УДК 00 – *004.89*

Царева М.В.
студент
Поволжский государственный университет телекоммуникаций и информатики

ПРОБЛЕМА ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА В ЧАТ-БОТАХ

Аннотация: В статье рассматриваются основные проблемы обработки естественного языка в современных диалоговых системах. В ходе работы сформулированы проблемы, которые ограничивают возможности чат-ботов распознавать печатную и устную речь.

Ключевые слова: Естественный язык, обработка естественного языка, чат-боты, диалоговые системы.

Tsareva M.V.
student
Volga State University of Telecommunications and Informatics

THE PROBLEM OF NATURAL LANGUAGE PROCESSING IN CHATBOTS

Abstract: The article deals with the main problems of natural language processing in modern dialog systems. In the course of the work, problems were formulated that limit the ability of chat bots to recognize printed and oral speech.

Key words: Natural language, natural language processing, chatbots, conversational systems.

Введение

В современном цифровом обществе одним из главных инструментов социального взаимодействия между людьми становятся смартфоны. И с ростом популярности мобильных устройств в жизни современного человека всё более значимую роль приобретают месседжеры с чат-ботами, которые помогают быстро найти информацию на сформулированный вопрос

пользователя. Чат-боты работают на основе набора правил и сценариев поведения, однако естественный язык нечеткий и неоднозначный, одна мысль может иметь много способов изложения, поэтому коммерческий успех диалоговых систем зависит от решения задач языкового процессинга.

Обработка естественного языка (Natural language processing) – область исследований, находящаяся на пересечении компьютерных наук, искусственного интеллекта и лингвистики. Предметом её исследований являются методы компьютерного анализа и синтеза текстов на естественном языке, т.е. вопросы, относящиеся к обработке и пониманию естественного языка для перевода текста и конструирования грамотных ответов на вопросы.

Основной проблемой обработки естественного языка является языковая неоднозначность.

Целью данной статьи является определение основных проблем, которые ограничивают возможности диалоговых систем(чат-ботов) распознавать печатную и устную речь.

Языковой барьер в чат-ботах.

Естественный язык - это то, что люди используют для общения друг с другом. Разговорные агенты(чат-боты) обрабатывают естественный язык и генерируют ответ на естественном языке на вводимые пользователем данные, тем самым имитируя человеческое общение.

Система обработки естественного языка работает по следующему алгоритму:

- Пользователь задает данные (в письменном или устном виде);
- В случае с устной речью программа записывает звук и конвертирует его в текст.
- NLP-система анализирует текст, разбивая его на составляющие и пытаясь понять запрос.

- На основе полученного результата программа определяет, какие действия нужно выполнить.

Но человеческая речь чрезвычайно сложный механизм, полный нюансов. Сленг, интонация, юмор, сарказм, синтаксис и орфографические ошибки – это особенности человечной речи, которые затрудняют обработку информации искусственным интеллектом. Обучить разговорного агента, который может работать каждый сценарий, учитывая все тонкости естественного языка на данный момент практически невозможно. Особенно проблема актуальна при распознавании устной разговорной речи.

С точки зрения специфики естественного языка, анализируемого системой, диалог человека и компьютера является особенным типом дискурса, в рамках которого структура речи и текста отличается от множества форм текста, которые обычно изучает компьютерная лингвистика (новости, отзывы, научные статьи и т.д.).

Выделим основные особенности человеческой речи, которые значительно осложняют обработку запроса в чат-ботах:

- Опечатки и ошибки в произношении слов.

Приложение при распознавании не всегда может получить корректные данные от пользователя из-за наличия орфографических, синтаксических ошибок или неправильного построения структуры диалога. В зависимости от типа диалоговой системы, предметной области и многих других особенностей, ошибки модуля распознавания речи могут достигать величины, которые в значительной степени могут влиять на корректность ответа системы.

- Особенности устной речи: спонтанность, сокращения, слэнг.

В подавляющем большинстве случае устная разговорная речь синтаксически более сложная и «шумная» в отличии от письменной, по

причине того, что она более быстрая, отрывистая, с большим количеством пропусков. В устной речи часто встречаются паузы, заполненные каким-либо выражением, краткие неформальные реплики, разговорные и сленговые слова. Пользователь может спонтанно дополнять только что сказанную фразу или говорить фрагментам по несколько слов, повторить одно и то же или внезапно начать заново, думая, что система не расслышала. Модуль обработки естественного языка должен уметь справляться со спецификой устной речи, чтобы минимизировать впоследствии ошибки понимания пользователя.

