Д. П. Уральский, А. А. Гончар, М. В. Аврискин
Тюменский государственный университет, г. Тюмень
УДК 004.94

ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ МЕТОДОВ И АЛГОРИТМОВ ДЛЯ КРАТКОГО АННОТИРОВАНИЯ АУДИОКОНТЕНТА С ИЗВЛЕЧЕНИЕМ ТЕКСТА

Аннотация. В статье рассматриваются подходы в аннотировании аудиоконтента и проводится оценка методов транскрибации и суммаризации, используемых для подхода аннотирования с извлечением текста. Рассматриваются Whisper, NeMo, mT5, mBart, а также TF-IDF, LSA, TextRank и LexRank.

Ключевые слова: суммаризация, аннотирование аудио, транскрибация, аудио в текст, обработка естественного языка.

Термины:

1. Транскрибация — это процесс перевода аудиозаписей или видеозаписей в текст [1].
2. Суммаризация — это процесс сжатия больших объемов текстовой информации до краткого и информативного содержания, охватывающего основные идеи и ключевую информацию текста [2].

Введение. Обработка естественного языка (Natural language processing, NLP) — направление в области искусственного интеллекта и математической лингвистики, которое занимается вопросами компьютерного анализа и синтеза естественных языков. Основная задача NLP заключается в разработке систем, способных выполнять языковые задачи на уровне, сравнимом с человеческим. Примером такой задачи в обработке естественного языка может быть аннотирование аудиоматериалов.

Аннотирование аудиоконтента — давно известная задача в области обработки естественного языка. В статье "Audio Summarization with Audio Features and Probability Distribution Divergence" [3] рассматривается проблема увеличения объема аудиоконтента, что препятствует быстрому ознакомлению с информацией. Методы автоматического аннотирования аудиофайлов могут стать одним из решений данной проблемы. В статье авторы выделяют следующие подходы в аннотировании аудио:

1. Использование только аудиопризнаков. Этот подход основан на анализе аудиосигнала для создания аннотации без текстовой информации.
2. Извлечение текста из аудиосигнала. Этот подход включает транскрибацию аудиофайла для извлечения текста, после чего следует суммаризация для получения краткой аннотации.
3. Гибридный подход. Этот подход комбинирует оба предыдущих подхода.

В рамках данного исследования будет рассмотрен второй подход для аннотирования, включающий в себя суммаризацию и транскрибацию.

В статье «Обзор задачи автоматической суммаризации текста» [4] выделяют два подхода в суммаризации:

1. Экстрактивный подход выбирает самые информативные части текста, обычно предложения, для составления краткого содержания.
2. Абстрактивный подход генерирует новый текст, содержащий основную информацию, выраженную не обязательно теми же словами или фразами, что и в исходном тексте.

Цель данного исследования заключается в сравнении методов и алгоритмов транскрибации и суммаризации, используемых для аннотирования аудиоконтента с подходом извлечения текста из аудиосигнала.

Для достижения цели были поставлены следующие задачи:

1. Исследовать имеющиеся подходы в области транскрибации аудиофайлов и суммаризации текста.
2. Реализовать алгоритмы и методы для транскрибации аудиофайлов и суммаризации текста.
3. Оценить эффективность реализованных алгоритмов и методов транскрибации и суммаризации.

Предобработка данных. Для обработки аудиофайлов необходимо преобразовать их к нужным значениям их свойств. Список свойств, следующий:

1. Дискретизация должна составлять 16kHz.
2. Звук должен быть в mono формате.
3. Формат файла должен быть wav.
4. Длина аудиофайла не должна превышать 10 минут.

Для предобработки текста перед экстрактивной суммаризацией были использованы следующие методы:

1. Нормализация текста — приведение всех слов в нижний регистр, удаление знаков пунктуации, чисел и пробельных символов.
2. Токенизация текста — разбиение каждого предложения в тексте на слова.
3. Удаление стоп-слов — слова, которые не несут смысловой нагрузки (союзы, предлоги и т. д.).
4. Лемматизация — приведение слова к словарной форме. Например:
   a. Для существительных — именительный падеж, единственное число.
   b. Для прилагательных — именительный падеж, единственное число, мужской род.
   c. Для глаголов, причастий, деепричастий — глагол в инфинитиве несовершенного вида.

Для абстрактивной суммаризации подаваемый на вход текст сначала разбивается на N частей, состоящих из неразрывных предложений. Каждая из этих частей по количеству символов не превышает L, где

$$N = math.ceil\left(\frac{\text{Длина исходного текста}}{3000}\right) \quad (1)$$

$$L = math.ceil\left(\frac{\text{Длина исходного текста}}{N}\right) \quad (2)$$

Функция math.ceil производит округление полученного результата в большую сторону.

