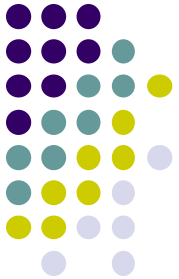


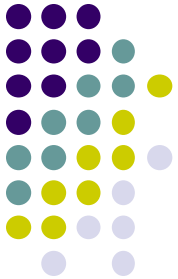
КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И АНАЛИЗ ТЕКСТА: ВВЕДЕНИЕ

Большакова Елена Игоревна

СОДЕРЖАНИЕ

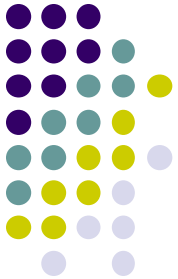
1. Компьютерная лингвистика (КЛ): истоки, междисциплинарность
2. Особенности естественного языка (ЕЯ)
   - Уровни и единицы языка и текста
   - ЕЯ и искусственные языки
   - Асимметрия знаков и смыслов
3. Моделирование в КЛ
4. Лингвистические ресурсы
5. Прикладные задачи КЛ
6. Заключение
7. Содержание курса, литература

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: ИСТОКИ



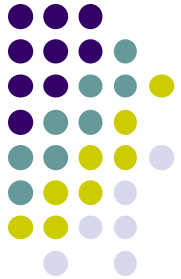
- Начало работ – 50-е годы,
  Потребности практики: машинный перевод
- Название научной области:
  - Автоматическая обработка тестов на *естественном языке* (ЕЯ)
  - Вычислительная/ Компьютерная лингвистика
    Computational Linguistics (CL)
  - *Natural Language Processing (NLP)*
- Междисциплинарное научное направление:
  - Лингвистика
  - Математика
  - Информатика (*Computer Science*)
  - Искусственный интеллект (*Artificial Intelligence*)

КЛ и ЛИНГВИСТИКА



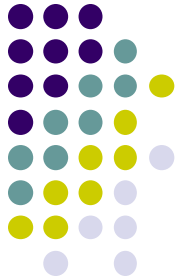
- Общая лингвистика
  - Фонология (звуки речи)
  - Морфология (структура и форма слов ЕЯ)
  - Синтаксис (структура и функции предложений)
  - Семантика (смысл языковых высказываний)
  - Прагматика (значение высказываний)
- Социолингвистика
- Психолингвистика
- Лексикография (описание лексикона ЕЯ)
- Прикладная лингвистика

КЛ: МАТЕМАТИКА, ИНФОРМАТИКА И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ



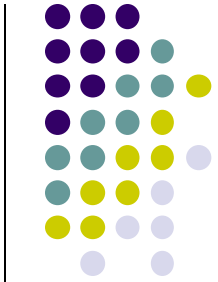
- Математика: применение результатов и аппарата
  - Математическая лингвистика:
    Теория формальных языков и грамматик – возникла из *порождающих грамматик* Н.Хомского (50-е гг.) для анализа синтаксических структур Е
- Информатика (*Computer Science*): общая методология с КЛ и ИИ – компьютерное моделирование
- Искусственный интеллект: компьютерное моделирование интеллектуальных функций
  - Методы моделирования: *эвристические*
  - Обработка ЕЯ – интеллектуальная функция
  - Междисциплинарный характер области

ЦЕЛИ КЛ, СЛОЖНОСТИ ЕЯ



- Разработка компьютерных моделей и программ для обработки информации, представленной на естеств. языке (неструктурированные тексты, речь)
- Основа – формальные модели, обычно зависящие от конкретного ЕЯ (редактор *Word*, но не *NotePad)*
- Естеств. язык – сложная система знаков (звуковых и письменных), возникшая в процессе человеческой деятельности как основное средство общения
- Функции ЕЯ:
  - коммуникация (обмен информацией)
  - мышление
  - познание (сохранение знаний)
- Сложность задач КЛ: сложность любого языка, многообразие естественных языков

ЯЗЫКОВЫЕ ЗНАКИ



Семиотика – теория знаковых систем
(язык жестов, знаки дорожного движения,
морская сигнализация флажками)

Три стороны языкового знака (*треугольник Фреге*)

