КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: МЕТОДЫ, РЕСУРСЫ, ПРИЛОЖЕНИЯ

Большакова Елена Игоревна
МГУ им. М.В. Ломоносова, Факультет ВМиК
bolsh@cs.msu.su

СОДЕРЖАНИЕ

1. Компьютерная лингвистика: истоки
2. Задачи компьютерной лингвистики (КЛ)
3. Особенности естественного языка
4. Моделирование в КЛ
5. Лингвистические ресурсы
6. Прикладные задачи КЛ

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: ИСТОКИ

- Междисциплинарная область: обработка ЕЯ
  - Автоматическая обработка тестов на ЕЯ
  - Машинная /Инженерная лингвистика
  - Вычислительная/ Компьютерная лингвистика
- Смежные области исследований
  - Лингвистика
    - Фонология (звуки речи)
    - Морфология (структура и форма слов ЕЯ)
    - Синтаксис (структура и функции предложений)
    - Семантика и прагматика (смысл и значение высказываний)
    - Лексикография (описание лексикона ЕЯ)
    - Психолингвистика
  - Математика
  - Информатика (Computer Science)
  - Искусственный интеллект

КЛ, МАТЕМАТИКА И ИСКУССТВЕННЫЙ ИНТЕЛЕКТ

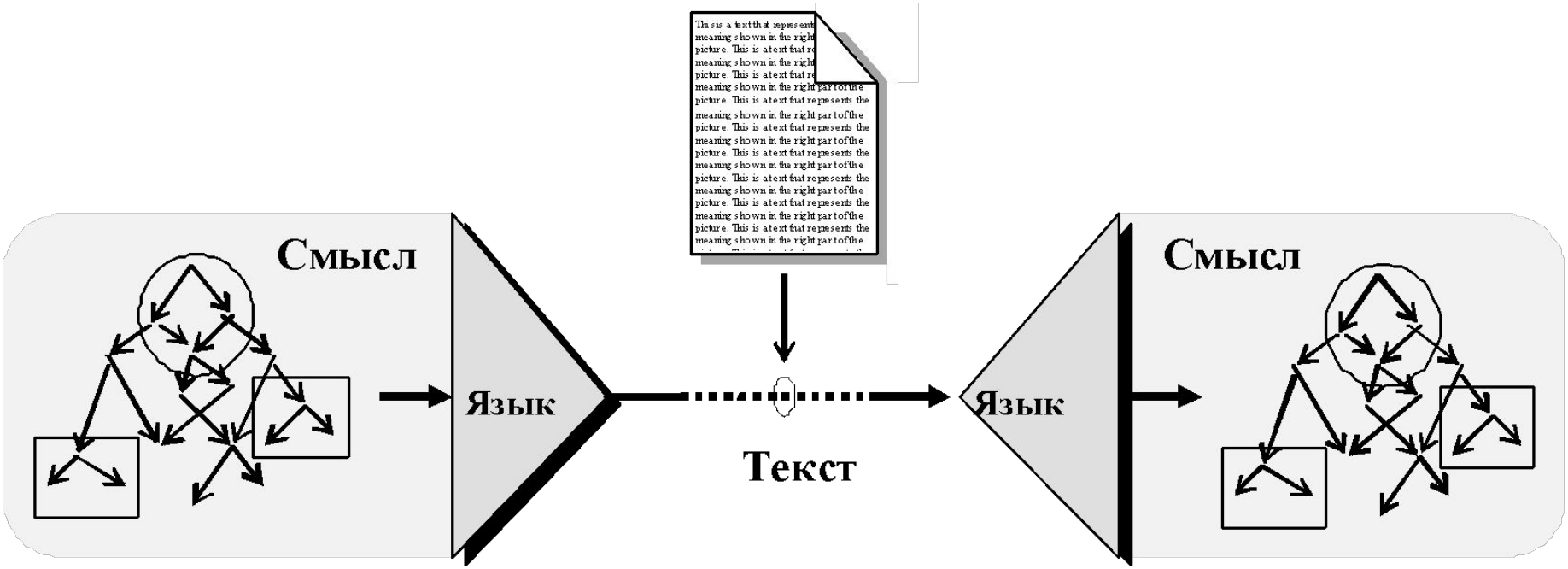
- Математика
  - Математическая лингвистика
    Порождающие (формальные) грамматики - Н. Хомский
    Квантитативная лингвистика
- Искусственный интеллект (ИИ)
  - Задача – компьютерные модели интеллектуальных функций
  - Часть *Computer Science,* пересечение (по задачам и методам) с КЛ
  - Первая известная работающая программа ИИ по обработке ЕЯ – система Т. Винограда (70-е годы); Пример диалога с системой:
    - Pick up a big red block. (человек)
    - OK
    - Is there a large block behind a pyramid?
    - Yes, Three of them.
    - Grasp the pyramid.
    - I don't understand, which pyramid you mean.

ЗАДАЧИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

- Разработка компьютерных программ для автоматической обработки текстов на ЕЯ – *лингвистических процессоров*
- Лигвистический процессор:
  Основа – формальная модель языка
  Зависимость от конкретного ЕЯ
  Пример: редактор *Word*, но не *NotePad*
- Сложность задач КЛ:
  - ЕЯ – сложная многоуровневая система знаков, возникшая для обмена информацией и постоянно изменяющаяся
  - Многообразие ЕЯ (способов выражения одного и того же смысла)

ОСОБЕННОСТИ ЕЯ: ПРЕОБРАЗОВАТЕЛЬ СМЫСЛ-ТЕКСТ

- Объект – *текст*
- Линейность текста
- Составлен из различных единиц
- Единицы принадлежат к разным уровням



ОСОБЕННОСТИ ЕЯ: УРОВНИ и ПОДУРОВНИ

- Синтаксический (предложения ЕЯ)
  - подуровень словосочетаний (*увидел лес*, *красивый закат*)
  - надуровень сверхфразовых единств (сложных синт. целых ≈ абзацев), объединяющихся по смыслу и лексико-грамматически (повторы слов, анафорические ссылки)
- Морфологический (слова ЕЯ, словоформы)
  - Подуровень морфем; м*орфема* – минимальная значимая часть слова (корень, приставка, суффикс…)
- Фонологический (звуки / символы)

? Уровни/ Срезы ?

- Семантический - набор элементарных единиц – *сем*
- Лексический: *лексема* – совокупность *словоформ* слова (*конь*, *коня*, *коню*, *коне*)
- *Дискурсивный* (связный текст) – схематические структуры текстов (патентные формулы, деловые письма и т.п.)

ЕЯ и ИСКУССТВЕННЫЕ ЯЗЫКИ

Искуств. языки, например: языки программирования
Близки по функциям, но

Принципиальные отличия:

- Открытость и изменчивость ЕЯ (на всех уровнях) ⇒ невозможность единожды разработать лингв. процессор
- Нестандартная сочетаемость (*синтактика*) единиц ЕЯ на всех уровнях, например, *лексическая* сочетаемость:
  крепкий чай, но не *тяжелый чай* (*heavy tea*)
- Большая системность (число уровней) и степень ассиметрии связи единиц и выражаемых ими смыслов
  - Полисемия (многозначность)
  - Синонимия (совпадение смыслов)
  - Омонимия (совпадение форм)

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 *круг* соответствует именительному и винительному падежам.
- *Лексико-морфологическая* омонимия – совпадение словоформ двух разных лексем, например, *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 – описание свойств обрабатываемого текста.

Особенности моделей КЛ:

