Министерство образования Республики Беларусь

Учреждение образования

“Белорусский государственный университет информатики и радиоэлектроники”

Кафедра информационных интеллектуальных технологий

Лабораторная работа 3

Вариант 7

“Синтаксический анализ текстов естественного языка”

Выполнил Мулярчик Д.С.

гр.121701 Лемантович Д.С

Проверил Крапивин Ю. Б

Минск 2024

Задание: Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического синтаксического анализа ТЕЯ.

Используемые инструменты: Python с PyQt и Natural Language Toolkit.

Структуры хранения: TXT-файлы

Структурно-функциональная схема

Структурно-функциональная схема приложения представлена черным ящиком:

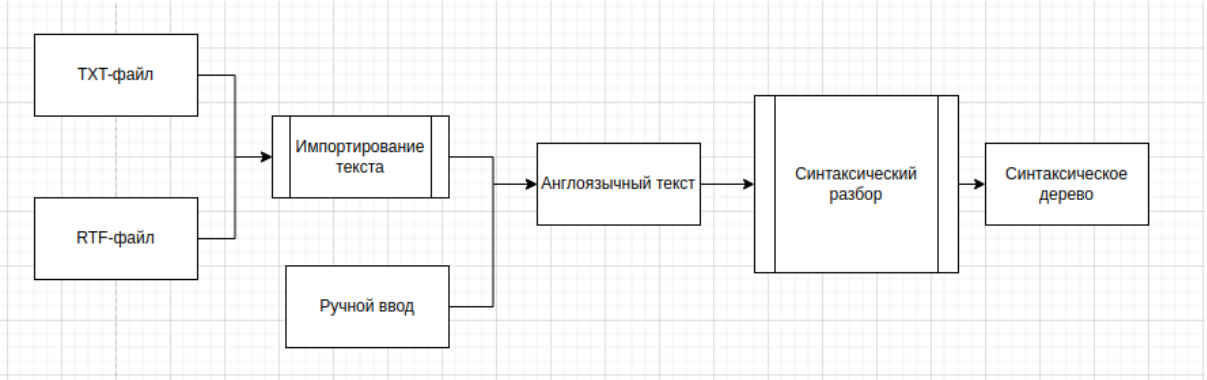

Рис.1 - Схема

Структуры хранения

После ввода текста, он разбивается на предложения. Предложения хранятся в виде списка. Синтаксический анализ выводит в виде дерева.

Алгоритм обработки

1. Вход: текст на естественном языке.
2. Вызов функции **main_analysis(text)**, где text - текст на естественном языке.
3. Разбиение текста на слова по пробелам.
4. Помещаем все слова, которые не являются знаками препинания в список tagged.
5. Передаем список tagged в функцию **parse()** класса **RegexpParser.**
6. Открываем окно с деревом.
7. Выход: синтаксическое дерево.

Алгоритм фильтрации и поиска

1. Вход: данные - словарь, который получен после обработки текста, запрос - в случае фильтрации по слову или части речи - строка.
2. Фильтрация предложений по слову, результат список предложений.
3. Возвращаем пользователю
4. Выход: список предложений.

Пример работы приложения

1. При запуске приложения нас встречает интерфейс:

