Краткая информация о проекте

Наименование	AP19677835 «Исследование моделей и разработка интеллектуальной вопросно-ответной системы на основе семантических подходов для государственного языка в сфере законодательства Республики Казахстан» (0123РК00861)
Актуальность	Современные системы и разработанные технологии в сфере обработке естественного языка далее -NLP (Natural language processing) и искусственного интеллекта (ИИ) удивляют своими достижениями и скоростью развития и внедрения различные профессиональные сферы деятельности и в повседневной жизни человека. Так же необходимо учитывать, что за полученными результатами и достижениями в NLP стоить большой научный труд и исследования, технологические решения и возможности. Одно из актуальных и востребованных направлении NLP диалоговые и вопросно-ответные системы. Данные системы включают в себя большой ряд подзадач, результаты которых могут применены в различных сферах искусственного интеллекта как речевые технологии, машинный перевод, поисковые системы и др. Существует различные научные подходы и методы для решения вопросно-ответных систем. Далее будут представлены некоторые из них. Программное обеспечение, предлагаемое различными производителями, такие как В настоящее время имеется довольно большое количество реализаций вопросно-ответных систем. Заслуживает внимания реализация вопросно-ответной системы START . Другие интересные системы, претендующие на статус открытых, являются системы Cyc , OpenEphyra и PIQUANT. Также исследователи семантической паутины (Semantic Web) предложили использовать технологии семантического поиска для английского языка, таких как OntoSearch, Semantic Wikis, Semantic Portals , мульти-агентная P2P семантическая система маршрутов(запросов) на основе онтологии, вопросно-ответные системы на основе онтологии . А так же известная смарт станция «Алиса» — виртуальный голосовой помощник, созданный компанией Яндекс. Распознаёт естественную речь, имитирует живой диалог, даёт ответы на вопросы пользователя и, благодаря запрограммированным навыкам, решает прикладные задачи. В данное время применяется различные чат-бот системы на казахском языке в государственном портале egov и различные банковские системы. Но к сожалению они ограничены тематикой и протоколами ответа, а так же не поддерживают голосового функционала. К сожалению, аналогов данных систем (открытого доступа и/или платных ресурсов) для казахского языка не существуют. Выше представленные программные продукты разработаны для много ресурсных (европейские, славянские

	группы) языков как английский, испанский, русский и др. К сожалению, для тюркских языков (казахский, кыргызский, турецкий, узбекский и др.) на данный момент в открытом доступе нет программной реализации. Представленные работы и системы не могут применены частично или полностью для казахского языка. Это связано с тем что для каждого языка собираются и обрабатываются отдельные ресурсы с учетом лингвистических свойств. Так же для каждого языка свойствен свой диалект, семантические (когнитивные) понятия и др.
Цель	Целью проекта является исследование задачи и разработка вопросно-ответной технологии (алгоритмы, модели, системы) на казахском языке в сфере законодательства Республики Казахстан, на основе современных семантических подходов и нейросетевом обучении.
Задачи	Основными задачами проекта являются: • Разработка подхода сбора и обработки текстовых данных из открытых электронных источников.; • Разработка матрицы вопросов-ответов взаимодействия; • Разработка метода синтеза ответа на казахском с учетом семантики на основе машинного обучения. • Разработка метода прогнозирования ответов путем построения модели классификации и нейросетевом обучении • Разработка и внедрение модуля поддержки голосового интерфейса системы на казахском языке
Ожидаемые и достигнутые результаты	Ожидаемые результаты проекта: исследование задачи и разработка прототипа интеллектуальной вопросно-ответной технологии (алгоритмы, модели, системы) на основе нейросетевом обучении с возможностью ведения диалога на казахском языке. Планируется издание статей в изданиях, индексируемых в Science Citation Index Expanded базы Web of Science и (или) имеющих процентиль по CiteScore в базе Scopus не менее 35 (тридцати пяти) и публикация) статьи в рецензируемом зарубежном или отечественном издании, рекомендованном КОКСНВО. Выполненные работы и полученные результаты на 2023 г. : - Осуществлен аналитический обзор по тематике исследования. Были исследованы и проанализированы современные зарубежные аналоги вопросно-ответных систем, чат-боты и веб порталы, связанные с юридическими и нормативными документами. Описаны своиства ChatGPT и его аналоги, такие как: «Alphachat», «Law ChatGPT», «DoNotPay», «LawDroid», «Ross Intelligence», «CaseMine».