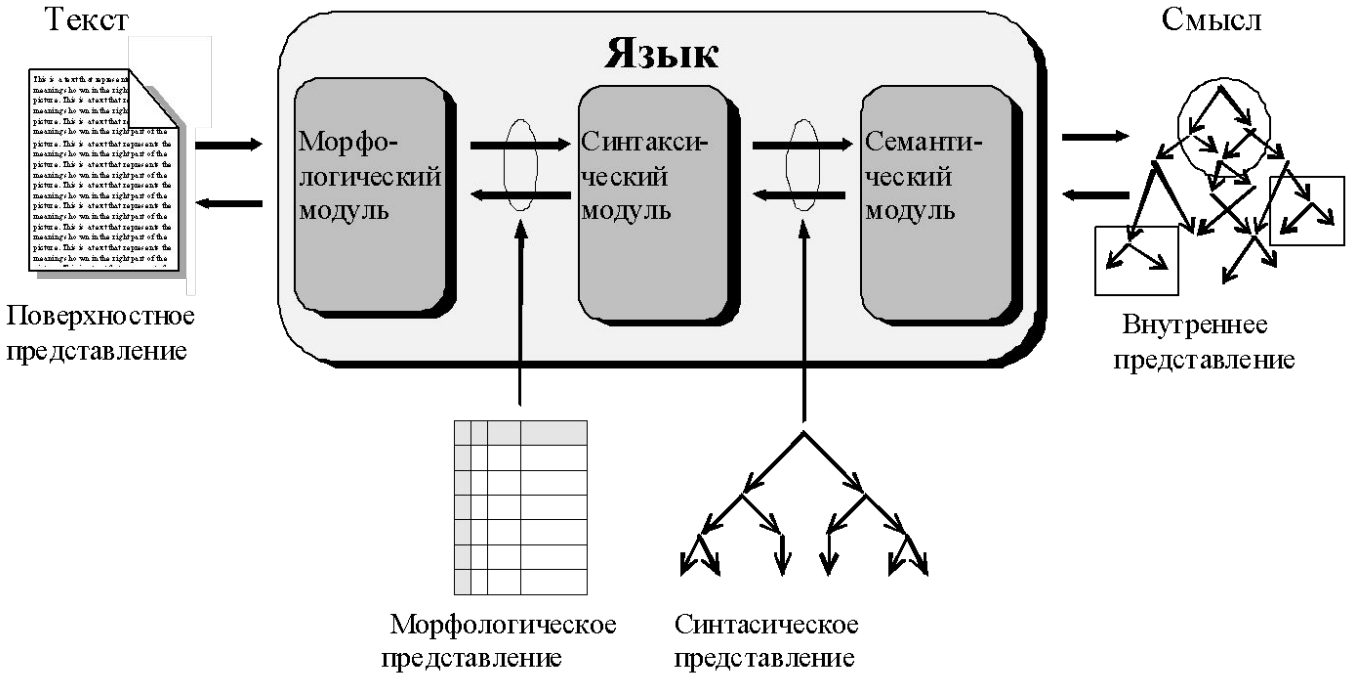
- Формальность и алгоритмизируемость;
- Функциональность: цель – воспроизведение функций языка как «черного ящика», а не моделирование языковой деятельности человека;
- Общность модели, т.е. покрытие ею довольно большого множества текстов;
- Экспериментальная обоснованность, предполагающая тестирование модели
- Опора на те или иные словари как обязательную составляющую модели.

МОДУЛЬНОСТЬ ЛИНГВ. ПРОЦЕСОРОВ

Сложность ЕЯ ⇒

лингвистический процессор – многоэтапный преобразователь

- Анализ текста: первичный модуль – графематический анализ
- Синтез теста: другое направление обработки



ВИДЫ И ОСОБЕННОСТИ МОДЕЛЕЙ

В зависимости от учета уровней ЕЯ:

- Структурные (несколько уровней)
- Редуцированные - Статистическая модель : статистика символов/букв, их биграмм и триграмм (уровень символов)
  или слов, их биграмм и триграмм
- Структурно-статистические

На разных уровнях ЕЯ:

- Модели морфологии (анализ: лемма или основа с морфологическими характеристиками исходной словоформы)
- Модели синтаксиса, анализ: синтаксическое дерево:
  - деревья непосредственно составляющих (валентности слов, например: *передать* - *кто? кому? что?* – subcategorization frame)
  - деревья зависимостей (валентности – модели управления слов)
- Модели семантики представление смысла (свойства, отношения, состояния, действия) – на основе моделей ИИ:
  формулы исчисления предикатов или семантические сети

МОДЕЛЬ «Смысл⇔Текст»

И. А. Мельчук, Ю. Д. Апресян (с 70-х годов)

Смысл – инвариант синонимичных преобразований текста.

- ориентация на синтез текстов
- многоуровневость модели, разделение основных уровней на поверхностный и глубинный уровень, например: *глубинный* (семантизированный) и *поверхностный* («чистый») синтаксис.
- Сохранение всей информации при переходе с уровня на уровень;
- *Лексические функции* для описания нестандартной синтактики, на их основе сформулированы правила синтаксического перифразирования;
- Упор на словарь, а не на грамматику; в словаре – информация для разных уровней языка (синтаксис: модели управления слов, описывающие их синтаксические и семантические валентности);
- Семантическое представление текста: семантический граф + коммуникативная организация смысла

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ

Лингвистические процессоры базируются на определенном представлении лингвистической информации:

- Компьютерные словари
- Грамматики ЕЯ
- Базы словосочетаний
- Тезаурусы и онтологии
- Коллекции и корпуса текстов

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: СЛОВАРИ и ГРАММАТИКИ

Словари для ЛП обычно разрабатываются специально .