Затем каждая часть разбитого текста проходит процесс токенизации и преобразования текстовой информации в числовую.

Материалы и методы. Для задачи транскрибации было проведено сравнение двух нейросетей: Whisper и NeMo. Сравнение этих моделей между собой производилось на датасете common_voice_10_0 [5]. Русскоязычная часть датасета состоит из 275 часов аудиозапи-

сей, из которых 84% уже были проверены вручную. Сам датасет представлет собой аудиозаписи голосов 3217 человек различного пола и возраста. Кроме того, набор данных включает в себя разнообразные речевые особенности, характерные для различных языковых групп и диалектов. Оценка эффективности данных нейронных сетей была проведена с использованием метрики WER (Word Error Rate) [6], которая рассчитывается по следующей формуле:

$$WER = \frac{S + D + I}{S + D + C} \quad (3)$$

где S — количество замен, D — количество удалений, I — количество вставок, C — количество правильных слов.

В рамках задачи экстрактивной суммаризации был проведен сравнительный анализ следующих методов: LexRank, TextRank, LSA, TF-IDF.

Для задачи абстрактивной суммаризации было проведено сравнение двух нейросетей: mBart и mT5.

Сравнение методов каждого подхода суммаризации производилось на датасете gazeta [7], состоящем из 74 126 записей. Для оценки использовались данные из двух колонок: одна содержала исходный текст, а другая — эталонный вариант суммаризации. Сравнение проводилось на 100 случайно выбранных записях датасета. Оценка эффективности методов для каждого подхода была проведена с использованием метрик ROUGE (-1, -2, -L) и METEOR.

ROUGE-1 оценивает точность генерации текста по количеству слов из оригинала, включенных в сгенерированный текст.

ROUGE-2 — метрика, измеряющая совпадение последовательностей из двух слов между оригинальным и сгенерированным текстом.

ROUGE-L измеряет схожесть между двумя текстами на основе их самой длинной общей последовательности, учитывая порядок слов, но не требуя их непрерывности.

Более подробное описание того, что оценивает метрика ROUGE, можно найти в статье "Mastering ROUGE Matrix: Your Guide to Large Language Model Evaluation for Summarization with Examples" [8].

METEOR оценивает сгенерированный текст на соответствие с эталонным текстом, устанавливая выравнивание слов и учитывая точное сравнение, стемминг и синонимы из WordNet. Выбирается наиболее подходящее соответствие, учитывая схожий порядок слов. Более подробное описание того, что оценивает метрика METEOR, можно найти в статье "Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output" [9].

Результат. Оценка методов для транскрибации и суммаризации приведена в таблицах ниже. Оценка методов производилась в Google Colaboratory, имеющим следующие характеристики: процессор Intel(R) Xeon(R) CPU @ 2.20GHz и видеокарта NVIDIA Tesla T4.

Таблица 1

Оценка методов транскрибации

	Whisper	*NeMo*
WER	17,26%	11,95%
Объем обработки аудиозаписи за секунду	6 сек.	25 сек.

Из результатов оценки этих моделей (см. табл. 1) видно что Whisper выполняет транскрибацию медленнее и его показать WER выше, что указывает на то, что в процессе транскрибации эта модель допускала больше ошибок, чем модель NeMo.

При выборе между этими двумя моделям также стоит учитывать и то, что модель NeMo ограничена использованием только на графических процессорах (GPU) от NVIDIA.

Таблица 2

Оценка методов экстрактивной суммаризации

С лемматизацией				
	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
TF-IDF	20,8%	14,1%	3,2%	12,7%
LSA	19,4%	13%	3%	11,8%
TextRank	19%	12,8%	2,4%	11,5%
LexRank	19,6%	14%	3%	12,7%
Без лемматизации				
	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
TF-IDF	20,6%	13,9%	3,3%	12,5%
LSA	20,1%	12,9%	3%	11,6%
TextRank	18.6%	12,7%	2,4%	11,3%
LexRank	19,2%	14,4%	3,1%	13,2%

Исходя из результатов оценки этих методов (табл. 2) видно, что лучше всего себя показали TF-IDF с использованием лемматизации и LexRank без использования лемматизации. Низкие показатели по выбранным метрикам можно объяснить тем, что результат работы методов сравнивался с эталонным вариантом, который представляет собой абстрактивный вариант суммаризации, состоящий из предложений, передающих тот же смысл и информацию, что и в исходном тексте, но не обязательно совпадающим с ними по структуре и форме. Поэтому для получения более надежного результата была проведена дополнительная ручная оценка двух методов, которые показали лучший результаты по метрикам. При ручной оценке лучше всего себя показал LexRank.