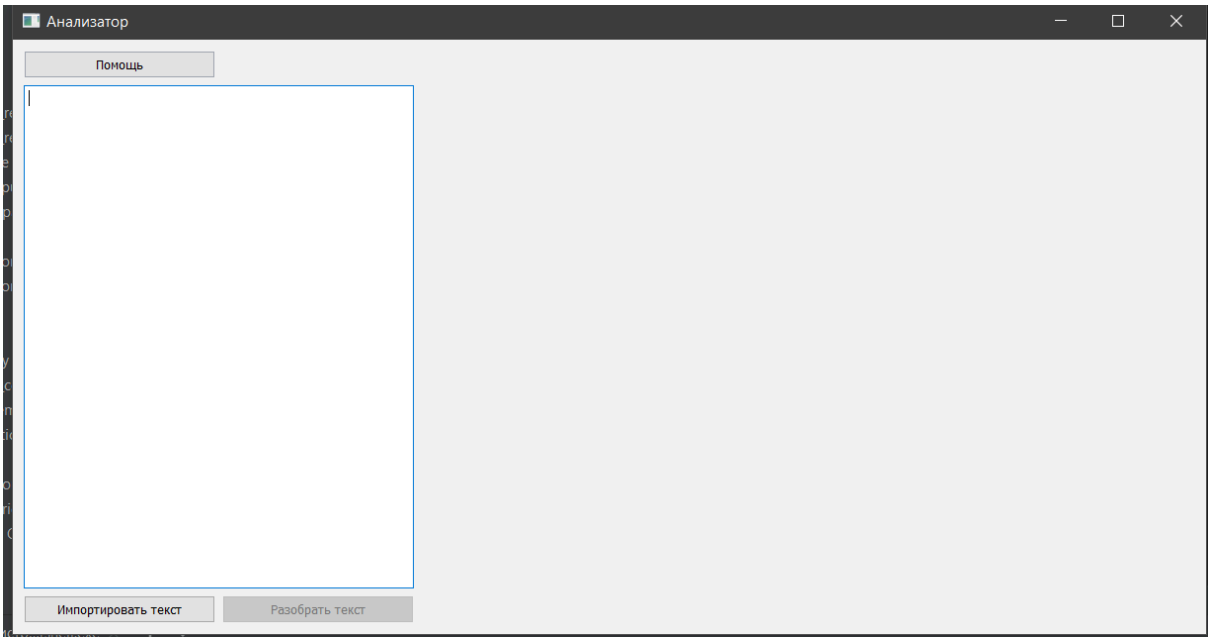

Рис.2 - Интерфейс

2. В него можно написать текст или импортировать текст из TXT- или RTF-файла, нажав на кнопку «Импортировать файл». Пример текста:

   Have you ever thought about what your future life is going to be like? What are you going to do when you finish school? It is never too early or late to start thinking about your future career. Maybe you enjoy some of the subjects at school more than others. If you do, this is a good sign, because they will guide you to your future profession.

Рис.3 - Пример текста

3. Пример разбора данного текста после нажатия кнопки «Разобрать текст»:

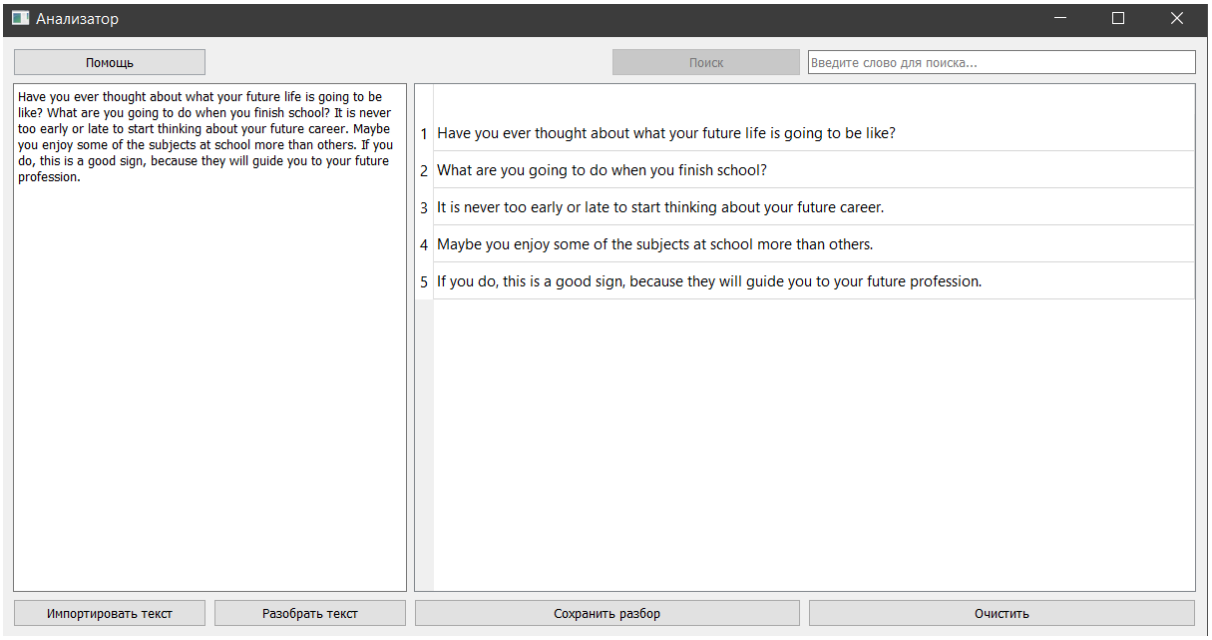

Рис.4 - Пример разбора

4. При двойном нажатии на ячейку таблицы, выводится дерево с результатом разбора предложения.