Различаются:

- Охватом лексики: общая/специальная
- Представленной информацией (в словарной статье):
  - морфологические словари
  - словари моделей управления
- Видом:
  - словари синонимов:
  - словари паронимов: *чужой* и *чуждый*, *правка* и *справка*
  - словари терминов некоторой предметной области

Грамматики – набор правил, описывающих структуру предложений:

Пример:

SUBJECT\|gender 1 ^, number 1 ^, case 1 ^\|<1:;SBJ1;gender 1 +,number 1 +,case 1 +>\|<1:;SPRE;gender 1 +,number 1 +, case 1 +>\|<1:;SPOST;gender 1 +,number 1 +, case 1 +>

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: БАЗЫ СЛОВОСОЧЕТАНИЙ

Сравнительно новый тип лексического ресурса,

Отражает стандартную и нестандартную сочетаемость слов ЕЯ

Обширная база словосочетаний РЯ – система **КроссЛексика**

- Примерно миллион словосочетаний общей лексики
- Словосочетания многих синтаксических типов:

  - определяемое слово → определитель (*полевая форма*, *вполне удачный*)
  - существительное → его дополнение (*рост возмущения*)
  - глагол → его дополнение (*заметить разницу*, *решить продать*)
  - прилагательное → его дополнение (*дошедший до ручки*)
  - сочиненная пара (*наземный и воздушный*, *орел и решка*)
- Семантические связи слов: синонимы, антонимы, гиперонимы, холонимы
- Пометы стиля слов (устарелый, разговорный, бранный, и т.п).

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: ТЕЗАУРУСЫ И ОНТОЛОГИИ

- Тезаурус – семантический словарь
  - **РуТез** – информационно-поисковый тезаурус, 52 тыс. понятий из общественно-политической области; связи: синонимия, род-вид (выше-ниже), ассоциация, онтологическая зависимость,
  - **КроссЛексика** (поскольку представлены смысловые отношения)
- Онтология – формальное описание определенного набора понятий, сущностей
  - **WordNet** – лингвистическая онтология на базе английских слов

    - Дж. Миллер, 1984 г., модель человеческой памяти
    - слова разбиты по частям речи
    - для слов каждой части речи выделены *синсеты* – наборы синонимов
    - версия 3.0 – 155 тыс. лексем, 117 тыс *синсетов* (понятий)
  - **EuroNet** – аналогичные лексические ресурсы для других европейских языков

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: КОРПУСА ТЕКСТОВ

Трудоемкость создания лингвистических процессоров и лексических ресурсов $\Rightarrow$

автоматизация их построения

- *Коллекция текстов:* представительный набор текстов, собранных по определенному принципу
- *Корпус текстов:* коллекция текстов с лингвистической разметкой: морфологической, лексической, синтаксической, дискурсивной

  - использование в лингвистических исследованиях

  - применение для машинного обучения моделей

  - для РЯ – Национальный корпус русского языка
- *Интернет-корпус:* тексты сети Интернет как корпус современной речи

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ

- Машинный перевод
- Информационный поиск
- Классификация и кластеризация текстов
- Реферирования и аннотирование текстов
- Формирование ответов на вопросы
- Автоматизация подготовки и редактирования текстов
- Извлечение информации из текстов
- Генерация текстов на ЕЯ
- Организация диалога с пользователем на ЕЯ
- Обучение ЕЯ
- Распознавание и синтез звучащей речи

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: МАШИННЫЙ ПЕРЕВОД

Самое раннее приложение, 50е годы

Большое количество исследований

- Простейшие модели – стратегия пословного перевода и ее модификации; дает приемлемое качество только для родственных языков (испанский-португальский)
- Концепция внутреннего *языка-посредника* для задач многоязыкового перевода (для европейских языков)
- Одна из наиболее полных лингвистических моделей перевода: отечественная система **ЭТАП** (языки - русский и французский, научно-технические тексты, основана на лингв. модели «Смысл⇔Текст»
- Современное направление – *статистическая трансляция* (переводчик поисковика Google)

