DOI: 10.24412/2413-2527-2023-436-19-27

Применение технологий обработки естественного языка для голосового управления на основе открытого словаря

С. А. Михайлова, К. Г. Аникеев
Военно-космическая академия имени А. Ф. Можайского
Санкт-Петербург, Россия
mikhaylova_sa@mail.ru

Аннотация. **Рассмотрена возможность применения технологий обработки естественного языка для решения задачи голосового управления путем выделения в естественной речи фраз, имеющих сходный смысл с командами управления. Рассмотрены возможности библиотек и сервисов распознавания речи. Предлагается структура системы голосового управления на базе библиотеки распознавания речи Vosk. Данная структура позволяет пополнять словарь фраз, соответствующих командам управления, на основе оценки семантической близости.**

Ключевые слова: **распознавание речи, голосовое управление, обработка естественного языка.**

Введение

В настоящее время существует проблема, связанная с упрощением управления программным обеспечением, например презентацией, во время выступления докладчиков. Одним из решений данной проблемы можно считать создание системы речевого управления. Это позволит докладчикам самостоятельно регулировать темп выступления и управлять переключением мультимедийного сопровождения, не отвлекаясь на специальные устройства, такие как клавиатура, мышь или пульт дистанционного управления, без привлечения человека-ассистента.

Широкое распространение мультимедийных устройств совместно с быстрым развитием технологий искусственного интеллекта и предоставлением доступа к моделям машинного обучения привели к стадии активного внедрения технологий распознавания речи для решения различных задач во многих отраслях: в банковской сфере, промышленности, логистике, медиасфере, медицине, науке и образовании [1, 2]. Имеется множество примеров использования голосового управления различными объектами на основе умных колонок Amazon Echo или Google Home, голосовых ассистентов, таких как Siri от Apple или Google Assistant, голосовых переводчиков, например Google Translate. Большинство подобных систем построено на основе четко определенного командного языка управления объектами, или ограниченного подмножества фразеологизмов [3], или закрытого словаря команд управления.

В случае отсутствия строго формализованных команд требуется учитывать наличие множества слов или фраз, являющихся смысловыми синонимами. Особенно ярко эта проблема проявляется в случае выделения неявных команд управления в естественной речи, например в речи докладчика для управления мультимедийным сопровождением. Для решения указанной задачи необходимо сочетание технологий распознавания речи и методов обработки естественного языка.

Целью работы является разработка программы распознавания в естественной речи человека фраз, семантически сходных с командами управления на основе открытого словаря.

Голосовое управление

Голосовое управление — это способ управления различными устройствами или системами с помощью голосовых команд. Эта технология позволяет пользователю взаимодействовать с устройствами, не используя физические кнопки или сенсорные экраны.

В основе голосового управления лежат технологии распознавания речи. Широко распространены технологии и сервисы на их основе от крупных игроков IT-индустрии, такие как Google Speech Recognition, Apple Siri, Amazon Alexa, Nuance Dragon, Kaldi, TensorFlow.

Google Speech Recognition — технология распознавания речи от Google, используемая в таких продуктах, как Google Assistant, Google Home и Google Translate. Она позволяет пользователям голосовых устройств взаимодействовать с устройствами на естественном языке. Технология Google доступна на нескольких языках, включая русский. Развитие технологии происходит с помощью машинного обучения.

Apple Siri — технология распознавания речи от Apple, довольно популярная и интуитивно понятная. Она используется для управления устройствами Apple. Siri использует технологию обработки естественного языка для понимания запросов пользователей и предоставления ответов.

Amazon Alexa — технология распознавания речи, используемая в продуктах Amazon, таких как Amazon Echo и Alexa-enabled устройства. Она позволяет пользователям управлять устройствами и задавать вопросы на естественном языке. Пользователи могут использовать Amazon Alexa, чтобы интегрировать его со своими устройствами, установить напоминания и делать покупки.

Яндекс.Облако Контур.Распознавание — технология распознавания речи, созданная компанией Яндекс. Она может быть использована для распознавания речи на нескольких языках, включая русский, а также используется для автоматического распознавания голосов в телефонных автоответчиках. Яндекс.Облако Контур. Распознавание также используется для распознавания речевых команд в умных домах и системах безопасности.

Nuance Dragon — технология распознавания речи для профессионального использования, включая медицинские и юридические сферы. Она позволяет пользователям диктовать текст и управлять компьютером через голос. Nuance Dragon доступен на нескольких языках и имеет возможность обучения для конкретного пользователя.

Kaldi — библиотека открытого исходного кода для распознавания речи на различных языках, включая русский. Она может быть использована для создания собственной системы распознавания речи и дообучения уже существующих моделей на конкретный голос. Kaldi используется в таких областях, как биометрия, помощь водителю, медицина и образование.

TensorFlow — открытая библиотека машинного обучения, которая используется для создания и обучения нейронных сетей для распознавания речи и других задач, связанных с голосом. TensorFlow может быть использован для создания встроенной в устройства системы распознавания речи, примерами таких устройств могут быть умные наушники или колонки.

Чаще всего при взаимодействии с голосовым устройством используется включение режима восприятия команды механически (с помощью специальной кнопки) или с помощью ключевого слова при непрерывном прослушивании. Для голосовой активации режима восприятия команд используются модели, обученные на распознавание пробуждающего слова. Для распознавания самих команд используется комбинация технологий преобразования речи в текст Speech-to-Text (STT) и понимания естественного языка Natural Language Understanding (NLU).

Speech-to-Text — компонент обработки голоса, получающий от пользователя входные данные в аудиоформате и преобразующий этот фрагмент в текст.

Natural Language Understanding — компонент, определяющий намерения пользователя и основные элементы информации в команде, необходимые для ее выполнения [4]. В зависимости от намерений активируется нужный класс в модуле исполнения команд, который выполняет требование пользователя. Элементы команды, необходимые для ее выполнения, выделяются с помощью технологии выделения именованных сущностей Named Entity Recognition (NER).

В случае идентификации команды компонент управления выполняет соответствующее действие, в противном случае возможно уточнение запроса или конкретизация команды пользователем [5, 6].

Подсистема голосового управления требует предварительного создания словаря распознавания, содержащего все возможные слова, которые могут встречаться в подаваемых командах, и их транскрипции. Команды содержат указанные слова в строго определенной последовательности, другой порядок слов не распознается в качестве команды. Словарь преобразуется в форму, необходимую для работы алгоритмов распознавания, путем составления файлов грамматики, содержащих описания всех возможных команд [7].

В качестве возможной реализации управления презентацией были рассмотрены голосовые модели и сервисы Vosk, Assembly AI, PocketSphinx, Speech Recognition, Api.ai, Kaldi (табл. 1). Тестирование моделей осуществлялось на персональном компьютере с процессором на базе 11th Gen Intel(R) Core(TM) (i3-1125G4@2,00 GHz, 8 Gb RAM, Win 10).

Все модели имеют возможность создания и управления собственным словарем, в том числе путем добавления специфичных фраз и соответствующих им действий.

Таблица 1

Сравнение голосовых моделей и сервисов

Модель	**Технология**	**Поддержка русского языка**	**Время распознавания голоса и вывода текста фразы в терминал**	**Время инициализации модели**	**Возможность локального подключения**
Vosk	Carpathian	Vosk-Model-RU	2 с	2,3 с	Да
Assembly AI	Глубокое обучение	Обучение собственной модели	Мгновенно	Менее 1 с	Нет
PocketSphinx	HMM	Обучение собственной модели	3,2 с	43 с	Да
Speech Recognition	Google Web Speech API	Сервисы Google Cloud и IBM Watson	Мгновенно	Мгновенно	Нет
Api.ai (Dialogflow)	Google	Имеется	3,2 с	Мгновенно	Нет
Kaldi	Carpathian	Имеется	-	-	Да

Модель Vosk предназначена для работы в реальном времени [8]. Для использования Vosk необходим доступ к определенным библиотекам и зависимостям в операционной системе, таким как Python, Kaldi, и другие. Хотя модель Vosk обеспечивает высокую точность распознавания речи, она все равно может иметь ограничения в определенных сложных сценариях, таких как сильный шум или различия в дикции. Vosk — это автономный инструмент для распознавания речи с открытым исходным кодом. Он позволяет использовать модели для 17 языков и диалектов. Для использования Vosk с новыми языками или диалектами может потребоваться предварительное обучение модели. Модели Vosk малы (50 Мб). Существуют и более точные модели, их размер достигает 2 Гб. Существует реализация библиотеки на Python, Java, Node JS, C#, C++ и др. Возможен запуск на операционных системах Windows, Linux, Android. На данный момент для русского языка доступны две модели: vosk-model-small-ru-0.4 (50 Мб) и vosk-model-ru-0.10 (2 Гб). Стоит отметить, что данная библиотека распознавания речи не обучена определять жаргонизмы и ненормативную лексику, но позволяет проводить дообучение моделей на пользовательской выборке.

Assembly AI предоставляет API для распознавания речи. Для создания голосовой модели необходим доступ к достаточному объему голосовых данных для обучения модели. Для облачного обучения и использования голосовой модели необходимо стабильное и быстрое интернет-соединение. Низкокачественные или зашумленные данные могут ухудшить производительность модели. При использовании облачного сервиса для создания голосовых моделей необходимо учитывать конфиденциальность голосовых данных и обеспечить их защиту.

