[добавить шапку]

Аннотация

При обработке естественного языка возникает потребность распознавать имеющиеся в тексте онтологические объекты , а также уметь выстраивать из них цепочку уровней общности. Система WordNet позволяет для конкретного понятия найти гипероним и гипоним, таким образом можно построить дерево вложенностей. В данной работе датасет вложенности понятий был собран путём присвоения словам базового текста одного из возможных уровней в дереве гипонимии, взятого из WordNet. Для собранных наборов данных с различным уровнем гипонимии меток проведены эксперименты по качеству возможного распознавания сущностей.

Введение

Одна из распространных задач машинного обучения - получение объектов и понятий из текстов. Общая задача - научить машину понимать естественный человеческий текст. Этим занимается раздел NLP,Natural Language Processing. Методы этого раздела имеют много применений, например при текстовом поиске, синтезировании речи, кластеризации текстов. Один их подходов –выделение структурированной системы из неразмеченного текста. Главная процедура при этом – выявление объектов и выяснение отношени[?] между ними.
Классическая задача в этом разделе – извлечение именованных сущностей, NER.Под термином именованная сущность понимаем объект конкретного типа, у которого есть название или идентификатор. Примером являются имена людей, названия компаний: Victoria, Samsung. Решение задачи NER даёт возможность извлечь информацию из текста.
Из этих же соображений можно решать задачу определения онтологических объектов. Главная цель – выделить объект и отнести его к конкретному типу. Так слово "Chair"можно отнести к классу *Furniture*, а слово "river"к классу *Geographic location* . При этом делать это можно с различной степенью подробности, слово "river"также можно отнести к классу *Reservoir*. Здесь полезно ввести связь между более и менее общими понятиями – гипонимами и гиперонимами. По этим связям можно выстроить многоуровневую структуру онтологических объектов[Ещё пример?]
Видим, что есть свобода в выборе уровня, к которому отнести рассматриваемый объект. Итак, наша цель - провести исследование зависимости качества получения информации от способа выбора уровня. Для этого мы создаём[создали?] множество датасетов с разметкой , фиксирующей онтологические объекты, но варьируя фиксацию уровней гипонимии. Полученное множество датасетов уже позволяет искать исследуемую зависимость. [актуальность. Задача выделния информации из текста очень востребована, популярны и основные её подзадачи. В свою очередь исследование зависимости качества позволяет повысить эффективность решения поставленной задачи]

Обзор литературы:

Есть много методов решения задачи распознавания именованных сущностей. Например популярны методы машинного убучения с учителем [пример] Одной из первых была работа 1991 Лизы Рау, где предлагался эвристический подход к решению задачи. $https://www.researchgate.net/publication/3507105_Extracting_company_names_from_text$

Хороший обзор по медодам прошлого десятилетия есть в статье A Survey (2007) [ссылка] Всё это методы распознавания именованных сущностей. мы же решаем задачу поиска онтологических объектов. Тема другая, но похожая.