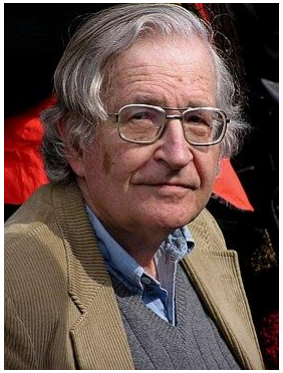
Автоматическая обработка текстов на естественном языке (АОТ) (ОЕЯ)

NATURAL LANGUAGE PROCESSING (NLP)

Введение



Аврам Ноам (Наум) Хомский (1928 год рождения) – американский лингвист

Статья «Синтаксические структуры» (1957)

60-е годы – системы, понимающие команды на естественном языке, системы машинного перевода

70-е годы – Марковские модели языка

90-е годы – активное развитие поисковых систем

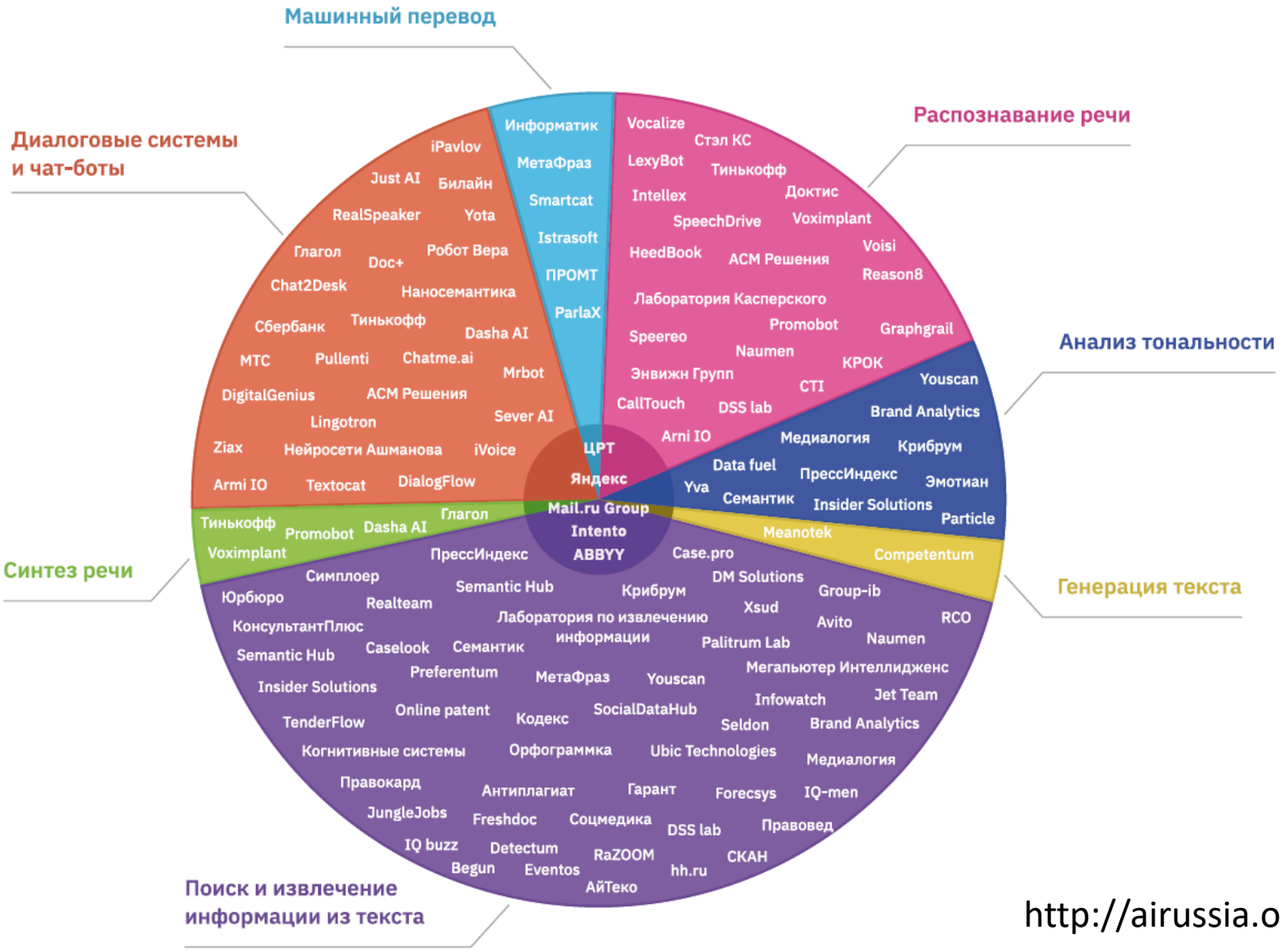
2010-е годы – глубокое обучение

2017 год – архитектура трансформера

Задачи, в которых необходима обработка текста

- Информационный поиск
- Машинный перевод
- Реферирование (аннотирование) текста
- Классификация и кластеризация:
  - Фильтрация спама
  - Анализ тональности
  - Рубрицирование текста
- Извлечение информации
- Диалоговые системы «Вопрос – ответ» (чат-боты)
- Автоматизация подготовки и редактирования текстов (проверка орфографии, подстановка следующего слова)
- Автоматическая генерация текста

Карта компаний и технологий



http://airussia.online/

Российские компании

Яндекс – информационный поиск, перевод, Алиса

ЦРТ – распознавание и синтез речи

ABBYY – работа с документами

Mail.ru – информационный поиск, Маруся, рекламный таргетинг

