Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

Кузнецов М. Д., Дочкина В., Резяпкин В., Русскин А., Ярмошик Д.
kuznetsov.md@phystech.edu

Данная работа посвящена изучению онтологических объектов и их гипонимий. Проведен сбор датасета, состоящий из текстов на английском языке, анализ и разметка данных текстов по нескольким уровням вложенности. Также (будет) проведена серия экспериментов для определения зависимости качества распознавания объектов от уровня вложенности. Обученные модели, полученые в результате данных экспериментов, могут использоваться в получении дополнительных признаков для решеня различных задач NLP. Исследованная методика разметки онтологических объектов в свою очередь может быть использована для автоматической разметки других текстов.

Ключевые слова: *гипонимия, онтология, онтологические объекты.*

Введение

Именованное распознавание объектов (NER - Named entity recognition) является важной задачей NLP (Natural Language Processing — обработка естественного языка), которая заключается в автоматическом распознавании объектов в тексте и классификация по определенным типам сущностей, такие как люди, организации, геополитические субъекты, места, события и т. д. К примеру, в предложении “Linux is my favourite OS” модель NER разметки определит слово “Linux” как “OS”, остальные же слова оставит “IR” (irrelevant). NER является фундаментальным компонентом многих методов извлечения информации, включая извлечение связей, , “ответов на вопросы” и др. В данной работе рассматривается немного отличная от NER задача. Мы рассматриваем не именнованные сущности, а так называемые онтологические объекты. Онтологическим объектом является любое слово, обозначающее какую-либо сущность, явление, предмет(?скажи, пожалуйста, правильно ли я это понимаю?) и др. Один из аспектов “связывания” объектов есть отношения обощений, или гипонимии и гиперонимии. К примеру, слово “seat” является гиперонимом слову “chair” и гипонимом слова “furniture” . Каждая сущность может быть обобщена до более высокого уровня вложенности понятий. Поставленная перед нами задача заключалась в том, чтобы создать некоторое количество датасетов, состоящих из одинаковых текстов, но имеющих разную глубину разметок по гипонимиям, а также проанализировать связи между уровнями вложенности обобщений для одинаковых онтологических объектов. Результаты проведенного анализа можно активно использовать для повышения качества выполнении задач NLP, а также разработки новых возможностей для уже существующих моделей. Данная модель (в будущем и возможно) способна анализировать контекст слова и давать именно ту ветвь вложенности обобщений, которая имеется в виду по смыслу в тексте/предложении. Для проведения эксперимента и исследования в качестве изначальных датасетов взяты (еще под вопросом, какие именно) тексты, для разметки взят словарь английский словать WordNet, имеющий информацию о гипонимах и гиперонимах.

Научный руководитель: Стрижов В. В. Задачу поставил: Бурцев М. С. Консультант: Баймурзина Д. Р.

Постановка задачи

Во-первых, для онтологических объектов были собраны и обработаны подходящие уровни гипонимии, на которые впоследствие эти объекты будут распределены, созданы словари на базе WordNet.
Во-вторых, проанализированы на оптимальность эти самые уровни.
В-третьих, подготовка датасетов, путем разметки выбранные тексты по разным уровням гипонимии.
В-четвертых, проведены некоторые эксперименты и анализ данных датасетов, выбор оптимальной глубины для различных задач.

Литература

[1] *Воронцов К. В.* LaTeX 2_ε в примерах. 2006. `http://www.ccas.ru/voron/latex.html`.

[2] *Львовский С. М.* Набор и вёрстка в пакете LaTeX. 3-е издание. Москва: МЦНМО, 2003. 448 с.

[3] *Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, Jiawei Han* Learning Named Entity Tagger using Domain-Specific Dictionary

[4] *Emily Sheng, Prem Natarajan.*A Byte-sized Approach to Named Entity Recognition

[5] *Bill Yuchen Lin, Wei Lu* Neural Adaptation Layers for Cross-domain Named Entity Recognition

[6] *Dilek Küçük* Joint Named Entity Recognition and Stance Detection in Tweets

[7] *Jian Ni, Radu Florian* Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping