Модуль I. Лекция 3.
Компьютерная лингвистика

План

1. Работа с текстом
2. Автоматическое аннотирование и реферирование текста
3. Аппаратное и программное обеспечение информационных технологий в лингвистике.

Что такое компьютерная лингвистика?

В условиях все возрастающего количества текстов в окружающем человека мире возникает проблема: как в этом море информации найти нужные документы и познакомиться с их содержанием? Решению данной проблемы может помочь составление рефератов и аннотаций полнотекстовых документов. Они дают читателю представление о содержании исходных документов и позволяют оценить степень необходимости обращения к полным текстам каждой работы. Кроме того, рефераты и аннотации акцентируют внимание читателя на новых сведениях, т.е. позволяют за небольшой промежуток времени узнать много новой информации.

Работа с текстом

Рефераты и аннотации составляются вручную, например самим автором исходного текста или библиографическим работником, или автоматически, с помощью специальных компьютерных программ. Наиболее качественным является первый вид рефератов и аннотаций, поскольку в этом случае создается новый текст, называющий основную мысль высказывания и отличающийся связным характером. Но для обработки большого массива текстов за минимальное количество времени требуется привлечение автоматических средств для решения задачи реферирования и аннотирования текстов.

Реферат определяется как связный текст, который кратко выражает

- центральную тему,
- предмет
- цель,
- методы,
- результаты исследования.

Рефераты обычно составляют к научно-техническим документам: научным монографиям, статьям, патентам на изобретение и др. В зависимости от жанра исходного текста (монография, статья, патенты др.) и от предметной области (медицина, химия, лингвистика и т.д.) заданные элементы реферата могут различаться. Так, для научных рефератов дополнительно к названным выше элементам реферата прибавляется краткое изложение сути, практической апробации перспектив исследования.

Автоматическое аннотирование и реферирование текста

Различают следующие виды рефератов [6, 89]:

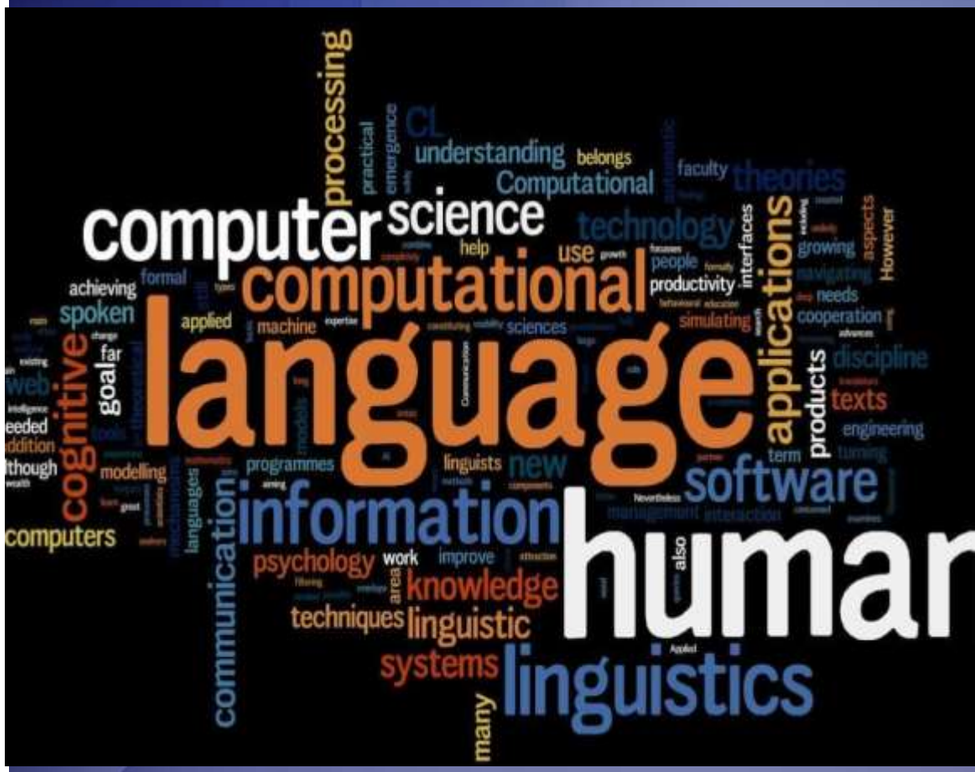
• связный текст — новое текстовое образование, порождаемое на основе логико-смыслового анализа исходного текста;

