Введение в обработку естественного языка
Introduction to natural language processing

Павел Браславский

Немного о себе

2001 – : Уральский федеральный университет
2005 – 2012: Яндекс
2012 – 2017: Kontur labs
2018 – : СПбАУ & JetBrains Research

Information Retrieval & Natural Language Processing

query log mining, evaluation, automatic summarization & snippet generation, CQA, wordnets, computational humor, …

RuSSIR, книга «Введение в информационный поиск»

http://kanas.ru/pb

Онлайн-курс

- https://stepik.org/course/1233/
- Класс: http://bit.ly/spbau2018

ИДЕОЛОГИЯ КУРСА

Основные идеи

- Вводный курс
- Курс на русском с учетом специфики русского языка
- Базовые инструменты + приложения
- Лекции + практические задания

Актуальность ОЕЯ

- большие объемы текстовых данных (1990-е – веб)
- мобильные технологии → системы общения (с 2015?)

Русский язык

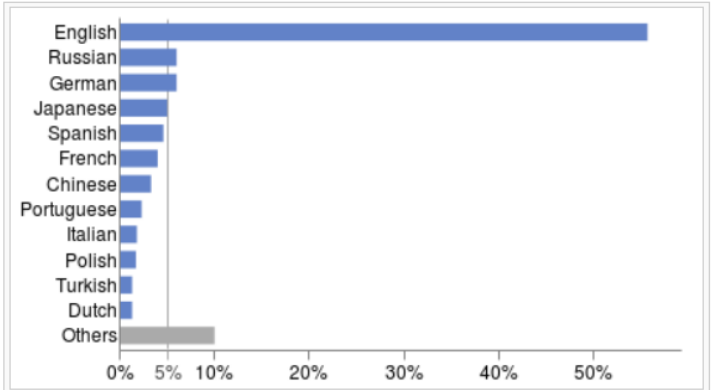
French	Indo-European, Romance	80 million in France, Belgium, Luxembourg, Switzerland and Canada (2014)~ Re: Victor	194 million - 134 as L2 in French-speaking countries - 60 as a foreign language in the world	274 million	One of the six official languages of the United Nations.
Russian	Indo-European, Slavic	180 million in Russia, Belarus, Kazakhstan, Kirghizstan (2014)	80 million - 65 in former USSR states but in decline (2014) - 15 as a foreign language in the world	260 million	One of the six official languages of the United Nations
Bengali	Indo-European, Indo-Aryan	250 million (150 in Bangladesh and 100 in east India) (2014)	20 million L2 s B c		

#8

Content languages for websites [edit]

Estimated percentages of the top 10 million websites using various content languages as of 18 March 2015:[2]

Rank	Language	Percentage
1	English	55.5%
2	Russian	5.9%
3	German	5.8%
4	Japanese	5.0%
5	Spanish	4.6%
6	French	4.0%
7	Chinese	2.8%
8	Portuguese	2.5%
9	Italian	1.9%



Content languages for websites as of 12 March 2014[2]

Особенности РЯ

- Флективный язык → более свободный синтаксис
- Мало ресурсов (!)

Чего нет в этом курсе?

- лингвистической теории
- обработки звучащей речи
- некоторых популярных приложений
  - классификация текстов
  - исправление опечаток
  - чат-боты
  - …
- нейронных сетей/глубокого обучения

Организация материала

- Инструменты
  - морфологический анализ
  - синтаксический анализ
  - языковые модели
  - лексическая семантика
- Приложения
  - информационный поиск
  - вопросно-ответный поиск
  - автоматическое реферирование
  - анализ тональности
  - извлечение информации
  - машинный перевод

Практические задания

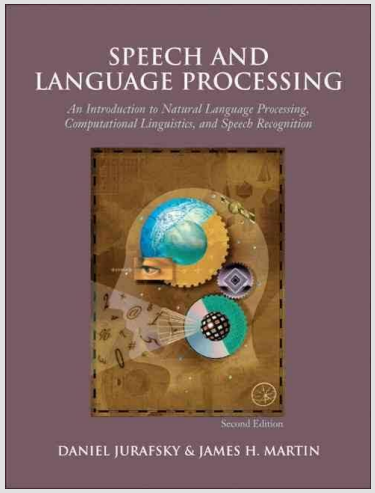
- Морфологический анализ
- Автоматическое реферирование
- Извлечение именованных сущностей
- Анализ тональности
- Машинный перевод*

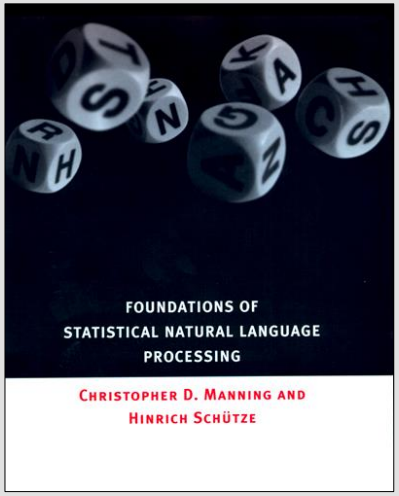
Что нужно знать и уметь?