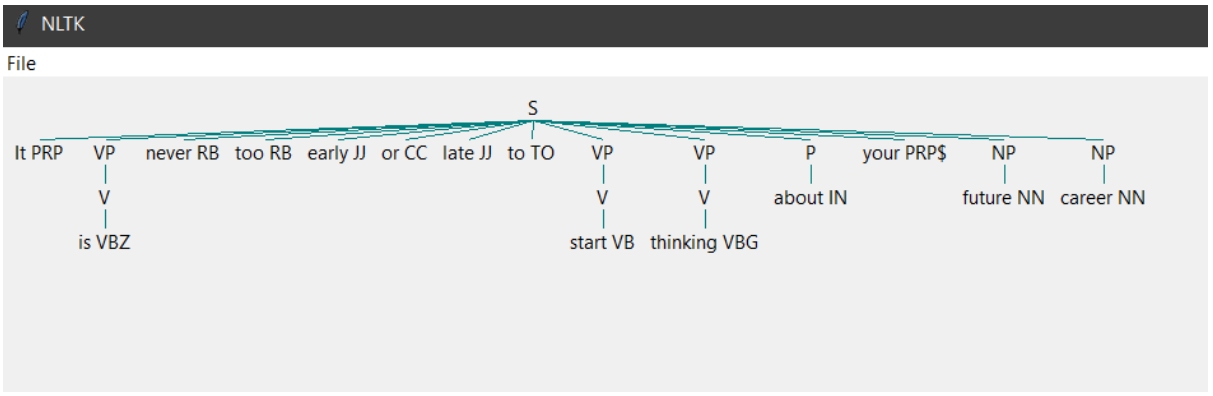

Рис. 6 - Пример дерева

Оценка быстродействия

При разработки программы был разработан модуль, обработки и изменения встроенного корпуса библиотеки nltk.

Запуск:

```
[03.10.2024 15:05:51 | INFO]: 0.04900252819061280
```

Вывод: В результате выполнения лабораторной работы была разработана программа для синтаксического анализа текста на естественном языке. В будущем эта программа может применяться для разбора крупных текстов как в целях машинного обучения, так и в целях использования в учреждения образования и у лингвистов.

Ответы на вопросы:

1. Естественно-языковой интерфейс: определение, характеристики, примеры.

Естественно-языковой интерфейс (Natural Language Interface) - это система, которая позволяет пользователю взаимодействовать с компьютерной программой или устройством, используя обычный человеческий язык. Вместо использования специфических команд или интерфейсов, пользователь может задавать вопросы, выражать свои намерения или давать инструкции на естественном языке, а система должна понимать и обрабатывать эти запросы.

Характеристики естественно-языкового интерфейса:

Распознавание и понимание языка: Естественно-языковой интерфейс должен иметь возможность распознавать и понимать различные языковые конструкции, включая фразы, предложения, вопросы и инструкции.

Семантический анализ: Система должна иметь способность анализировать смысл и интенцию, выраженные в высказываниях пользователей. Это может включать выделение ключевых слов, определение сущностей и отношений между ними.

Генерация ответов: После понимания запроса пользователя система должна быть способна генерировать ответы, которые соответствуют заданному вопросу или инструкции. Ответы могут быть текстовыми или могут включать выполнение определенных действий.

Диалоговая поддержка: Естественно-языковой интерфейс может поддерживать диалоговую форму взаимодействия, позволяя пользователю задавать серию вопросов или уточнять свои запросы в рамках одной сессии.

Примеры естественно-языковых интерфейсов включают:

Голосовые помощники: Такие системы, как Siri от Apple, Google Assistant и Amazon Alexa, позволяют пользователям задавать вопросы и давать команды на естественном языке для выполнения различных задач, таких

как поиск информации, установка напоминаний, управление умным домом и другие.

Чат-боты: Чат-боты могут использоваться для общения с клиентами в онлайн-магазинах, банках или других организациях. Они позволяют пользователям задавать вопросы и получать информацию в текстовой форме, а также выполнять определенные действия, такие как заказ товаров или бронирование билетов.