Just AI – чат-боты

PROMT – перевод

Тинькофф – распознавание и синтез речи, голосовой помощник Олег

Наносемантика – чат-боты

Примеры приложений

ПАО “Сбербанк” внедрило в систему онлайн-мониторинга новостей интеллектуальные технологии

Amazon разработал сервис для извлечения сложных медицинских данных из неструктурированного текста

Компания «Гарант» разработала новый автоматизированный сервис по подбору судебной практики

Brand Analytics разработал систему мониторинга и анализа упоминаний в социальных медиа в режиме реального времени

Диалоговые ассистенты

Действия над текстом

Сегментация – выделение в тексте лексических единиц (слов (графематический анализ), предложений)

Морфологический анализ – определение грамматических характеристик слова (определение частей речи)

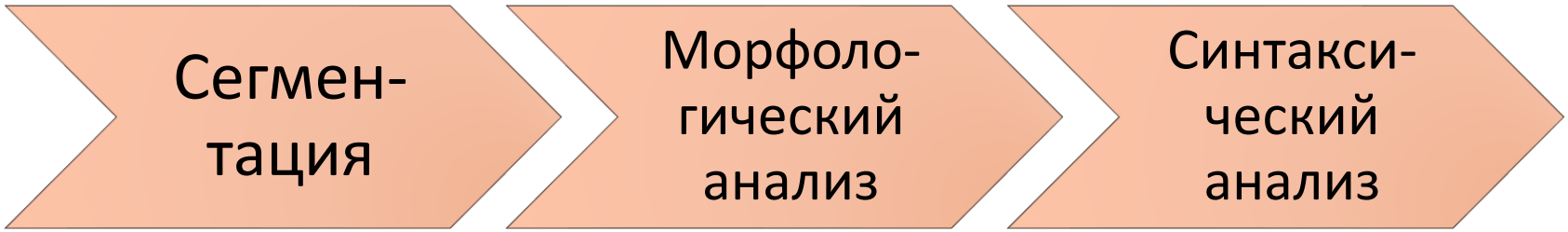
Лемматизация – определение начальной формы слова (леммы)

Синтаксический анализ – выявление грамматической структуры предложения

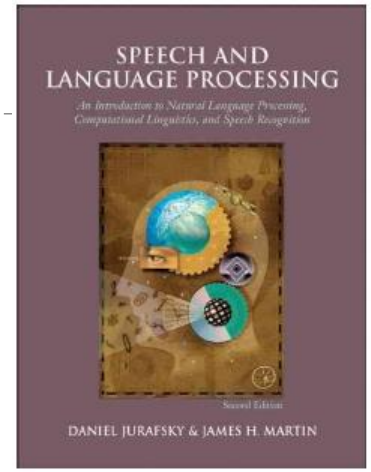
Семантический и прагматический анализ –определение смысла фраз

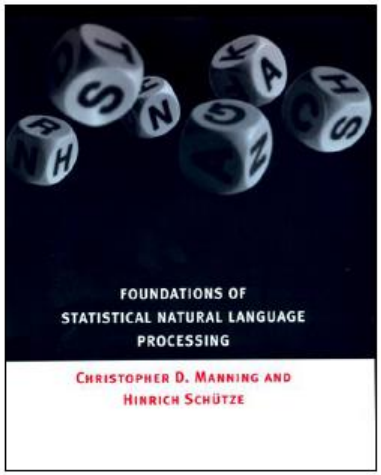
Лингвистический конвейер

Примененные последовательно действия над текстом образуют лингвистический конвейер (pipeline)

