

УДК 004.896

РАЗРАБОТКА ГИБРИДНОГО ДИАЛОГОВОГО АССИСТЕНТА С ПРИМЕНЕНИЕМ ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Александров Д. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – к.т.н., доцент ИДУ, старший научный сотрудник НЦКР

Бутаков Н.А.

(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе разработан гибридный диалоговый ассистент, состоящий из двух вопросно-ответных систем: система выдачи ответов из текстовой документации и диалоговой системы с применением технологии заполнения слотов. Такой тип гибридизации позволяет объединить преимущества обоих подходов, особенно в части использования гетерогенных полуструктурированных данных.

Введение. В последнее десятилетие наблюдается устойчивая тенденция к автоматизации рутинных процессов. В частности, это затронуло область обработки текстовых данных и извлечения смыслов. Интерфейсом взаимодействия человека и системы обработки информации может служить диалоговая система. Современные диалоговые системы можно классифицировать по способу применения следующим образом: вопросно-ответные системы: агент выдаёт чёткий ответ на вопрос пользователя, используя знания, полученные из различных источников, таких как коллекции текстовых документов, веб-документы, базы знаний; исполнители задач: агент выполняет задания пользователя начиная бронированием столика в ресторане и заканчивая планированием путешествия; собеседники: агент ведёт дружескую непринуждённую беседу с пользователем на общие темы.

Для имплементации текстовой вопросно-ответной системы требуется большое количество размеченных данных в виде пар вопрос-ответ. Однако в большинстве случаев таких структурированных датасетов нет в наличии. Для пользователей определённых типов продукции или услуг обычно предоставляют инструкции или мануалы, которые можно использовать в качестве слабоструктурированных данных для создания вопросно-ответной системы.

Основная часть. Предлагаемый гибридный подход в создании доменно-ориентированной вопросно-ответной системы заключается в объединении архитектуры модуля поиска текстовых документов с помощью задачи ранжирования и архитектуры заполнения предопределённых слотов для выполнения желания пользователя.

Улучшенный механизм поиска текстовых документов реализован с помощью сочетания методов поиска информации, основанных на метриках и на данных. С помощью скоринговых функций TF-IDF и BM-25 оценивалось сходство вопросов и ответов из тренировочной выборки в 200 пар вида вопрос - релевантный документ по типу каждый с каждым. Таким образом образовалось 40000 пар типа вопрос-документ с соответствующими скорями. Эти скоры использовались как признаки для обучения ранжирующей модели. С качестве алгоритмов оптимизации использовались байесовская оптимизация, генетический алгоритм и алгоритм имитации отжига. Эксперименты показали улучшение результатов до 9% по сравнению с эвристическими методами.

Диалоговый ассистент на основе заполнения слотов реализован в виде модуля выделяющего ключевые слова из запроса пользователя и осуществляющего поиск среди заранее предопределённых слотов. При этом, в случае несоответствия запроса пользователя ни одному из примеров датасета (недостаток или избыток выделенной информации) ассистент задаёт уточняющий вопрос пользователю.

Использование гибридной архитектуры стало возможным благодаря применению классификатора домена, к которому относится пользовательский запрос. В качестве данных

для обучения классификатора использовались вопросы из датасета вида вопрос-документ, а также индексы в конце документаций. Точность классификации равна 90.41% для домена А и 93.74% для домена В.

Выводы. Имплементация гибридного диалогового ассистента выполнена на языке python. Гибридная архитектура является масштабируемой, т. е. для включения в нее дополнительного модуля типа текстового поисковика или диалогового ассистента с заполнением слотов требуется небольшое количество данных и переобучение классификатора доменов. Использование такого типа диалоговых ассистентов возможно для поиска информации по текстовым документациям для консультирования пользователей по вопросам использования технических или бытовых приборов, а система с заполнением слотов поможет направлять пользователя на формулирование более корректных запросов с помощью уточняющих вопросов.