## Слайд 1. Тема 2.3. Обработка естественного языка
## или
## Компьютерная лингвистика и анализ текста

В компьютерном мире всё чаще появляются цифровые помощники, позволяющие взаимодействовать с другими людьми, и разработанные для этого информационные ресурсы. Привлекательность этих умных устройств отчасти обусловлена тем, что они не просто передают информацию, но также понимают её до некоторой степени, облегчая коммуникацию высокого уровня, комбинируя, фильтруя и обобщая данные в легкоусвояемую форму. Такие приложения как машинные переводчики, системы «вопрос-ответ», инструменты транскрипции голосовой информации и обобщения текста, а также чат-боты, становятся неотъемлемой частью нашей жизни в компьютерном мире.

**Слайд 2**. Компьютерная лингвистика и анализ текстовых данных – востребованное направление в IT. Перечисленные приложения: машинные переводчики, чат-боты, системы «вопрос-ответ» и другое прикладное ПО строятся на методах и подходах анализа естественного языка. Компоненты анализа языка основаны на современной инфраструктуре анализа текстовой информации: коллекции приемов и методов, объединяющей инструменты для работы со строками; лексических ресурсах; компьютерной лингвистике; алгоритмах машинного обучения, преобразующих данные на естественном языке в машинную форму и обратно.

**Слайд 3**. **Цель раздела**: Освоение методов и подходов компьютерной лингвистики и анализа текстовых данных.

**Задачи**:

- Изучить методы и подходы компьютерного анализа и обработки естественного языка.
- Овладеть навыками компьютерной лингвистики и анализа текстов, необходимыми для разработки современного программного обеспечения.

**Слайд 4**. **Содержание раздела**

**Часть 1. Введение в компьютерную лингвистику**

- Основные понятия компьютерной лингвистики. Лингвистические данные. Сложности компьютерной обработки и анализа естественного языка.
- Инструменты для анализа текста. Модули и библиотеки для NLP.

**Часть 2. Корпусная лингвистика**

- Создание и применение корпусов текста.

- Обработка и преобразования корпуса текста: сегментация, лексемизация, промежуточный анализ корпуса.
- Лемматизация, векторизация, семантический анализ текстов. Распознавание именованных сущностей и извлечение отношений. Метод TF-IDF, косинус сходства, коэффициент Отиаи.

**Слайд 5**. **Часть 3. Прикладной анализ текстовых данных**

- Классификация в анализе текстовых данных.
- Кластеризация как инструмент выявления сходств в тексте.
- Контекстно-зависимый анализ текста.
- Визуализация текста. Графовые методы анализа текста.

**Часть 4. Создание чат-бота**

- Проектирование основного «тела» чат-бота.
- Интеграция чат-бота в социальные сети (чаты). Обработка запросов.

**Слайд 6**. **Тема 1. Введение в компьютерную лингвистику**

- Основные понятия компьютерной лингвистики. Лингвистические данные. Сложности компьютерной обработки и анализа естественного языка.
- Инструменты для анализа текста. Модули и библиотеки для NLP.

Рассмотрим **Основные понятия компьютерной лингвистики**

**Слайд 7.** Компью́терная **лингви́стика** (также: математи́ческая или вычисли́тельная лингви́стика, англ. *computational linguistics*) – научное направление в области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллекта, которое ставит своей целью использование математических моделей для описания естественных языков.

Математическая лингвистика является ветвью науки искусственного интеллекта. Её история началась в Соединённых Штатах Америки в 1950-х годах. С изобретением транзистора и появлением нового поколения компьютеров, а также первых языков программирования, начались эксперименты с машинным переводом, особенно русских научных журналов. В 1960-х годах подобные исследования проводились и в СССР (например, статья о переводе с русского на армянский в сб. «Проблемы кибернетики» за 1964 год). Однако качество машинного перевода до сих пор сильно уступает качеству перевода, произведённого человеком.

**Слайд 8.** Существует еще несколько определений компьютерной лингвистики:

**Компью́терная лингви́стика** – деятельность по формализации знаний о естественном языке на разных его уровнях с целью использования в компьютерных технологиях.

