УДК 004.912

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ТЕКСТОВ ДЛЯ ОРГАНИЗАЦИИ ЧАТ-БОТОВ В ДИСТАНЦИОННОМ ОБУЧЕНИИ

© **Братищенко В. В., 2024**
Байкальский государственный университет, г. Иркутск

В данной статье рассматриваются возможности применения технологий обработки текстов на естественном языке для организации диалога с обучающимися в виде чат-бота. Обосновывается применение в учебных чат-ботах распознавания семантики вопроса. Предлагается использовать методы классификации текстов для обработки ответов на тестовые задания открытого типа, а оценку ответа формировать на основе оценки вероятности отнесения ответа правильному классу. Описана

технология формирования набора заданий, обучения классификатора и проведения тестирования. Приведены результаты эксперимента по обучению классификатора и обработке ответов.

Ключевые слова: дистанционное обучение, чат-бот, text mining, тестовые задания открытого типа, классификация текстов, мешок слов

Информационные технологии оказывают значительное влияние на процессы обучения. Поиск информации стал существенно проще. Практически по любому вопросу можно получить справку в Интернете. На учебную задачу можно найти готовое решение.

Можно получить доступ к учебным графикам и расписаниям. Вузы выкладывают в Интернет программы дисциплин, практик, а также требования к выпускным квалификационным работам. Доступны коммуникации в разных формах, начиная с электронной почты и заканчивая занятиями в форме видеоконференций. Ключевым в обучении, тем не менее, остается взаимодействие обучаемого и преподавателя. В статье обсуждается применение чат-ботов для организации такого взаимодействия.

Информационные технологии влияют на способы организации учебного процесса. Все чаще применяются методики «смешанного» обучения [1], в которых очные занятия сопровождаются дистанционным обучением. При этом актуальной является задача мотивации студентов к изучению учебных материалов дистанционно. Для этого изучение лекции, прохождение контрольного опроса и/или тестирования в дистанционном курсе становятся допуском к очному занятию. Такая методика получила название «Перевернутый класс», в нем основные положения теории изучаются дистанционно, а дискуссии и применение теории для решения практических задач организуются в очной форме. Эта методика требует создания дистанционных курсов со множеством интерактивных элементов и иллюстративных материалов. В технологическом плане такие курсы требуют значительных трудозатрат коллектива авторов. Все это окупается снижением затрат в процессе обучения и мотивацией обучаемых к самостоятельной работе.

Другой особенностью дистанционного обучения становится организация общения обучаемых с преподавателем. Общение в режиме видеоконференции или чата предполагает наличие строгого временного регламента таких мероприятий, что противоречит самой идее дистанционного обучения, а именно, использованию учебных ресурсов в любое удобное обучаемому время. Поэтому актуальной является задача замены преподавателя искусственным интеллектом для ответа на простые вопросы, отсылки к справочным и методическим материалам, проведения контрольных опросов и тестирования. В диалоге с искусственным интеллектом можно выполнять диагностику пробелов в знаниях и выделять учебные элементы для дополнительного изучения.

Цель применения компьютерных технологий, в принципе, не меняется — рутинные, формальные процессы выполняются компьютерными программами, оставляя на долю человека неформальные, сложные и творческие задачи. Современной тенденцией является постоянное увеличение доли процедур, выполняемых компьютерами. Искусственный интеллект начинают применять в процессах, традиционно выполняемых человеком. Примером таких процессов является компьютерная обработка обращений в различных call-центрах. В них применяют несколько уровней обработки обращения. На первом происходит регистрация и идентификация наиболее простых проблем. Сложные задачи передаются более квалифицированным специалистам на следующих уровнях. Анализ текстов с помощью технологий Text Mining может успешно применяться для классификации обращений, заменяя, таким образом, операторов первого уровня. Применение такой технологии для организации чата в дистанционном обучении студентов [2, 3] поможет решить несколько задач. Студенты смогут получать ответы на типовые вопросы в любое время. Такой инструмент поможет им преодолеть боязнь «глупых» вопросов. Преподаватели будут избавлены от ответов на типовые вопросы.

Для реализации таких чатов доступно множество инструментов. Технологически просто создать чат-бот в среде Telegram или ВКонтакте. Основную трудность представляет наполнение чат-бота: как организовать распознавание вопросов и генерацию ответов. Известные чат-системы ChartGPT, GigaChat или YandexGPT генерируют слишком общие, нестрогие ответы. Например, GigaChat дает следующее определение: «Производная функции — это функция, которая определяется через другую функцию. Другими словами, производная функции f(x) относительно x равна f'(x). Производная функции показывает, как изменяется значение функции при изменении ее аргумента на небольшое количество единиц». Очевидно, что преждевременно включать такие ответы в учебный процесс. Тем не менее, обработка текстов на естественном языке может применяться для организации диалогов с обучаемыми для решения некоторых задач.