PocketSphinx подходит для встроенных устройств и систем с ограниченными ресурсами. Для точного распознавания речи голосовая модель PocketSphinx требует четкого произношения. Звуковое окружение должно также быть достаточно тихим, чтобы предотвратить ошибки. Выполнение голосовой модели PocketSphinx требует больших вычислительных мощностей, что может быть вызвано проблемами при использовании на медленных или устаревших устройствах. Для работы голосовой модели PocketSphinx необходимо наличие дополнительного программного обеспечения для обработки звуковых и текстовых данных, например для работы с акустическими моделями и словарями. Голосовая модель PocketSphinx ограничена объемом словаря, который может быть использован при распознавании речи. Это означает, что словарь должен быть заранее создан и загружен в приложение, в него можно добавлять только ограниченное количество новых слов. Хотя голосовая модель PocketSphinx обеспечивает точное распознавание речи в большинстве случаев, она не всегда может давать точные результаты при сложных условиях, таких как шумное звуковое окружение или нечеткое произношение слов.

Speech Recognition использует веб-интерфейс Google для распознавания речи. Требует подключения к интернету для работы. Голосовая модель Speech Recognition способна обрабатывать аудиозаписи различного качества, включая записи с низким уровнем шума и различными акцентами.

Api.ai предоставляет возможности обработки естественного языка для создания чат-ботов и интерфейсов разговора. Голосовая модель Api.ai работает через интернет, поэтому требуется постоянное подключение к сети. Для использования голосовой модели Api.ai пользователь должен иметь доступ к современным веб-браузерам, таким как Google Chrome, Mozilla Firefox, Safari или Microsoft Edge. Для доступа к Api.ai требуется создание аккаунта Google. Это позволит пользователю сохранять и управлять своими голосовыми моделями. Api.ai имеет ограничения на количество запросов, которые могут быть отправлены в течение определенного периода времени. Существует ограничение на количество символов, которое может содержать один запрос к голосовой модели. Api.ai может не обрабатывать некоторые виды аудиофайлов из-за их качества или формата.

Kaldi — мощный и гибкий инструмент для распознавания речи, часто используется в исследованиях и проектах с большими объемами данных. Kaldi требует большого объема свободного места на жестком диске для обработки аудиофайлов и обучения моделей. Kaldi требуется достаточно быстрый процессор для обработки аудиофайлов и обучения моделей. Низкопроизводительный процессор может снизить скорость обработки. Kaldi требует обученной акустической модели для обработки аудиофайлов. Если модель плохо обучена, это может снизить точность распознавания речи.

Одни из главных критериев проекта — автономность и производительность. Этим критериям отвечала только одна из вышеупомянутых голосовых моделей, Vosk. Благодаря ей стала возможной обработка голоса в автономном режиме, а ее модели весили гораздо меньше, чем ее соперники.

Определение семантической близости текстов

Понятие семантической близости изучалось многими авторами [9, 10]. Под мерой (степенью) смысловой (семантической) близости и похожести (далее — «близость», «сходство») понимается показатель семантического сходства пары рассматриваемых слов или пары наборов слов естественного языка.

Для выделения запроса из распознанного текста производится разбиение текста на отдельные фразы для дальнейшего определения семантического сходства с командами управления. Каждую команду управления презентацией можно представить в виде метки класса семантически сходных слов/фраз, активирующих ее. Класс может пополняться на основе ключевых слов. Ключевые слова (теги) класса команд управления — набор слов естественного языка или терминов, обозначающих действие, выполняемое командой.

Семантическая близость пар слов/фраз может быть определена с помощью готовых тезаурусов, словарей или других семантических сетей [11–15], либо на основе векторного представления для текущего слова (Word Embeddings) [16]. Одними из самых известных методов определения семантической близости слов являются модели word2vec[17], GloVe[18], StarSpace [16] и языковая модель BERT [19].

Обученная модель строит векторное пространство, позволяющее определить семантическую близость понятий, векторные представления которых похожи. Таким образом, любую пару слов из словаря можно сравнить, используя, например, косинусное расстояние между векторами [20].

Насколько два текста близки по содержанию можно определить, путем вычисления косинусной меры близости двух векторов, представляющих тексты:

$$\text{similaruty} = \cos(\theta) = \frac{x \times y}{|x| \times |y|},$$

где x – запрос пользователя, y – категория запроса.

Матрица сходства S представляет собой значения сходства между извлеченными векторными признаками для всех пар корпуса текстов. Таким образом, *Sij* является величиной степени сходства текстов i, j.

Управление презентацией

Проект с голосовым управлением возможно запустить множеством способов. Например, при помощи работы с периферийными устройствами, подключенными к ЭВМ, такими как мышь, клавиатура, устройствами для манипулирования работы курсора на экране ЭВМ, например пульт дистанционного управления, а также при помощи сенсорного экрана. Последняя технология упрощает работу пользователя с манипуляцией курсором в самой операционной системе, но требует для себя дополнительного дорогостоящего аппаратного дополнения.