- Означающее – *signifier*
  - последовательность звуков или графических знаков
- Означаемое (понятие, смысл) – *signified*
  - представление этого предмета, явления в сознании человека
- Денотат (референт) – *referent*
  - обозначаемый предмет, явление действительности

Соответственно, три аспекта системы знаков:

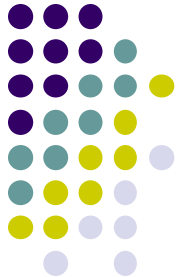
синтактика – семантика – прагматика

ОСОБЕННОСТИ ЕЯ

Большая многоуровневая
комбинаторная система языковых знаков:

- Несколько сот тысяч языковых знаков
- Многоуровневость системы языка и текста
  - каждый уровень (подсистема): правила сочетания *единиц* (знаков) этого уровня (грамматика),
  - взаимосвязь, иерархия уровней: разложимость единиц на меньшие
- Открытая коммуникативная система
  (постоянная изменчивость языка)
- Избыточность, гибкость ЕЯ (разные способы
  выражения одного и того же смысла)
- Многозначность, неопределенность смысла знаков
- Универсальность ЕЯ (выражение разных смыслов)

ЯЗЫК и РЕЧЬ (ТЕКСТ)



Разграничение в лингвистике:

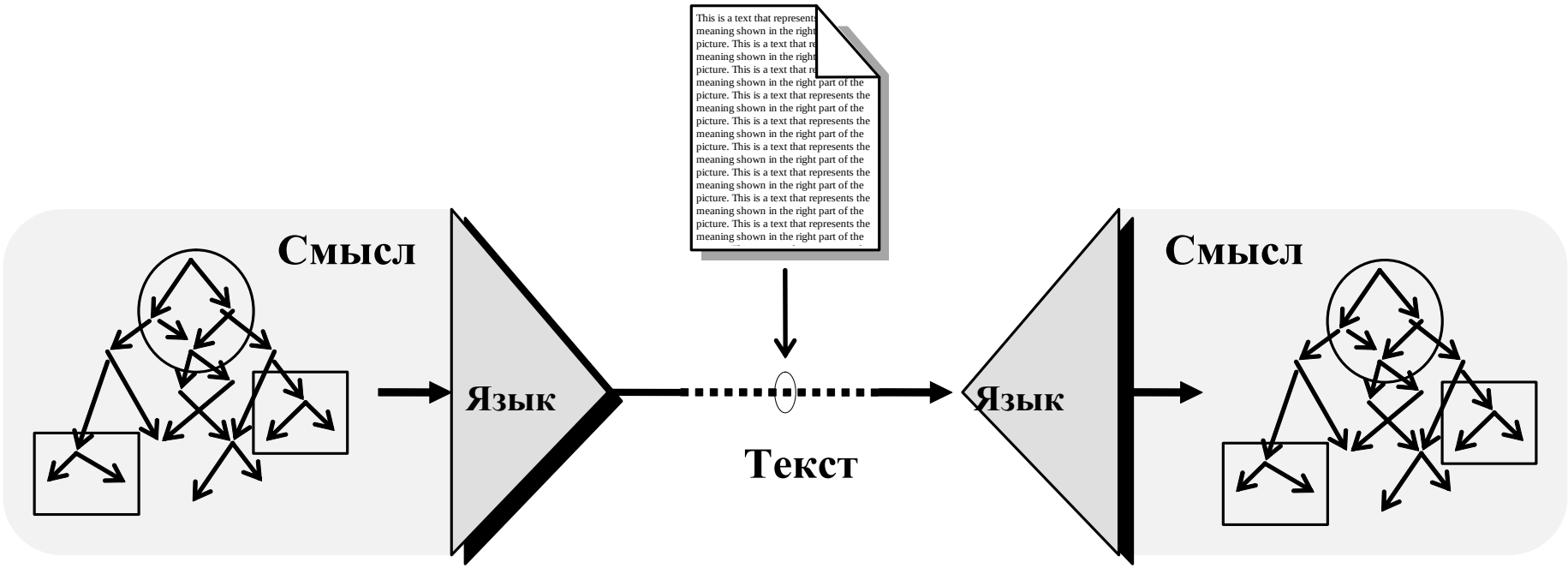
- Язык: система знаков ЕЯ
- Речь (устная, письменная): линейная последовательность знаков, построенная в процессе общения, в соответствии с принятыми правилами