• реферат-клише — модификация заданной клишированной структуры, пустые ячейки которой заполняются после анализа заданного текста;

• квазиреферат — перечень наиболее информативных предложений текста.

Очевидно, что для автоматического создания рефератов — связных текстов требуются более сложные компьютерные программы, чем для создания рефератов-клише и квазирефератов.

Некоторые исследователи считают реферат и аннотацию синонимами [6, 88], а некоторые предлагают разводить эти понятия, определяя аннотацию как краткое изложение содержания документа, дающее общее представление о его теме [2, 55]. Согласно этому определению в отличие от реферата, знакомящего читателя с сутью излагаемого в документе содержания, аннотация выполняет лишь сигнальную функцию.



Аппаратное и программное обеспечение информационных технологий в лингвистике

Изучите важную роль аппаратного и программного обеспечения в компьютерной лингвистике, используемой для анализа английского языка.

Применение компьютерной лингвистики в английском языке

Для выполнения объемных расчетов над лингвистическими данными, а также для лингвистического моделирования удобно использовать электронные вычислительные машины (или компьютеры).

Компьютер — это электронное устройство, служащее для автоматического создания, обработки, передачи и воспроизводства информации по созданным человеком алгоритмам (программам), написанным на понятном для машины языке. Как следует из приведенного определения, в использовании компьютеров сочетается аппаратное (hardware) и программное обеспечение (software) информационных технологий.

К аппаратному обеспечению относится сам компьютер (стационарный или переносной), а также периферийные устройства, служащие для ввода/вывода информации в компьютер пользователем(клавиатура, мышь, монитор, принтер и т.д.) или для соединения компьютера с другими устройствами (например, модем).

Машинный перевод

Наряду с аппаратным и программным обеспечением (ПО) информационных технологий некоторые исследователи используют также понятие lingware (или linguware), которым обозначаются все лингвистические компьютерные ресурсы (грамматические справочники, словари, энциклопедии, лингвистические базы данных и т.п.).

Совокупность аппаратных, программных и лингвистических средств, необходимых для автоматической обработки лингвистических данных, обозначим понятием автоматическое рабочее место(АРМ) лингвиста [22,258]. АРМ лингвиста будет включать сам компьютер, операционное и базовое прикладное ПО, а также всевозможные лингвистические компьютерные ресурсы, касающиеся родного и изучаемых иностранных языков.

Анализ тональности и эмоциональной окраски текста

Программное обеспечение — это компьютерные программы, представляющие собой последовательность написанных на машинном языке команд, служащие для управления аппаратными средствами или для выполнения различных операций над информацией, и соответствующая документация.

В зависимости от назначения программных средств различают системное и прикладное программное обеспечение. Системные программы служат управлению работой аппаратных средств и включают операционные системы, утилиты, драйверы и некоторые другие виды программ. Прикладные программы предназначены для конечного пользователя и позволяют ему выполнять различные операции над информацией: создавать и обрабатывать текст (текстовые редакторы), обрабатывать графические изображения (графические редакторы), работать над звуковой и видеоинформацией (мультимедийные программы), создавать электронные таблицы для обработки статистических данных (электронные таблицы) и т.д. Для лингвиста особенно полезными являются такие виды прикладных программ, как электронные переводчики и словари, а также мультимедийные обучающие программы.