Общение на естественном языке может применяться для распознавания семантики вопроса. Для выдачи ответа на запрос обучаемого в [3] предлагается следующий сценарий:

- ввод пользователем текстового вопроса по теме онлайн-курса на естественном языке;

- проведение лексической обработки (разбиение на конструкции и слова, а также их обработка — токенизация и лемматизация);
- перевод текстовых данных в математическую модель (векторизация с использованием модели word2vec);
- семантический анализ обработанного вопроса (классификация темы вопроса и поиск релевантных ответов).

Выдача пользователю ответов в виде топ-3 ответов наиболее релевантных заданному вопросу.

Используя эту методику, система выдает точный, заранее подготовленный ответ. Методическое обеспечение такой технологии потребует структуризации сведений в виде наборов тем, типовых вопросов и соответствующих готовых ответов. Кроме этого, по этим данным потребуется выполнить лексическую обработку, построение модели и настройку классификатора. Распознавание семантики вопроса всегда носит стохастический характер. Для верификации распознавания в формируемый ответ целесообразно включить типовую формулировку вопроса, чтобы пользователь мог сопоставить типовой вопрос со своей формулировкой.

Другим применением чат-ботов может быть тестирование. Тестирование по сути является чатом, в котором вопросы задает система (бот). Задания в тесте могут быть закрытого или открытого типов. К заданиям закрытого типа относятся задания пяти видов: оценка истинности утверждения, выбор одного правильного ответа, выбор нескольких правильных ответов, определение правильной последовательности вариантов, определение соответствия между вариантом ответа и категорией. Задания открытого типа требуют от тестируемых самостоятельно сформулировать ответ. В компьютерных системах тестирования этот ответ сравнивается с одним из правильных вариантов, определенных на этапе разработки задания.

Компьютерное тестирование часто критикуют за излишнюю «жесткость», так как результатом выполнения тестового задания является, чаще всего, оценка в дихотомической шкале «верно» или «неверно». В вопросах закрытого типа применяется выбор одной или нескольких альтернатив. Для того, чтобы выбор не был слишком очевиден, ошибочные альтернативы должны быть правдоподобны. Выбор таких альтернатив тестируемым часто свидетельствует не столько о недостатках в знаниях, сколько о его невнимательности. К другим недостаткам заданий закрытого типа относят достаточно большую вероятность угадывания, а также распространение правильных ответов среди тестируемых. В заданиях открытого типа ответ вводится в виде текста, что делает угадывание маловероятным. Тем не менее, задания этого типа также являются достаточно «жесткими», так как предполагают один или несколько правильных ответов.

В работе [4] рассматривается технология «мягкого тестирования». Для этого предлагается «Ввести в тестовые задания многозначные логические отношения, создать критериально-ориентированную технику оценки выполнения заданий, включающую не только полюсные («верно» и «не верно») варианты оценки, но и более широкий спектр, в том числе двумерную, матричную шкалу». Предложенная методика, очевидно, требует трудоемкой методической подготовки, а «критериально-ориентированная техника оценки» будет отражать пристрастия автора.

Применение искусственного интеллекта открывает новые возможности в области компьютерного тестирования. Предлагается реализовать «мягкое тестирование» на основе методов искусственного интеллекта по обработке текстов на естественном языке. «Мягкое тестирование» целесообразно применять для проверки знаний по сложным объектам и явлениям, для которых характерны различные определения. Рассмотрим в качестве примера определение базы данных. Из разных источников собраны следующие варианты:

- поименованная, целостная, единая система данных, организованная по определенным правилам, которые предусматривают общие принципы описания, хранения и обработки данных [5],
- совокупность данных, организованных в соответствии с некоторой концептуальной моделью данных, которая описывает характеристики этих данных и взаимоотношения между соответствующими им реалиями, и которая предназначена для информационного обеспечения одного или нескольких приложений [6],
- совокупность данных, хранимых в соответствии со схемой данных, манипулирование которыми выполняют в соответствии с правилами средств моделирования данных,
- совокупность данных, организованных в соответствии с концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, которая поддерживает одну или более областей применения.