В некоторых случаях, например при контроле аппаратной конфигурации ПЭВМ, отсутствует возможность подключения дополнительных специализированных устройств для обеспечения взаимодействия пользователя с программами. В подобных ситуациях возможно применение стандартных микрофонов, встроенных в акустические системы вычислительной техники.

Рассмотрим применение голосовых моделей на примере управления презентацией Microsoft PowerPoint.

К системе голосового управления презентацией предъявляются следующие требования:

– распознавание нестандартных формулировок команд в разговорном языке;

– динамическое пополнение словаря;

– независимость от скорости произношения, особенностей дикции, диалектов и посторонних шумов;

– локальная обработка голосовых команд;

– режим постоянного отслеживания команд.

В настоящее время разработаны системы распознавания речи, не зависящие от диктора, доступные для применения в виде сервисов и библиотек. Для реализации голосового управления презентацией в ходе изложения, например, учебного материала на естественном языке необходимо обеспечить выполнение первых двух требований. Для этого необходимо решить задачу выделения и идентификации команд управления презентацией в естественной речи человека, что соответствует построению системы управления на основе бесконечного словаря команд управления с динамическим пополнением.

С целью голосового управления презентацией сформирована система команд управления, представленная в таблице 2.

Фразы, семантически сходные с командами управления, могут быть различной длины. Например, для перехода к следующему слайду могут использоваться фразы: «далее», «следующий слайд», «на следующей схеме», «перейдем к следующему слайду», «перейдем к следующему вопросу». Также возможны нелинейные переходы: «на слайде 5», «на схеме вариантов использования», которые должны осуществлять переход по номеру или названию слайда.

С другой стороны, некоторые фразы, которые могут указывать на необходимость применения управляющей команды, при более детальном рассмотрении таковыми не являются. Например, фразы «в следующем вопросе рассмотрим…», «далее рассмотрим…», «на следующем занятии…» могут быть распознаны как команды перехода к следующему слайду или слайду с определенным названием, но не являются семантически эквивалентными ей.

Приведенные примеры можно отнести к индивидуальным особенностям построения речи, что позволяет адаптировать виртуального ассистента к речи конкретного докладчика путем формирования персонализированного словаря и/или словаря исключений. В случае выполнения неверно распознанной команды возможна ее отмена с добавлением исходной фразы, рассматриваемой как семантический эквивалент команды, в словарь исключений.

Таблица 2

Система команд

Ключевые слова/фразы	**Назначение команды**
Следующий слайд / Вперед	Переключение на следующий слайд
Предыдущий слайд / Назад	Переключение на предыдущий слайд
Кисть	Выбрать инструмент «кисть»
Стереть	Выбрать инструмент «ластик»
Отмена	Убирает инструмент «кисть» или «ластик», автоматическое переключение на стандартный курсор
Запуск	Запускает файл формата .pptx, путь к которому был указан при запуске программы
Открой слайд (выбери страницу) [номер_страницы]	Автоматическое переключение на слайд, порядковый номер которого был назван в команде
Открой слайд (выбери страницу, открой рисунок, открой список) [название]	Автоматическое переключение на слайд, содержащий заголовок, который был назван в команде

Программа управления презентацией

Программа управления презентацией построена на основе сочетания технологий распознавания речи и обработки естественного языка (рис. 1). Программа написана на языке Python 3.10 с использованием библиотеки PyQt5.

Захват аудио выполняется в режиме постоянного прослушивания на основе библиотеки PyAudio. Голосовой поток передается в модуль распознавания речи и преобразования ее в текст, построенный на базе моделиVosk. Распознанный текст поступает на вход модуля поиска команд в тексте, в котором разбивается на отдельные фразы, для последующей оценки семантической близости фраз с ко-

мандами управления. Диаграмма деятельности модуля поиска команд в тексте (рис. 2) более подробно описывает его работу.

В случае выявления команд, подлежащих выполнению, они передаются в модуль исполнения команд, который реализует управление презентаций или функциями операционной системы, такими как открытие файла.