Единицы разного уровня:

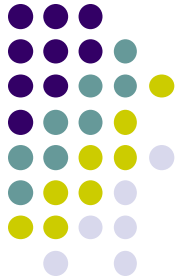
- Языка:
  - фонемы / графемы(буквы)
  - морфемы
  - слова (лексемы)
- Речи / текста:
  - слова (словоформы)
  - словосочетания
  - предложения (фразы)

ЯЗЫК КАК ПРЕОБРАЗОВАТЕЛЬ СМЫСЛ ⇔ ТЕКСТ

- Центральный объект – *текст*, линейность текста
- Текст составлен из различных *единиц*, относящихся к разным *уровням*
- Единицы: незначащие и значащие (языковые знаки)



ОСОБЕННОСТИ ЕЯ: УРОВНИ



- Синтаксический – предложения (*фразы*) ЕЯ
  - подуровень *словосочетаний*
    синий цвет, смотрю кино, чай с сахаром
  - надуровень *сверхфразовых единств* (≈ абзацев)
    – предложений, объединяющихся по смыслу
- Морфологический – слова (*словоформы*)
  - подуровень *морфем*: *по-стро-ен, за-гвозд-ка*

    Морфема – минимальная значащая единица (*корень*, *приставка*, *суффикс*, *окончание*…)
    смысловое содержание + звуковое выражение
- Фонологический/буквенный: звуки (*фонемы*)/ буквы – <u>незначащие</u> единицы речи, служат для различения значащих, смысловых (морфемы, слова, фразы)

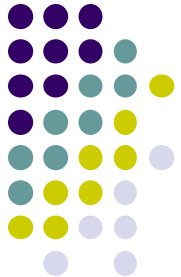
ДРУГИЕ УРОВНИ ЕЯ

- Семантический (смысловой) – набор элементарных единиц (*сем*)
- Лексический – множество слов-лексем (*лексикон*)
  Лексема – совокупность *словоформ* слова
  например: *лист*, *листа*, *листу*, *листе*
- Дискурсивный – уровень *связного текста*: схематические структуры текстов

Важно: взаимосвязь всех уровней

Сложность, открытость системы ЕЯ $\Rightarrow$
ее универсальность, т.е. возможность построить практически бесконечное число высказываний (смыслов)

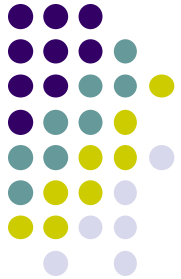
ЕЯ и ИСКУССТВЕННЫЕ ЯЗЫКИ



Искусственные языки:
языки программирования (ЯП), языки логики,
близки к ЕЯ по функциям, но есть принцип. отличия:

- Открытость и изменчивость ЕЯ (на всех уровнях) ⇒ невозможность один раз и навсегда создать лингвистический процессор
- Нестандартная сочетаемость (*синтактика*) единиц ЕЯ на всех уровнях, например, лексическая:
  крепкий чай, но не *сильный чай* (*strong tea*),
  однако: *сильный акцент* – *heavy accent*
  бить тревогу – *sound the alarm*
- Большая системность (число уровней) и степень асимметрии связи единиц и выражаемых ими смыслов:
  полисемия , *синонимия*, *омонимия*

ЕЯ: АСИММЕТРИЯ ЗНАК⇔СМЫСЛ



Асимметрия связи:

Означающее (единица ЕЯ) ⇔ Означаемое (ее смысл)

- *Полисемия* – многозначность языковой единицы

  например, для слова *земля*:

  Земля, суша, почва, страна, территория

- *Синонимия* – совпадение единиц по основному смыслу (обычно: различия в смысл. оттенках и стиле)

  синонимия слов: *горячий – жаркий*

  синонимия предлогов: *о поездке – про поездку*

  синонимия приставок, суффиксов, союзов и др.

