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

Распознование именнованных сущностей (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 “Jim bought 300 shares of Acme Corp. in 2006.” модель NER разметки определит слово “Jim” как “Person”, “Acme Corp.” как “Organization”, а “2006” как “Time”, остальные же слова оставит “Out” (out-of-tag). 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. Рассматриваются не именнованные сущности, а так называемые онтологические объекты. Онтологическим объектом является любое слово, обозначающее какую-либо сущность, явление, предмет и др. Один из аспектов “связывания” объектов есть отношения обобщений, или гипонимия – иерархическая организация элементов, основанная на родо-видовых отношениях. Выделяют два определения: гипоним – понятие, выражающее частную сущность по отношению к другому, более общему понятию; и гипероним – понятие, выражающее общее, родовое понятие, название класса (множества) предметов (свойств, признаков). К примеру, слово “seat” является гиперонимом слову “chair” и гипонимом слова “furniture” . Каждая сущность может быть обобщена до более высокого уровня вложенности понятий. Поставленная задача заключалась в том, чтобы создать некоторое количество датасетов, состоящих из одинаковых текстов, но имеющих разную глубину разметки по гипонимиям, а также проанализировать связи между уровнями вложенности обобщений для одинаковых онтологических объектов. Результаты проведенного анализа можно 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 или предложении. Для проведения эксперимента и исследования в качестве изначальных датасетов

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

взяты (еще под вопросом, какие именно) тексты, для разметки взят английский словарь WordNet, включающий информацию о гипонимах и гиперонимах.

Обзор литературы

В решении задачи NER применялось множество различных методов и моделей. В сентябре 2018 года были представлены две модели для решения проблемы необходимости большого количества вручную размеченных данных, включающие в себя нейронную модель AutoNER с "Tie or Break scheme" [3]. В USC были проведы исследования, направленные на разметку не только слов, но и частей слов [4]. Также, результаты решения NER задач были использованы для определения эмоциального окраса сообщений в социальных сетях [5]. В том числе, были исследования, направленные на увеличение точности, а также многоязычности посредством использования Википедии [6]. Задача, поставленная в данной работе, несколько отличается от задачи NER, но модели и наработки можно использовать и/или модернизировать для решения.

Постановка задачи

Необходимо рассмотреть различные сущности, их гиперонимы; собрать и обработать подходящие друг другу для различных сущностей уровни гипонимии для онтологических объектов, на которые впоследствие эти объекты будут распределены, создать словари на базе WordNet. Следующий шаг – анализ выбранных уровней на оптимальность . Необходимо также провести обработку датасетов, путем разметки выбранных текстов по разным уровням гипонимии. Завершающим этапом исследования является проведение некоторых экспериментов и анализ данных датасетов, выбор оптимальной глубины для различных задач.

Литература

[1] *Воронцов К. В.* LaTeX 2$_\varepsilon$ 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] *Dilek Küçük* Joint Named Entity Recognition and Stance Detection in Tweets

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

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