**Компью́терная лингви́стика** – область знаний, решающая проблемы общения человека и компьютера на естественном языке.

**Компью́терная лингви́стика –** широкая область использования компьютерных инструментов – программ, компьютерных технологий организации и обработки данных – для моделирования функционирования языка в тех или иных условиях, ситуациях, проблемных областях.

**Слайд 9.** Процесс становления и формирования современной лингвистики как науки о естественном языке представляет собой длительное историческое развитие лингвистического знания. В основе лингвистического знания лежат элементы, формирование которых происходило в процессе деятельности, неразрывно связанной с освоением структуры устной речи, появлением, дальнейшим развитием и совершенствованием письма, обучением письму, а также толкованием и расшифровкой текстов.

Естественный язык как объект лингвистики занимает центральное место в этой науки. В процессе развития языка менялись и представления о нем. Если раньше не придавалось особого значения внутренней организации языка, и он рассматривался, прежде всего, в контексте взаимосвязи с внешним миром, то, начиная с конца XIX – начала XX вв., особая роль отводится внутреннему формальному строению языка. Именно в этот период известным швейцарским лингвистом Фердинандом де Соссюром были разработаны основы таких наук, как семиология и структурная лингвистика, и подробно изложены в его книге «Курс общей лингвистики» (1916 г.).

**Слайд 10.** Ученому принадлежит идея рассмотрения языка как единого механизма, целостной системы знаков, что в свою очередь дает возможность описать язык математически. Соссюр первым предложил структурный подход к языку, а именно: описание языка посредством изучения соотношений между его единицами. Под единицами, или «знаками» он понимал слово, которое объединяет в себе и смысл и звучание. В основе концепции, предложенной швейцарским ученым, лежит теория языка как системы знаков, состоящей из трех частей: языка (от фр. langue), речи (от фр. parole) и речевой деятельности (от фр. langage).

**Слайд 11.** Лингвистические данные: **лексемы** и **слова**

Чтобы полностью использовать данные, закодированные в языке, мы должны научить наш разум думать о языке не как по понятном и естественном,

но как о произвольном и неоднозначном. Единицей анализа текста является лексема.

**Лексемы** – атомарные единицы данных в анализе текста. Это строки, закодированные байтами и представляющие семантическую информацию, но не содержащие никакой другой информации (например, значения слова).

Слова, напротив, – это символы, представляющие смысл и отображающие текстовые или вербальные конструкции как звук или изображение. Лексемы не являются словами (хотя нам трудно смотреть на лексемы и не видеть слов).

**Слайд 12.** Рассмотрим лексему «коса», изображенную на рисунке 1. Эта лексема представляет смысловое слово *коса-n1* – первое определение существительного, в котором используется лексема, обозначающие сплетённые в одну прядь волосы, чаще всего у девочек, за которую частенько любят дергать мальчишки.

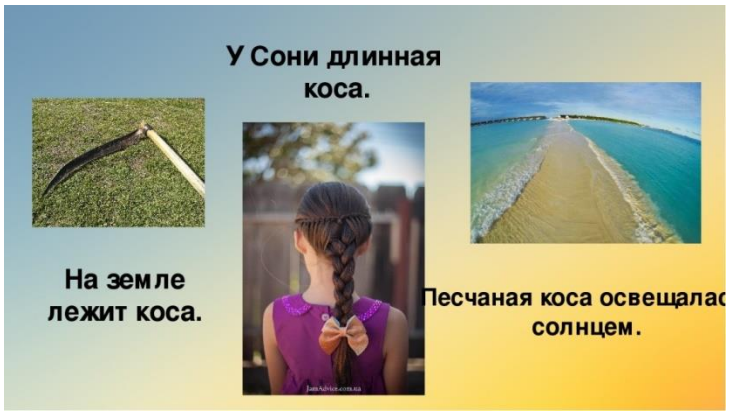


Рис. 1. Пример слова с несколькими различными значениями

Все другие понятия некоторым способом связаны с символом, и все же символ совершенно произволен; у другого человека, например, данная лексема вызовет иные ассоциации с другим смыслом. Допустим, подумает о косе, которой косят траву. Это происходит потому, что слова не имеют универсального фиксированного смысла, не зависящего от таких контекстов, как культура и язык. Англоговорящие читатели используют адаптивные формы слов с приставками и суффиксами, изменяющими время, род и т.д. Читатели, говорящие на китайском языке, напротив, распознают множество пиктографических изображений, смысл которых определяет порядок их следования.