- *Омонимия* – звуковое совпадение или совпадение на письме (по форме) двух или более языковых единиц;

  отличие от полисемии: обычно это случайное совпадение, нет смысловой связи между совпавшими единицами

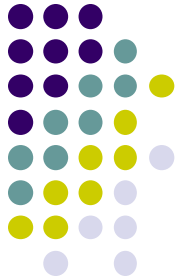
ЕЯ : ОМОНИМИЯ

Совпадение по форме двух разных по смыслу единиц

Наиболее частые виды:

- *Лексическая омонимия* - одинаково звучащие/пишущиеся слова, не имеющие общих элементов смысла,
  например: *рожа* – лицо и вид болезни.
- *Морфологическая омонимия* – совпадение форм одного и того же слова (лексемы), например, словоформа *стол* соответствует именительному и винительному падежам.
- *Лексико-морфологическая омонимия* – совпадение словоформ двух разных лексем, например, словоформа *стих* – глагол в единств. числе мужского рода и существительное в единств. числе, именит. падеже
- *Синтаксическая омонимия* – неоднозначность синтаксической структуры (и соответствующего смысла):
  Студенты из Львова поехали в Киев
  Flying planes can be dangerous (пример Хомского).

МОДЕЛИРОВАНИЕ в КЛ

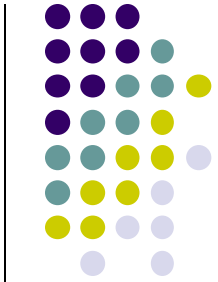


Модель языка/текста должна обладать структурным и/или функциональным подобием

Особенности моделей КЛ :

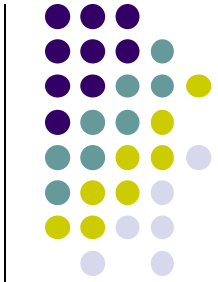
- Формальность (в отличие от лингвистических)
- Функциональность:
  цель – воспроизведение функций языка,
  а не моделирование языковой деятельности человека
- Общность модели, т.е. покрытие ею довольно большого множества текстов
- Ориентация на конкретные прикладные задачи КЛ
- Экспериментальная обоснованность (тестирование)
- Опора при создании модели или ее работе на те или иные лингвистические ресурсы

ПОДХОДЫ К ПОСТРОЕНИЮ МОДЕЛЕЙ КЛ



- Основанный на правилах, или инженерный:
  rule-based, knowledge-based
  - Модель – набор закодированных лингвистических правил
  - Правила создаются экспертами-специалистами
- Основанный на статистике и машинном обучении
  - Обычно необходим размеченный текстовый корпус
  - Виды обучения: – обучение без учителя (*unsupervised*)
    – обучение с учителем (*supervised*)
    – частичное обучение с учителем (*semi-supervised*)
  - Современный тренд: нейронные сети, глубокое обучение
  - Обученные модели плохо интерпретируемы
- Комбинирование подходов в промышленных решениях (гибридные модели)

