Автоматическая обработка естественного языка
Введение

Екатерина Владимировна Еникеева
protoev@yandex.ru
https://github.com/named-entity/hse-nlp

3 сентября 2021

О чем этот курс

АОЕЯ / AOT = Natural Language Processing (NLP)

- ➢ особое внимание к русскоязычным данным

- 3 курс:
  - базовые статистические модели
  - методы АОТ по уровням (NLP pipeline)
  - оценка качества
  - спойлеры к 4 курсу
- 4 курс: прикладные задачи АОТ («семантика», АОТ на уровне целого текста)

Основные активности

- Все занятия практические:
  - Лекция + обсуждение + код
  - Семинар с самостоятельной работой
- Чтение статей (на английском)
- Домашки
- Финальный проект

Оценка

- домашки (3) 40% ~10 дней
- квизы по статьям (3) 20% ~неделя
- проект (1) 40% ~месяц

Что нужно для оценок 9/10:

- задания со * в домашках и квизах
- оформление проекта

Финальный проект (1)

Корпус-менеджер (поиск по корпусу с лингвистической разметкой):

- Поэтапная разметка корпуса
- Организация поиска
- Представление результатов
- Более подробное ТЗ – в начале октября

Финальный проект (2)

По одному / командой до 4 человек

Возможно распределение ролей в команде:

- разработчик(и)
- аналитик
- менеджер проекта / продуктовый менеджер

Ранний период NLP

Машинный перевод и идея AI (искусственного интеллекта)

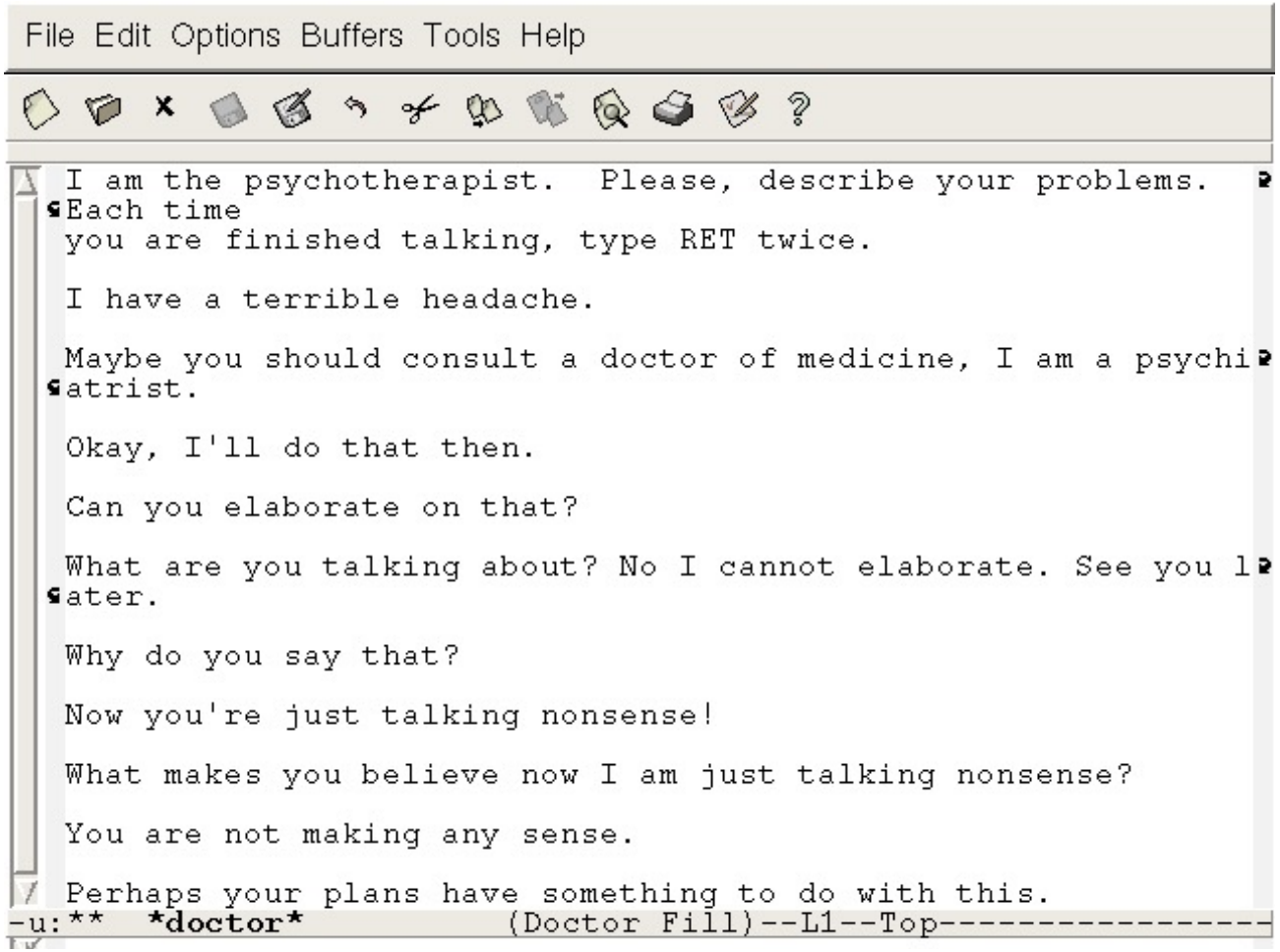
- **1940-е** – тест Тьюринга
- **1947** – Warren Weaver – идея статистического перевода
- **1954** – Джорджтаунский эксперимент – перевод по правилам
- **1958** – первая Всесоюзная конференция по МП
- **1966** – доклад ALPAC, AI Winter

