Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.78

Беленькая
Анна Ильинична

МЕТОДЫ СБОРА И АНАЛИЗА ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ДИАЛОГОВЫХ СИСТЕМ

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 02 «Системный анализ, управление и обработка информации (по отраслям)»

Научный руководитель

Захарьев Вадим Анатольевич,
кандидат технических наук,
доцент

Минск 2021

ВВЕДЕНИЕ

Диалоговые системы - неотъемлемая составляющая интеллектуальных компьютерных систем, служащих для переработки информации. Именно с их помощью осуществляются практически все процессы внутри системы. Основным требованием диалоговых систем является обеспечение более удобной и естественной формы взаимодействия интеллектуальных систем с пользователями. Задача построения виртуального собеседника является центральной в области обработки естественного языка. В последнее время диалоговые системы снова набирают популярность. Многие крупные IT-системы создают диалоговые версии, такие как Siri (Apple), Cortana (Microsoft), Alexa (Amazon) и Алиса (Яндекс).

Целью диссертационной работы является исследование методов сбора и анализа данных для обучения диалоговых систем. Также будут рассмотрены различные подходы к построению диалоговых систем, применяемые архитектуры и открытые наборы данных для обучения таких систем.

Объект исследования: диалоговые системы.

Предмет: процесс сбора и анализа обучающих данных (текстовых корпусов) для обучения диалоговых систем.

Задачи:

1. Рассмотреть виды диалоговых систем, их классификации и основные характеристики;
2. Проанализировать существующие методы построения диалоговых систем;
3. Рассмотреть виды диалоговых корпусов, дать их классификацию и проанализировать существующие наборы данных для обучения диалоговых систем, находящиеся в открытом доступе;
4. Провести исследование архитектур, использующихся для построения диалоговых систем.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования

Тема диалоговых систем является актуальной, так как с каждым годом в мире возрастает потребность в автоматизированных интеллектуальных диалоговых системах, которые смогли бы заменить человека при выполнении рутинных задач. В настоящее время диалоговые системы начинают активно применяться для оказания информационно-аналитических услуг во многих сферах нашей жизни, например чат-боты для изучения иностранных языков, различные системы для оказания медицинских услуг, помощники в интернет магазинах, интеллектуальные помощники в банках и так далее.

Цель исследования

Целью диссертационной работы является исследование методов сбора и анализа данных для обучения диалоговых систем. Также будут рассмотрены различные подходы к построению диалоговых систем, применяемые архитектуры и открытые наборы данных для обучения таких систем.

Задачи исследования

1. Рассмотреть виды диалоговых систем, их классификации и основные характеристики;
2. Проанализировать существующие методы построения диалоговых систем;
3. Рассмотреть виды диалоговых корпусов, дать их классификацию и проанализировать существующие наборы данных для обучения диалоговых систем, находящиеся в открытом доступе;
4. Провести исследование архитектур, использующихся для построения диалоговых систем.

Новизна полученных результатов

Научная новизна заключается во всестороннем анализе и систематизации информации по большому количеству источников данных для обучения диалоговых систем (лингвистических корпусов), разработке методических рекомендаций по применению данных корпусов, а также оценки их использования для обучения диалоговых систем на основе современных архитектур.

Личный вклад соискателя.

Соискателем выполнены все изложенные в работе разработки и исследования. Постановка задач и обсуждение результатов проводились совместно с научным руководителем и сотрудниками факультета информационных технологий и управления Белорусского государственного университета информатики и радиоэлектроники. Обработка, интерпретация данных, а также выводы сделаны автором самостоятельно.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях:

- 56-я научная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» (Минск БГУИР 2020);

СОДЕРЖАНИЕ РАБОТЫ

В первом разделе диссертации рассматриваются различные виды диалоговых систем, их характеристики. Показана стандартная структура диалоговой системы и описаны ее основные компоненты, описаны основные требования к диалоговым системам и практические аспекты их реализации.

Во втором разделе рассматриваются основные подходы к реализации диалоговых систем. Далее рассматриваются методы генерации ответов: на основе правил, на основе поисковой модели, на основе порождающей модели. Также описаны различные виды диалоговых корпусов, использующихся для обучения диалоговых систем и дана их классификация. Далее рассматриваются существующие наборы данных для обучения диалоговых систем, находящиеся в открытом доступе.

В третьем разделе рассматриваются основные особенности нейронных сетей, базовая рекуррентная нейронная сеть, а также особенности различных архитектур, использующихся для построения диалоговых систем: Sequence-to-sequence, Long-Short Term Memory (LSTM), Hierarchical Neural Network Model (HRED) и Hierarchical Latent Variable Encoder-Decoder Model (VHRED).

В четвертом разделе рассматриваются экспериментальные результаты. В качестве обучающих корпусов выбраны «Twitter Corpus» и «Ubuntu Dialogue Corpus». Для улучшения качества результата был использован метод «word2vec». В ходе эксперимента было выявлено, что архитектура Hierarchical Latent Variable Encoder-Decoder Model является более удачной по сравнению с моделями HRED и LSTM

ЗАКЛЮЧЕНИЕ

В работе были рассмотрены классификация диалоговых систем, их основные характеристики, структура стандартной диалоговый системы и ее основные компоненты. Диалоговые системы достаточно актуальны в наше время, поскольку их можно встретить во многих сферах нашей жизни, например помощники в интернет магазинах, чат-боты для изучения иностранных языков, диалоговые системы для оказания медицинских услуг и т.д.

Далее были рассмотрены актуальные подходы к построению диалоговых систем. Из них можно выделить генеративный подход, который является наиболее актуальным в настоящее время. Отличие систем, основанных на порождающей модели, заключается в том, что они не используют никакого предопределенного репозитория, а генерируют ответы на сообщения пользователя сами по себе. Такие системы могут быть обучены при помощи языковых корпусов различного вида. Далее были рассмотрены виды диалоговых корпусов и их классификация, а также открытые наборы данных, созданные из устного или письменного английского языка и находящиеся в открытом доступе.

Диалоговые системы на основе порождающих моделей строятся с использованием аппарата искусственных нейронных сетей. Были рассмотрены основные особенности нейронных сетей, базовая рекуррентная нейронная сеть, а также особенности архитектур sequence-to-sequence, LSTM, HRED и VHRED. Архитектура sequence-to-sequence наиболее часто используется для создания диалоговых систем. Чтобы избежать проблемы долговременной зависимости используется архитектура Long-Short Term Memory. Архитектура HRED (Hierarchical Neural Network Model) является расширением классической RNN с использованием LSTM. Архитектура VHRED (Hierarchical Latent Variable Encoder-Decoder Model) отличается от HRED дополнительными скрытыми переменными.

Далее было произведено проектирование модифицированных диалоговых систем. За основу была взята библиотека Tensorflow и Keras, позволяющая реализовывать глубокие нейронные сети. В качестве обучающих корпусов были выбраны «Twitter Corpus» и «Ubuntu Dialogue Corpus». Для улучшения качества результата был использован метод «word2vec», который на основе входных данных формирует словарь, состоящий из векторных представлений слов, где каждое слово имеет ряд признаков, описывающих его.

По результатам эксперимента можно сделать вывод, что архитектура VHRED является более удачной по сравнению с моделями HRED и LSTM.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Беленькая, А. И. Методы построения информационных диалоговых систем/ А. И. Беленькая // Материалы 56-ой научной конференции аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» БГУИР, Минск, 21-24 апреля 2020 г./ – с. 32.