохранения необходимой информации, представленной определенным образом на страницах ресурсов);
- Произвести обзор и сравнение существующих систем, способных осуществлять подобные операции;
- Произвести разработку специализированного программного обеспечения, которое отвечало бы требованиям к сбору и сохранению содержимого веб-страниц в необходимом виде;
- Произвести тестирование разработанного программного обеспечения на материале выбранных ресурсов.

Достижение намеченной цели обусловило последовательное решение поставленных задач.

2.1. Поиск источников данных для формирования речевого корпуса

В ходе работы по улучшению языковой модели для распознавания новостных выпусков было решено остановиться на изучении новостных сайтов теле-и радиоканалов, осуществляющих свое вещание на территории Российской Федерации и имеющих сайты в сети Интернет.

Анализ ресурсов в сети Интернет и их систематизация осуществлялись следующим образом. Был произведен поиск и обработка информации о существующих теле- и радиоканалах, вещающих по всему миру и на территории Российской Федерации. Каналы далее были изучены на предмет рейтинга и распространения эфирного вещания. Предпочтение отдавалось радиостанциям и телеканалам, вещающим на всей территории Российской Федерации. Далее были отобраны каналы, тематика которых соответствует новостной, либо информационно-культурологической. Из отобранных сайтов были выделены те, на страницах которых присутствует как текст, так и аудио или видео выпусков передач. Далее была проведена процедура оценки соответствия содержимого текстового подстрочника медиафайлам. Решение об использовании данных сайта телерадиовещания для создания речевого корпуса принималось на основании соответствия орфографического подстрочника содержимому медиафайла. Было решено остановиться на тех сайтах, подстрочник новостей которых полностью или в достаточной степени соответствует содержимому аудио- или видеофайла.

Содержимое веб-страниц каждого сайта из списка телерадиоканалов было тщательнейшим образом изучено. Для определения ссылок на страницы выпусков определенных телерадиопередач осуществлялся просмотр нескольких однотипных новостных страниц телерадиоканала (например, страниц с выпуском утренних новостей). Данная процедура проводилась для того, чтобы установить, на страницах с какими адресами находится необходимая информация. При подобном анализе сайта также определяется шаблон, по которому можно найти остальные ссылки. В шаблоне была найдена общая для всех адресов часть и та часть, которая их отличает (например, номера выпусков).

Было установлено, что на страницах телеканалов и в исходном коде страниц присутствуют теги для оформления текста, ссылки для загрузки медиафайлов, также в том или ином виде присутствуют данные относительно разделения контента на рубрики, названия передачи и ее выпуска и дикторе, произносящем реплику или читающем текст. Данная информация была найдена в тексте HTML-разметки с помощью регулярных выражений — инструмента для создания шаблонов поиска содержимого в тексте [1, 2]. Таким образом оказалось возможным создавать шаблоны для поиска необходимой информации на странице выпуска передачи, а также в тексте орфографической расшифровки.

Метод представления ссылок на медиафайлы на страницах разных каналов также различался: иногда данные представлялись не в виде гиперссылок, а в виде текста. Для загрузки данных по таким ссылкам необходимо вручную копировать текст из исходного кода страницы и вставлять его в поисковую строку. Некоторые ссылки являлись относительными и для их копирования приходилось дополнительно вручную включать в имя ссылки необходимую дополнительную часть (например, адрес телеканала). В результате анализа html-разметки кода страницы оказалось возможным установить способы представления орфографической расшифровки выпуска передачи.

На основании такого детального анализа нескольких сайтов телерадиоканалов оказалось возможным установить закономерности представления информации, которые могут быть использованы в дальнейшем для анализа любого источника текстовых данных. Выделенные закономерности и вышеописанный технологический процесс создания алгоритма поиска информации на сайте телерадиоканала легли в основу создания правил и справочников системы сбора текстовых и медиафайлов, необходимых для формирования речевого корпуса.

2.2. Требования к системе сбора текстов и медиафайлов

После определения вида представления информации веб-ресурсов были определены требования к системе сбора текстовой и медиа информации.

Предполагается, что система сбора текстовой и медиа информации должна осуществлять автоматический поиск веб-страниц выпусков программ телерадиоканала по заданным шаблонам с помощью регулярных выражений. При этом, следует обратить внимание, что данное требование подразумевает определение пользователем лишь шаблона для поиска на основании анализа адресов нескольких страниц. Анализ сайтов телерадиоканалов показал, что адреса страниц с выпусками определенных передач являются однотипными и поддаются типизированию с помощью регулярных выражений. Поиск программой существующих страниц должен осуществляться методом автоматического перебора ссылок по шаблону и занесения в список для скачивания только существующих адресов.

Для формирования речевого корпуса также важно, чтобы программа автоматически по заданному шаблону могла извлекать из HTML-документов ссылки для сохранения медиафайлов, а также преобразовывала гиперссылки, если они представлены в виде текста. Помимо загрузки медиафайлов, программа должна извлекать текст из HTML-документов страниц, с которых были

загружены медиафайлы, автоматически по заданному алгоритму на основании списка пар тэгов. При загрузке медиафайлов и текстового подстрочника крайне важно обеспечить сохранение информации о том, с какой страницы был загружен тот или иной медиафайл. Данное требование также является обязательным для системы сбора.

