Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

Дочкина В., Кузнецов М., Резяпкин В., Русскин А., Ярмошик Д.

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.

Ключевые слова: *Natural language processing (NLP), named entity recognition (NER).*

1 1 Введение

Одна из распространных задач машинного обучения - получение объектов и понятий из текстов. Общая задача - научить машину понимать естественную речь. Этим занимается раздел NLP,Natural Language Processing. Методы этого раздела имеют много применений, например при текстовом поиске, синтезировании речи, кластеризации текстов. Одним из возможных подходов является выделение структурированной системы из неразмеченного текста. Главная процедура при этом – выявление объектов и выяснение отношени[?] между ними.

Одной из классических задача в этом разделе считается распознавание именованных сущностей (named entity recognition, NER).Под термином именованная сущность понимаем объект конкретного типа, у которого есть название или идентификатор. Примером являются имена людей, названия компаний. 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?] несколько датасетов с разметкой , обозначающей одни и те же онтологические объекты метками разного уровня 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]

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Бурцев М. С. Консультант: Баймурзина Д. Р.

2 2 Постановка задачи

Для улучшения качества решения задачи выделения онтологических объектов строится набор датасетов на основе произвольнях английских текстов. В тексте для каждой сущности с помощью системы Wordnet строится восходящая по уровню общности цепочка гипонимов. Каждый датасет отличается конкретным выбором уровня гипонимии. Далее для каждого датасетапроводится обучение одной из моделей для решения рассматриваемой задачи. Для этого можно использовать модели, аналогичные по строению моделям для задачи поиска именованных сущностей ТОгда,используя для выбранной модели соответствующую метрику качества, можно выделить наиболее эффективные по результату уровни гипонимиии.

3 3 список литературы

Методы решения задач распознавания именованных сущностей активно развиваюнся,начиная с девяностых годов прощлого века. Одной из первых была работа [1], где предлагался эвристический подход к решению задачи. Дальнейшее развитие методов в течение десяти лет хорошо описыается в [2].алее исследования продолжаются, часто уходя в узкие области для улучшения эффективности в частных случаяях. Так в [3] рассматриваются подходы, основанные на применении нейросетей,где размер метки накладываются на часть слова. И наоборот, использование нейронных сетей позволяет по сочетанию соседних слов. а не только последнему слову. предсказывать, какая метка будет дальше, как этоделалось в [4] Также хорошо показали себя модели на вероятностной основе, например CRF, описанная в [5]

Рассматриваемая здесь задача поиска онтологических объектов я вляется более общей. чем задача выявлеия именованных сушностей(?), номногие принципы и модели можно использоватьт в обеих задачах

Литература

[1] *Rau L. F.* Extracting names from text Proc. of the SeventhConference on ArtificialIntelligence Applications CAIA-92

[2] *Nadeau D., Sekine S.,* A survey of Named Entity Recognition and classification, 2007 `www.researchgate.net/publication/44062524_A_Survey_of_Named_Entity_Recognition_and_Classification`

[3] `arxiv.org/pdf/1809.08386v1.pdf`

[4] arxiv.org/pdf/1809.08730v2.pdf

[5] `repository.upenn.edu/cgi/viewcontent.cgi?referer=https://en.wikipedia.org/&httpsredir=1&article=1162&context=cis_papers`