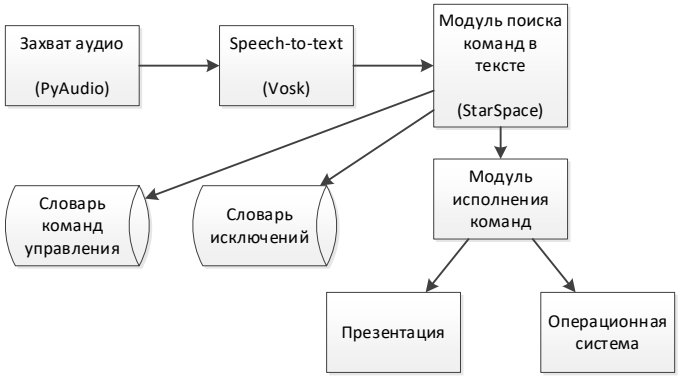

Рис. 1. Структура программы голосового управления

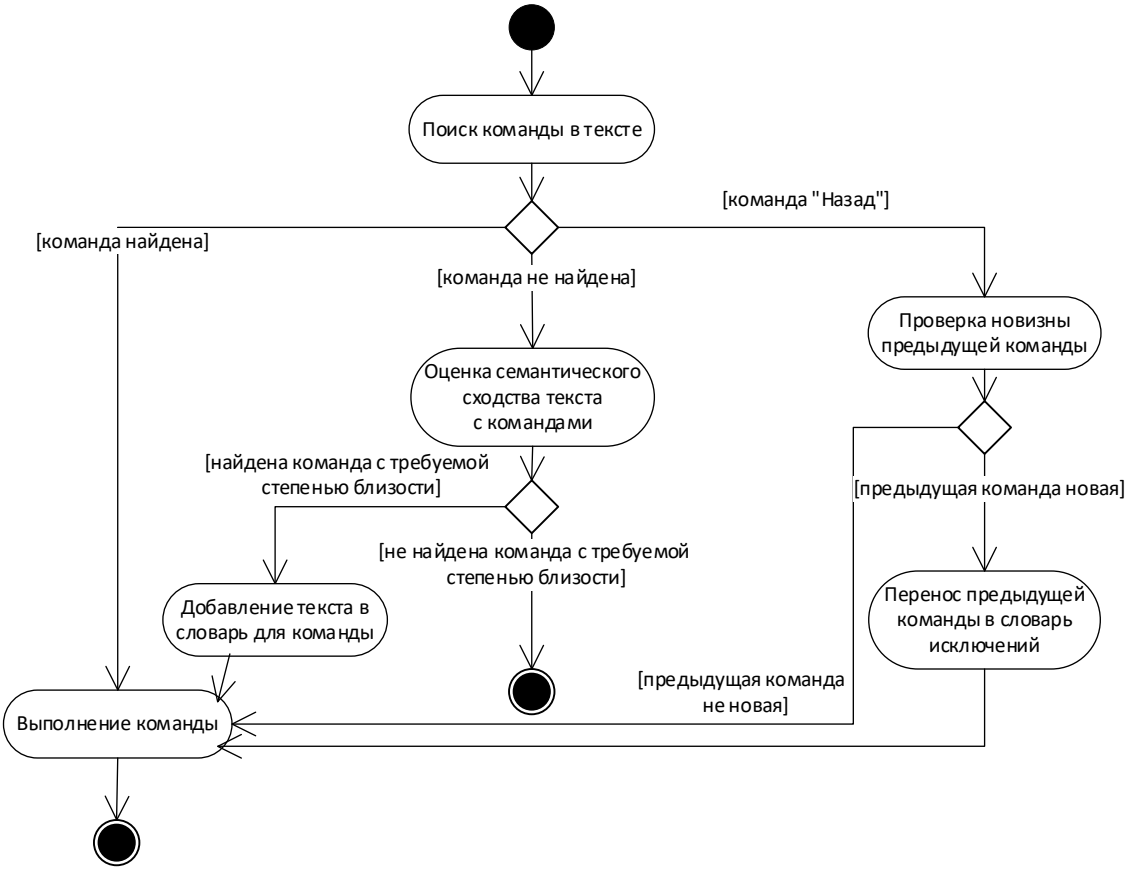

Рис. 2. Диаграмма деятельности модуля поиска команд в тексте

Работа с голосовым ассистентом проста и интуитивно понятна. Пользователь при помощи голоса подает системе команду, которая воспринимается и выполняет определенный алгоритм с презентацией (например, запускает файл формата .pptx).Далее, изучив документацию к программе, пользователь способен выполнять различные манипуляции с презентацией.

При запуске программы на экране появляется интерфейс взаимодействия с ней, содержащий:

- поле ввода информации;
- кнопку «Файловая система»;
- кнопку «Старт!»;
- кнопку «Информация».

Интерфейс программы показан на рисунке 3.

Кнопка «Файловая система» связана с голосовой командой «Открыть» и открывает проводник, в котором выбирается нужная презентация. После выбора элемента и подтверждения команды «Открыть» в поле ввода информации появляется путь к данному файлу.

