МЕТОДИКА СОЗДАНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ФОРМИРОВАНИЯ РЕЧЕВОГО КОРПУСА

Пискунова Виктория

Piskunova.victoria@gmail.com

ФГУП НИИ «Квант»

Цели и задачи разработки ПО

- Цель: сбор данных для формирования речевого корпуса «большого» (достаточного) объема.

- Задачи:
  - поиск источников «аудио/видео+текст»;
  - анализ методов представления данных;
  - создание ПО;
  - решение задачи систематического приращения данных для формирования корпуса
  - на основании данных создать речевой корпус;

Источники данных

Сайты теле- и радиоканалов:

- архив выпусков передач или фрагменты;
- медиаролик — аудио или видео;
- текстовая расшифровка;
- возможность скачать медиаролик (ссылка в html-коде);
- периодическое пополнение архива передач.

Система сбора текстов и медиа

Правила сбора данных:

- Шаблоны поиска страниц выпусков передач;
- Шаблоны поиска медироликов на страницах;
- Теги, оформляющие текст.

Созданное ПО:

- ищет страницы выпусков;
- загружает медиафайлы;
- создает текстовый файл;
- создает файл связи между загруженными файлами;
- запускает сбор по расписанию (без участия оператора).

Дальнейшая обработка собранного материала

- Внесение в базу данных для распознавания (единица базы данных – текст+аудио/видео);
- Извлечение аудиодорожки из видеоролика;
- Распознавание речи;
- Соотнесение распознанного текста с эталонным (загруженным со страницы выпуска), выравнивание по времени;
- Экспорт в виде набора аудиофайлов с эталонной текстовой расшифровкой.

Собранные корпусы

- Новостной корпус теле- и радиопередач (**видео+тексты**) – 725 часов;
- Корпус дикторской речи проекта «Война и мир. Читаем роман» (**видео+текст+фотографии чтецов**) – 233 часа, 4905 дикторов;
- Корпус чтения художественных произведений (аудиокниги) (**аудио+тексты**) – 23 часа, 72 произведения, 72 диктора;
- Корпус текстов новостных сообщений (**тексты**) – ок.4,5 млн. текстовых документов (ок. 1 млрд словоформ);
- Текстовый корпус субтитров к художественных фильмам (**тексты**) – ок. 9 тыс. документов, 15 млн словоформ.

Спасибо за внимание!