Следует отметить, что система сбора данных должна работать в многопоточном режиме, а также поддерживать сбор HTML-документов с нескольких сайтов одновременно с разделением сохраненных файлов по источнику. В связи со спецификой задачи многопотоковой обработки ресурсов одного или нескольких сайтов оказалось необходимым контролировать минимальный интервал запросов к сайту для предотвращения блокировки доступа к нему.

Актуальным также явилась возможность сохранения списка ссылок для поиска HTML-документов для возобновления процесса скачивания, например, прерванного по каким-то причинам, выполнять сбор документов по расписанию, то есть автоматически, без участия пользователя системы, но по заданным им заранее критериям, отображать текущее состояние сбора документов для осуществления контроля закачки, то есть иметь пользовательский интерфейс. В случае если предполагается работа с программой на виртуальных машинах с сохранением информации на сервере, необходимо, чтобы пользователь имел возможность осуществлять вышеперечисленные процессы в системе виртуализации.

После сбора необходимых данных из сети Интернет необходимо систематизировать собранные данные в единую базу данных, вычистить тексты и нормализовать их в зависимости от дальнейших лингвистических задач, преобразовать аудиофайлы в необходимый формат, при необходимости, выделить и сохранить аудиодорожку из видеофайла, записать в таблицу всю метаинформацию относительно акустических свойств загруженных медиафайлов.

Таким образом, единица базы данных для дальнейшего преобразования должна представлять собой видеофайл, созданный на его основании аудиофайл (извлеченная звуковая дорожка), либо только аудиофайл (в случае, если, например, велась загрузка радиопередач) с прилагающимся к нему текстовым файлом, содержащим орфографический подстрочник, который был загружен со страницы медиафайла.

2.3. Система сбора текстов и медиаданных

В связи с определением требований к системе сбора текстовой и медиа информации встал вопрос относительно возможности использования уже существующих систем сбора данных в сети Интернет. С этой целью было проведено изучение программ по загрузке содержимого веб-страниц сайтов. На сегодняшний день существует достаточно много приложений, позволяющих скачивать содержимое веб-страниц, однако ни одно из них не удовлетворила требованиям, выделенным в ходе разработки алгоритмов сбора данных для формирования речевых корпусов.

Так, например, программа Download Master требует указания полного списка ссылок для скачивания и не может самостоятельно определить количество

страниц на сайте, не поддерживает возможности сохранение текста в виде отдельного файла и запуска по расписанию.

Программа Teleport Pro осуществляет загрузку всего содержимого вебстраницы без возможности фильтрации (загрузки только аудио или видео, без загрузки картинок или дополнительных файлов), предоставляя возможность сохранения текста с HTML-разметкой (невозможно, например, указать теги для скачивания только произносимого в новости текста). Также ни одна программа не может выделять текст из HTML-разметки страницы.

Реализация процедуры скачивания содержимого сайтов возможна с помощью сценариев в PowerShell, однако она требует знания основ программирования и написания сценариев на соответствующем языке, а также умения создавать скрипты для запуска программы из консольного приложения, что существенно ограничивает круг пользователей, которые могли бы создавать корпуса подобным способом.

Таким образом, оказалась очевидной необходимость создания собственного программного продукта, отвечающего требованиям к системе сбора текстов и медиаданных. Специальное программное обеспечение — система сбора текстовых и медиаданных (ССТиМ) было разработано на языке C# в среде разработки Microsoft Visual Studio 2015. Система состоит из нескольких компонентов — программы для создания и запуска оператором задач сбора, программы последующей обработки данных для создания речевого и текстового корпусов и части для настройки запуска задач сбора и процессов по дальнейшей обработке данных по расписанию. Разработка отдельных компонентов программного обеспечения осуществлялась в соответствии с требованиями, установленными ранее. Таким образом, система сбора текстовых данных и медиафайлов представляет собой программный продукт для создания задачи сбора и последующей обработки собранных файлов.

Задача сбора — это файл, созданный в ССТиМ, в котором указаны все настройки, необходимые для скачивания данных с определенного адреса в сети Интернет:

- Корневые ссылки — набор начальных страниц Интернет-ресурса;
- Прописанные при помощи регулярных выражений шаблоны ссылок, обнаруженных на страницах, открытых через корневые ссылки, сохранение которых необходимо осуществить (шаблоны ссылок на текстовые документы или медиафайлы);
- Адрес сохранения корпуса в файловом хранилище пользователя программой;
- Рубрики, на которые, согласно указанным шаблонам, возможно произвести разбиение собираемого корпуса;
- Шаблоны извлечения текста по заданным тегам;
- Шаблоны преобразования текста в ссылки, который позволяет задать список регулярных выражений для извлечения ссылок из HTML-разметки страниц, если они представлены в виде текста и не могут быть опознаны программой, как ссылки;
- Время минимального интервала обращения к сайту для предотвращения блокировки;

- Количество потоков одновременной обработки для осуществления многопотокового скачивания;
- Расписание запуска процесса сбора в случае, если загрузку необходимо осуществлять по графику.