Кнопка «Информация» связана с соответствующей голосовой командой, открывающей документацию к программе в формате html-файла. В ней описаны поддержива-

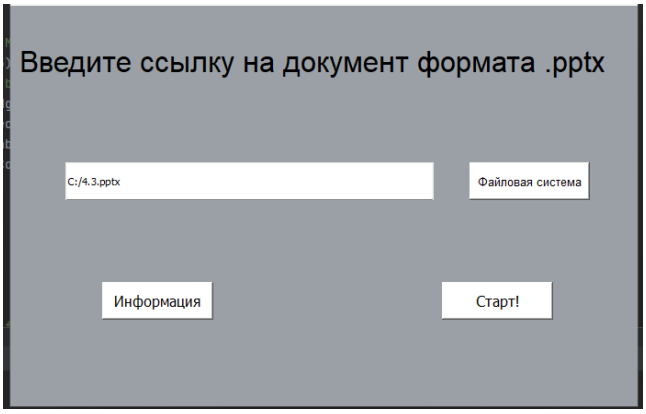

Рис. 3. Интерфейс программы голосового управления

емые в проекте голосовые команды и перечислены заранее добавленные в словарь этих команд фразы.

Кнопка «Старт!» и соответствующая голосовая команда запускает проект. Обязательное условие использования данного элемента интерфейса — наличие и правильное написание абсолютного пути к файлу презентации. Если данное условие соблюдено — программа запускается.

Сразу после запуска идет инициализация голосовой модели Vosk. Осуществляется подгрузка словаря и всех других элементов, необходимых для нормальной работы программы.

Следующий шаг — запуск приема голосового потока. Звуковая информация, попадающая в программу, преобразуется в формат файла JSON, после чего парсится и обрабатывается с целью поиска словосочетания команды или семантически сходной фразы в нем. Если команда в тексте определена, программа начинает выполнять алгоритм, соответствующий данной команде.

Заключение

Предлагаемое решение позволяет сократить размерность закрытого словаря команд управления за счет введения классов семантически сходных фраз, а также добавления для каждого класса словаря исключений. Подобный подход позволяет снизить количество вызовов алгоритма определения семантически эквивалентных фраз, что повышает скорость обработки речи при задействовании меньших вычислительных ресурсов.

Решение построено на базе свободно распространяемых библиотек, содержащих предобученные модели, доступных для локальной установки, что позволяет использовать данный продукт без доступа к интернет-сервисам.

Дальнейшее развитие приложения возможно за счет повышения точности распознавания команд в естественной речи, удобства использования и интеграции с другими подходами управления, например жестами, что сделает его еще более полезным и привлекательным для широкого круга пользователей.

Литература

1. Кипяткова, И. С. Автоматическая обработка разговорной русской речи: Монография / И. С. Кипяткова, А. Л. Ронжин, А. А. Карпов; Санкт-Петербургский ин-т информатики и автоматизации РАН. — Санкт-Петербург: ГУАП, 2013. — 314 с.

2. Обработка естественного языка, распознавание и синтез речи. Применения // Искусственный интеллект. 2019. № 2. С. 67–100.

3. Ротов, А. П. Внедрение средств распознавания речи в тренажерные комплексы управления воздушным движением / А. П. Ротов, А. Ю. Княжский // Аэрокосмическое приборостроение и эксплуатационные технологии: Сборник докладов Четвертой Международной научной конференции (Санкт-Петербург, Россия, 04–21 апреля 2023 г.). Часть 1. — Санкт-Петербург: ГУАП, 2023. — С. 72–76.

4. Petraitytė, J. Cómo construir un asistente de voz con herramientas de código abierto como Rasa y Mozilla // Planeta Chatbot. — 2019. — 30 September. URL: http://planetachatbot.com/tutorial-como-construir-asistente-voz-con-herramientas-de-codigo-abierto-rasa-y-mozilla (дата обращения 01.12.2023).

5. On the Record. Exploring the ethical, technical and legal issues of voice assistants (CNIL White Paper Collection No. 1). — Commission Nationale de l'Informatique et des Libertés, 2020. — 84 p. URL: http://www.cnil.fr/sites/cnil/files/atoms/files/cnil_white-paper-on_the_record.pdf (дата обращения 03.12.2023).

6. Campoy, A. Voice Assistants 101: A Look at How Conversational AI Works / A. Campoy, S. Sassi // Sophilabs. — 2019. — 28 August. URL: https://sophilabs.com/blog/voice-assistants-101 (дата обращения 28.11.2023).

7. Топорин, А. А. Подсистема голосового управления системы интеллектуального управления мобильным роботом // Вестник науки и образования. 2020. № 14-4 (92). С. 9–13.

8. Vosk. Распознавание речи без сети // Alpha Cephei Speech Recognition. URL: http://alphacephei.com/vosk/index.ru (дата обращения 28.11.2023).

9. Frawley, W. Linguistic Semantics. — New York: Routledge, 1992. — 552 p.

10. Semantic Similarity from Natural Language and Ontology Analysis / S. Harispe, S. Ranwez, S. Janaqi, J. Montmain. — Cham: Springer Nature, 2015. — 252 p. — (Synthesis Lectures on Human Language Technologies).
DOI: 10.1007/978-3-031-02156-5.