**Слайд 13.** Избыточность, неоднозначность и зрительные ассоциации делают естественные языки динамичными, быстро развивающимися, способными отражать текущий опыт. Так, например, лексема «батарея» изменила свой смысл в результате развития электроники и означает «хранилище,

преобразующее химическую энергию в электрическую». Однако, согласно службе Google Books Ngram Viewer, в XIX и в начале XX вв. слово «батарея» чаще использовалось для обозначения «артиллерийского и минометного подразделения из нескольких орудий и минометов, а также позиции, которую занимает такое подразделение». Контекст понимания языка зависит не только от окружающего текста, но также от периода времени. Для четкого определения смысла слов требуется больше вычислений, чем простой поиск в словаре.

**Слайд 14.** Компьютерная лингвистика частично пересекается с обработкой естественных языков. Однако в последней акцент делается не на абстрактные модели, а на прикладные методы описания и обработки языка для компьютерных систем. Под обработкой естественного языка понимается комплекс вычислительных приемов для отображения между формальным и естественным языками.

Полем деятельности компьютерных лингвистов является разработка алгоритмов и прикладных программ для обработки языковой информации. Приложения, использующие приемы обработки естественного языка для анализа текстовых и аудиоданных, становятся неотъемлемой частью нашей жизни.

**Слайд 15.** Разработанные компьютерными лингвистами приложения от нашего имени просматривают огромные объемы информации в Сети и предлагают новые и персонализированные механизмы взаимодействия человека с компьютерами. Эти приложения настолько распространены, что мы привыкли к широкому спектру закулисных инструментов: от спам-фильтров, следящих за нашим почтовым трафиком, до поисковых систем, которые ведут нас прямо туда, куда мы хотим попасть, и виртуальных помощников, всегда готовых выслушать и ответить.

Информационные продукты с поддержкой анализа естественного языка находятся на пересечении экспериментальных исследований и практической разработки ПО. Приложения, анализирующие речь и текст, взаимодействуют непосредственно с пользователем, чьи ответы обеспечивают обратную связь, которая оказывает влияние и на приложение, и на результаты анализа. Этот благотворный цикл часто начинается с самого простого, но с течением времени может перерасти в мощную систему, возвращающую ценные результаты.

**Слайд 16. Сложности компьютерной обработки и анализа естественного языка**

Естественные языки определяются не правилами, а контекстом использования, который требуется реконструировать для компьютерной обработки. Часто мы сами определяем значения используемых слов, хотя и совместно с другими участниками беседы. Словом «коса» мы можем обозначить прядь волос или инструмент для скашивания травы, но при этом оба – говорящий/автор и слушатель/читатель – должны согласиться с общим пониманием этого слова в ходе диалога. Поэтому язык обычно ограничивается обществом и регионом – передать смысл часто намного проще людям, имеющим жизненный опыт, похожий на ваш.

**Слайд 17.** В отличие от формальных языков, которые всегда являются предметными, естественные языки намного более универсальны. Мы используем одно и то же слово при заказе морепродуктов на обед, в поэме, описывающей уныние и недовольство, и для обозначения астрономической туманности. Для поддержания множества смыслов язык должен быть избыточным. Избыточность представляет серьезную проблему, потому что мы не можем (и не делаем этого) указать буквальный смысл для каждой ассоциации, каждый символ по умолчанию является неоднозначным. Лексическая и структурная неоднозначность является основным достижением человеческого языка; она не только дает нам возможность генерировать новые идеи, но также позволяет общаться людям с разным опытом и культурой, несмотря на почти гарантированные случайные недоразумения.

**Слайд 18. Что такое обработка естественного языка**

Обработка естественного языка (далее NLP – Natural language processing) – область, находящаяся на пересечении computer science, искусственного интеллекта и лингвистики. Цель заключается в обработке и “понимании” естественного языка для перевода текста и ответа на вопросы.

С развитием голосовых интерфейсов и чат-ботов, NLP стала одной из самых важных технологий искусственного интеллекта.