По окончании загрузки данные о расположении файлов сохраняются в таблице, которая создается автоматически и содержит информацию относительно вида и типа загруженных данных, а также наличии связей данных между собой (связи текста с видео или аудиофайлом). Далее с помощью имеющихся модулей системы пользователь может настроить извлечение аудиодорожки, запустить процесс нормализации текстов, распознавания звука и соотнесение результатов распознавания с текстом, сохраненным с веб-страницы, а также формирование отчетной таблицы для изучения статистики собранных материалов.

Созданное ССТиМ было разработано, создано и передано для тестирования. Система состоит из клиентской части — программы с пользовательским интерфейсом, и серверной — с помощью которой возможно осуществлять сбор данных по расписанию, указанному пользователем. Программный продукт был реализован на языке C# для среды Microsoft, поскольку данная среда используется лингвистами ФГУП НИИ «Квант».

Тестирование СПО проводилось на материале сайтов телерадиоканалов путем создания тестового речевого корпуса медиафайлов с текстовым подстрочником, а также путем расчета приращения текстового корпуса новостной тематики посредством загрузки текстов одновременно с более чем 10 новостных сайтов (без поиска видеофрагментов к новостям).

Кроме того, была апробирована система создания расписания, которая позволила осуществлять загрузку данных в указанные промежутки времени, тем самым позволив проводить сбор без участия оператора. Также оказалось возможным своевременно актуализировать корпус текстов, то есть ежедневно загружать новости с новостных сайтов без участия оператора.

Всего в результате тестовой работы программы было собрано более полутора тысяч гигабайт текстовой и медиа информации (аудиозаписей передач, видиозаписей выпусков новостей и т. п.) с образцами речи более 2000 дикторов. Очевидно, что собранные данные необходимо проверять на предмет соответствия речи тексту новостного выпуска, поскольку ССТиМ выполняет лишь функцию первоначального сбора большого объема данных для последующей обработки. Данные, собранные с помощью ССТиМ явились основой для формирования речевого корпуса, который используется сотрудниками ФГУП НИИ «Квант» для работ по улучшению качества распознавания речи и на данный момент не может быть предоставлен в открытый доступ. Также с помощью ССТиМ был создан отдельный постоянно увеличивающийся (благодаря функции загрузки текстов по расписанию) текстовый корпус новостных текстов объемом 4 млн. документов (на момент окончания тестовых испытаний), материалом для которого стали тексты сайтов ведущих новостных газет и телеканалов Российской Федерации. Данный текстовый корпус был использован в качестве материала для построения языковой модели для распознавания новостных

текстов. Следует отметить, что созданная программа имеет пользовательский интерфейс, что позволяет создавать базы данных пользователям, которые не знакомы с программированием и имеют лишь знания относительно работы с регулярными выражениями.

Для дальнейшего использования в целях улучшения качества распознавания и построения акустических и языковых моделей собранный материал был систематизирован и нормализован с помощью созданных специальных программ и алгоритмов обработки.

Заключение

В результате работы по созданию текстовых и речевых корпусов было разработано специальное программное обеспечение, позволяющее в режиме автоматизации осуществлять сбор текстовых и медиаданных, находящихся в открытом доступе в сети Интернет. Как было показано, разработка данного продукта явилась следствием необходимости в многопоточной автоматизированной обработке большого количества текстовых и медиаданных. Тестирование работоспособности программного продукта, созданного в ФГУП НИИ «Квант», подтвердило возможность его использования для создания речевых и текстовых баз различной тематики, данные которых могут быть успешно использованы при разработке систем распознавания речи, анализа текстов, построения акустических и языковых моделей, а также для применения в исследовательских целях.

Литература

1. *Friedl J.,* (2008), Mastering regular expressions, 3rd Edition, O'Reilly Media
2. *Goyvaerts J., Levithan S.,* (2012), Regular Expressions Cookbook, Detailed Solutions in Eight Programming Languages, 2nd Edition, O'Reilly Media
3. *Kachkovskaia T., Kocharov D., Skrelin P., Volskaya N.,* (2016), CoRuSS — a New Prosodically Annotated Corpus of Russian Spontaneous Speech, URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/144_Paper.pdf
4. *Krivnova O., Zakharov L., Strokin G.,* (2001), Speech corpora (experience of creation and usage) [rechevyie korpusa (opyt razrabotki i ispol'zovania)], URL: http://www.dialog-21.ru/digest/2001/articles/krivnova/
5. *Stepanova S. B., Asinovsky A. S., Bogdanova N. V., Rusakova M. V., Sherstinova T. U.,* (2008), Speech corpus of the russian everyday communication "one speaker's day": basic conception and current state [Zvukovoy korpus russkogo yazika povsednevnogo obsh'eniya "Odin rechevoy den'": koncepciya i sostoyanie formirovaniya], URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/76.htm