- Неоднозначность реплик.

Наличие лингвистической неоднозначности (лексической, синтаксической, контекстной) – ещё одна проблема для корректного распознавания человеческой речи. Ею пропитаны задачи обработки естественного языка, ежедневно возникающие в различных прикладных областях, в том числе в диалоговых системах.

- Синонимия, избыточность формулировок.

Избыточность информации в репликах пользователя и их синонимичность является проблемой, которую необходимо разрешать, т.к. идентичные по смыслу команды, сформулированные различными способами, должны приводить систему к одинаковому результату.

- Зависимость фраз от контекста.

Диалог на естественном языке состоит из реплик, связанных между собой текущим контекстом. Одинаковые реплики пользователя в различные моменты времени могут иметь различный смысл и должны приводить систему к различным ответам.

- Сложность выбора оптимального семантического представления.

Значение, которое пользователь вкладывает в свою речь, может быть очень сложным и включать в себя множество нюансов. Для того чтобы понимать эти нюансы и поддерживать с пользователем эффективный диалог, при создании диалоговой системы важно найти универсальную форму семантического представления поступающей от пользователя информации. Семантическое представление текста внутри современных диалоговых систем может отличаться друг от друга и зависит от специфики задач, для решения которых она создавалась. Кроме того, разработчики диалоговых систем создают собственные уникальные семантические представления для поддержания более свободного диалога с пользователем или диалога в специфической предметной области.

- Кросс-языковая морфология и малоресурсные языки

Системам обработки естественного языка намного проще работать с языками, имеющими четкую структуру построения предложения, такими как: английский, немецкий, французский и т.д. Несколько сложнее работать с языками славянской группы, где, простая перестановка слов может поменять, а может и не поменять смысл. К тому же во многих языках нет такого разнообразия склонения слов, как в русском или немецком. И совсем тяжело внедрить распознавание естественного языка для малораспространенных и исчезающих языков, местных диалектов и языков, не имеющих письменности.

Вывод

В статье были рассмотрены основные проблемы обработки естественного языка в современных диалоговых системах. Основной проблемой обработки естественного языка является языковая неоднозначность. Были сформулированные основные проблемы, с которыми сталкиваются чат-бот ассистенты при попытке распознать поданную пользователем информацию:

- Опечатки
- спонтанность, сокращения, слэнг
- Неоднозначность реплик
- Синонимия, избыточность формулировок
- Зависимость фраз от контекста
- Сложность выбора оптимального семантического представления.

Использованные источники:

1. Алымов, А. С. Детектирование бот-программ, имитирующих поведение людей в социальной сети «Вконтакте» / Алымов, А. С., Баранюк, В.В., Смирнова, О.С. // International Journal of Open Information Technologies. – 2016. – Том 4, № 8. – С. 55 – 60.

2. Жеребцова Ю.А. Проблемы обработки естественного языка в диалоговых системах (91 с.) [Электронный ресурс]//Электронный журнал Наука и технологии (дата публикации: 01.10.2019).- URL: http://www.nait.ru/(дата обращения: 15.11.2022)

3. Заханевич, Д.Ю. Исследовательские подходы и методы применения искусственных интеллект и машинное обучение в социально-экономических процессах // Вестник ОмГУ. Серии: экономика. – 2020. – № 2.-с. 66–79.

4. Лебедев, Б. Д. Рекомендательные системы с применением машинного обучения для интернет-ресурсов // Modern Science. – 2019. – №. 5(3). – С. 265–268.

5. Максутов Р. Правильный NLP: как работают и что умеют системы обработки естественного языка (дата публикации: 23.01.2020).- URL: https://tproger.ru/articles/naturallanguage-processing/ (дата обращения: 15.11.2022)

6. Обломова О. Natural language processing (NLP): не то НЛП, про которое вы подумали (15 с.) [Электронный ресурс]//Электронный журнал 4brain (дата публикации: 01.06.2022).- URL: https://4brain.ru/blog/natural-language-processing/ (дата обращения: 15.11.2022)

7. Песцова И.И. Основные проблемы при работе с естественными языками (9 с.) [Электронный ресурс]//Справочник (дата публикации: 29.10.2020).- URL: https://spravochnick.ru/informatika (дата обращения: 15.11.2022)