

УДК 004.8

РАЗРАБОТКА АЛГОРИТМА АВТОМАТИЗИРОВАННОГО ПРОТОКОЛИРОВАНИЯ СОБРАНИЙ

Медведева О.И. (Университет ИТМО)

Научный руководитель – доцент, к.ф.-м.н., Михайлова Е.Г.

(Университет ИТМО)

Введение. В современном мире сотрудники компаний со всего мира проводят большое количество рабочего времени на онлайн-собраниях и встречах. Одной из важнейших составляющих собраний является ведение протокола. В протоколах собраний фиксируются главные темы обсуждения, принятые решения и выполненные задачи. Запись протокола - это трудоемкий процесс, который отнимает много времени и сил. Именно поэтому в последнее время существует спрос на разработку систем автоматического протоколирования собраний. Многие исследования основаны на англоязычных открытых наборах данных, например, AMI [1] и ICSI [2], но для русского языка такие наборы данных отсутствуют, что затрудняет развитие методов генерации протоколов собраний.

Основная часть. Входные данные представляют собой видеозаписи онлайн-собраний в формате MP4. В рамках научной работы собран набор данных видеовыступлений экспертов с видеохостинга YouTube общей длительностью более 5 часов. В качестве разметки для тестирования взята сгенерированная стенограмма видеозаписей.

Разработку алгоритма протоколирования собраний можно разделить на этапы: распознавание текста из речи и реферирование текста. Решение задачи распознавания текста из речи осуществляется при помощи модели машинного обучения на предварительно извлеченных из видеозаписей аудиоданных.

Особенностью онлайн-собраний является то, что во время него могут возникать паузы и обрывы связи, ситуации, когда говорят сразу два или более человек, а также шумы и помехи. Для обработки таких моментов применяется сегментация звука, диаризация аудио и детекция голосовой активности при помощи инструмента PyAnnote для разделения звуковой дорожки на временные отрезки в соответствии с принадлежностью аудиопотока тому или иному говорящему. Для распознавания речи и преобразования ее в текст существует несколько инструментов, которые могут быть использованы для русского языка, например, API от Google или Vosk. В данной работе используется инструмент Whisper от OpenAI, так как модель самостоятельно определяет язык речи, транскрибирует тексты, восстанавливая пунктуацию и капитализацию. Распознанный текст проходит оценку качества по четырем метрикам:

- 1) WER (word error rate) – показатель, насколько распознанный текст отличается от оригинала;
- 2) MER (match error rate) – это доля совпадений слов ввода-вывода, которые являются ошибками;
- 3) WIL (word information lost) – это приближение к доле потерянной словесной информации;
- 4) WIP (word information preserved) – это доля сохраненной словесной информации.

При вычислении метрик оригинальный и распознанный текст сравниваются на наличие совпадений, удалений, замен и добавлений слов. Во время тестирования модуля распознавания текста из речи среднее значение метрики WER составило 26 %, метрики MER – 23 %, WIL – 33%, WIP – 66%.

После распознавания текста производится его реферирование для извлечения наиболее значимой информации. Выделяют два подхода: экстрактивное реферирование и абстрактное реферирование. В первом подходе основная идея состоит в том, чтобы извлечь топ-к значимых предложений входного текста и объединить их для получения сводки в качестве вывода. При втором подходе конечная цель состоит в том, чтобы понять значение

наиболее важных идей, развиваемых во входном тексте, а затем создать на выходе краткое изложение этих идей с использованием слов или предложений, которых может не быть во входном тексте. Для онлайн-собраний более приемлемым является экстрактивное реферирование, так как важно передать оригинальные формулировки говорящих. В данной работе используется тематическое экстрактивное реферирование. С помощью обученной модели латентного размещения Дирихле [3] моделируется тема входного текста, которая представляет собой список ключевых слов текущего текста. Модель обучалась на корпусе слов, которые прошли предварительную токенизацию, лемматизацию и разбиение на группы слов. Далее оригинальный текст делился на предложения. Между вектором каждого предложения и вектором ключевых слов рассчитывалось косинусное сходство. Топ-k наиболее схожих предложений вошли в итоговый реферат текста, качество которого оценивалось по метрике ROUGE (Recall-Oriented Understudy for Gisting Evaluation/ Оценка реферирования, ориентированная на полноту). В данной работе используются несколько разновидностей ROUGE:

- 1) ROUGE-1 – это доля совпадений отдельных слов в оригинальном и предсказанном рефератах;
- 2) ROUGE-2 – это доля совпадений пар слов в оригинальном и предсказанном рефератах;
- 3) ROUGE-L – оценивает самую длинную последовательность совпавших слов.

При сравнении полученных рефератов со сгенерированными F-мера метрики ROUGE-1 достигла 51%, ROUGE-2 – 35%, ROUGE-L – 49%.

Выводы. В данной работе проведен анализ существующих методов распознавания текста из речи и реферирования текста, разработан алгоритм автоматизированного протоколирования собраний. При распознавании текста из речи было достигнуто сходство с оригиналом в 74%, а доля совпадения отдельных слов в модуле реферирования составила 51%. Таким образом, применяя методы машинного обучения без учителя к русскоязычным данным можно достичь высоких результатов, не прибегая к сбору большого набора данных на русском языке.

Список использованных источников:

1. Kraaij W., Hain T., Lincoln M., Post W. The AMI Meeting Corpus // 5th International Conference on Methods and Techniques in Behavioral Research. – 2005. – С. 137–140.
2. Devroey P., Van Steirteghem A., A review of ten years experience of ICSI // Human Reproduction Update. – 2004. – С. 19–28.
3. David M., Andrew Y., Michael I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – С. 993-1022.

Медведева О.И. (автор)

Подпись

Михайлова Е.Г. (научный руководитель)

Подпись