ВИДЫ МОДЕЛЕЙ В КЛ



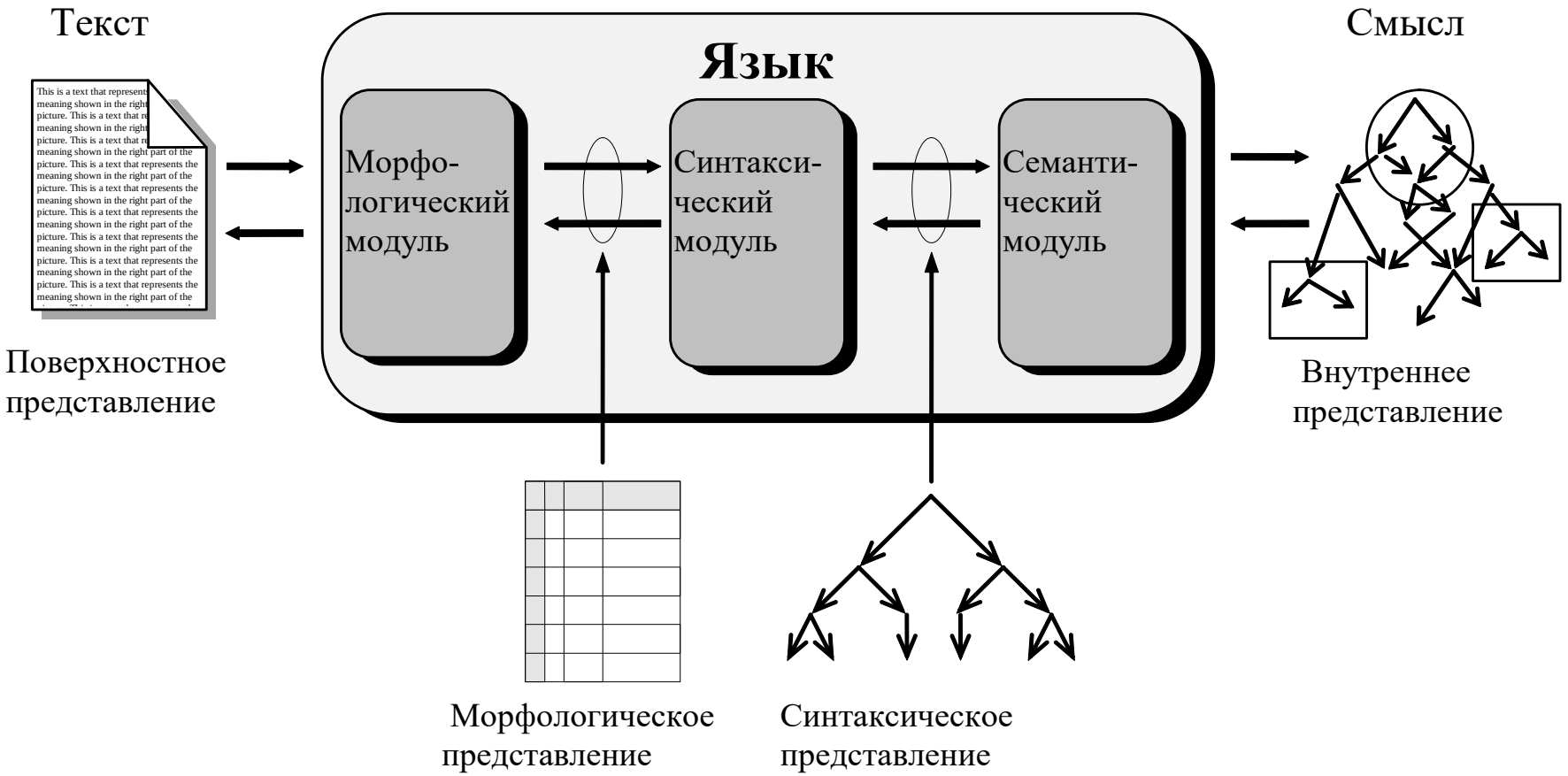
- Структурные, многоуровневые, многомодульные:
  multi-component, pipelined

  Модули могут быть созданы в рамках разных подходов и относятся к разным задачам/уровням/этапам:
  - модели морфологии, синтаксиса, семантики
  - современная тенденция: объединение уровней

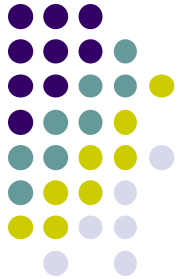
- Модели языка или Модели текста
  - *Language Model*: модель всего языка, строится по коллекции текстов, в частности: *статистическая языковая модель* опирается на статистику слов (или символов/букв) и их последовательностей – *N-грамм*
  - *Признаковая модель текста* учитывает набор лингвистических и статистических характеристик (признаков) текста, обычно в рамках текст.коллекции

ЭТАПЫ ОБРАБОТКИ ТЕКСТА В МНОГОУРОВНЕВЫХ МОДЕЛЯХ

Лингв. процессор – многоэтапный/многомодульный преобразователь (два направления: анализ и синтез)



АНАЛИЗ ТЕКСТА В МНОГОУРОВНЕВЫХ МОДЕЛЯХ

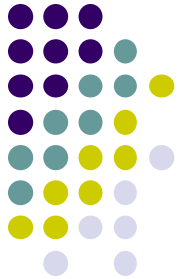


Уровни / Этапы анализа ~ Уровни языковой системы

- Графематический анализ и сегментация текста
- Морфологический анализ
  - Постморфологический анализ: разрешение морфологической омонимии
  - Предсинтаксис: сегментация на предложения, выделение синтаксических групп
- Синтаксический анализ предложений
- Семантический и дискурсивный анализ

❖ *глубина обработки* текста: количество уровней

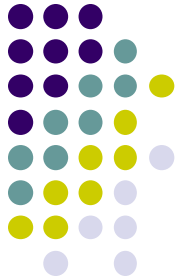
ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ



Источники лингвистической информации:

- Компьютерные и текстовые **словари**, различаются представленными единицами ЕЯ
  - словари синонимов: *бродить / шататься*
  - словари паронимов: *чужой / чуждый*
  - словари терминов предметной области: *интеграл*
  - словари устойчивых словосочетаний: *острая нехватка*, *задать вопрос*
- Тезаурусы и онтологии
- Грамматики ЕЯ – набор правил, описывающих синтаксическую структуру предложений: $S \Rightarrow NP\ VP$
- Коллекции и корпуса текстов, датасеты (*dataset*)

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: ТЕЗАУРУСЫ И ОНТОЛОГИИ



- Тезаурус – семантический словарь
  - *РуТез* – информационно-поисковый тезаурус: 52 тыс. понятий из общественно-политической области; семантич. связи: синонимия, род-вид, ассоциация
- Онтология – формальное описание некоторого набора понятий/сущностей (предметной области / всего мира)
  - *WordNet* – лингвистическая онтология на базе англ. слов: слова разбиты по частям речи, для каждой части речи выделены *синсеты* (синонимы), между которыми установлены семантич. связи
    Версия 3.0 – 155 тыс. лексем, 117 тыс. синсетов
  - *EuroNet* – аналоги для других европейских языков

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: КОРПУСА ТЕКСТОВ

Применение:

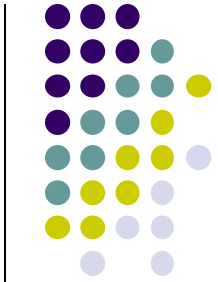
– машинное обучение моделей ЕЯ и текста

– автоматизация построения словарных ресурсов

- *Коллекция текстов* – набор объединенных по некоторому признаку текстов
  (например: нормативно-правовые документы)
- *Корпус текстов* – представительный массив текстов:
  - собран с учетом определенных свойств/принципов
  - обладает *лингвистической разметкой*: (морфологической, лексической, синтаксической,…)

  РЯ*: Национальный корпус русского языка, OpenCorpora*
- *Dataset* – набор данных для обучения, часто с разметкой

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ



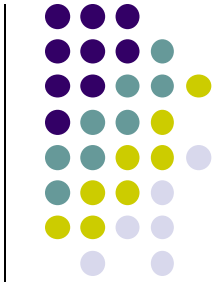
Традиционные направления:

- Машинный перевод
- Информационный поиск
- Реферирование и аннотирование текстов
- Автоматизация создания и редактирования текстов
- Формирование ответов на вопросы
- Генерация текстов на ЕЯ
- Организация диалога на ЕЯ, чат-боты

Text Mining:

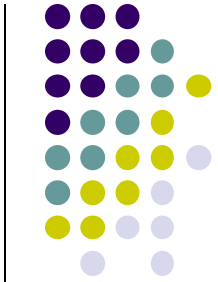
- Извлечение информации из текстов
- Классификация и кластеризация текстов
- Извлечение терминов и ключевых слов
- Анализ мнений и оценка тональности текстов

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: МАШИННЫЙ ПЕРЕВОД



- Драйвер КЛ, начало – Джорджтаунский эксперимент,1954 г.: перевод с русского на английский (словарь – 250 слов)
  Простейшая лингвист. модель: *пословный* перевод
- 50-60 гг. – двуязычные системы,
  пословный и *пословно-пооборотный* перевод
- 60-70 гг. – *пофразный* перевод, пред- и пост-редактирование человеком, промышленные системы
- 70-80 гг. – *многоязычные* системы
- 80-90 гг. – идея *интерлингвы* (семантического языка-посредника)
- 2000-2015 гг. – применение статистики, корпусов текстов: *статистическая трансляция*
- 2015-16 гг. – появление обученных *нейронных моделей*

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: ИНФОРМАЦИОННЫЙ ПОИСК

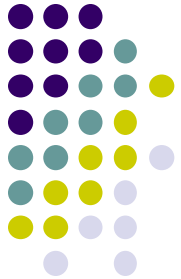


- Поиск в коллекциях текстовых документов – с 50-х гг.
  - *Индексирование* документа, т.е. выделение *ключевых слов и словосочетаний*, выполнялось вручную
  - *Поисковый образ* документа – *ключевые слова* (отражают основное содержание документа)
  - Поиск документа по запросу в виде набора ключ. слов