**Слайд 19.** Полное понимание и воспроизведение смысла языка – чрезвычайно сложная задача, так как “живой” язык имеет ряд особенностей:

- Человеческий язык – специально сконструированная система передачи смысла сказанного или написанного. Это не просто экзогенный сигнал, а осознанная передача информации. Кроме того, язык кодируется так, что даже маленькие дети могут быстро выучить его.
- Человеческий язык – дискретная, символьная или категориальная сигнальная система, обладающая надежностью.

- Категориальные символы языка кодируются как сигналы для общения по нескольким каналам: звук, жесты, письмо, изображения и так далее. При этом язык способен выражаться любым способом.

Поскольку методы анализа текста применяются в первую очередь в машинном обучении, необходим язык программирования с богатым набором научных и вычислительных библиотек. На эту роль как нельзя лучше подходит язык Python, включающий в себя набор мощных библиотек, таких как Scikit-Learn, NLTK и т.д.

**Слайд 20. Язык как данные**

Язык – это неструктурированные данные, которые используются людьми для общения между собой. Структурированные или полуструктурированные данные, в свою очередь, включают поля или разметку, позволяющие компьютеру анализировать их. Но, несмотря на отсутствие машиночитаемой структуры, неструктурированные данные не являются случайными. Напротив, они подчиняются лингвистическим правилам, которые делают эти данные понятными для людей.

Специалисты по обработке и анализу данных создают приложения данных, основанные на анализе естественного языка, поэтому наша первейшая задача – создать модель, описывающую язык и способную делать выводы на основе этого описания.

**Слайд 21.** Формально модель языка должна принимать на входе неполную фразу и дополнять ее недостающими словами, наиболее вероятными для завершения высказывания. Этот тип моделей языка сильно влияет на аналитическую обработку текста, потому что демонстрирует основной механизм приложений обработки языка – использование контекста для угадывания смысла. К примеру, компания «Яндекс» запустила «Балабобу» – сервис, который с помощью нейросетей и методов анализа текста дописывает любое предложение или текст. «Балабоба» генерирует текст в разных стилях: современные интернет цитаты, рекламные слоганы, короткие истории, подписи в Instagram, синопсисы фильмов, гороскоп, народные мудрости, тосты, теории заговора, ТВ-репортажи, народные мудрости и многое другое.

**Слайд 22.** Модели языка также раскрывают базовую гипотезу машинного обучения на текстах: текст предсказуем. Фактически, механизм оценки моделей в академическом контексте, связность, измеряет предсказуемость текста вычислением энтропии (степень неопределенности или неожиданности) распределения вероятностей модели языка.

Рассмотрим следующие незаконченные фразы: «собака – друг ...» и «ведьма летит на ...». Эти фразы имеют низкую энтропию, и модели языка с высокой степенью вероятности будут угадывать продолжения «человека» и «метле» соответственно (более того, мы удивились бы, если бы эти фразы завершались как-то иначе). С другой стороны, фразы с высокой энтропией, такие как «сегодня я собираюсь поужинать с ...», предполагают множество вариантов продолжения («другом», «родителями» и «коллегами» выглядят одинаково вероятными). Человек, услышавший такую фразу, может использовать свой опыт, воображение и память, а также ситуационный контекст, чтобы восполнить пробел. Компьютерные модели не обязательно имеют одинаковый контекст и в результате оказываются более ограниченными. Формально, модель использует контекст для определения узкого пространства решений, в котором есть небольшое количество вариантов.

Это понимание дает нам возможность генерализировать формальную модель в другие модели языка, работающие, например, в приложениях машинного перевода или анализа настроений. Чтобы воспользоваться предсказуемостью текста, мы должны определить ограниченное числовое пространство решений, на котором может действовать модель. Благодаря этому мы можем использовать методы статистического машинного обучения, с учителем и без учителя, для построения моделей языка, извлекающих смысл из данных.

На первом шаге в машинном обучении выявляются характерные признаки данных, помогающие предсказать цель. Текстовые данные предоставляют массу возможностей для извлечения поверхностных признаков, простым разбиением строк, или более глубоких, позволяющих извлекать из текста морфологические, синтаксические и даже семантические представления.

**Слайд 23.**

**Лекция окончена.**

Приглашаем к выполнению первого практического задания, посвященного **основным понятиям компьютерной лингвистики, лингвистическим данным и сложности обработки и анализа естественного языка.**