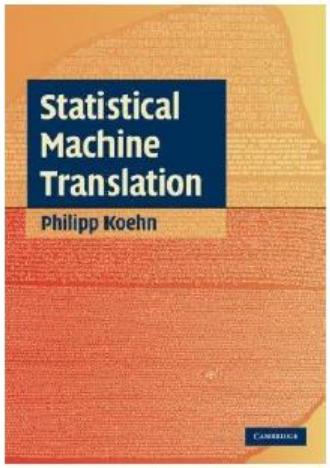


Литература

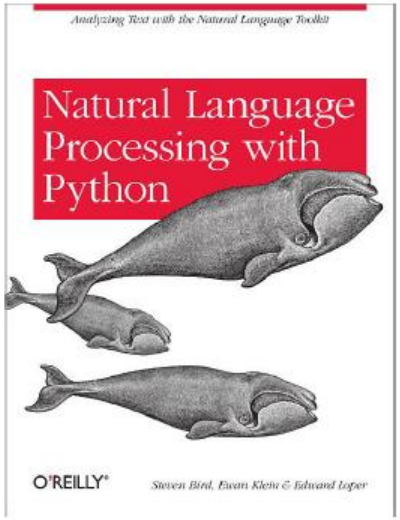

SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


FOUNDATIONS OF
STATISTICAL NATURAL LANGUAGE
PROCESSING
CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE


Кристофер Д. Маннинг,
Прабхакар Рагхаван,
Хайнрих Шютце
Введение
в
информационный
поиск
Яндекс


Statistical
Machine
Translation
Philipp Koehn
CAMBRIDGE


Искусственный
интеллект
Современный подход
ВТОРОЕ ИЗДАНИЕ
Стюарт Рассел • Питер Норвиг


Analyzing Text with the Natural Language Toolkit
Natural Language
Processing with
Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper


ПРИКЛАДНАЯ
И
КОМПЬЮТЕРНАЯ
ЛИНГВИСТИКА
• Компьютерная морфология
• Компьютерный синтаксис
• Компьютерное представление значений
• Распознавание и синтез речи
• Машинное обучение в лингвистике
• Корпусная лингвистика
• Машинный перевод
• Информационный поиск
• Извлечение информации
• Диалоги и чат-боты
• Анализ тональности
• Компьютерная текстология
• Квантитативная лингвистика: что можно сосчитать в языке и речи?
• Речевое воздействие и манипулирование

Литература

Daniel Jurafsky and James H. Martin. Speech and Language Processing, 2009. (SLP) Черновики готовящегося 3-го издания.

Chris Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, 1999. (FSNLP)

Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск, 2011. (IIR) Английский оригинал + слайды.

Philipp Koehn. Statistical Machine Translation, 2010. (SMT)

Стюарт Рассел, Питер Норвиг. Искусственный интеллект: современный подход, 2-е издание, 2015. Сайт оригинальной книги (3-е издание).

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. Прикладная и компьютерная лингвистика. Николаев И.С., Митренина О.В., Ландо Т.М. (Ред.), 2016.

Прикладная и компьютерная лингвистика / Под ред. И.С.Николаева, О.В.Митрениной, Т.М.Ландо. — М.: ЛЕНАНД, 2017. — 320 с.

Русский язык

Языки мирового значения [править | править код]

Современными международными языками можно считать[6] (в порядке убывания общего количества владеющих языком):

Ранг	Язык	Родной	Второй	Общее число носителей
1	Китайский язык[8]	1,2 миллиарда	до 300 миллионов	до 1,5 миллиарда
2	Английский язык[6]	500 миллионов	до 1 миллиарда	до 1,5 миллиарда
3	Испанский язык[9]	425 миллионов	до 125 миллионов	до 550 миллионов
4	Арабский язык[10]	300 миллионов	до 120 миллионов	до 420 миллионов
5	Португальский язык[11]	230 миллионов	до 30 миллионов	до 260 миллионов
6	Русский язык	160 миллионов	до 100 миллионов	до 260 миллионов
7	Немецкий язык[12]	120 миллионов	до 80 миллионов	до 200 миллионов
8	Французский язык[13]	75 миллионов	до 195 миллионов	до 270 миллионов