  Применяется в современных корпоративных информационных системах

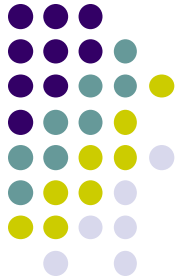
- Полнотекстовый поиск – с 90-х гг. (в сети Интернет)
  - *Автоматическое индексирование* текстов
  - Несколько моделей обработки текста и поиска, современная – с использованием нейросетей

ИНФОРМАЦИОННЫЙ ПОИСК: СМЕЖНЫЕ ЗАДАЧИ



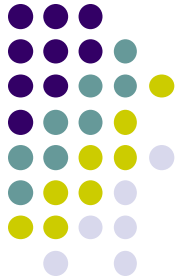
- Классификация текстов – отнесение к классам с заданными свойствами/параметрами
- Рубрицирование текстов – классификация, соотнесение с иерархической системой классов
- Кластеризация текстов – создание подмножеств тематически близких документов
- Построение *вторичных документов*:
  - Реферирование текста – построение краткого реферата для одного или нескольких текстов
  - Аннотирование текста – краткое описание содержания текста (упрощенно: список ключевых слов)

ПРИМЕНЕНИЕ КЛАССИФИКАЦИИ и КЛАСТЕРИЗАЦИИ



- Упорядочивание и навигация по набору документов
  - составление интернет-каталогов
- Информационный поиск
  - ограничение области поиска
  - «интеллектуальная» группировка результатов
- Фильтрация потока документов
  - фильтрация спама
  - выявление «искусственных» текстов (боты)
  - определение дубликатов документов
- Персонализированный подбор информации
  - контекстная реклама
  - новости об определенном событии и т.п.

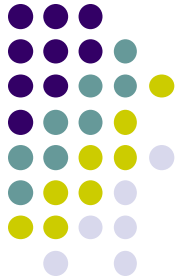
ПРИКЛАДНЫЕ ЗАДАЧИ: *INFORMATION EXTRACTION*



Извлечение информации (знаний) из текстов:

- Специфика задачи – выявление в текстовой коллекции информации, релевантной определенной проблеме, теме:
  - конкретных объектов (имен лиц, названий фирм и т.п.)
  - их отношений , связанных с ними событий и фактов:
    …прошла встреча…, …выдан кредит..
  - терминов и их связей, ключевых слов: *адресная шина*
- Извлеченные данные структурируются и визуализируются
- Приложения:
  - мониторинг новостных лент
    Сколько кораблей затонуло в текущем году?
  - аналитика экономической, производственной, финансовой деятельности

ПРИКЛАДНЫЕ ЗАДАЧИ: *OPINION MINING*

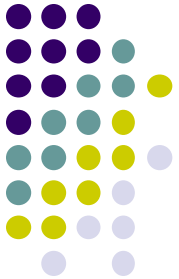


- Близко по целям и методам к направлению *Information Extraction*
- *Opinion Mining*
  - извлечение из текстов, в том числе сети Интернет (форумы, блоги и т.п.), мнений, отзывов, суждений (о персоналиях, товарах, услугах, фильмах, книгах и проч.)
  - их последующий анализ и классификация (например, по источнику/ тональности)
- *Sentiment Analysis* – анализ тональности текстов, т.е. определение их общей эмоциональной оценки:

  положительная, *отрицательная*, *нейтральная*

  о политиках, партиях, фирмах и компаниях и пр.

ПРИКЛАДНЫЕ ЗАДАЧИ: *QUESTION ANSWERING*



Ответы на вопросы –
сравнительно новая задача, актуальная

(но и забытое старое направление ИИ, 70 гг.)