«Мягкое тестирование» можно было бы применять для проверки ответа на соответствие любому из множества внесенных в тестовую систему определений. Очевидно, что проверки на совпадение текста были бы в этом случае малоэффективными. Предлагается для проверки таких ответов использовать методы и модели Text Mining [7].

Доступной основой такого применения является классификация текстов на основе частотных характеристик слов. Текст представляется в виде «мешка слов» — параметрами каждого текста будут частотные характеристики слов, входящих в текст. В качестве частотных характеристик можно

использовать абсолютные или относительные частоты (TF — Term Frequency), а также меткику TF-IDF — Term Frequency-Inverse Document Frequency, используемую для оценки важности слова в контексте документа, являющегося частью коллекции документов. Вес некоторого слова в этой метрике пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Для исключения использования в параметрических описаниях текстов разных форм одного и того же слова, слова заменяются леммами (словарными формами). Для уменьшения размерности параметрического описания коллекции текстов исключают слова, не несущие смысловой нагрузки (так называемый стоп-список). Параметрическое описание коллекции текстов используют для настройки классификаторов. Для тестирования в качестве классов текстов можно выбрать определяемые понятия, а в качестве текстов — определения понятий. Эти данные можно использовать для обучения классификатора. Настроенный классификатор применяется для определения класса ответа тестируемого. Результатом классификации ответа является оценка вероятности принадлежности ответа каждому классу. Оценкой задания можно считать вероятность принадлежности ответа определяемому понятию. Для точных определений эта вероятность будет близка к единице.

Для апробации предлагаемой технологии был использован корпус текстов, включающий определения различных понятий и объектов, применяемых в обработке информации. В качестве лемматизатора использовался Mystem. В среде Orange3 была выполнена загрузка корпуса определений и настройка классификаторов. В качестве классов текстов выступали определяемые объекты. Результатом классификации утверждения (ответа тестируемого) является оценка вероятности принадлежности утверждения каждому классу. Среди классификаторов наибольшую точность продемонстрировала модель логистической регрессии. В таблице 1 приведены результаты предсказания классов для некоторых вариантов ответов. Первые два ответа соответствуют разным определениям базы данных, использованным для обучения классификатора, им соответствует высокая вероятность распознавания. Третий ответ содержит совсем неточное определение, которое классификатор с небольшой вероятностью отнес к СУБД. Четвертое утверждение соответствует определению системы управления базами данных. Приведенный пример демонстрирует возможность применения предложенной технологии оценки тестовых заданий открытого типа.

Таблица 1. Предсказание класса по тексту ответа

№	Леммы определения	Предсказанный класс	Вероятность класса «База данных»	Вероятность класса «СУБД»
1	поименованный, целостный, единый система данные, организовывать по определенный правило, который предусматривать общий принцип описание, хранение и обработка данные	База данных	0,916	0,028
2	совокупность данный, организованный в соответствие с некоторый концептуальный модель данный, который описывать характеристика этот данные и взаимоотношение между соответствующий они реалия, и который предназначать для информационный обеспечение один или несколько приложение	База данных	0,958	0,014
3	набор информация, который храниться упорядоченный в электронный вид	Система управления базами данных	0,092	0,414
4	совокупность программный и лингвистический средство общий или специальный назначение, обеспечивать управление создание и использование база данные	Система управления базами данных	0,008	0,952

Итак, предложенная технология «мягкого тестирования» имеет очевидные ограничения — она применима для характеристик сложных объектов и явлений. Её преимуществом является значительное снижение трудоемкости подготовки тестовых заданий. Фактически нужно подготовить достаточно большой корпус определений и выполнить лемматизацию и обучение классификатора. Хорошие перспективы в плане распознавания имеет применение эмбеддинговых моделей [8], которые создают векторные описания текстов с учетом контекстов употребления слов. ■

1. Блинов В.И., Сергеев И.С. Модели смешанного обучения в профессиональном образовании: типология, педагогическая эффективность, условия реализации // Профессиональное образование и рынок труда. 2021. №1 (44). URL: https://cyberleninka.ru/article/n/modeli-smeshannogo-obucheniya-v-professionalnom-obrazovanii-tipologiya-pedagogicheskaya-effektivnost-usloviya-realizatsii (дата обращения: 11.11.2023).

2. Киуру К.В., Попова Е.Е., Маковецкая Ю.Г. Новые технологии дистанционного обучения в системе высшего и дополнительного профессионального образования // Проблемы современного педагогического образования. 2022. №75-3. URL: https://cyberleninka.ru/article/n/novye-tehnologii-distantsionnogo-obucheniya-v-sisteme-vysshego-i-dopolnitelnogo-professionalnogo-obrazovaniya (дата обращения: 02.11.2023).

