Автоматическая обработка текстов на естественном языке.

Евгений Борисов

Введение в NLP

язык как средство познания мира

сегментация на объекты

классификация наблюдаемых явлений

обобщения

упорядочение окружающей действительности

Введение в NLP

о языке и задачах его автоматической обработки

обработка текстов на естественном языке (ЕЯ)
Natural Language Processing (NLP)

NLU / natural language understatnding

NLG / natural language generation

SP / speech processing (recognition/generation)

Введение в NLP

обработка текстов на естественном языке

NLP : NLU / NLG

- машинный перевод (MT)
- диалоговые системы (чат-боты)
- извлечение именованных сущностей, (named-entity recognition, NER)
- извлечения фактов и отношений (relation extraction)
- реферирование (summarization)
- поиск обоснования в тексте (argumentation mining)
- классификация текстов (оценка тона и т.п.)

Введение в NLP

Сложности автоматической обработки текстов - неоднозначности в языке

«эти типы стали есть на складе»

омонимия - случайное совпадение слов

ключ, лук, замок, печь

полисемия - несколько значений, связанных исторически

стол <организация или объект>

местоименная анафора — ссылки на контекст

Прискакал принц на белом коне. Принцесса выбежала ему навстречу и поцеловала его <...принца> .

эллипсис - пропуски в тексте

Он не может решить задачу, а я знаю как <...решить задачу>.

https://habr.com/ru/company/abbyy/blog/437008/

Введение в NLP

История развития компьютерной лингвистики (NLP MT)

перевод как расшифровка (1949, немецкий язык это зашифрованный английский)

язык как система логического вывода (индуктивный подход)

подход основанный на данных: большая таблица соответствий фраз

рационалистический подход: формальные грамматики Хомского (1957)

переход на уровень семантики, формальные онтологии, контекст (1970)

ML, корпусная лингвистика, статистические языковые модели (1980)

WWW, задача информационного поиска (1990)

Deep Learning, рекуррентные нейросети (2005)

Word2vec, семантические пространства (2013)

Attention, механизм внимания (2017)

Введение в NLP

Уровни сложности при автоматической обработке текстов

Лексика и морфология - список слов и речевых оборотов языка

Синтаксис - правила формирования последовательностей слов

Семантика - смыслы последовательностей слов.

Введение в NLP

Аналитический подход:
наборы грамматических правил

Подход основанный на данных:
корпус размеченных текстов и методы ML

Введение в NLP

Аналитический подход:
наборы грамматических правил

Подход основанный на данных:
корпус размеченных текстов и методы ML

методы решения задач NLP

частотный анализ (мешок слов, TF-IDF)

морфологический/синтаксический разбор

семантические пространства (Word2Vec)

языковые модели

Введение в NLP

Литература

```
git clone https://github.com/mechanoid5/ml_nlp
```