- линейная алгебра
- теория вероятностей, статистика
- машинное обучение
- лингвистика на «школьном уровне»
- навыки программирования

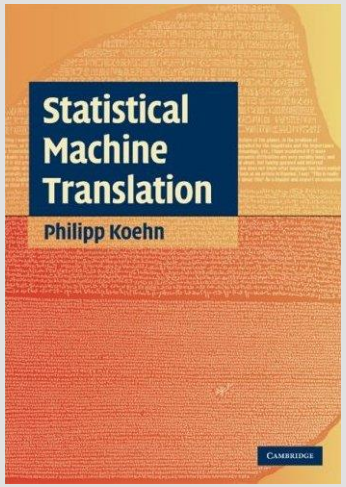
РЕСУРСЫ

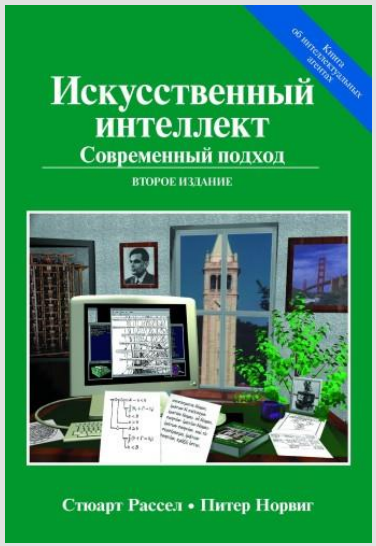
Книги

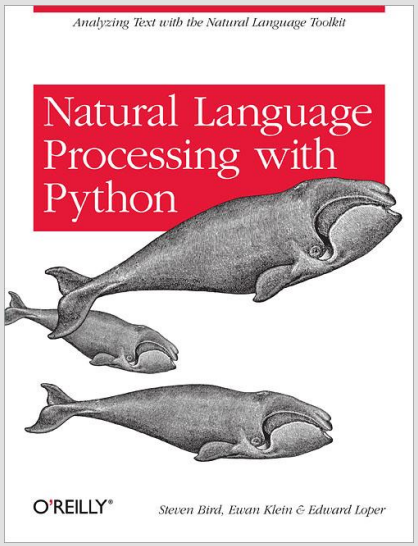

SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


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
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
URSS
URSS

Видеокурсы

- Dan Jurafsky & Chris Manning (2012)
- Michael Collins (2013)
- Dragomir Radev (2015)
- Chris Manning & Richard Socher (2017)

Ресурсы

- ACL, COLING, EACL, NAACL, EMNLP, RANLP… (http://aclweb.org/anthology/)
- AAAI, IJCAI
- SIGIR, CIKM, WSDM, ECIR
- INTERSPEECH
- Диалог (http://www.dialog-21.ru/), AINL
- Яндекс, Mail.Ru
- habrahabr по тегу «лингвистика»
- ПостНаука

ЧТО ТАКОЕ ОЕЯ?

