INTERNATIONAL JOURNAL OF FORMAL EDUCATION

Volume: 3 Issue:11 | Oct–2024 ISSN: 2720-6874
http://journals.academiczone.net/index.php/ijfe

Разработка корпуса узбекской речи для обучения автоматического распознавания речи

И.А. Расулова
(PhD), доцент
кафедры узбекского языка и литературы
Самаркандского государственного
архитектурно-строительного университета
e-mail: istoda.87@gmail.com

Аннотация. Данная статья посвящена актуальной проблеме развития технологий автоматического распознавания речи (АРР) для узбекского языка. Отсутствие достаточного количества корпусов речи является основным препятствием для создания эффективных систем АРР, способных удовлетворять растущий спрос на обработку естественного языка для узбекского языка. В статье анализируется существующая ситуация с ресурсами для обработки естественного языка для узбекского языка, выявляются ключевые проблемы, связанные с недостатком корпусов речи. Предлагается концепция разработки нового корпуса узбекской речи, включающая описание его структуры, методов сбора и аннотирования данных. Рассматриваются перспективы использования разрабатываемого корпуса для обучения систем автоматического распознавания речи и других приложений компьютерной лингвистики. Особое внимание уделяется выбору и описанию конкретных технологий АРР, которые будут использоваться для обучения на разрабатываемом корпусе, таких как Kaldi и HTK.

Keywords: корпус речи, узбекский язык, автоматическое распознавание речи, обработка естественного языка, машинное обучение, Kaldi, HTK.

Введение

В современном мире автоматическое распознавание речи (АРР) стало неотъемлемой частью многих технологий, включая виртуальных помощников, системы диктафона, машинный перевод и др. [1] Развитие АРР для неевропейских языков, включая узбекский, отстает из-за недостатка достаточного количества корпусов речи. [2] Корпусы речи являются важным ресурсом для обучения и тестирования алгоритмов АРР. [4] Они содержат большое количество записей речи, аннотированных транскрипцией и другими метаданными.

Актуальность исследования:

Разработка корпуса узбекской речи для обучения АРР является актуальной задачей по следующим причинам:

Отсутствие достаточных ресурсов: В настоящее время доступных корпусов узбекской речи для обучения АРР недостаточно для создания высококачественных систем распознавания. [5]

Увеличение спроса на технологии АРР: С развитием мобильных устройств и интернета возрастает спрос на технологии АРР для узбекского языка[2], что требует создания более эффективных систем распознавания. [6]

Развитие компьютерной лингвистики: Разработка корпусов речи способствует развитию компьютерной лингвистики для узбекского языка и открывает новые возможности для исследований в области обработки естественного языка. [7]

Структура корпуса:

Volume: 3 Issue:11 | Oct–2024 ISSN: 2720-6874
http://journals.academiczone.net/index.php/ijfe

Разрабатываемый корпус узбекской речи будет содержать следующие данные:

Аудиозаписи: Записи речи в разных диалектах узбекского языка, включая разные голосовые характеристики (пол, возраст, акцент). [6]

Транскрипция: Текстовая транскрипция аудиозаписей, соответствующая орфографическим нормам узбекского языка.

Метаданные: Дополнительная информация об аудиозаписях, включая информацию о говорящем, теме речи, контексте записи, и др.

Методы сбора и аннотирования данных:

Для сбора данных будут использоваться следующие методы:

Запись речи в реальном времени: Запись речи от носителей узбекского языка в разных контекстах, включая естественную речь, чтение текстов и др. [5]

Использование открытых данных: Использование доступных в открытом доступе записей узбекской речи, например, из библиотек аудио книг.

Создание искусственных данных: Использование синтеза речи для генерации искусственных данных, что позволяет увеличить разнообразие корпуса и улучшить качество обучения систем APP. [6]

Аннотирование данных будет осуществляться следующими методами:

Ручная аннотация: Ручная транскрипция аудиозаписей с использованием специальных инструментов для аннотирования.