Content languages for websites
03.03.2021

Rank	Language	Percentage
1	English	60.6%
2	Russian	8.3%
3	Turkish	3.9%
4	Spanish	3.8%
5	Persian	3.3%
6	French	2.7%
7	German	2.3%
8	Japanese	2.1%

Проблемы

Язык постоянно изменяется

Есть правила, но много исключений

Полисемия – многозначность

Синонимия – полное или частичное совпадение значений разных единиц

Омонимия – совпадение по форме двух разных по смыслу единиц

Омонимия

Синтаксическая омонимия –

- *Студенты из Львова поехали в Киев*
- *Только рупор капитана их к отплытью призовет*

Лексико-морфологическая омонимия – *стих* – глагол в единственном числе мужского рода и существительное в единственном числе, именительном падеже

Морфологическая омонимия – словоформа *круг* соответствует именительному и винительному падежам

Лексическая омонимия – *ключ* – источник воды и объект для открытия замка (тоже омонимия)

Схема решения задач

Каждый шаг – это исследовательский поиск

Для решения необходимы: Коллекция текстов (датасет), методика, метрики качества

Датасет размечается

Датасет разбивается на Тренировочный и Тестовый

Лингвистические ресурсы

Корпусы текстов

Морфологические словари

Тезаурусы и онтологии

Грамматики

Корпус текстов

Корпус текстов – это коллекция текстов, собранная по определенному принципу представительности (по жанру, авторской принадлежности и т.п.), в которой все тексты размечены, т.е. снабжены некоторой лингвистической разметкой (аннотациями) – морфологической, акцентной, синтаксической и т.п.

Национальный корпус русского языка – http://www.ruscorpora.ru/

Национальный корпус русского языка

В проекте участвуют специалисты Института русского языка им. В. В. Виноградова РАН (ИРЯ РАН), Национального исследовательского института «Высшая школа экономики» (ВШЭ), Института проблем передачи информации РАН (ИППИ РАН), Института лингвистических исследований РАН (ИЛИ РАН) в Санкт-Петербурге, Воронежского государственного университета.

Программную и организационную поддержку проекту с его основания оказывает компания «Яндекс».

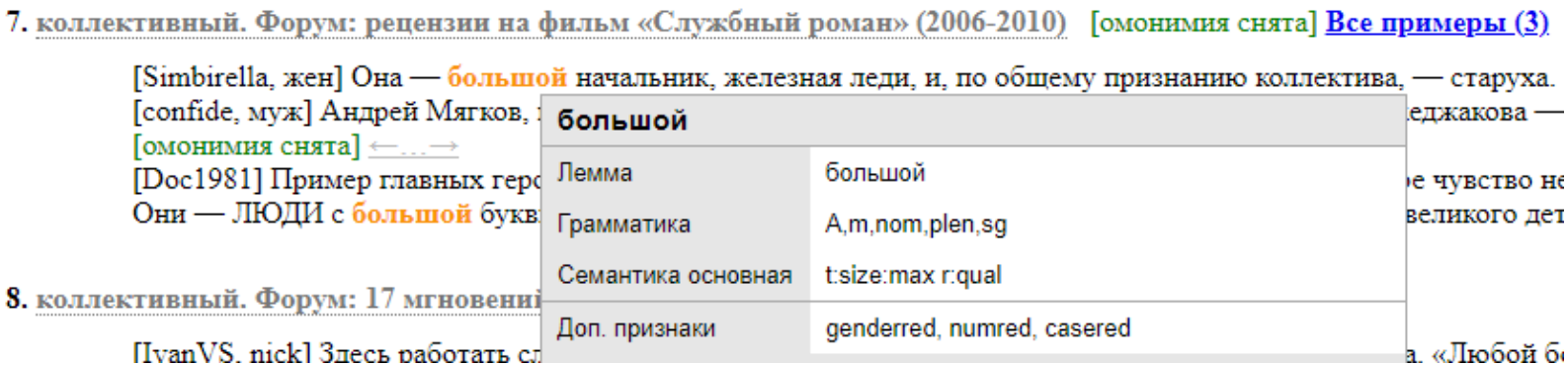
Число текстов – 4 304 893*

Число предложений – 117 204 099

Число словоупотреблений – 1 513 115 610

* на дату 05.09.2022

Национальный корпус русского языка



OpenCorpora

предложений: 110304

токенов: 1989538

слов: 1539979