Чатботы

они же:
chatterbots,
виртуальные собеседники,
диалоговые системы

1964-1966 – ELIZA

1972 – PARRY



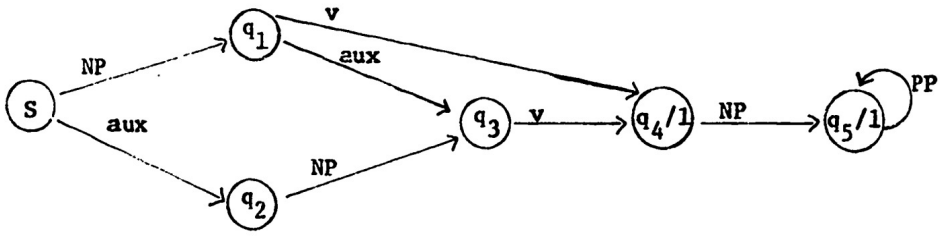
Natural Language Understanding

- **1970-e** – Conceptual Dependency Theory (R. Schank)

Figure 2. Basic form of a conceptual dependency graph.

Figure 3. Representation of "John gave Mary a book."

- Augmented Transition Network (W.A. Woods)



Появление статистических методов

- **конец 1980х-1990е –** внедрение статистических методов в различные направления NLP:
  - распознавание речи (speech recognition)
  - морфологический анализ (POS-tagging)
  - коллокации
  - классификация текстов