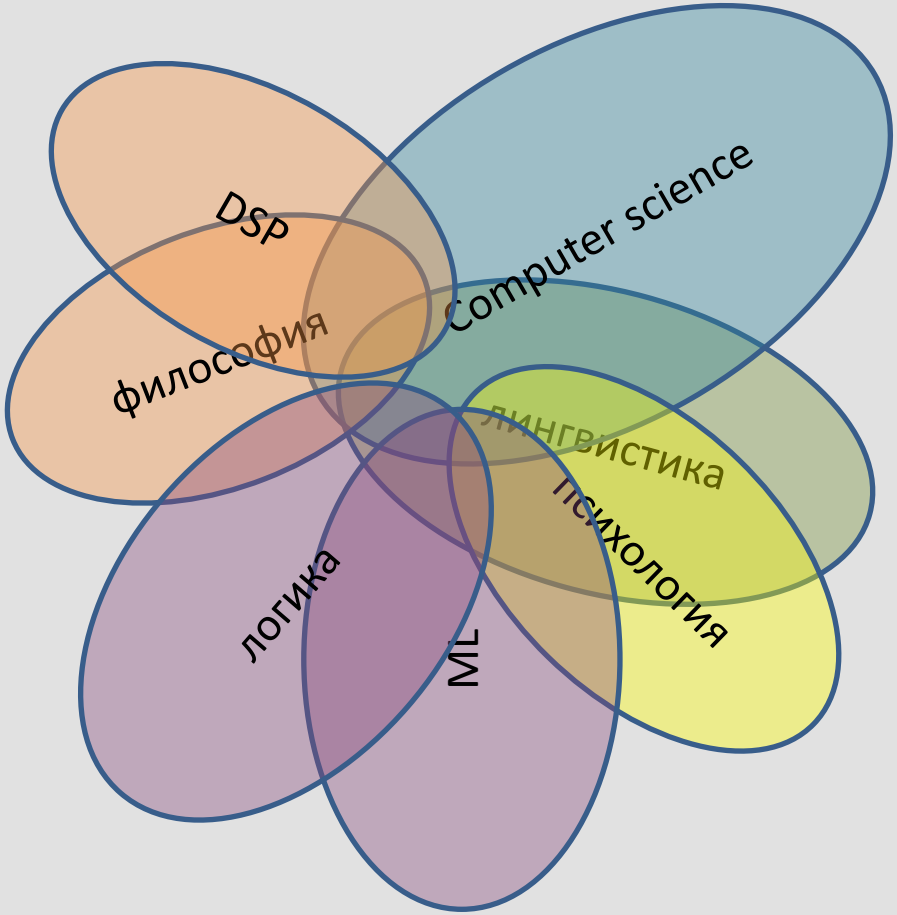
Термины

- Computational linguistics / математическая/компьютерная лингвистика
- **Natural language processing / обработка естественного языка** /автоматическая обработка текстов
- Natural language engineering, human language technology
- Прикладная лингвистика
- Speech and language processing
- Speech recognition and synthesis / распознавание и синтез речи

Чем занимается

Автоматическая обработка речи
с использованием знаний о языке

Междисциплинарная область



Значение ОЕЯ

- Связь языка и сознания
- Объем текстовых /речевых данных
- Мобильные технологии
- Многоязычие

Разделы знаний о языке

- Фонетика
- Морфология
- Синтаксис
- Семантика
- Прагматика

Инструменты vs. Приложения

Приложения

- Диалоговые системы / системы общения
  conversational agents / dialog systems
- Вопросно-ответные системы
  question answering
- Информационный поиск
  information retrieval
- Машинный перевод
  machine translation

Приложения – 2

- Извлечение информации
  information extraction
- Анализ тональности
  sentiment analysis
- Автоматическое реферирование
  automatic summarization
- Обучение языку
  language learning

Почему ОЕЯ – это сложно

- Неоднозначность (ambiguity)
- Многие задачи ОЕЯ можно рассматривать как задачи снятия неоднозначности (disambiguation):
  - *Печь* – существительное или глагол?
  - *Лук* – овощ, оружие или фотография?
  - Только рупор капитана //их к отплытью призовет. – *призовет капитана* или *рупор капитана*?
  - *Скрипка, лиса* или *скрип колеса*?

Методология ОЕЯ

- Правила ←→ статистика
- Основные модели: конечные автоматы, системы на основе правил, логика, вероятностные модели, векторное представление
- Основные методы: поиск в пространстве состояний (динамическое программирование), машинное обучение

КРАТКАЯ ИСТОРИЯ

1940-е и 1950-е

- Язык изучают разные науки: радиотехника, информатика/кибернетика, лингвистика, психология, философия
- Теория формальных языков, КСГ (CFG)
- Теория информации
  - Канал с помехами (noisy channel), теория кодирования

1957-1970

- Ноам Хомский
- ОЕЯ в рамках ИИ
  - «игрушечные» системы на правилах
- Байесовские методы (определение авторства)
- Брауновский корус (Brown corpus)
  - 1M слов (1964)

1970-1983

- Статистический подход (HMM, noisy channel, …)
- Логика
- Понимание языка (от синтаксиса к семантике)
- Моделирование дискурса

1983-1993

- Конечные автоматы (FSA) в морфологии, фонологии и снтаксисе
- Подходы «от данных» (data-driven), новые стандарты в оценке (evaluation)
- Генерация речи

1994-1999

- Вероятностные методы ++
  - Частеречная разметка (POS tagging), синтаксический анализ (parsing), разрешение анафоры (anaphora resolution), …
- Приложения +
- Веб

2000-2008

- Доступные данные
- Мероприятия по оценке
- Взаимодействие с сообществом ML
- Высокопроизводительные системы
- Подходы «без учителя» (unsupervised) – topic modeling
- Статистический машинный перевод (SMT)

2008 –

- Deep learning ++
- Приложения ++
- Индустрия

ОЕЯ в СССР и России

- Теория «Смысл <=>Текст» (Игорь Мельчук, 1960-е гг.)
  - Машинный перевод (ЭТАП), синтаксический анализ
- Информационный поиск
  - Рамблер, Яндекс, Mail.Ru
- Распознавание речи
  - ЦРТ, …
- Information Extraction
  - Интегрум, Медиалогия, Крибрум, …
- Машинный перевод
  - ЭТАП, ПРОМТ, Яндекс, ABBYY
- Корпусы
  - НКРЯ, opencorpora, ГКРЯ
- Инструменты
  - АОТ, pymorphy
- Инициативы по оценке
  - РОМИП, Dialog evaluation

Академические центры

- Отделение теоретической и прикладной лингвистики МГУ
- Институт лингвистики РГГУ
- Школа лингвистики ВШЭ
- Кафедра математической лингвистики СПбГУ