Автоматическая аннотация: Использование методов машинного обучения для автоматической транскрипции, с последующей ручной коррекцией.

Выбор технологий APP:

Для обучения систем APP на разрабатываемом корпусе будут использоваться две популярные технологии:

Kaldi: Kaldi является открытой платформой для обучения систем APP. [10] Она предоставляет богатый набор алгоритмов и инструментов, что делает ее удобной для экспериментов и разработки новых моделей.

HTK: HTK является другой популярной платформой для обучения систем APP. Она известна своей стабильностью и широким набором инструментов для акустической обработки речи.

Выбор Kaldi и HTK определяется их широкой применимостью, открытостью и наличием большого количества документации и поддержки в сообществе разработчиков.

Перспективы использования корпуса:

Разрабатываемый корпус узбекской речи может быть использован для следующих целей:

Обучение систем APP: Корпус будет использоваться для обучения алгоритмов APP для узбекского языка с использованием Kaldi и HTK.

Разработка ресурсов для обработки естественного языка: Корпус может быть использован для разработки словарей, морфологических анализаторов и других ресурсов для обработки естественного языка. [11]

Проведение научных исследований: Корпус предоставит исследователям ценные данные для анализа речевых особенностей узбекского языка и разработки новых моделей обработки естественного языка[3].

Методы и инструменты обработки лингвистических данных зависят от целей и задач исследования. Основные методы и инструменты включают:

Volume: 3 Issue:11 | Oct–2024 ISSN: 2720-6874
http://journals.academiczone.net/index.php/ijfe

Лексикографические инструменты: такие как словари, тезаурусы, конкордансы и глоссарии, которые помогают анализировать и классифицировать лексические единицы, а также уточнять значения слов и их использование.

Морфологические анализаторы: программы, которые определяют форму слова и его грамматические характеристики. Они могут автоматически разбирать текст на слова и выявлять их морфологические свойства.

Синтаксический анализ: методы и инструменты для анализа структуры предложений и отношений между компонентами. Они помогают выявить синтаксические единицы, такие как подлежащее, сказуемое и дополнение, а также их взаимосвязи.

Статистические методы: позволяют измерять частотность слов и выражений, а также выполнять статистический анализ текстовых корпусов для выявления закономерностей.

Машинное обучение: методы, основанные на использовании компьютерных алгоритмов для обработки и анализа текстов. Они позволяют автоматически классифицировать и кластеризовать тексты, извлекать ключевые слова и информацию из больших объемов данных.

В зависимости от задачи, исследователи могут сочетать различные методы и инструменты для достижения наилучших результатов.

Примеры успешного применения информационных технологий в лингвистических исследованиях включают:

Определение авторства текстов: методы, такие как стилистический и лексический анализ, а также анализ графем, используются для идентификации авторства. Инструмент LIWC выделяет стилистические и лексические особенности текстов, что позволяет точно определить автора.

Анализ тональности текстов: процесс определения эмоциональной окраски текста с помощью информационных технологий, таких как машинное обучение и анализ данных. Примеры включают анализ отзывов о товарах и услугах в социальных сетях.[12]

Анализ социолингвистических феноменов: использование информационных технологий для анализа лексических и грамматических особенностей различных диалектов и региональных вариантов языка. Например, можно проводить сопоставительный анализ использования прецедентных феноменов в заявлениях высокопрофильных политиков.

Компьютерный анализ дискурса: анализ разговорных данных для выявления лингвистических паттернов с помощью методов машинного обучения. Примером может служить анализ дискуссий в интернет-форумах или социальных сетях.[13]]

Автоматический перевод: процесс перевода текста между языками с использованием компьютерных алгоритмов и методов машинного обучения. Это исследование включает оценку точности и качества машинного перевода различных языков. [14]

Заключение:

Разработка корпуса узбекской речи для обучения APP является важным шагом в развитии компьютерной лингвистики для узбекского языка. Он позволит создать более эффективные системы APP и открыть новые возможности для использования узбекского языка в современных технологиях. [15]