Применение методов АОТ

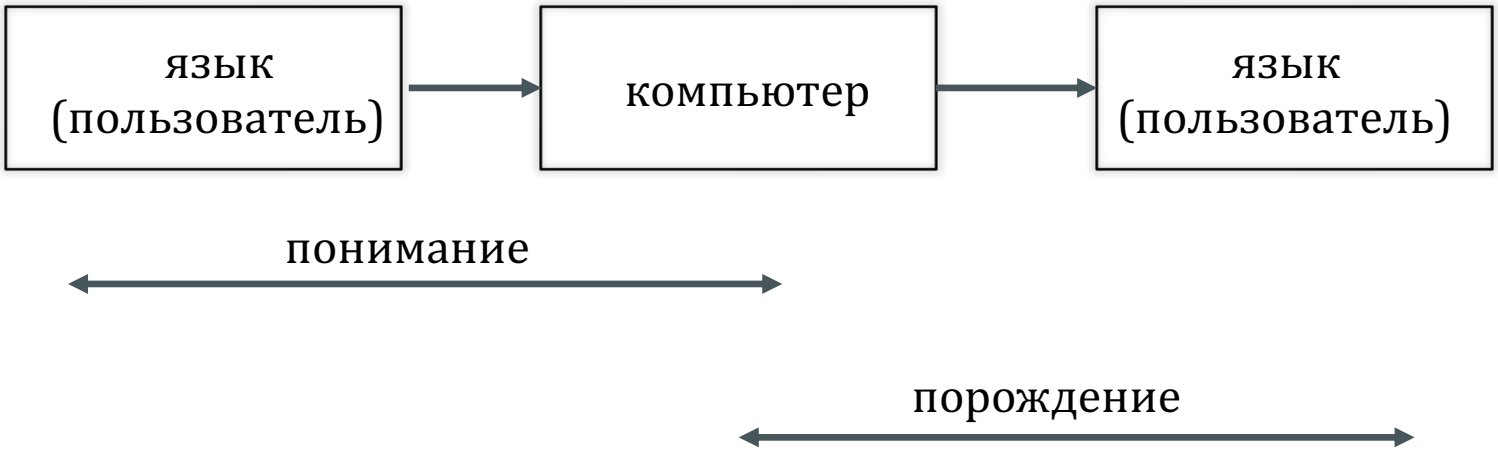
- информационный поиск
- автодополнение (клавиатуры, email); исправление опечаток
- извлечение именованных сущностей, фактов
- автоматическое реферирование; антиплагиат
- оценка тональности, извлечение мнений
- классификация текстов; выделение подтем в документе
- вопросно-ответные системы
- распознавание и синтез речи
- …

Задачи моделирования (1)

- синтаксис
  - supervised и unsupervised
  - rule-based — формальные грамматики
- семантика
  - онтологии, тезаурусы
  - дистрибутивные модели — классические ДСМ, word embeddings
  - семантические роли, фреймы

Задачи моделирования (2)

- понимание (NLU) vs. порождение (NLG) текста



-> связь АОТ и искусственного интеллекта (ИИ, AI – Artificial Intelligence)

NLP challenges

- **неоднозначность** языка на всех уровнях (linguistic ambiguity): 1 форма – N значений
- **синонимия** всех уровнях: 1 значение может выражаться N разними способами
- стилистическое разнообразие
- продуктивность (неологизмы)
- идиоматичность, некомпозициональность

Методы

- rule-based (основанные на правилах, требуют экспертизы)
- **статистические** (требуют данных)
  - классические
  - основанные на машинном обучении
- гибридные

Почти во всех задачах state-of-the-art (SOTA) – нейронные сети

Этапы обработки текста

Сегментация
(тексты, абзацы, предложения)

Mr. Smith bought ticket to San Francisco.
Мистер Смит купил билет до Сан-Франциско.

Этапы обработки текста

Токенизация
(слова, токены, стоп-слова)

Mr. Smith bought ticket to **San Francisco**.
Мистер Смит купил билет до **Сан-Франциско**.

??? Аналитические формы, компаунды, коллокации

Этапы обработки текста

Лемматизация / стемминг

Mr. Smith **bought** ticket to San Francisco.
Мистер Смит **купил** билет до Сан-Франциско.

Лемма ~ лексема ~ начальная форма

Стем ~ основа ~ усеченная словоформа

Этапы обработки текста

Морфологический анализ (~POS-tagging)

Mr./NNP Smith/NNP bought/VBD ticket/NN to/TO San/NNP Francisco/NNP ./.