3. Рожкин П.А., Нехаев И.Н., Маркин К.А. Конструирование системы интеллектуального поиска ответов на вопросы обучающихся на онлайн-курсе на основе word2vec // IJAS. 2018. №1. URL: https://cyberleninka.ru/article/n/konstruirovanie-sistemy-intellektualnogo-poiska-otvetov-na-voprosy-obuchayuschihsya-na-onlayn-kurse-na-osnove-word2vec (дата обращения: 11.11.2023).

4. Морев И. А. Образовательные информационные технологии. Часть 2. Педагогические измерения: Учебное пособие. – Владивосток: Изд-во Дальневост. ун-та, 2004. – 174 с.

5. Отраслевой стандарт «Информационные технологии в высшей школе. Термины и определения. ОСТ ВШ 01.002-95».- М., 1995 г., 24 с.

6. Марков А.С., Лисовский К.Ю. Базы данных. Введение в теорию и методологию: Учебник.- М.: Финансы и статистика, 2004.- 512с.:ил.

7. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

8. Спивак А.И., Лапшин С.В., Лебедев И.С. Классификация коротких сообщений с использованием векторизации на основе ELMo // Известия ТулГУ. Технические науки. 2019. №10. URL: https://cyberleninka.ru/article/n/klassifikatsiya-korotkih-soobscheniy-s-ispolzovaniem-vektorizatsii-na-osnove-elmo (дата обращения: 10.11.2023).

СПИСОК ЛИТЕРАТУРЫ:

Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

Блинов В.И., Сергеев И.С. Модели смешанного обучения в профессиональном образовании: типология, педагогическая эффективность, условия реализации // Профессиональное образование и рынок труда. 2021. №1 (44). URL: https://cyberleninka.ru/article/n/modeli-smeshannogo-obucheniya-v-professionalnom-obrazovanii-tipologiya-pedagogicheskaya-effektivnost-usloviya-realizatsii (дата обращения: 11.11.2023).

Киуру К.В., Попова Е.Е., Маковецкая Ю.Г. Новые технологии дистанционного обучения в системе высшего и дополнительного профессионального образования // Проблемы современного педагогического образования. 2022. №75-3. URL: https://cyberleninka.ru/article/n/novye-tehnologii-distantsionnogo-obucheniya-v-sisteme-vysshego-i-dopolnitelnogo-professionalnogo-obrazovaniya (дата обращения: 02.11.2023).

Марков А.С., Лисовский К.Ю. Базы данных. Введение в теорию и методологию: Учебник.- М.: Финансы и статистика, 2004.- 512с.:ил.

Морев И. А. Образовательные информационные технологии. Часть 2. Педагогические измерения: Учебное пособие. – Владивосток: Изд-во Дальневост. ун-та, 2004. – 174 с.

Отраслевой стандарт «Информационные технологии в высшей школе. Термины и определения. ОСТ ВШ 01.002-95».- М., 1995 г., 24 с.

Рожкин П.А., Нехаев И.Н., Маркин К.А. Конструирование системы интеллектуального поиска ответов на вопросы обучающихся на онлайн-курсе на основе word2vec // IJAS. 2018. №1. URL: https://cyberleninka.ru/article/n/konstruirovanie-sistemy-intellektualnogo-poiska-otvetov-na-voprosy-obuchayuschihsya-na-onlayn-kurse-na-osnove-word2vec (дата обращения: 11.11.2023).

Спивак А.И., Лапшин С.В., Лебедев И.С. Классификация коротких сообщений с использованием векторизации на основе ELMo // Известия ТулГУ. Технические науки. 2019. №10. URL: https://cyberleninka.ru/article/n/klassifikatsiya-korotkih-soobscheniy-s-ispolzovaniem-vektorizatsii-na-osnove-elmo (дата обращения: 10.11.2023).

Application of intellectual text processing for organizing chat bots in distance learning

This article discusses the possibilities of using natural language text processing technologies to organize a dialogue with students in the form of a chat bot. The use of question semantics recognition in educational chat bots is substantiated. It is proposed to use text classification methods to process answers to open-type test tasks, and to form an assessment of the answer based on an estimation of the probability of assigning the answer to the correct class. The technology for generating a set of tasks, training a classifier, and testing is described. The results of an experiment on training a classifier and processing responses are presented.

Keywords: distance learning, chat bot, text mining, open-type test, text classification, bag of words