- Нужен не документ или *сниппет*,
  а ответ на конкретный вопрос , например:
  Кто придумал вилку? ⇒ метапоиск
- Примерная стратегия построения ответа:
  – определение типа вопроса
  – построение запроса к интернет-поисковику
  – извлечение из найденных документов нужной информации
  – построение фразы ответа

ПРИКЛАДНЫЕ ЗАДАЧИ : *WRITING SUPPORT*

Автоматизация подготовки и редактирования текстов

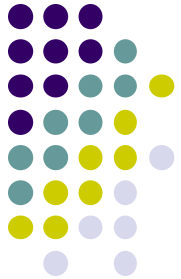
- Первые прикладные программы:
  - автоматическая простановка переносов слов
  - проверка орфографии (спеллеры, автокорректоры)
- Коммерческие системы: проверка орфографии (*Spelling*), выявление неправильного употребления предлогов и артиклей, частично проверка синтаксиса, иногда оценка сложности стиля

  английский – *Grammarly*, русский – *Орфограммка*
- Более сложные проверки, когда в результате описок или плохого знания языка возникают лексические ошибки:

  овальный/оральный; болотный/болотистый

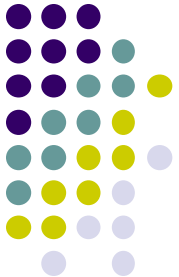
  По сути: неправомерная замена слов

ДРУГИЕ ПРИКЛАДНЫЕ ЗАДАЧИ



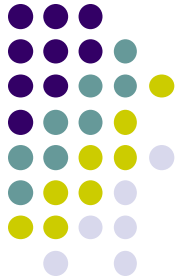
- Генерация текстов
  - Ранние системы на правилах: *FoG* (Канада) – двуязычная генерация текстов метеосводок (на английском и французском языках)
  - Современные системы генерации на базе обучения нейросетевых языковых моделей (GPT-2 и др.)
- Чат-боты, или разговорные агенты (виртуальные собеседники): *ELIZA (1965), A.L.I.C.E (2000-04)* и др. *Евгений Гусман (2016), Алиса* (Яндекс,2017)
- Компьютерная текстология: сравнение текстов, анализ изменений, определение авторства
- Автоматизация построения словарей
- Распознавание и синтез звучащей речи (учет *фонологического* уровня)

ЗАКЛЮЧЕНИЕ



- Появляются все новые прикладные задачи, требующие методов КЛ и анализа текста.
- В разных приложениях используются модели языка/текста, различающиеся по сложности.
- Современная тенденция – широкое применение машинного обучения, хотя подход на правилах также актуален (дает лингвистическую интерпретацию)
- Прорыв в качестве анализа текста связан с применением нейросетевых моделей
- Цель курса – дать интегральное представление о базовых и современных подходах КЛ, методах и инструментах для решения задач, а также понимание границ применимости разных моделей

СОДЕРЖАНИЕ КУРСА «КЛиАТ»



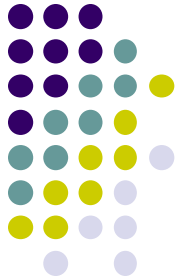
- Лекционный, теоретический материал (слайды):
  - ✓Основные модели и процессоры ЕЯ
  - ✓Статистические методы, машинное обучение
  - ✓Лингвистические ресурсы (обзорно)
  - ✓Методы разработки прикладных задач КЛ (не всех)

Контрольная работа – конец 3-го модуля/начало 4-го

Экзамен в конце 4 модуля (письменный)

Суммарно-балльная система определения оценки за курс

- Семинары: разбор задач по моделям, консультации, доклады по современным работам КЛ
- **Домашние задания** по темам курса (на 2-3 недели):
  - ✓ 4 основные, много вариантов на выбор (тестирование инструментов, программирование экспериментов и др.)
  - ✓ индивидуальное задание: презентация доклада или разработка прикладной программы КЛ



СПАСИБО ЗА ВНИМАНИЕ!

ЛИТЕРАТУРА

- Автоматическая обработка текстов на ЕЯ и анализ данных: учеб. пособие – М.: НИУ ВШЭ, 2017. https://miem.hse.ru/clschool/the_book
- Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Third Edition Draft, Prentice Hall, 2020. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- Генедара Т. Обработка естественного языка с помощью TensorFlow – М.: ДМК Пресс, 2019, 382 с.
- Риз Р. Обработка естественного языка на Java – М.: ДМК Пресс, 2016, 623 с.
- Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
- Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие – М.: МИЭМ, 2011.

http://clschool.miem.edu.ru/uploads/swfupload/files/98e8cdfb0288b275a3197626ffe06e277a03d43d.pdf