

История (1)

Ранний период: машинный перевод

- **1947** – Warren Weaver – идея статистического перевода
- **1954** – Джорджтаунский эксперимент – перевод по правилам
- **1958** – первая Всесоюзная конференция по МП
- **1966** – доклад ALPAC

История (2)

Другие направления

- **1964-1966** – ELIZA, первые чатботы
- **1970е** – онтологии; Conceptual Dependency Theory (R. Schank)
- **конец 1980х-1990е** – внедрение статистических методов (распознавание речи, POS-tagging)

Этапы обработки текста

Сегментация

Mr. Smith bought ticket to San Francisco.

Мистер Смит купил билет до Сан-Франциско.

Этапы обработки текста

Токенизация

Mr. Smith bought ticket to **San Francisco**.

Мистер Смит купил билет до **Сан-Франциско**.

Этапы обработки текста

Лемматизация / стемминг

Mr. Smith **bought** ticket to San Francisco.

Мистер Смит **купил** билет до Сан-Франциско.

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

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)

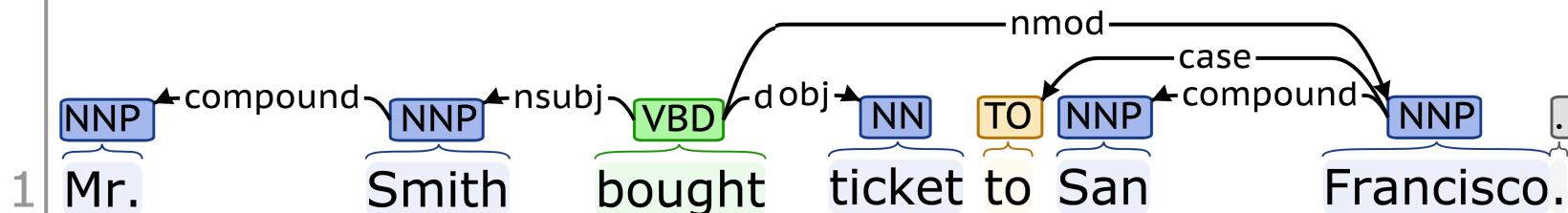
билет/(NOUN,inan,masc sing,nomn |

NOUN,inan,masc sing,accs)

...

Этапы обработки текста

Синтаксический анализ (parsing)



Этапы обработки текста

Семантический анализ? (Semantic Role Labeling)

bought: [ARG0: Mr. Smith]

[V: bought]

[ARG1: ticket to San Francisco]

Задачи моделирования

Синтаксис

- КС-грамматики
- обучение по размеченному корпусу
- grammar induction

Задачи моделирования

Моделирование значения

- онтологии, тезаурусы
- дистрибутивная семантика, word embeddings
- семантические роли; фреймы

Задачи моделирования

NLG vs. NLU

- понимание текста
- порождение текста

приложения:

- голосовые ассистенты, чатботы

Прикладные задачи

- извлечение фактов (named entity recognition, fact extraction, relation extraction)
- автоматическое рефериование
- оценка тональности
- классификация текстов; выделение подтем в документе
- вопросно-ответные системы

Прикладные задачи

Не такие очевидные:

- распознавание и синтез речи
- информационный поиск
- антиплагиат
- вопросно-ответные системы

Методы

- rule-based (основанные на правилах)
- **статистические**
- гибридные

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

Universal Dependencies

<http://universaldependencies.org>

более 70 языков; более 100 корпусов

Идея – однообразная разметка морфологии и
синтаксиса для разных языков

UD Pipe

<http://ufal.mff.cuni.cz/udpipe>

все этапы обработки текста в формате UD

анализатор:

<https://github.com/ufal/udpipe/releases/tag/v1.2.0>

модели:

<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2898>

UD Pipe

онлайн-демо:

<http://lindat.mff.cuni.cz/services/udpipe/>