Natural Language Processing

Яковенко Ольга

Natural Language Processing – NLP – Автоматическая обработка естественного языка

- Распознавание речи
- Поисковые системы
- Автоматическое исправление опечаток
- Обнаружение спама…

NLP ⟷ NLU ⟷ NLG

Natural Language Processing

Natural Language Understanding

Natural Language Generation

Уровни языка

Он закрыл дверь.

- Фонетический

[О Н З А К Р Ы Л Д В` Э Р`]

- Морфемный

он за-кры-л дверь-[]

- Морфологический

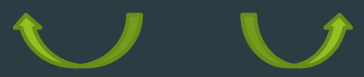
… закрыл – Г, сов.в.; пр.вр., м.р., ед.ч. …

- Лексический

… закрыл – «затворить, задвинуть двери или другие подобные преграды, …»

- Синтаксический

он закрыл дверь



NLP (basic)

Стемминг

Сокращение слова до его основы/корня (stem – корень)

умывался -> умыв
дверьми -> двер

Лемматизация

Приведение слова к его нормальной форме

умывался -> умываться
дверьми -> дверь

NLP (basic)

Стемминг

Сокращение слова до его основы/корня (stem – корень)

умывался -> умыв
дверьми -> двер

Лемматизация

Приведение слова к его нормальной форме

умывался -> умываться
дверьми -> дверь

Морфемно-морфологический уровень

NLP (basic)

Part-of-speech tagging (POS tagging/тэггинг)

По цепочке слов (предложению) определить соответствующую цепочку частей речи

Я ем мороженое -> Сущ., Г., Сущ.

Syntactic parsing (построение дерева разбора)

Построение дерева зависимостей предложения

он закрыл тяжёлую дверь

NLP (basic)

Part-of-speech tagging (POS tagging/тэггинг)

По цепочке слов (предложению) определить соответствующую цепочку частей речи

Я ем мороженое -> Сущ., Г., Сущ.

Syntactic parsing (построение дерева разбора)

Построение дерева зависимостей предложения

он закрыл тяжёлую дверь

Морфолого-синтаксический уровень

NLU (basic)

Language modeling

Построение языковой модели языка

Я -> играю в
играю в -> баскетбол
играю в -> волейбол
играю в -> футбол

Word sense disambiguation

Разграничение смыслов слова в зависимости от контекста

Открыть ключом != Бьёт ключом

NLU (basic)

Language modeling

Построение языковой модели языка

Я -> играю в
играю в -> баскетбол
играю в -> волейбол
играю в -> футбол

Word sense disambiguation

Разграничение смыслов слова в зависимости от контекста

Открыть ключом != Бьёт ключом

Лексико-синтаксический уровень

NLU (complex)

Sentiment recognition (распознавание тональности)

Определение эмоциональной окраски сообщения

Я ненавижу Сбербанк!!! -> негативное

Topic recognition (вопросно-ответные системы и чат-боты)

Определение тематики высказывания

Я ненавижу Сбербанк!!! -> Сбербанк

NLU (complex)

Sentiment recognition (распознавание тональности)

Определение эмоциональной окраски сообщения

Я обожаю ЦФТ!!!-> позитивное

Topic recognition (вопросно-ответные системы и чат-боты)

Определение тематики высказывания

Я обожаю ЦФТ!!!-> ЦФТ

Лексико-синтаксический уровень

NLG (complex)

Machine translation (машинный перевод)

Автоматический перевод текста с одного языка на другой

-//- -> Automatic translation of text from one language to another

Question answering & Chat-bots (вопросно-ответные системы и чат-боты)

По заданному вопросу (утверждению) подобрать корректный ответ

Какой город является столицей России? -> Москва

NLG (complex)

Machine translation (машинный перевод)

Автоматический перевод текста с одного языка на другой

-//- -> Automatic translation of text from one language to another

Question answering & Chat-bots (вопросно-ответные системы и чат-боты)

По заданному вопросу (утверждению) подобрать корректный ответ

Какой город является столицей России? -> Москва

Лексико-морфолого-синтаксический уровень + общие знания о мире

Задание

1) Named entity recognition (распознавание именованных сущностей)

По цепочке слов определить цепочку «сущностей», категорий объектов

Петя обратился в колл-центр Золотой Короны -> Name, o, o, o, Org

2) Automatic summarization (автоматическая суммаризация текста)

По тексту сформировать основную идею

В тридевятом царстве, в тридесятом государстве … -> Иван поверг злодея и женился

3) Optical character recognition (OCR)

По письменному/напечатанному тексту распознать слова, которые были написаны

4) Speech recognition (распознавание речи)

По аудио дорожке определить слова, которые были произнесены