**Конспект всех лекций по Естественному Языку:**

**Что такое Естественный Язык**

Естественный язык является знаковой системой передачи и хранения информации, а также средством создания и приема текста. Основным элементом языковой системы является знак, который передает знания и может быть устным или письменным. Различные уровни языка, такие как звуки, слова, фразы, выражения и предложения, выполняют различные функции и образуют структуру языка.

Проблема автоматической обработки текстовой информации на естественных языках заключается в выделении значений, тональностей и прагматических отношений между ними. Основной задачей такого текстового анализа является выявление намерений, тональностей и определение связей между ними. Другими задачами являются разработка машинного перевода и системы поиска информации.

**В чём задача Компьютерной лингвистики**

В компьютерной лингвистике основные направления включают гипергенеративные технологии представления текста, компьютерное моделирование процесса понимания содержания документов, автоматизированные системы поиска и анализа данных и автоматическое реферирование текстов.

Современные разработки включают в себя моделирование процесса понимания смысла текста, разработку морфологических анализаторов, лингвистических баз данных, информационные поисковые системы и системы интеллектуальной обработки текстов.

Системы распознавания речи призваны упростить и автоматизировать процессы общения, поиска информации и создания текстов. Кроме того, они могут быть использованы в таких областях, как телефония и создание систем для инвалидов.

**Алгоритмическая обработка речи**

Алгоритмическая обработка речи включает в себя распознавание речи, понимание речи и синтез речи. Уровни анализа речи включают фонетический, фонологический, лексический и семантический уровни. Также проводится тестирование систем управления голосом. Лекция о автоматическом синтезе речи. Этот процесс заключает в себе генерацию человеческого голоса из текста, при этом обеспечивается плавное звучание, своя интонация, естественный голос и тембр, а также правильное расставление ударений. Для достижения этого используются различные методы и подходы.

Одним из них является директный подход, который применяется от середины одной фонемы до середины другой. Вторым подходом является аллофонный подход, где фонема окружается выбранными визуальными элементами. Однако оба подхода имеют некоторые недостатки, такие как нестабильный тембр и слишком механическое звучание. Также возникают проблемы с произношением некоторых слов, включая искажения или выпадение звуков из слова.

**Подходы к синтезу речи**

Для выбора подхода к синтезу речи необходимо учитывать задачу, тип голосового аппарата и структуру языка. Технологии современных компьютеров позволяют успешно моделировать и создавать устройства, основанные на автоматическом синтезе речи.

**Обработка письменного текста**

Автоматическая обработка письменного текста, включая автоиндексирование и морфологический анализ структур словоформ. Решаемые задачи включают выделение словоформ из текста, распознавание слов и их нормализацию. Для этого используются автоматические словари и информационно-поисковые тезаурусы, которые содержат информацию о отношениях между словами.

Синтаксис — это анализ и описание структуры предложений. В рамках автоматического реферирования это составление краткого содержания документов. Квазиреферирование является одним из направлений этого процесса.

**Методы обработки текста**

Статические позиционные методы генерации краткого содержания не годятся для позиционных метрик в квазиреферировании. Негод необходимо учесть связь с левым и правым контекстом, выделение тривиальных фрагментов. Квазиреферирование имеет преимущество в простоте. Метод синтаксического дерева представляет собой технику понимания естественного языка. Существуют общедоступные программы для квазиреферирования, например, в составе сервисов Yandex.

Шаблонные технологии, относительно простые и надежные, находят широкое промышленное применение. Главная особенность этих технологий заключается в том, что содержимое исходного текста представлено в виде фрагментов действия. Метод фокусировки внимания полезен для практики реферирования.

Нейросети, корпусная лингвистика, статистические методы — все это основные принципы автоматического реферирования. Автоматическая обработка и интерпретация данных возможна благодаря электронному архиву. Электронный архив является мощным инструментом для синхронизации и совместной работы исследователей.

Диахронический подход к индексации представляет различные подпроекты, которые включают в себя разметку текстов на разные темы. Национальный корпус русского языка предоставляет лингвистическую информацию. Формализмы ограничены, но предоставляют полезное представление лингвистических данных.