Мистер/(NOUN,anim,masc sing,nomn)
Смит/(NOUN,anim,masc,Name sing,nomn | **...**)
купил/(VERB,perf,tran masc,sing,past,indc)
билет/(NOUN,inan,masc sing,**nomn** | NOUN,inan,masc sing,**accs**)
...

Этапы обработки текста

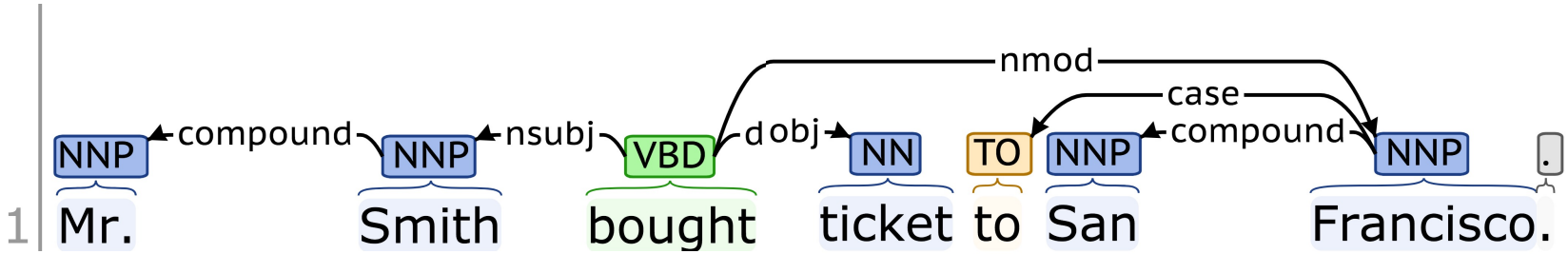
Разрешение неоднозначности (лемм / тегов)

Mr./NNP Smith/NNP bought/**VBD** ticket/NN to/TO San/NNP Francisco/NNP ./.

Мистер/(NOUN,anim,masc sing,nomn)
Смит/(**NOUN,anim,masc,Surn sing,nomn** | **…**)
купил/(VERB,perf,tran masc,sing,past,indc)
билет/(NOUN,inan,masc sing,nomn | **NOUN,inan,masc sing,accs**)
…

Этапы обработки текста

Синтаксический анализ (parsing)