Таблица 3

Оценка методов абстрактивной суммаризации

	METEOR	*ROUGE-1*	*ROUGE-2*	*ROUGE-L*
mT5	22,5%	14,1%	3,8%	13%
mBart	22,8%	15,4%	4,2%	14,3%

По результатам оценки этих моделей (табл. 3) видно, что mBart проявил себя лучше по отобранным метрикам. Низкие показатели по метрикам можно объяснить тем, что mT5 и mBart генерируют текст, который по своей структуре может отличаться от эталонного варинта, но нести ту же смысловую нагрузку. Поэтому было решено провести еще и ручную проверку работы нейросетей, чтобы убедиться в качестве получаемой суммаризации. В ходе ручного тестирования mBart также проявил себя лучше, чем mT5.

Заключение. По результатам проведенной оценки методов и алгоритмов, применяемых в аннотировании аудиоконтента с извлечением текста, выяснилось, что для задачи транскрибации стоит выбирать модель NeMo. Для задачи суммаризации при использовании экстрактивного подхода рекомендуется применять алгоритм LexRank, а для абстрактивного подхода — модель mBart. При этом стоит отметить, что использование данных, отличающихся от примененных в статье, может привести к иному результату и потребовать дополнительных исследований.

В перспективе планируется провести сравнительную оценку методов, применяемых для аннотирования аудиоконтента, с подходом извлечения только аудиопризнаков, и последующее сравнение полученных результатов с представленными в данной статье. Сравнение результатов двух разных подходов позволит установить, какой из них справляется с задачей аннотирования аудиоконтента лучше всего.

СПИСОК ЛИТЕРАТУРЫ

1. Каменская А.С. Адаптация Google Cloud Speech-to-text API для автоматической транскрибации веб-конференций в реальном времени / А.С. Каменская — Текст: электронный // Автоматика и программная инженерия. — 2019. — № 2 (28). — URL: https://cyberleninka.ru/article/n/adaptatsiya-google-cloud-speech-to-text-api-dlya-avtomaticheskoy-transkribatsii-veb-konferentsiy-v-realnom-vremeni/viewer (дата обращения: 17.02.2024).
2. Luís Gonçalves Automatic Text Summarization with Machine Learning — An overview / Luís Gonçalves. — Текст: электронный // Medium: сайт. — URL: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25 (дата обращения: 18.02.2024).
3. Audio Summarization with Audio Features and Probability Distribution Divergence / Carlos-Emiliano Gonz´alez-Gallardo, Romain Deveaud, Eric SanJuan, Juan-Manuel Torres-Moreno. — Текст: электронный // Arxiv: сайт. — URL: https://arxiv.org/pdf/2001.07098.pdf (дата обращения: 19.02.2024).
4. Белякова А.Ю. Обзор задачи автоматической суммаризации текста / А.Ю. Белякова, Ю.Д. Беляков. — Текст: электронный // Инженерный вестник Дона. — 2020. — № 10. — С. 2-8. — URL: https://cyberleninka.ru/article/n/obzor-zadachi-avtomaticheskoy-summarizatsii-teksta/viewer (дата обращения: 19.02.2024).
5. Mozilla Foundation Dataset Card for Common Voice Corpus 10.0. — Текст: электронный // HuggingFace: сайт. — URL: https://huggingface.co/datasets/mozilla-foundation/common_voice_10_0 (дата обращения: 20.02.2024).
6. Metric: wer. — Текст: электронный // HuggingFace: сайт. — URL: https://huggingface.co/spaces/evaluate-metric/wer (дата обращения: 20.02.2024).
7. Ilya Gusev Dataset Card for Gazeta / Ilya Gusev. — Текст: электронный // HuggingFace: сайт. — URL: https://huggingface.co/datasets/IlyaGusev/gazeta (дата обращения: 20.02.2024).
8. Marawan Mamdouh Mastering ROUGE Matrix: Your Guide to Large Language Model Evaluation for Summarization with Examples / Marawan Mamdouh. — Текст: электронный // Dev: сайт. — URL: https://dev.to/aws-builders/mastering-rouge-matrix-your-guide-to-large-language-model-evaluation-for-summarization-with-examples-jjg (дата обращения: 20.02.2024).
9. Abhaya Agarwal Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Outpu / Abhaya Agarwal, Alon Lavie. — Текст: электронный // Statmt: сайт. — URL: https://statmt.org/wmt08/pdf/WMT12.pdf (дата обращения: 21.02.2024).