Распознавание речи

Распознавание речи (Speech Recognition) - это область компьютерной лингвистики, которая занимается разработкой технологий и алгоритмов для преобразования аудиосигнала, содержащего человеческую речь, в текстовую форму, понимаемую компьютером.Вот основные аспекты распознавания речи в компьютерной лингвистике:

- **Звуковой сигнал**: Процесс распознавания речи начинается с записи аудиосигнала с микрофона. Этот звуковой сигнал содержит речь, которую необходимо распознать.
- **Предварительная обработка**: Звуковой сигнал обычно подвергается предварительной обработке, которая включает в себя фильтрацию шума, усиление сигнала и разбиение на аудиофрагменты.
- **Функции признаков**: Для анализа речи необходимо извлечь характеристики из аудиосигнала, такие как частота, мел-кепстральные коэффициенты (MFCC), и другие. Эти функции признаков используются для представления речи в числовой форме.
- **Моделирование звука**: С использованием функций признаков, компьютер создает статистические модели, которые описывают, какие фонемы, слова или фразы могут соответствовать звуковым признакам.
- **Сравнение и распознавание**: Звуковой сигнал сравнивается с моделями, и компьютер пытается найти наилучшее соответствие. Этот процесс включает в себя сопоставление и классификацию фонем, слов или фраз.
- **Коррекция и адаптация**: Важной частью систем распознавания речи является коррекция ошибок и адаптация к конкретному говорящему или контексту. Это может включать в себя использование языковых моделей, контекстуального анализа и машинного обучения.
- **Выдача текста**: После успешного распознавания речи система выдает текстовую версию произнесенной фразы.

Современные достижения компьютерной лингвистики

Современная компьютерная лингвистика продолжает быстро развиваться и привносить множество инноваций. Вот некоторые из современных достижений и технологий в компьютерной лингвистике:

1. **Глубокое обучение (Deep Learning)**: Использование нейронных сетей, в частности, глубоких нейронных сетей, позволило значительно улучшить качество задач компьютерной лингвистики, такие как машинный перевод, распознавание речи, анализ тональности и многое другое.
2. **Машинный перевод**: Модели глубокого обучения, такие как Transformer, сделали современные системы машинного перевода более точными и способными работать с различными языками.
3. **Обработка естественного языка (NLP)**: Технологии NLP, включая BERT и GPT, позволяют компьютерам понимать и генерировать естественный язык, что приводит к созданию лучших систем вопросно-ответных систем, суммаризации текста и анализа тональности.
4. **Распознавание речи**: Системы распознавания речи, такие как Google Speech-to-Text и Amazon Transcribe, становятся все более точными и широко используются в голосовых помощниках, автоматизированных транскрипциях и других приложениях.
5. **Обработка больших данных**: Современные методы обработки больших данных позволяют анализировать огромные объемы текстовой информации и извлекать ценные знания из текстовых корпусов.
6. **Биометрическая аутентификация**: Использование технологий компьютерной лингвистики для биометрической аутентификации, такой как распознавание голоса, голосовые пароли и распознавание лиц, становится все более распространенным.
7. **Прикладные приложения**: Компьютерная лингвистика используется во многих областях, включая медицину (клиническая нотация), юриспруденцию (поиск и анализ юридических текстов), финансы (анализ новостей и финансовых отчетов) и многое другое.
8. **Социальные медиа и аналитика**: Анализ текстовых данных в социальных медиа позволяет компаниям и организациям следить за обратной связью клиентов, а также анализировать публичные мнения и тренды.
9. **Обработка естественного языка для чат-ботов**: Создание чат-ботов с возможностью понимания и взаимодействия на естественном языке становится все более популярным для обслуживания клиентов и автоматизации задач.
10. **Развитие языковых моделей и корпусов данных**: Создание больших корпусов данных и разработка мощных языковых моделей позволяют улучшать качество и эффективность приложений компьютерной лингвистики.

Вопросы для обсуждения

1. Что такое реферат?
2. Назовите виды реферата. Приведите примеры
3. Что является главными смысловыми единицами исходного текста?
4. Назовите цели логико-семантического метода
5. Какие системы автоматического реферирования и аннотирования текстов вы знаете?

Спасибо за внимание!!!