Этапы обработки текста

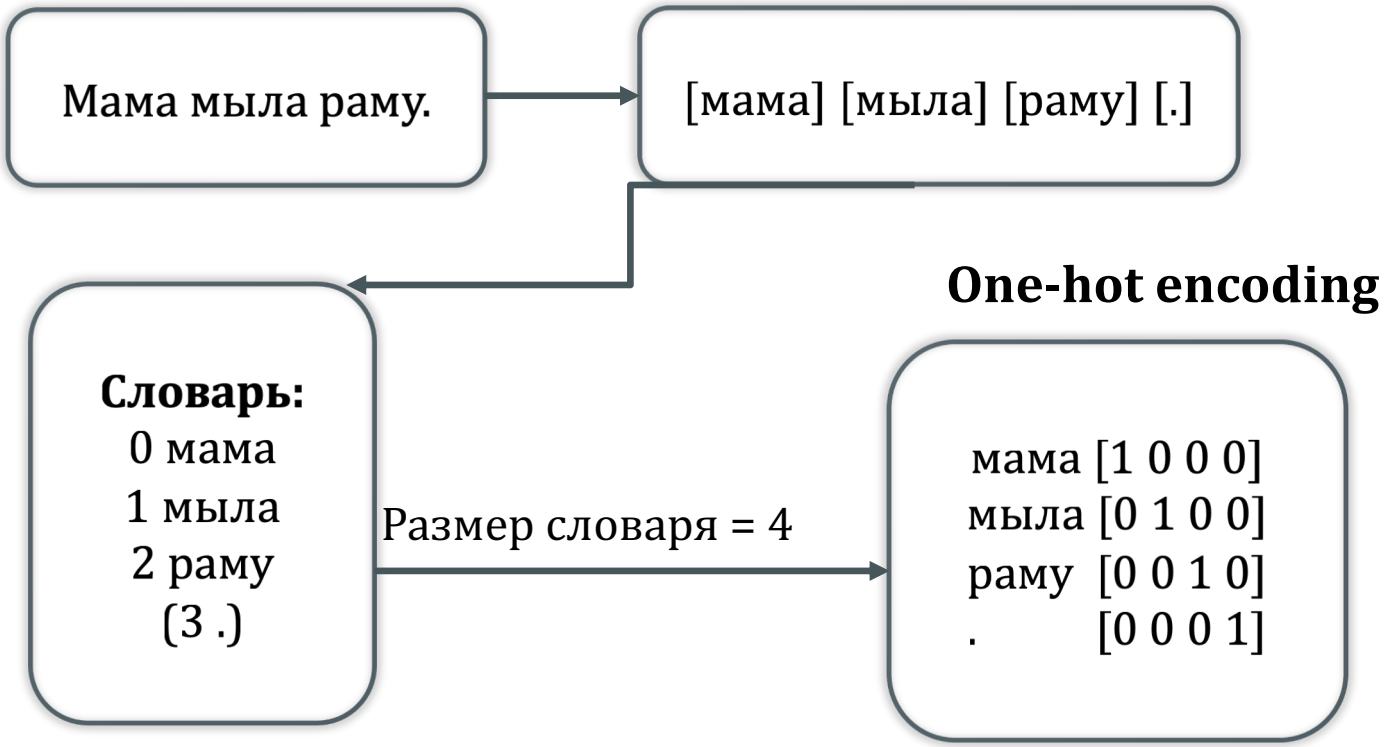
Семантический анализ? (Semantic Role Labeling)

купить: [ARG0: Мистер Смит]

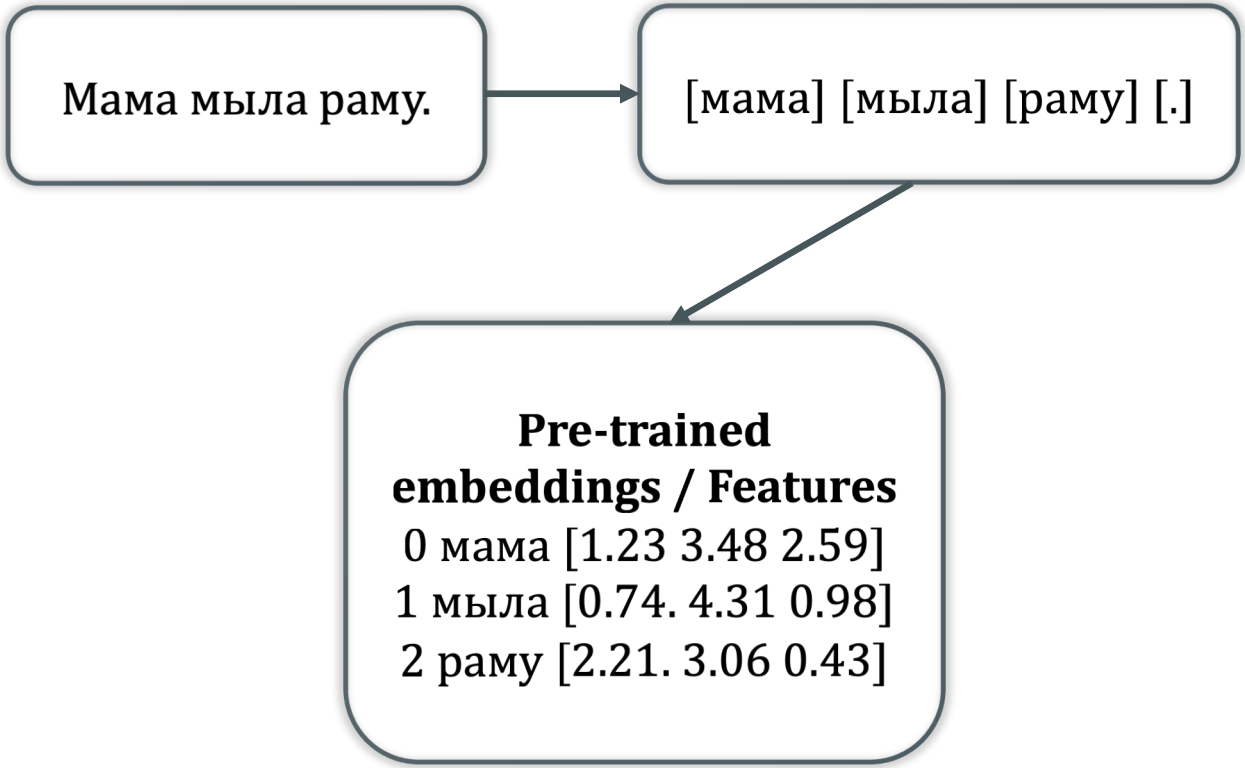
[V: купил]