11. RussNet: Building a Lexical Database for the Russian Language / I. Azarova, O. Mitrofanova, A. Sinopalnikova, [et al.] // Third International Conference on Language Resources and Evaluation (LREC 2002): Proceedings of Workshop on WordNet Structures and Standardisation, and How this Affect Wordnet Applications and Evaluation (Las Palmas, Spain, 29–31 May 2002). — Pp. 60–64.

12. Braslavski, P. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus / P. Braslavski, D. Ustalov, M. Mukhin // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (Gothenburg, Sweden, 26–30 April 2014). — Association for Computational Linguistics, 2014. — Pp. 101–104. DOI: 10.3115/v1/E14-2026.

13. YARN: Spinning-in-Progress / P. Braslavski, D. Ustalov, M. Mukhin, Y. Kiselev // Proceedings of the Eighth Global WordNet Conference (GWC 2016), (Bucharest, Romania, 27–30 January 2016) / V. B. Mititelu, [et al.] (eds.). — Global WordNet Association, 2016. — Pp. 58–65.

14. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска. — Москва: Изд-во Московского ун-та, 2011. — 512 с.

15. Creating Russian WordNet by Conversion / N. V. Loukachevitch, G. Lashevich, A. A. Gerasimova, [et al.] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, Россия, 01–04 июля 2016 г.). Выпуск 15 (22). — Москва: Изд-во РГГУ, 2016. — С. 405–415.

16. StarSpace: Embed All the Things! / L. Wu, A. Fisch, S. Chopra, [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. 2018. Vol. 32. Pp. 5569–5577. DOI: 10.1609/aaai.v32i1.11996.

17. Distributed Representations of Words and Phrases and Their Compositionality / T. Mikolov, I. Sutskever, K. Chen, [et al.] // Advances in Neural Information Processing Systems 26 (NIPS 2013): Proceedings of the 27th Annual Conference on Neural Information Processing Systems (Stateline, NV, USA, 05–10 December 2013) / C. J. C. Burges, [et al.] (eds.). — Curran Associates, 2013. — Pp. 3111–3119.

18. Pennington, J. Glove: Global Vectors for Word Representation / J. Pennington, R. Socher, C. D. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (Doha, Qatar, 25–29 October 2014). — Association for Computational Linguistics, 2014. — Pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

19. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), (Minneapolis, MN, USA, 02–07 June 2019). Volume 1 / J. Burstein, [et al.] (eds.). — Association for Computational Linguistics, 2019. — Pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

20. Бермудес, С. Х. Г. Метод измерения семантического сходства текстовых документов // Известия Южного федерального университета. Технические науки. 2017. № 3 (188). С. 17–29. DOI: 10.23683/2311-3103-2017-3-17-29.

DOI: 10.24412/2413-2527-2023-436-19-27

Application of Natural Language Processing Technologies for Voice Control Based on an Open Dictionary

S. A. Mikhaylova, K. G. Anikeev
Military Aerospace Academy
Saint Petersburg, Russia
mikhaylova_sa@mail.ru

Abstract. **The possibility of using natural language processing technologies to solve the problem of voice control by identifying phrases in natural speech that have a similar meaning to control commands is considered. The capabilities of speech recognition libraries and services are considered. The structure of a voice control system based on the Vosk speech recognition library is proposed. This structure allows you to replenish the dictionary of phrases corresponding to control commands based on an assessment of semantic proximity.**

Keywords: **speech recognition, voice control, natural language processing.**

REFERENCES

1. Kipyatkova I. S., Ronzhin A. L., Karpov A. A. Automatic processing of spoken Russian speech: Monograph [Avtomaticheskaya obrabotka razgovornoy russkoy rechi: Monografiya]. Saint Petersburg, St. Petersburg State University of Aerospace Instrumentation, 2013, 314 p.

2. Natural Language Processing, Speech Recognition and Synthesis. Applications [Obrabotka estestvennogo yazyka, raspoznavanie i sintez rechi. Primeneniya], *Artificial Intelligence [Iskusstvennyy intellekt]*, 2019, No. 2, Pp. 67–100.

3. Rotov A. P., Knyazhsky A. Yu. Implementation of Speech Recognition Tools in Air Traffic Control a Training Complexes [Vnedrenie sredstv raspoznavaniya rechi v trenazhernye kompleksy upravleniya vozdushnym dvizheniem], *Aerospace Instrumentation and Operational Technologies: Proceedings of the Fourth International Scientific Conference [Aerokosmicheskoe priborostroenie i ekspluatatsionnye tekhnologii: Sbornik dokladov Chetvertoy Mezhdunarodnoy nauchnoy konferentsii], Saint Petersburg, Russia, April 04–21, 2023. Part 1*. Saint Petersburg, St. Petersburg State University of Aerospace Instrumentation, 2023, Pp. 72–76.