**Разметка текста**

Основные подходы к разметке включают основанные на актуациях и на абстракциях. Они используют разные формализмы для представления данных. 16 ноября 2023 года, модель в компьютерной нейролингвистике (ЛП), предполагает описание статических свойств, обрабатываемого языка. Это описание организуется на модель языка, используемые в данной модели языка, обычно строятся на основе теорий, создаваемых лингвистами путем изучения различных текстов и на основе своей лингвистической интуиции (интроспекции). Сложность ее описания и обработки ведет к разбиению этого отдельного процесса на отдельные этапы, соответствующие уровням языка. При анализе текста отдельные модули, в свою очередь, полностью объединяются.

1. Графематический анализ
2. Морфологический анализ
3. Синтаксический анализ
4. Семантический и прагматический анализ

Таким образом, анализируя текст, можно рассматривать наиболее характерные и проблемные этапы его обработки и разбиения на отдельные уровни языка. При вводе текста отдельные модули ЛП требуют учета всех особенностей, в них уровнях и наличия соответствующих модулей. При решении определенных прикладных задач можно обойтись без развернутых отдельных уровней, к примеру, в ранних экспериментальных программах ЛП, обрабатываемые тексты относились к очень узким проблемным областям.

Если одним примером редуцированной модели, которой моделируется недостаточно часто встречаемая языковая модель частотности символов и имен + составляющие. Такая статистическая модель отображает лингвистическую информацию на уровне символов текста и ее достаточно, например, для выявления ошибок в тексте, или для распознавания его языковой принадлежности.

Отметим, что возможны структурно-статистические модели, в которых, при представлении отдельных уровней, их унифицируется та или иная статистическая информация - слов. При морфологическом синтезе, в качестве таких данных являются: лексема и конкретные морфологические характеристики запрашиваемой словоформы. Для моделирования синтаксиса в рамках предложено было большое число разных идей и методов, отличающихся способом они осуществляют синтаксиса языка.

**Подходы и созданию моделей**

1. Генеративный подход Хомский
2. Модель смыслов
3. Модель текстов
4. Теория синтаксических групп
5. Математическая языковая модель

В рамках данного подхода, для представления синтаксической структуры предложения используется более наглядное и распространенное дерево зависимостей, где в узлах дерева расположены слова предложения (в том числе глагол-сказуемое, обычно), а каждая дуга дерева, связывающая пару узлов, интерпретируется как синтаксическая подчинительная связь между ними.

**Деревья описания языка**

Деревья составляющих больше подходят для описания языков, с простым и сильным словородным слов, представление с их помощью разоруженных и непроизводственных конструкций, требует расширение используемого грамматического формообразования. Валентность - способность слова, присоединять другие слова-антонты, слова-связки. В соответствии с этой теорией, ее рассматривается, как особый род преобразователь, выполняющий переработку из заданных смыслов в соответствующие им смыслы:

1. Ориентация на синтез текста
2. Многоуровневый, модульный
3. Интегральная модель
4. Специальные средства
5. Четырехпортовый словарь

**Инструменты исследования и описание языка**

Модель Толково-комбинаторного словаря И.А. Мельника является основным инструментом исследования и описания языка. И.А. Мельник признает действующие элементы языка, такие как мысль (=) текст. Подобная модель конкретного языка представляет собой сложную организацию правил, чисто механическое применение которых должно обеспечивать:

**Различные модели обработки языка**

Переход от данного текста, рассматриваемого на языке, к формальному описанию его смысла, то есть к его семантическому представлению (= записи смысла). Разные, но интуитивно схожие (синонимичные) тексты должны получать одинаковое, или хотя бы схожее, семантическое представление. В случае текста, имеющего более одного значения, И.А. Мельник исходит из следующего тезиса: язык - это особый вид преобразователя, понимающего переработку заданных смыслов в соответствующие им тексты, и наоборот.

Смысл = текст. Модель также должна быть действующей. Она представляет собой устройство, имеющее вход и выход, преобразующее входные данные (знание) в что-то другое, что выводится на выходе (результат). Модели бывают лингвистическими и алгоритмическими. Модель мышления должна быть абсолютно формальной, основанной на однозначных логических последовательностях, не требующих дополнительной информации. В качестве контрольного принципа выделяется принципиальная осуществимость модели или любого её фрагмента на вычислительной машине.