[ARG1: билет до Сан-Франциско]

Представление данных (1)



Представление данных (2)



Данные: корпуса / датасеты

Тесты + сегментация + метаданные + разметка

- Корпуса одного языка
  - Brown corpus, British National Corpus, Penn Treebank
  - Национальный корпус русского языка (НКРЯ)
- Параллельные и многоязычные:
  - Europarl, UN Corpus, Opus
- Под специфические задачи
  - Hate speech identification, Twitter US Airline Sentiment …

Данные: тегсеты

Английский и мультиязычные:

- Stanford NLP
- Universal Dependencies

Русский

- Соревнования «Диалога» (Ru-Eval)
- НКРЯ (Mystem), pymorphy / OpenCorpora

Данные: подготовка

Разметка

(спец. инструменты – BRAT, …)

- Согласованность разметчиков (Cohen’s kappa)
- Краудсорсинг

Отбор данных

- Dataset augmentation / distillation

Оценка качества

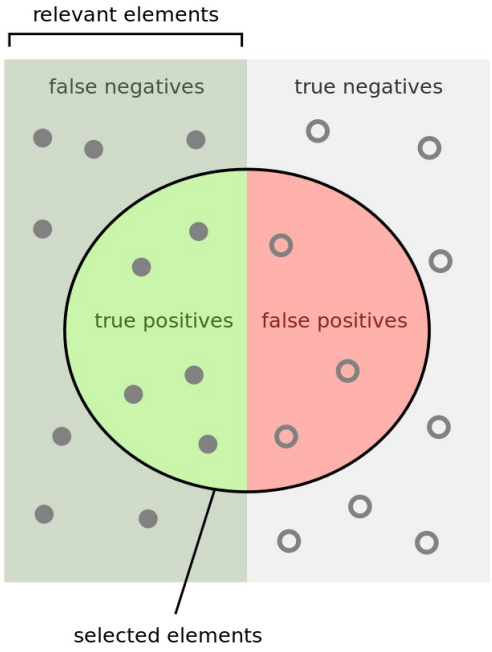
- **внутренняя**

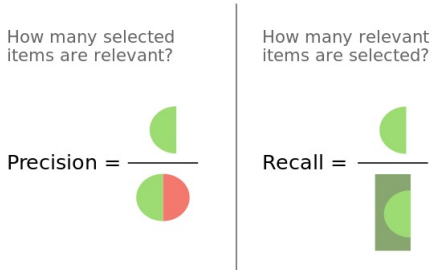
(из IR) точность, полнота, accuracy; специфические метрики

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **внешняя**

в более высокоуровневых приложениях





Золотой стандарт

= golden standard / benchmark

Проблемы:

- репрезентативность, сбалансированность
- выбор экспертов
- приближенность к реальным данным

Спасибо!

Вопросы?