4. Petraytite J. Cómo construir un asistente de voz con herramientas de código abierto como Rasa y Mozilla, *Planeta Chatbot*. Published online at September 30, 2019. Available at: http://planetachatbot.com/tutorial-como-construir-asistente-voz-con-herramientas-de-codigo-abierto-rasa-y-mozilla (accessed 01 Dec 2023).

5. On the Record. Exploring the ethical, technical and legal issues of voice assistants (CNIL White Paper Collection No. 1). Commission Nationale de l'Informatique et des Libertés, 2020, 84 p. Available at: http://www.cnil.fr/sites/cnil/files/atoms/files/cnil_white-paper-on_the_record.pdf (accessed 03 Dec 2023).

6. Campoy A., Sassi S. Voice Assistants 101: A Look at How Conversational AI Works, Sophilabs. Published online at August 28, 2019. Available at: http://sophilabs.com/blog/voice-assistants-101 (accessed 28 Nov 2023).

7. Toporin A. A. Voice Control Subsystem of Intelligent Mobile Robot Control System [Podsistema golosovogo upravleniya sistemy intellektualnogo upravleniya mobilnym robotom], *Vestnik Nauki i Obrazovaniâ*, 2020, No. 14-4 (92), Pp. 9–13.

8. Vosk. Offline Speech Recognition API [Vosk. Raspoznavanie rechi bez seti], *Alpha Cephei*. Available at: http://alphacephei.com/vosk/index (accessed 28 Nov 2023).

9. Frawley, W. Linguistic Semantics. New York, Routledge, 1992, 552 p.

10. Harispe S., Ranwez S., Janaqi S., Montmain J. Semantic Similarity from Natural Language and Ontology Analysis. Cham, Springer Nature, 2015, 252 p. DOI: 10.1007/978-3-031-02156-5.

11. Azarova I., Mitrofanova O., Sinopalnikova A., et al. RussNet: Building a Lexical Database for the Russian Language, *Third International Conference on Language Resources and Evaluation (LREC 2002): Proceedings of Workshop on WordNet Structures and Standardisation, and How this Affect Wordnet Applications and Evaluation, Las Palmas, Spain, May 29–31, 2002*, Pp. 60–64.

12. Braslavski P., Ustalov D., Mukhin M. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus, *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30, 2014*. Association for Computational Linguistics, 2014, Pp. 101–104. DOI: 10.3115/v1/E14-2026.

13. Braslavski P., Ustalov D., Mukhin M., Kiselev Y. YARN: Spinning-in-Progress. In: *Mititelu V. B., et al. (eds.) Proceedings of the Eighth Global WordNet Conference (GWC 2016), Bucharest, Romania, January 27–30, 2016*. Global WordNet Association, 2016, Pp. 58–65.

14. Lukashevich, N. Thesauruses in information retrieval problems [Tezaurusy v zadachakh informatsionnogo poiska], Lomonosov Moscow State University, 2011, 512 p.

15. Loukachevitch N. V., Lashevich G., Gerasimova A. A., et al. Creating Russian WordNet by Conversion, *Computational Linguistics and Intellectual Technologies: Proceedings of the 2016 Annual International Conference «Dialogue» [Kompyuternaya lingvistika i intellektualnye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog»], Moscow, Russia, July 01–04, 2016. Issue 15 (22)*. Moscow, Russian State University for the Humanities, 2016, Pp. 405–415.

16. Wu, L., Fisch A., Chopra S., et al. StarSpace: Embed All the Things!, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, Vol. 32, Pp. 5569–5577. DOI: 10.1609/aaai.v32i1.11996.

17. Mikolov T., Sutskever I., Chen K., et al. Distributed Representations of Words and Phrases and Their Compositionality. In: *Burges C. J. C., et al. (eds.) Advances in Neural Information Processing Systems 26 (NIPS 2013): Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Stateline, NV, USA, December 05–10, 2013*. Curran Associates, 2013, Pp. 3111–3119.

18. Pennington J., Socher R., Manning C. D. Glove: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29, 2014*. Association for Computational Linguistics, 2014, Pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

19. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Burstein J., et al. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, June 02–07, 2019. Volume 1*. Association for Computational Linguistics, 2019, Pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

20. Bermudez S. J. G. Method for Measuring the Semantic-Similarity of Textual Documents [Metod izmereniya semanticheskogo skhodstva tekstovykh dokumentov], *Izvestiya of the Southern Federal University. Engineering Science [Izvestiya Yuzhnogo federalnogo universiteta. Tekhnicheskie nauki]*, 2017, No. 3 (188), Pp. 17–29. DOI: 10.23683/2311-3103-2017-3-17-29.