Автоматизированные системы обработки запросов: Некоторые компании используют естественно-языковые интерфейсы для обработки запросов клиентов. Например, системы поддержки клиентов могут анализировать электронные письма или сообщения в социальных сетях и генерировать автоматические ответы на основе понимания содержания сообщений.

2. Задачи создания естественно-языкового интерфейса.

Распознавание и понимание языка: Одна из основных задач - разработка алгоритмов и моделей, способных распознавать и понимать различные языковые конструкции. Это может включать распознавание речи, синтаксический анализ предложений, определение смысла и интенции высказываний.

Семантический анализ: Для более глубокого понимания запросов пользователей, система должна иметь возможность анализировать семантику предложений и текстов. Это может включать выделение ключевых слов, определение сущностей, распознавание отношений между ними и выявление контекста.

Генерация ответов: Естественно-языковой интерфейс должен быть способен генерировать понятные и информативные ответы на запросы пользователей. Это может включать синтез речи или генерацию текстовых ответов, которые соответствуют заданному запросу.

Диалоговая поддержка: Создание интерфейса, который может поддерживать диалоговую форму взаимодействия, является одной из задач. Это может включать умение системы запоминать предыдущие сообщения и уточнять запросы, а также поддерживать последовательность вопросов и ответов в рамках одной сессии.

Управление ошибками и неоднозначностями: Естественно-языковой интерфейс должен быть способен обрабатывать ошибки, неоднозначности и неточности в запросах пользователей. Это может включать обнаружение и исправление ошибок в синтаксисе или семантике запросов, а также предоставление уточняющих вопросов для уточнения намерений пользователя.

Адаптация к контексту: Система должна быть способна учитывать контекст взаимодействия и адаптироваться к предыдущим запросам и ответам. Это может включать сохранение состояния диалога, чтобы обеспечить последовательность и связность взаимодействия.

Оценка качества и обратная связь: Для создания естественно-языкового интерфейса также важно проводить оценку качества системы и получать обратную связь от пользователей. Это позволяет оптимизировать работу интерфейса, исправлять ошибки и улучшать пользовательский опыт.

3. Отличия естественно-языкового интерфейса от других видов интерфейса.

Взаимодействие на естественном языке: Основное отличие естественно-языкового интерфейса заключается в возможности взаимодействия с системой на естественном языке, таком как русский, английский и др. Вместо использования специализированных команд или нажатия кнопок, пользователь может задавать вопросы, выражать намерения или давать инструкции на естественном языке.

Более гибкое и интуитивное взаимодействие: Естественно-языковой интерфейс предоставляет более гибкое и интуитивное взаимодействие с системой. Пользователь может выразить свои запросы и инструкции в свободной форме, не ограничиваясь заранее определенными командами или шаблонами. Это делает взаимодействие более естественным и удобным для пользователей.

Понимание контекста: Естественно-языковой интерфейс способен учитывать контекст взаимодействия. Он может использовать предыдущие вопросы и ответы для более точного понимания запросов пользователя. Например, если пользователь задает вопрос "Какая погода сегодня?" и затем задает вопрос "И вчера было тепло?", система может использовать

информацию о текущей дате и предыдущем запросе, чтобы понять, что пользователь интересуется погодой вчерашнего дня.

Более широкий спектр задач: Естественно-языковой интерфейс может быть применен для решения широкого спектра задач. Он может использоваться для поиска информации, выполнения операций в умном доме, заказа товаров, бронирования билетов, общения с чат-ботами и многого другого. В то время как графический пользовательский интерфейс (GUI) обычно предназначен для конкретных задач и операций, естественно-языковой интерфейс более гибок и универсален.

Необходимость обработки естественного языка: Естественно-языковой интерфейс требует разработки специальных алгоритмов и моделей для распознавания и понимания естественного языка. Это включает в себя задачи, такие как распознавание речи, синтаксический и семантический анализ, классификация и генерация текста. В отличие от других видов интерфейсов, где взаимодействие основано на нажатии кнопок или выборе опций, естественно-языковой интерфейс требует более сложной обработки текстовой информации.

4. Компьютерная лингвистика: определение, задачи, направления.