Список литературы:

1. Тампель И.Б. Автоматическое распознавание речи - основные этапы за 50 лет // Научно-технический вестник информационных технологий, механики и оптики. 2015. №6. URL:

Volume: 3 Issue:11 | Oct–2024 ISSN: 2720-6874
http://journals.academiczone.net/index.php/ijfe

https://cyberleninka.ru/article/n/avtomaticheskoe-raspoznavanie-rechi-osnovnye-etapy-za-50-let (дата обращения: 13.09.2024).
2. Musaev, M., Khujayorov, I., Ochilov, M., Khassanov, Y., Mussakhojayeva, S., and Varol, H.A. (2021). USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. arXiv preprint arXiv:2107.14419.
3. Тоирова Г.И. Общие принципы создания национального корпуса узбекского языка // Наука и образование: актуальные вопросы теории и практики. - 2021. - С. 284-288.
4. Чилингарян, К. П. (2021). Корпусная лингвистика: теория vs методология. Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика, 12(1),196-218. https://www.researchgate.net/publication/353273193_Korpusnaa_lingvistika_teoria_vs_metodologia1
5. Продеус, А. Н. (2013). Речевые корпуса: создание и проблемы. Электротехнические и компьютерные системы, (9), 118-126.
6. Медетов, Б., Нурланкызы, А., Кулакаева, А., Жетписбаева, А., & Намазбаев, Т. (2024). ОЦЕНКА ВЛИЯНИЯ ЯЗЫКА НА ТОЧНОСТЬ РАСПОЗНАВАНИЯ ЧЕЛОВЕЧЕСКОГО ГОЛОСА С ПОМОЩЬЮ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ. Вестник КазАТК, 131(2), 456-466.
7. НУРИМОВ, П., & НИЁЗМАТОВА, Н. (1992). РАСПОЗНАВАНИЕ КАРАКАЛПАКСКОЙ РЕЧИ С ПОМОЩЬЮ CMU SPHINX. 1· 2019_, 90.
8. Продеус, А. Н. (2013). Речевые корпуса: создание и проблемы. Электротехнические и компьютерные системы, (9), 118-126.
9. Ахмедиярова, А. Т., Нурланкызы, А., Кулакаева, А. Е., & Медетов, Б. Ж. (2024). АНАЛИЗ ЭФФЕКТИВНОСТИ НЕЙРОННЫХ СЕТЕЙ ПО РАСПОЗНАВАНИЮ ЧЕЛОВЕЧЕСКОГО ГОЛОСА. Вестник Ауэс, 1(64).
10. Ангапов, В. Д. (2023). АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ГОЛОСА В ГОЛОСОВЫХ ПОМОЩНИКАХ. Проблемы современной науки и образования, (8 (186)), 8-14.
11. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding.
12. Полюшина Д.В., Андронов А.Н. Анализ тональности комментариев, оставленных в социальной сети // L Огарёвские чтения. - 2022. - С. 745-752.
13. Распопина Е.Ю. Дифференциальные и жанровые особенности компьютерного интернет-дискурса // Вестник Иркутского государственного лингвистического университета. - 2010. - №. 1 (9). - С. 125-132.
14. Salohitdinovna, B. S. (2023). IQTIDORLI BOLALARNI O ‘QITISHGA IXTISOSLASHGAN MAKTABLAR FAOLIYATI SAMARADORLIGINI OSHIRISH. Ta'limning zamonaviy transformatsiyasi, 2(1), 785-789.
15. Жаббарова Р.У., Бурнашев Р.Ф. Инструментарий обработки лингвистической информации // Science and Education. - 2023. - Т. 4. - №. 4. - С. 654-664.
16. Young, S., Hwang, Y., Acero, A., Badr, I., Beal, R., Bellegarda, J.R., Chen, L., Chu, S., Cole, R., and Droppo, J. (2006). The HTK book (for HTK version 3.4). Cambridge University Engineering Department.