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: ИНФОРМАЦИОННЫЙ ПОИСК

Полнотекстовый поиск

- *Поисковый образ* документа – *ключевые слова* (отражают основное содержание документа)
- *Запрос на поиск* документов – набор ключевых слов
- Результат поиска – релевантные документы
- ***Индексирование*** документа

  - выделение ключевых слов и словосочетаний (вручную человеком или автоматически, для этого – статистические и лингвистические критерии)

  - выделение всех знаменательных слов (поиск в сети Интернет)
- *Векторная модель* текста – набор слов (*bag of words*)

ИНФОРМАЦИОННЫЙ ПОИСК: СМЕЖНЫЕ ЗАДАЧИ

Используется, как правило, векторная модель текста

- Классификация текстов – отнесение к классам с заданными свойствами/параметрами
- Рубрицирование текстов – классификация, соотнесение с иерархической системой классов
- Кластеризация текстов – создание подмножеств близких тематически документов

- Для решения – методы машинного обучения
- Приложения: выявление спама и др.

Эти задачи относят к научному направлению

Text Mining – часть *Data Mining*

ИНФОРМАЦИОННЫЙ ПОИСК: РЕФЕРИРОВАНИЕ, АННОТИРОВАНИЕ

- Реферирование текста – построение краткого реферата для одного или нескольких тематически связанных текстов
  - основная стратегия – отбор наиболее значимых предложений
  - сложности: учет анафорических ссылок
- Аннотирование текста
  - *аннотация* – вторичный документ, еще более краткий, чем реферат
  - в простейшем случае – перечень основных тем/ключевых слов документа

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: QUESTION ANSWERING

Ответы на вопросы – сравнительно новая задача ИП и КЛ, очень актуальная

(но и хорошо забытое старое направление ИИ)

- Нужен не документ или сниппет, а ответ на конкретный вопрос , например: *Кто придумал вилку?*
- Примерная стратегия построения ответа:

- определение типа вопроса, и запрашиваемого понятия
- построение запроса к интернет-поисковику
- извлечение из найденных документов нужной информации
- построение фразы ответа

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ : WRITING SUPPORT

Автоматизация подготовки и редактирования текстов

- Первые программы:
  - автоматическая простановка переносов слов
  - проверка орфографии (спеллеры, автокорректоры)
- Коммерческие системы: проверка орфографии , частично – синтаксиса, а также – сложности стиля
- Исследовательские разработки:
  - выявление неправильного употребления предлогов (использование моделей управления)
  - обнаружение сложных лексических ошибок (описки, приводящие к другим словам: *овальный/оральный,* паронимические ошибки*: болотный/болотистый*

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ : INFORMATION EXTRACTION

Извлечение информации (знаний) из текстов:

- Специфика задачи – распознавание и выявление в тексте определенной значимой информации:

  - именованных сущностей: имен лиц, названий фирм и учреждений, географических названий, дат и т.п.

  - отношений (связей) выделенных сущностей, например: *работать в*, *давать кредит*

  - связанных с ними событий и фактов

- Частичный синтаксический анализ и лингвистические шаблоны, например: *N работать в NP*
- Близкая задача – выявление терминов-понятий и их определений: *число с плавающей точкой*

ДРУГИЕ ПРИКЛАДНЫЕ ЗАДАЧИ КЛ

- Opinion Mining и Sentiment Analysis :

  - выделение мнений (о товарах, фильмах и проч.) в форумах, блогах и т.п.

  - оценка тональности текста (контент-анализ)

- Автоматическая генерация текстов на ЕЯ

  - многоязыковая генерация инструкций, руководств пользователя, патентных формул

- Диалог с пользователем на ЕЯ

  - запросы к специализированной базе данных (язык ограничен лексически и грамматически)

- Обучение ЕЯ (отдельные уровни и модели)

- Распознавание и синтез звучащей речи:

  учет фонологического уровня, использование моделей морфологии

ЗАКЛЮЧЕНИЕ

- Расширяющийся круг прикладных задач КЛ, рассмотренные приложения: осязаемые результаты
- В основном используются простые и редуцированные модели языка

  Причина: трудоемкость разработки сложных моделей, неэффективность соответствующих алгоритмов
- Современная тенденция - применение машинного обучения, которое дополняет
- Традиционный подход – *rule-based* (основанный на правилах, имеющих лингвистическую интерпретацию)

СПАСИБО ЗА ВНИМАНИЕ!