Компьютерная лингвистика - это область исследования, которая сочетает методы компьютерных наук и лингвистики с целью разработки компьютерных систем, способных обрабатывать и анализировать естественный язык.

Задачи компьютерной лингвистики включают:

Распознавание и понимание речи: Разработка алгоритмов и моделей для распознавания и транскрипции речи. Это включает в себя задачи, такие как автоматическое распознавание речи (ASR), определение фонем, сегментация аудио и т.д.

Морфологический анализ: Определение базовых форм слова и его грамматических характеристик, таких как число, род, падеж и т.д. Морфологический анализ помогает различать разные словоформы и строить корректные синтаксические структуры.

Синтаксический анализ: Анализ структуры предложений и определение связей между словами в предложении. Это включает в себя парсинг предложений, построение деревьев зависимостей и определение синтаксических ролей слов.

Семантический анализ: Определение значения слов и фраз, а также выявление семантических отношений между ними. Это может включать задачи, такие как определение смысла слов в контексте, распознавание именованных сущностей и извлечение информации.

Генерация текста: Создание текстов на естественном языке на основе определенных правил или шаблонов. Это может включать генерацию текстовых ответов, описания данных, автоматическое создание текстовых статей и т.д.

11. Моделирование диалога.

Моделирование диалога - это область исследования и разработки компьютерных моделей и систем, которые способны смоделировать и воспроизводить естественные диалоги между человеком и компьютером или между компьютерами.

Задачи моделирования диалога включают:

Разработка диалоговых агентов: Создание компьютерных программ или роботов, которые могут участвовать в диалоге с пользователем. Это может быть реализовано с помощью чат-ботов, виртуальных ассистентов или систем автоматического ответа на вопросы.

Понимание и генерация реплик: Разработка моделей и алгоритмов, которые способны понимать реплики пользователя и генерировать соответствующие ответы. Это включает в себя задачи распознавания и классификации намерений пользователя, анализа тональности, извлечения информации из реплик и генерации естественно звучащих ответов.

Моделирование контекста: Учет контекста в диалоге для более точного понимания и генерации реплик. Это может включать использование предыдущих реплик пользователя, информации о предметной области или контекста выполнения задачи. Моделирование контекста помогает создать более когерентный и информативный диалог.

Управление диалогом: Разработка алгоритмов и стратегий для эффективного управления диалогом. Это включает принятие решений о том, когда задавать вопросы, как задавать уточняющие вопросы, как поддерживать вовлеченность пользователя и как завершать диалог.

Оценка и оптимизация диалоговых моделей: Разработка метрик и методов для оценки качества диалоговых моделей. Это может включать сравнение с моделями, основанными на правилах или с другими моделями машинного обучения. Также проводится оптимизация моделей для достижения лучшей производительности в диалоге.

Основные подходы к моделированию диалога включают правила и шаблоны, статистические модели, машинное обучение и глубокое обучение. Развитие моделей диалога исследуется в академическом сообществе и применяется в различных индустриях, включая клиентскую поддержку, образование, медицину, развлечения и другие сферы.

1.1.1 Синтаксический анализ текста.

Задачей синтаксического анализа текста ЕЯ является распознавание в каждом его предложении синтаксических отношений и представление их, как правило, в виде функционального или синтаксического дерева, в котором словам предложения указывается их грамматическая функция и определяется тип синтаксической связи между ними. При этом может иметь место синтаксическая многозначность, т. е. анализируемому предложению может соответствовать несколько вариантов синтаксического дерева.

Задача синтаксического анализа включает в себя:

Разделение текста на предложения: Входной текст может содержать одно или несколько предложений. Синтаксический анализатор должен способен разделить текст на отдельные предложения для более детального анализа.

Токенизация: Текст разбивается на отдельные слова или токены. Токенизация является важным шагом для создания последовательности элементов, с которыми синтаксический анализатор будет работать.

Разбор предложений: Самая важная часть синтаксического анализа - это разбор предложений, то есть определение синтаксической структуры предложения и отношений между его компонентами. Это может включать определение субъекта, объекта, глагола и других частей речи, а также определение синтаксических связей, таких как зависимости или союзы.

Построение дерева разбора: Результатом синтаксического анализа является дерево разбора или синтаксическое дерево, которое представляет синтаксическую структуру предложения. В дереве разбора каждый узел представляет слово или фразу, а дуги отражают синтаксические отношения между ними.