	Описаны их отличия и применяемые методы обучения с учителем, и обучения с подкреплением. - Исследованы виды нормативно-правовых документов, а также произведена их классификация на основные и производные виды, в зависимости от разработки и порядка принятия административных и правовых актов Республики Казахстан. ;- Описана иерархия законодательных документов РК. На основе иерархии производится сбор данных из законодательных документов для формирования БД. - Определена структура казахского языка законодательных документов на основе формальных грамматик. Для этого были проведен анализ текстов, предложений разных видов из документов правовых актов на казахском языке. - Был произведен сбор данных из статей законодательных сайтов Республики Казахстан. Главным источником информации сайт https://adilet.zan.kz/kaz/ . Выгрузка данных выполнялась в html формате с последующим преобразованием в текстовый формат (txt). Для сбора данных был применен универсальный многопоточный парсер сайтов для Windows с большим количеством функций и возможностью полного контроля и настройки всех этапов работы с WEB контентом - Content downloader.
Имена и фамилии членов исследовательской группы с их идентификаторами (Scopus Author ID, Researcher ID, ORCID, при наличии) и ссылками на соответствующие профили	Руководитель проекта: PhD. Рахимова Диана Рамазановна h-индекс: 4, Scopus author ID: 55682794500 Web of Science ResearcherID D-8421-2012 Члены исследовательской группы 1. Д.т.н., проф. Тукеев Уалшер Ануарбекович h-индекс: 5. Scopus author ID 55701639900 https://orcid.org/0000-0001-9878-981X 2. PhD Шормакова Асем *h-индекс*:2, ORCID ID: 0000-0002-1637-4643, Author ID Scopus: 55786942400, Researcher ID Web of Science: D-5836-2015 3. PhD Кәрібаева Айдана h-index: 3 orcid id: 0000-0002-2023-1573, Scopus id: 57196004542, ResearcherID: AAR-4134-2020 4. PhD Карюкин Владислав, h-1 , https://orcid.org/0000-0002-8768-0349 , Scopus Author ID: 57218952479 SciProfiles: 2410440 5. Магистр Турарбек Асем h-1, Scopus id 57031944900, https://orcid.org/0000-0002-4793-0446 6 Магистр Амирова Дина

	h-индекс: 1, ResearcherID: P-5668-2017, Orcid id: 0000-0002-0728-905X, Scopus ID: 57196009653
Список публикаций со ссылками на них	Публикации за 2023 г. : 1.Ualsher Tukeyev and etc. Kazakh-Tatar Machine Translation on the Base of Complete Set of Endings Model. Chapter of the book «Machine Learning». Editors: Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci ©Yeditepe University Press, pp. 73-90. ISBN: 978-975-307-139-0 Istanbul, 2023 2.Diana Rakhimova, Yntymak Abdrazakh. Study and development of an approach for identifying incorrect words for the Kazakh language in semi-structured data. Chapter of the book «Machine Learning». Editors: Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci ©Yeditepe University Press, pp. 23-39. ISBN: 978-975-307-139-0 Istanbul, 2023. 3.Тукеев У.А. Реляционные модели обработки тюркских языков. VIII — Международную научно-практическую конференцию «Информатика и прикладная математика» с 26 по 27 октября 2023 г. Алматы. 85-89 с.
Информация о патентах	-
-	